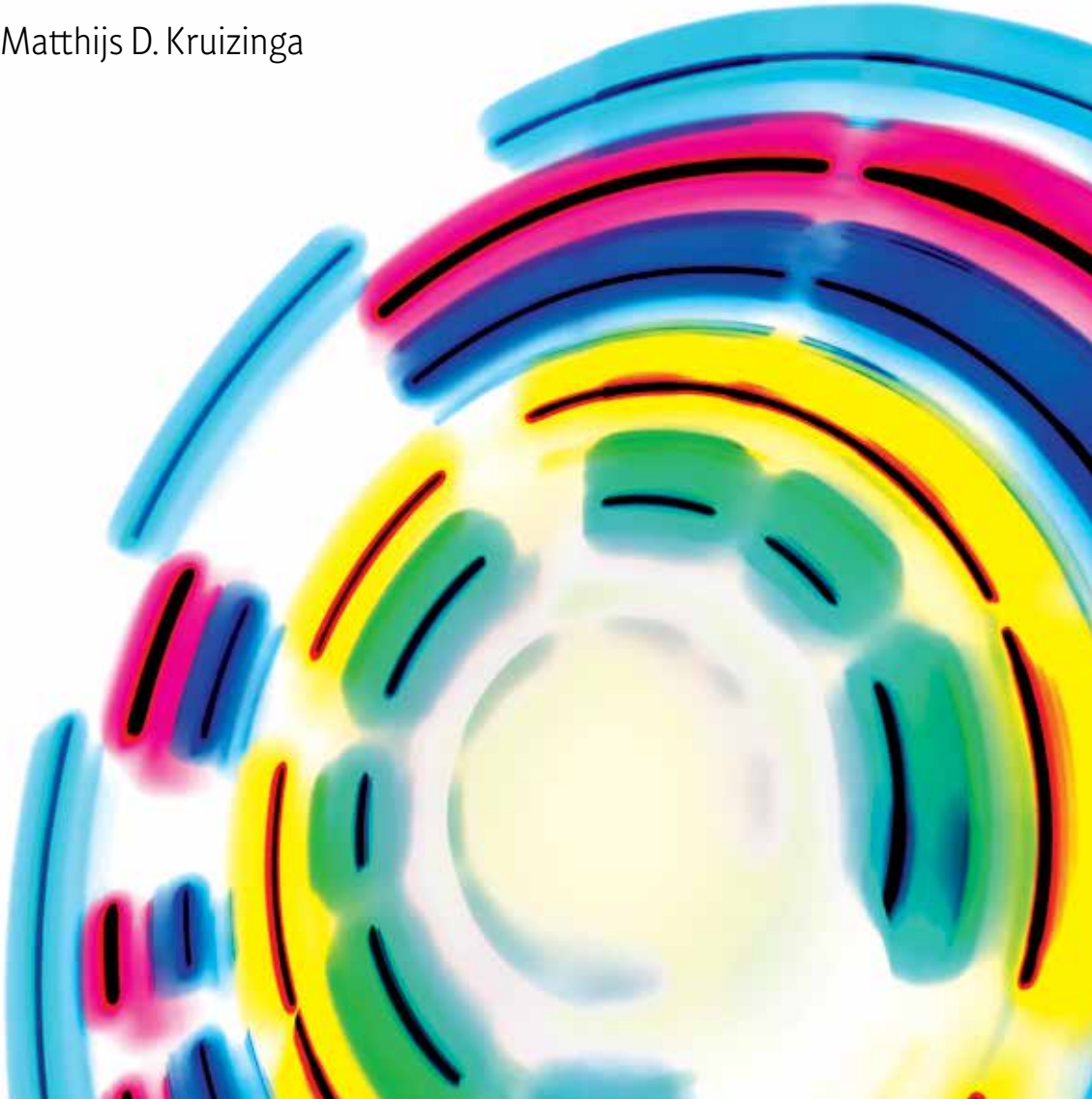
Trial@home for children

Novel non-invasive methodology for the pediatric clinical trial of the future

Matthijs D. Kruizinga



TRIAL@HOME FOR CHILDREN

NOVEL NON-INVASIVE METHODOLOGY FOR THE PEDIATRIC CLINICAL TRIAL OF THE FUTURE

Trial@home for children

Novel non-invasive methodology for the pediatric clinical trial of the future

PROEFSCHRIFT

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van rector magnificus prof.dr.ir. H. Bijl,
volgens besluit van het college voor promoties
te verdedigen op donderdag 10 februari 2022
klokke 16:15 uur

DOOR

Matthijs Derk Kruizinga
geboren te 's-Gravenzande
in 1991

PROMOTORES

Prof. dr. A.F. Cohen
Prof. dr. G.J.A. Driessen, Maastricht UMC

COPROMOTOR

Dr. F.E. Stuurman

LEDEN PROMOTIECOMMISSIE

Prof. dr. E. Dompeling, Maastricht UMC
Prof. dr. C.A.J. Knibbe, St. Antonius Ziekenhuis
Dr. P.J.M. van der Boog
Dr. M.G.H. van Oijen

Design Caroline de Lint, Voorburg (caro@delint.nl)

Publication of this thesis was financially supported by the foundation Centre for Human Drug Research (CHDR) in Leiden, the Netherlands

PART I – INTRODUCTION

PART II – TECHNICAL VALIDATION OF DIGITAL ENDPOINTS

PART III – CLINICAL VALIDATION OF DIGITAL ENDPOINTS

PART IV – NON-INVASIVE PHARMACOKINETICS

PART V – DISCUSSION

APPENDICES

PART I

INTRODUCTION

CHAPTER 1

The future of clinical trial design: The transition from hard endpoints to value-based endpoints

Handb Exp Pharmacol. 2019;260:371-397. doi:10.1007/164_2019_302. PMID:31707472

Matthijs D. Kruizinga,[1,2] Frederik E. Stuurman,[1] Geert Jan Groeneveld,[1,3]
Adam F. Cohen[1,3]

1 Centre for Human Drug Research, Leiden, the Netherlands
2 Juliana Children's Hospital, Haga Teaching Hospital, The Hague, the Netherlands
3 Leiden University Medical Center, Leiden, the Netherlands

Abstract

Clinical trials have been conducted since 500BC. Currently, the methodological gold standard is the randomized controlled clinical trial, introduced by Austin Bradford Hill. This standard has produced enormous amounts of high-quality evidence, resulting in evidence-based clinical guidelines for physicians. However, the current trial paradigm needs to evolve because of the ongoing decrease of the incidence of hard endpoints and spiraling trial costs. While new trial designs, such as adaptive clinical trials, may lead to an increase in efficiency and decrease in costs, we propose a shift towards value-based trial design: a paradigm that mirrors value-based thinking in business and health care. Value-based clinical trials will use technology to focus more on symptoms and endpoints that patients care about, will incorporate less research centers and will measure a state or consequence of disease at home or at work. Furthermore, they will measure the subjective experience of subjects in relation to other objective measurements. Ideally, the endpoints are suitable for individual assessment of the effect of an intervention. The value-based clinical trial of the future will have a low burden for participants, allowing for the inclusion of neglected populations such as children and the elderly, will be data-rich due to a high frequency of measurements and is possible to conduct with technology which is already available.

Introduction

Clinical trials are prospective research studies on human participants designed to answer specific questions about biomedical or behavioral interventions. This includes new pharmacological and non-pharmacological interventions (such as drugs, dietary supplements, vaccines, diets, medical devices, behavior change and surgical procedures), but also known interventions that warrant further study and comparison. The methodological gold standard is the randomized controlled clinical trial, which was introduced in the 1940s. The standard has served the medical profession well and has led to enormous amounts of high-quality evidence guiding our health care decisions. However, there are more and more reasons that changes are needed to the current trial paradigm. Spiraling costs for well-designed clinical trials preclude their application for all but the most expensive interventions. Trials focus on hard endpoints, like mortality and major clinical events, which are, thanks to the achievements made in the last 70 years, increasingly rare and therefore require increasing patient numbers to reach sufficient statistical power. Additionally, these events have decreasing consequences for the majority of the patients. The attrition rate, which is the percentage of new compounds that fail in clinical trials, has been 87-90% since the turn of the century, so a large amount of this effort is not benefitting patients directly[1,2]. Moreover, the therapeutic applications are evolving from single compound interventions that are widely usable, to precisely directed treatments and combinations of molecules and other interventions like devices or surgery. Modern technology and trials designs are therefore essential to sustain the evidence base for the increasing number of possible interventions.

Historical overview of clinical trial design

The first clinical trial occurs in the book of Daniel in the Old Testament (*Text box 1*) and was conducted by king Nebuchadnezzar of Babylon (500BC). The king ordered his servants to only consume meat and wine, a diet he believed to be superior. However, Daniel and several of his followers opted to only eat vegetables and drink water. They eventually gained authorization to do so for 10 days, after which they looked healthier than the servants of the king and were given their choice of food in perpetuity[3]. While this investigation does not concern a treatment and the quality of this trial and the resulting evidence is questionable (although probably correct), it is the first documented health care decision based on evidence gathered via a controlled experiment.

Text Box 1. The first documented clinical trial (3).

Daniel said to the guard whom the chief official had appointed (...):
"Please test your servants for ten days: Give us nothing but vegetables to eat and water to drink. Then compare our appearance with that of the young men who eat the royal food, and treat your servants in accordance with what you see."

So he agreed to this and tested them for ten days. At the end of the ten days they looked healthier and better nourished than any of the young men who ate the royal food.

DANIEL 1:12-16

The first novel therapy was not investigated in a trial until 1557, although completely by accident. When Ambroise Paré, a French surgeon working on the battlefield, ran out of the standard oil used to cauterize and treat wounds, he resorted to a surprising alternative. He documented the following: 'at length my oil lacked and I was constrained to apply in its place a digestive made of yolks of eggs, oil of roses and turpentine.' While he feared the worst for his patients, the alternative treatment appeared to be a big improvement compared to cauterization. The patients were 'feeling but little pain, their wounds neither swollen nor inflamed. The others to whom I had applied the boiling oil were feverish with much pain and swelling about their wounds'. This revelation led him to 'never again burn thus so cruelly, the poor wounded by arquebuses'[4].

However, this trial was uncontrolled and still extremely anecdotal. In 1747, the first controlled clinical trial took place, and it was on the open sea. James Lind was a surgeon on a ship and was greatly dismayed by the toll scurvy had on his fellow seafarers. He decided to conduct a trial with no less than 6 treatment arms with two patients each. They consisted of the most promising treatments adopted by physicians until then. Patients were administered a quart of cyder, elixir vitriol, vinegar, sea-water, an electuary recommended by another surgeon and finally: two oranges and one lemon a day. The patients who received fruit were found to be significantly better off compared to their crewmates[5], although Lind was hesitant to recommend the preventative treatment for all sailors in service because of the costs of lemons.

During this trial and the various investigations conducted in the centuries afterwards, treatment was allocated at the discretion of the investigator or physician, which meant that both patient and doctor were aware of the novel treatment. This changed in 1943, when the Medical Research Council (MRC) in the United Kingdom carried out the first double blind controlled trial investigating the efficacy of the antimycotic patulin as treatment for the common cold[6]. During this trial, patulin or placebo was allocated by a nurse using the method of alternation (or rotation) in an isolated room. While alternation was common during that time, this usually meant that the physician and patient were able to discern the used treatment arm. Besides the specifically designated room, two control groups and two treatment groups were used in this trial to reduce the possibility of advance knowledge of the allocations among the physicians who were also responsible for recruitment.

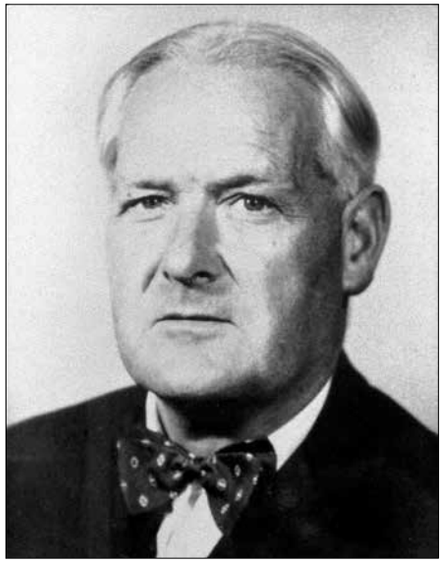


Figure 1. Sir Austin Bradford Hill, the statistician and researcher credited with the invention of modern randomization methods.

The strict double blinded treatment allocation was a big step forward in the prevention of observer and selection bias. However, the alternation method used was not truly random and therefore vulnerable to contamination of the study by mentioned bias. Austin Bradford Hill (*Figure 1*), a statistician and researcher based in London introduced a revolutionary random process of treatment allocation. He incorporated randomization in 1946 in the study design of a trial investigating the efficacy of streptomycin in the treatment of tuberculosis. In the resulting paper, the authors state the following: "the details of the (allocation) series were unknown to any of the investigators or to the coordinator and were contained in a set of sealed envelopes, each bearing on the outside only the name of the hospital and a number"[7]. Interestingly, it was deemed unnecessary to employ blinding in this landmark trial, as there was no possibility of bias when determining the presence of the primary endpoint: death.

During the decades following the patulin and streptomycin trials, randomization and double blinding became international standards, except for a small rebellion of researchers opposed to the burdensome processes which were the result of blinding and randomization in the 1970s[8,9]. Both were enthusiastically propagated by Bradford Hill, who can truly be considered the father of the modern clinical trial, and his colleagues[9]. Since 1948, more than 200,000 interventional clinical trials have been registered in international trial registries, many of them using randomization[10]. Little has changed in general trial design since then.

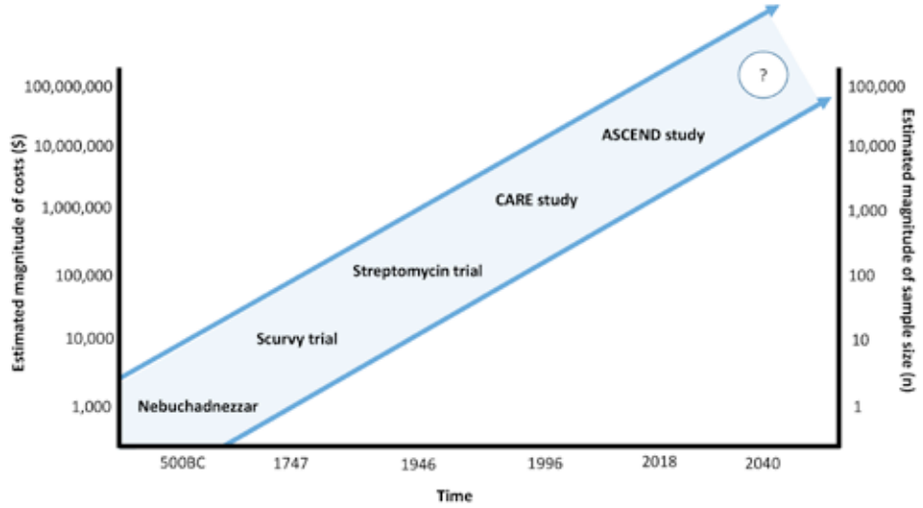
Increasing complexity and obstructions to clinical trials

While the general design has been relatively stable, clinical trials have become increasingly complex. Complexity of clinical trials is at least partly a natural result of the increasing complexity of health care. The first well designed clinical trials investigated a single antibiotic, but over the subsequent decades this evolved to drug combinations for tuberculosis or HIV, lifestyle interventions and finally, to combinations with medical devices such as drug-eluting stents for coronary artery disease and electrodes for deep brain stimulation in Parkinson's disease. The combination of multiple drugs, lifestyle programs and medical devices leads to the conclusion that we should no longer simply speak of investigational product in clinical trials, but rather of a new health care intervention.

A second factor is an increase in the size of clinical trials and the accompanying regulation. During the landmark streptomycin trial, participants were not aware they were included in a medical research study which is unthinkable in our current research practice. The circumstances and underlying rationale regarding the ratification of the declaration of Helsinki and the introduction of Good Clinical Practice guidelines have been well documented and are beyond the scope of this article. The regulations have undoubtedly increased trial quality, subject safety, and data integrity. However, overinterpretations of the guidelines has irreversibly led to the consequence that it is now very difficult to conduct trials the exact same way in multiple research locations. As a result, costly and burdensome monitoring procedures and bureaucracy have made clinical trials much more difficult and expensive to conduct.

Both the increasing complexity of the health care interventions and the increase in bureaucracy led to increased costs and loss of efficiency (*Figure 2*). Where the scurvy trial and the landmark streptomycin trial would probably cost no more than 100,000 dollars to conduct in the present day (not considering the current costs of a seaworthy wooden ship), current average costs of a phase II or phase III trial are approximately 8.6 and 21.4 million dollars respectively, but can be much higher and levels of 75 million dollars have been quoted[11]. The clinical trial of the future will have to counteract the spiraling costs because these contribute to the uncontrolled rise in cost of health care. Several advancements have already been made in this area.

Figure 2. The exponential rise of funds and sample sizes needed to conduct a clinical trial.



For example, in an attempt to improve efficiency and flexibility, some more recent adaptive trial designs utilize the results gathered in the trial to modify the trial's course according to pre-specified rules[12,13]. This is a requirement sometimes forgotten by proponents of adaptive clinical trials, which carries the risk of undermining the trial validity and integrity[14]. In addition, some experiment with umbrella and basket designs in cancer trials[15], but most of the designs that use treatment allocation or modification based on earlier findings increase the potential for bias. Although there are some good examples, ultimately these designs may not be sufficient for most situations and are only an evolution of the current paradigms (*Text box 2*)[16-18].

Besides changes in the general trial design, utilization of the principles of question-based clinical trial design can aid investigators in designing trials that answer the most pressing questions involved with a particular health care intervention[19]. This ideology (*Figure 3*)

Text box 2. Overview of innovative trial designs since 1946 (12,13,15–18)

Since the introduction of randomization and blinding in the 1940s, clinical trial design has seen only few new innovations. Designs of consequence were, among others, the sequential design, crossover design, factorial design and adaptive design. The designs are meant to improve efficiency, reduce the number of participants or improve the chances of finding clinically relevant outcomes. Some more specific designs utilized in oncology may improve clinical trial efficiency in the field as well. However, the several innovations come with flaws, such as the introduction of bias and preclusion of use in common situations. Still, some of the newer clinical trials designs, such as the adaptive trial, may significantly improve trial efficiency. Here, we will discuss the advantages and disadvantages of a select number of designs.

CROSSOVER DESIGN Crossover designs allocate each participant to a sequence of interventions. A simple randomized example is an 'AB/BA' design in which participants are randomized initially to intervention A or intervention B, and then 'cross over' to intervention B or intervention A, respectively. The major advantage is that subjects are used as their own, perfectly matched, 'controls'. This improves the statistical power of the trial and therefore the efficiency. However, there are important conditions to be met regarding the treatment before a crossover design can be utilized. First, the disease should be chronic and stable and the first treatment should not cure the disease. Second, a washout period must be implemented to allow for complete reversibility of drug effects. Besides the washout period, the investigator must be absolutely certain there is no carry-over effect. Also, treatment effects should be quickly observable in order to prevent natural progression of the disease to influence trial results.

FACTORIAL DESIGN As time progressed, more and more treatments became available. Subsequently, investigators also needed a method to research the effects of combinations of treatments. Factorial designs enable efficient simultaneous investigation of two or more interventions by randomizing participants in a treatment group receiving one, multiple or no intervention. The simplest form is the 2x2 factorial design investigating 2 interventions (A and B). In this design, participants either receive A alone, B alone, both A and B or neither A nor B (control). A major advantage is the option to investigate both the individual treatment benefits and effects and interactions of receiving multiple interventions together. When there are no treatment interactions, this design greatly increases the efficiency of a trial. However, interactions usually cannot be reliably excluded during trial design. Sample size then becomes a major factor in the trial, for if the trial is to have adequate power to detect an interaction, the sample size increases dramatically. This makes the factorial design still inefficient for many health care intervention studies and therefore relatively rare.

SEQUENTIAL DESIGN In the late 1950s, clinical trials started to adopt sequential designs. Here instead of a predefined sample size, a pair of statistical borders are drawn: one to decide the rejection of the null hypothesis and the other to accept. Trial results are analyzed continuously or during planned interim analyses and after each analysis an accept, reject or continue decision is made. This allows for early discontinuation of futile trials, but more importantly, also for a quicker and more efficient road towards acceptance of a new health care intervention. However, the design was the subject of several critical opinion pieces. It was argued that clinical trials were not about making 'accept' or 'reject' decisions, but rather about estimating the range effects between treatments and therefore it should remain necessary to conduct and complete adequately powered trials. While the sequential design has since then become a rarity, concepts have been integrated in some adaptive trial designs.

ADAPTIVE TRIAL DESIGN Adaptive designs add a review-adapt loop to the regular, linear paradigm. This way, scheduled interim analyses are conducted in order to apply and allow pre-specified changes to the trial's protocol or sample size. All changes have to be applied while retaining the validity and integrity of the trial. Common adaptive designs are the interim sample size reassessment, adaptation of the allocation ratio towards the superior treatment and the dropping of inferior treatments, addition of new treatment arms to save time and resources, population "enrichment" to narrow scope of the clinical trial. While the potential of adaptive clinical trials looks good on paper, the planning and implementation of the design requires a lot of effort and time, which may trump the potential gains in efficiency. Furthermore, specialized statistical knowledge is necessary to conduct simulations and gain insight in all possible consequences of the possible adaptations. Therefore, adaptive designs are still relatively rare, despite being available for more than 25 years.

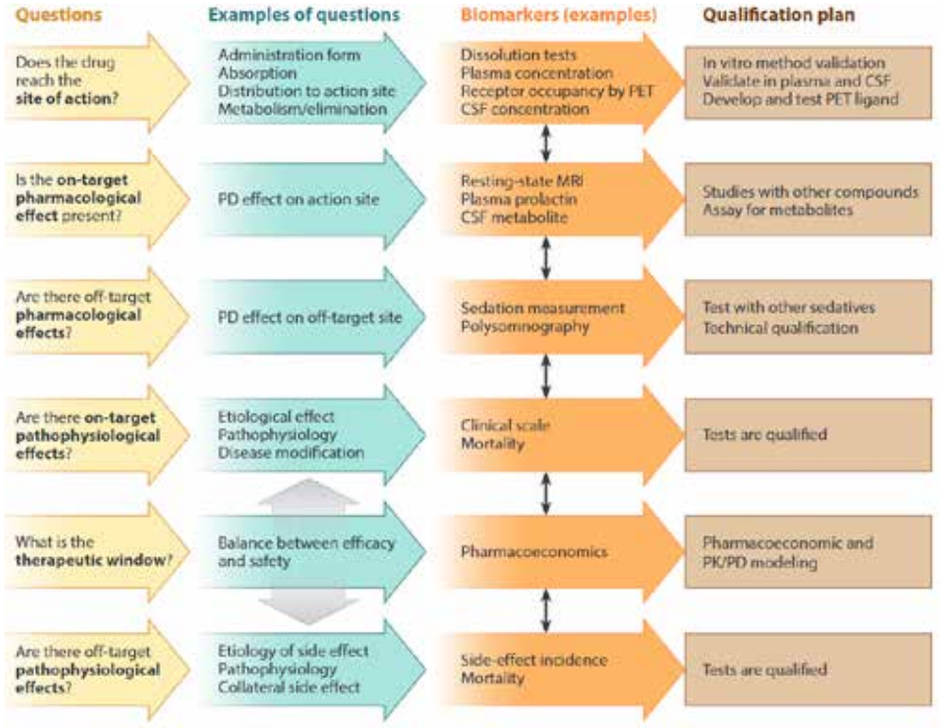
UMBRELLA & BASKET DESIGN In oncology, new trial designs have been developed to combine the principles of individualized medicine with histology and specific genomic changes of the tumor. Umbrella trials and platform trials maintain the single histology focus of traditional clinical trials but stratify treatment evaluation based on prespecified genomic biomarkers. In contrast to umbrella designs, basket designs do not focus on disease histology, but on patients with a specific genomic change. Patients are assigned a regimen that is expected to be active for tumors containing that alteration. Often this expectation is based on knowledge of the target of the drug and its role in the progression of the disease as well as previous approval of the drug, or a similar drug, for patients with the same genomic alteration in some specified histology.

PARALLEL GROUPS REMAINS THE STANDARD In the end, the classic parallel group design, introduced around the time of Austin Bradford Hill, is used the most often and this design will most likely remain the cornerstone of health care intervention research. New, innovative trial designs may make trials more efficient, smaller and thereby reduce costs. However, all new trial designs introduce limitations or some sort of bias and therefore may never be enough to solve the current problems facing the process of introducing new health care interventions. In the end, the concepts of blinding and randomization are strong concepts that well designed clinical trials may never go without.

encourages investigators to answer questions covered in six distinct domains during the developmental process. The domains cover the fundamentals of mechanistic health care intervention research: important pharmacokinetic aspects of new compounds such as absorption and excretion, but also the pharmacological, physiological and clinical effects a health care intervention could induce[19]. The domains may state the obvious, but research indicates that in almost half of drug development projects, mechanistic aspects are not investigated thoroughly enough and therefore have a considerable chance of failure[20]. Following the question-based principles could therefore increase efficiency in developmental processes while ensuring drug attrition occurs as early in the developmental process as possible and sharply reducing costs.

Furthermore, the health care intervention development process could be made more efficient by incorporating the principles of Health Technology Assessments (HTA) in early phase clinical trials. HTA is the systematic evaluation of properties, effects, and/or impacts

Figure 3. Domains of question-based drug development (19).



of health technology. The main purpose of conducting an assessment is to inform a policy decision making regarding reimbursement and decide on incorporation in treatment guidelines[21]. Usually, HTA is conducted at the end of the clinical drug development process, partly while using data that was gathered at an early phase. Utilization of the data as it becomes available at an early stage may identify compounds that are doomed to fail, while also allowing for the allocation of more resources towards health care intervention that show promise in early assessments[22].

While incorporation of innovative and question-based designs and early HTA in clinical trials will undoubtedly lead to efficiency gains and cost savings, outcome parameters obviously are an important factor. The role and importance of correct endpoint measurements may have remained a relatively underemphasized area, perhaps because of the importance that has been given to the so-called hard endpoints. Trial outcomes that evaluate the incidence of major health events, such as mortality or vascular or neurological events are doubtlessly important. However, as their incidence has become lower with better health care there is a requirement for even more patients in a trial and an increasing number that would never have experienced the event, whatever the treatment.

A good example is the recent ASCEND study investigating whether aspirin is of additional value for the primary prevention of cardiovascular disease in patients with type 2 diabetes. For this trial, a sample size of 15,000 subjects followed for 7.5 years was necessary on the basis of an event rate of 1.2 to 1.3% per year. At the end of the study, a barely significant effect (odds ratio 0.88 [0.80-0.97]) was reported on the composite endpoint 'Any serious vascular event including TIA'[23]. While statistically significant, the slightly lower chance of a group of complications does not represent great value for the individual patient who is part of a majority that will never experience any of these events whatever the treatment. On the contrary, the composite endpoint 'Any adverse event' is rarely included in clinical trials[24]. This imbalance invariably skews results when comparing advantages and disadvantages of new health care interventions. Furthermore, the sample- and effect size of the ASCEND study are in stark contrast with sample sizes in the early clinical trials. In the streptomycin trial of Austin Bradford Hill, only 109 subjects were included in the study and the death rate was halved from a control mortality of 45%[7].

To cope with the increased complexity, size and costs of the current clinical trials, a more radical change in general clinical trial design is needed, particularly in the way we choose trial endpoints: the value-based clinical trial in opposition to the event-based trial. In this approach we follow the principles of value-based health care[25].

Event based versus value-based endpoint trials

The idea of a value-based clinical trial is born out of the introduction of value-based thinking in business and health care. This concept of shared value in business was first introduced in 2006 by Michael E. Porter, professor of economics at Harvard, as a way of developing profitable business strategies that deliver tangible social benefits[26]. The paradigm was further expanded in a 2012 report regarding measurement strategies of shared value[27]. Porter and his colleagues concluded that the measurement of shared value strategies is as important as the implementation, since this allows quantification of value, provides insight in areas for improvement of the strategy and allows scaling towards larger implementation of the strategy in the organization.

After proving the benefits of shared value concepts in business, Porter turned his attention to introducing value-based thinking in health care[25]. Here, value is captured in the formula 'health outcomes that matter to patients/costs of delivering the outcomes'. The simple formula indicates that one can create value in health care either by improving health outcomes or by lowering costs during the care for a patient. This approach encourages to focus more on collaboration between health care providers and on sharing data to measure outcomes easily. Furthermore, the approach encourages health care providers to stop asking themselves how a patient fits in a specific treatment strategy but rather how the provider can help the individual patient sitting before them in the best way. Finally, it encourages health care providers to use big data and modern technology to assist in decision making and to evaluate the health care results. During the last 9 years, integration of value-based health care has accelerated, partly due to the accompanying incorporation of financial incentives in some countries to embrace the concept[28].

The value-based concepts in business and health care focus on the measurement of outcomes that matter to consumers or patients, e.g., social improvement and health benefits and on the analysis of costs. What is odd is that value-based thinking has not reached clinical trial design yet, as trials should generally focus on measuring the effect, and hopefully improvement of health care interventions in patients' lives (*Table 1*). Concurrently, the incorporation of modern technology and big data in clinical trials is slow[29]. Instead, clinical trials stay focused on the measurement of events and (composite) hard endpoints. While this approach was feasible and preferable during the streptomycin trial of Bradford Hill and the introduction of other radical new therapies, such as aspirin and metformin, medical research has now made our treatments quite advanced. So advanced that, in the

developed world, patients generally do not die of pulmonary tuberculosis anymore, cardiovascular mortality following a myocardial infarction is as low as 3%, and glycemic control is quite good with the current arsenal of anti-diabetic medication[30,31].

Table 1. Characteristics of value-based concepts in business and health care and our proposal for value-based clinical trial design.

	Business	Health care	Clinical research
Priority & focus	Generate economic and social value	Improve outcomes patients care about	Conduct trials with endpoints patients care about
Determination of value	Calculation of the relationship between social improvement and economic value creation of a process.	Health outcomes that matter to patients costs of delivering the outcomes	Value-based + hard endpoint / Costs of conducting the trial
Main distinct goals	Improve social outcomes and increase economic value of a business	Improve health care by improving important patient outcomes while staying cost-effective	Improve clinical trials by using value-based outcomes combined with hard endpoints, while improving efficiency to save costs.
Personalized approach	No	Yes	Yes
Utilize technology	Yes	Yes	Yes
Collect more data	Yes	Yes	Yes

Areas of opportunity for value-based trials

The current paradigm has also led to gaps in knowledge regarding the real-life impact and actual value our health care interventions for patients. This could underestimate the negative effects of treatments. For example, we know statins have a generalized negative effect on cellular mitochondria and consequently can lead to muscle complaints[32]. A larger than expected proportion of patients that suffer from muscle aches was reported already in 1991[33]. Nevertheless, the adverse event was not even mentioned during the landmark CARE trial, which demonstrated significant survival benefit for patients with coronary artery disease[34]. The effects of statins on general physical activity, an endpoint which directly measures the value of the therapy and the general wellbeing of all patients on the treatment, was not investigated in a randomized controlled trial until 2012[35,36]. They demonstrated a negative treatment effect on mobility, particularly in older patients. In a value-based system, this endpoint would have been included in the very first clinical trials.

Several other research fields have gaps in knowledge regarding the real-life impact of disease and could benefit from a more value-based approach, such as psychiatry. The cornerstones of current trials investigating depression are (validated) questionnaires and depression scales, such as the Hamilton Depression Scale. While used frequently, one could debate the value of scales and questionnaires that, at best, are a subjective measurement of only 40% of depression symptoms[37]. While it is easy to criticize the objective value of the response to statements such as 'my life is pretty full', included in the Zung Self-Rating Depression Scale[38], a value-based trial would focus less on asking questions like these every other visit. Instead, it would focus more on objectively measurable parameters that directly impact a patients' wellbeing. Examples include measuring the amount of social interaction a subject engages in by using their phone, monitor a patient's radius of action around their home, or monitoring the properties of the sound of their voice. Current technology allows an easily and cheaply implementation[39,40].

The introduction of value-based thinking could also improve the execution of trials, including those in vulnerable populations like children. Pediatric trials are notoriously difficult to conduct because of a difficult ethical approval process, slow recruitment, outcome measures that cannot be directly derived from adult trials and subjective data obtained from parents which introduces recall and respondent bias that clouds trial results[41]. Here, value could be generated by the introduction of technology in the home situation. This could reduce the burden and visits for the patient, generate more useful data and -in the process- ease the process of obtaining informed consent. Also, value-based trials would focus more on the individual child and objectively measurable behavior, such as the effect a certain asthma intervention has on general physical activity and sleep quality, which we believe to be important indicators of general health[42,43]. This would allow to answer questions which have been present in the field for decades and enables reevaluation of interventions that have provided conflicting results in conventional clinical trials, such as the efficacy of montelukast in pediatric asthma and recurrent wheezing[44,45].

The road towards new clinical endpoints

The value-based endpoints are opposed to the current hard endpoints that evaluate aspects of disease which are reliable to measure but lost their immediate relevance with the improvement of modern medicine. We propose the following guidelines for new value-based endpoints, on which we will further elaborate in the following pages (*Table 2*).

Table 2. Guidelines for the value-based endpoints

Evaluate symptoms and consequences of disease patients care about.
Focus on home measurement to increase real-life relevance and decrease burden for participants.
Employ continuous monitoring to elucidate day-to-day variability in symptoms.
Increase the role of EPROS in order to adequately value the subjective experience of patients.
Allow for individual assessment of the effect of an intervention.

First, the basis of value-based trial design should be that clinical trials evaluate symptoms and endpoints that patients care about. They should be assessments that directly or indirectly measure an aspect of the disease that, if relieved, improved, or prevented would be meaningful to patients. Second, value-based trials should trend towards incorporating less research centers in their trial and incorporate value-based endpoints which ideally measure a state or consequence of disease in the natural environment of the participant: home and work. While clinical research units allow standardized measurements by trained personnel, the environment usually does not induce the usual behavior of patients. Third, assuming that none of these endpoints are stable over time, endpoints should allow much more frequent measurements than weekly or monthly assessments. Ideally, the endpoints are suitable for individual assessment of the effect of an intervention. Finally, while the patients are in their natural environment, measurement of the subjective experience with the use of electronic patient reported outcomes (EPROS) is essential, especially in relation to other objective measurements.

Movement towards monocenter studies

The large sample sizes needed for the hard endpoints in event-based trials, as well as regulatory directives requiring local sites in pivotal studies, invariably leads to a multi-center study in multiple countries. As a consequence, the accompanying administrative burden, monitoring procedures and costs have risen enormously. We expect incorporation of value in trial design will lead to smaller sample sizes and therefore will reduce the need for large multi-center studies, when accompanied with a relaxation of regulations. Eventually, this could lead to the return of the monocenter study as the standard in clinical trial design. When a trial holds value for the patient, a short travel time will not deter participation. This will benefit trials by standardizing the conduct of complex measurements and procedures. Reductions in costs due to more streamlined logistics and efficient data management procedures are other advantages of the monocenter study. Evidently, when a sample size necessary for the trial exceeds the capacity of the service area of a clinical research unit, patients will have to travel longer for participation in the trial, which will inevitably lead to a multi-center approach. However, this effect could be countered by the embrace of technology to allow for home-measurements, either in a hybrid form with few visits at the research unit or in the form of a completely home-based trial.

Technology allows home-based trials

Home-measurements include use of technological advances to employ wearable devices and other *@home* devices to measure disease-activity outside of the clinical research unit. This has several advantages. First, directly measuring symptoms in daily life allows for more realistic data capture and will aid in the determination of real-world effects of health care interventions. Furthermore, the non-invasive nature of most devices for home use will reduce the burden for study participants, while automated data streams reduce the administrative burden and data entry errors at the research sites. The uncontrolled environment where study assessments will take place and the reduced threshold to drop out of a study are pitfalls of home-based trials, although there is little data available to substantiate this. Wearable devices in health care and research are generally hyped and a popular topic for publication (29,40,46-49). While a PubMed search in 2010 for the word 'wearable' would yield a mere 1,092 results, this number has increased to 8,827 in March 2019[50]. Nevertheless, incorporation of wearables in clinical trials has lagged behind the hype. A review regarding mobile device related endpoints in clinical trials identified only 22 interventional studies using mobile data as an endpoint[51]. Of these, 10 trials used device data as their primary endpoint. Still, there is no doubt home-based measurement and wearables hold promise in improving clinical trials. Exciting anecdotal reports exist of wearable technologies assisting in the diagnosis of Lyme disease and the tracking of inflammatory disease[48]. A simpler, more obvious and already widely used example is the measurement of blood pressure at home, which negates the well-known issue of white-coat hypertension[52].

An interesting development is the integration of smartwatches and other wrist worn sensors in clinical trials[47]. They allow for the continuous measurement of parameters such as physical activity, sleep and heart rate and some are also capable of measuring blood pressure and environmental factors such as temperature and altitude. The array of possibilities could be expanded in the near future with sensors capable of reliably measuring blood glucose, oxygen saturation and a variety of environmental factors such as air

pollutants, background sounds and ambient light levels[46]. However, the validity of measurements is a factor that should be investigated in individual watch models, particularly in the case of heart rate analysis. There appears to be some discordance[53], and devices are generally not medical grade or meant for use by patients. Furthermore, measurement devices may also be used incorrectly by patients as they are no longer assisted by extensively trained trial staff. However, when the frequency of measurements is high enough, occasional results that do not correspond to the gold standard are suboptimal but not disqualifying. With further progression of our technological capabilities, accuracy of wrist worn sensors will improve as well.

In addition, several devices have been developed for home use that can easily measure vital signs and other outcomes like ECG. Smartphone-based ECG recording systems are already validated for the evaluation of rhythm disorders and outperform conventional Holter monitoring in specific populations in both accuracy and patient satisfaction[54]. Another example is the use of spirometry. Where research participants used to come to a clinical research unit to perform a spirometry test to evaluate treatment, patients with respiratory disease can now easily perform a complete spirometry maneuver by connecting a mobile spirometer to their phone[55]. This could be combined with big data regarding environmental exposures such as pollution, pollen counts and general weather in the vicinity of a patient in order to create a personalized profile of what exposure leads to reduced pulmonary function in a specific patient: an approach that could join the concepts of value-based thinking with the hard endpoint FEV1. Other examples that could add valuable home-monitoring to clinical trials are the Abbot Freestyle Libre for glucose monitoring in diabetes and the biometric shirt system Hexoskin for continuous monitoring of biometric parameters.[56,57].

Finally, the device that may show the most promise for incorporation in clinical trials is a device virtually all participants already own: their smartphone. Every smartphone has a range of sensors that could collect data continuously, such as an accelerometer, light sensor, GPS, microphone and a variety of apps which frequency of use may indicate how a patient is feeling. Some of this data is already being collected by tech companies and could also be used for clinical research, with the caveat that any privacy concerns should be adequately covered. Customized apps could provide participants with simple but valuable test assessments. Furthermore, video-observed administration of medication by study subjects could be superior compared to directly observed administration, which is standard practice in research units[58]. One of the first studies that solely used the smartphone in clinical research is the mPower study[59]. In this study, patients with Parkinson's disease downloaded the study app on their iPhone. They were asked to perform a memory, tapping and voice activity on their phone, as well as a performing a walking activity and questionnaire. Patients could complete each activity three times a day but were allowed to skip assessments as they saw fit. Study designs like these could elucidate the day-to-day variability of symptoms and effects of health care interventions in several, if not all, diseases while being extremely non-invasive.

A second novelty of the mPower study was the implementation of electronic consent (eConsent). The concept of eConsent ideally allows patients to obtain all relevant information regarding the trial, ask questions to the responsible investigators and provide adequate written consent. The concept could make the mobile interventional clinical trial a reality, but there are several limitations that should be overcome before widespread implementation. For example, investigators should be wary of a 'Facebook effect', where subjects are barely aware of the consequences of their decision and the way their personal and medical data are treated. This is a realistic fear, as research indicates only 0.1-0.2% of consumers access end user license agreements at all[60]. Investigators in the mPower study successfully implemented eConsent after experimenting with several ways to test subject comprehension, such as forcing potential participants to obtain a perfect score on a series of questions regarding the study. Eventually, an assessment in which every incorrect answer led to more education appeared to be the most sensible approach[61]. While this is a promising approach for the implementation of eConsent, one should also not underestimate the proportion of patients who do not completely comprehend the current paper consent form format, which a recent study found to be a mere 54%[62]. In fact, the development of eConsent applications and accompanying formats should lead investigators and medical ethical committees to reappraise the necessity of the boilerplate language present in current informed consent forms and to focus on presenting concise and relevant information for potential study participants.

Electronic patient reported outcomes (EPROS) represent value for study participants

A bigger role of the subjective experience of patients in clinical trials also has potential to add value, as this directly reflects the value patients allocate to their treatment. It also helps investigators to define what it is their patients care about. This is already done by

the reintroduction of questionnaires as a patient-relevant endpoint in clinical trials, possibly in the form of an EPRO. Daily questionnaires were largely abandoned in trial design, and rightly so, after studies showed that actual compliance for completing a paper symptom diary is as low as 11%, much lower than participants tend to report voluntarily[63]. Concurrently, questionnaire assessments using a larger interval between measurements generally suffer from recall bias[64]. However, with the introduction of electronic diaries and EPROS, higher compliance up to 94% can be reached[63]. While study participants historically received a device from investigators for electronic data capture in clinical trials, the emergence of *bring your own device* (BYOD) in EPRO design has basically made every subject's smartphone a digital diary. This has obvious advantages, such as reduced costs, reduced administrative burden for clinical site and a reduced burden for study participants[65]. PROS can clearly demonstrate priorities of patients. A 2012 study comparing rheumatoid arthritis disease activity scores reported by patients and their physicians showed significant discordance[66]. The authors demonstrated that priorities for patients were general health outcomes such as fatigue and pain, where physicians relied more on sedimentation rates and joint counts, which are endpoints that regularly feature in rheumatoid arthritis trials. Outcomes gathered via EPROS also have additional value when they are combined with objective data that is gathered concurrently. For example, it is tempting to assume that studies investigating statin therapy would have caught an effect on objectively measured physical activity in those patients complaining of muscle aches via a daily questionnaire. We expect future studies will utilize combined assessments such as these more often.

Frequent measurements may allow for precision and personalized medicine

There is now ample evidence that not all patients benefit from health care interventions in the same manner. This could be due to variability in the patient, for instance in pharmacokinetics or in presentation of the disease. The precision medicine approach requires individualized treatment[67,68]. In practice, precision medicine has mainly been utilized in oncology research and pharmacogenomics. Other fields could also benefit from a more personalized approach, but, for most, individualized treatment is only possible if there are individual treatment outcomes. However, the probability of a major health event occurring cannot be used for individual treatment decisions in many diseases, considering the fact that the event will not occur for the majority of patients. When such endpoints are the sole basis of the evidence, it is impossible to individualize treatments. Precision medicine therefore requires parametric endpoints that can in some manner be related to treatment success or failure.

The innovations described in this chapter have in common that they allow investigators to increase the frequency of measurements without significantly increasing the burden for participants. Wearables and smartphones allow for continuous monitoring, which basically leads to investigators obtaining a high-resolution overview of the variability and day-to-day activities of patients. This will make it possible to create a profile of interindividual differences between patients, which could be an important factor for the introduction of precision medicine. With a such a detailed individual profile, a deviation from the normal individual pattern of several sensor, device and EPRO measurements may lead to detection of treatment benefits and early detection of health care problems and events. This may not only improve clinical trials, but also health care in general.

Validation of new endpoints

It is tempting to incorporate innovative new endpoints in clinical trials. However, all new value-based endpoints, including but not limited to endpoints generated by devices, should undergo proper validation procedures. First, there is the analytical validation: an analyte or device sensor is compared against the technical gold standards to investigate whether the endpoint actually measures what is claimed. This should always be included in the validation procedure and, considering the state of our current technological capabilities in wearable technology, may reveal discrepancies compared to gold standards. Then, it will be up to the investigator to interpret whether the discrepancies are disqualifying or not. As mentioned, the advantages of continuous monitoring may very well outweigh the disadvantages of small measurements errors.

The next step of validation should focus on demonstrating an association between the device output and disease-activity, disease severity, or another area of interest. This is a process of clinical validation comparable to the fit-for-purpose validation in laboratory biomarker research[69]. Pilot studies should be conducted where the relationship of the new endpoint with existing measures of disease activity. When pilot studies conducted during this process yield positive results, e.g. a correlation between new endpoint and old endpoint, the new endpoints could be incorporated in existing trials for comparison against clinical gold standards in larger groups.

When the novel endpoint does not correspond perfectly with existing disease activity scores, one may assume this is because of underwhelming performance of the novel endpoint. However, one should be open to the possibility that new endpoints may capture the disease severity better or in another way than the existing scores and scales estimating disease severity. They are often based on subjective clinical observations or questionnaires, subject to interrater variability[70]. Therefore, it is often difficult to speak of a clinical 'gold' standard. Investigators should critically review the underlying hypotheses behind the novel endpoint and compare this against the evidence of the reliability and practical usability of the old endpoints in clinical practice. This may lead to the development of new clinical gold standards.

Hype, hope and the drawbacks of continuous innovation

While innovative new wearables and devices can revolutionize clinical trial design and health care in general, innovation should always focus on questions arising from the field itself. In the last 10 years, many wearables and small devices have been announced that never reached clinical research or practice. While developers may have hoped to become one of these new gold standards, they died a quiet death not long after their unveiling or are being kept in development indefinitely (*table 3*)[71–78]. This may be because some are solutions without a problem or because developers claim exciting unproven health benefits in order to woo potential investors. Other devices have actually been released with exciting health claims but without proper validation, like the Owlet baby monitor. The manufacturer of this smart sock claimed that it was able to alert parents if their infant stops breathing. However, no independent clinical research had been performed at that point and a subsequent validation study found worrying accuracy[79,80].

A more recent example is the Urgonight, which is a wearable headband promising to improve sleep by using a method based on 'EEG neurofeedback', and already a winner of at least one innovation award[81]. However, a recent randomized controlled clinical trial indicated the employed method holds little promise in improvement of sleep quality[82]. Furthermore, home-based EEG measurements have generally been extremely difficult to carry out reliably, making us doubt the claims of efficacy even further. Devices such as these should make all researchers primed to maintain a critical approach towards the claims made by developers and the data captured by new, exciting devices.

Table 3. Examples of devices that did not (yet) live up to the hype (71–78)

Name	Device	Health claim	Last appearance
ADAMM	Chest- or back-worn device capable of cough counting and respiration, wheeze and heart rate monitoring.	Predicts asthma attack before onset of symptoms.	In development since 2015
Ampstrip	3.5-inch-long adhesive with single-lead ECG sensor, accelerometer, and a temperature sensor.	Measures several fitness related parameters and provides info to smartphone using a.	Cancelled in 2015
Bodytrak	In-ear device with several sensors measuring parameters such as body temperature, heart rate, VO2 and motion.	Measures continuously and in real-time using proprietary algorithms and machine learning libraries to provide health and wellbeing alerts via a simple and configurable user interface, in order to enable early intervention to improve outcomes and reduce injury.	In development since early 2017
iTBra™	Temperature sensor incorporated in breast patch/bra.	Identifies and categorizes abnormal circadian patterns in otherwise healthy breast tissue for early detection of breast cancer.	No news since 2015
K'Watch Glucose	Wrist-worn glucose monitor	Allows for painless and discreet continuous glucose monitoring using microneedle cassette in the watch.	In development since early 2017
Motio HWTM	Wrist-worn device with variety of sensors.	Diagnoses and monitors sleep apnea	No news since 2017
S-Skin	A microneedle patch and separate LED device.	Penetrates the skin to deliver effective ingredients and enhance absorption. Measures the hydration, redness, and melanin of the skin to provide customized skincare using LED light.	No news since 2016
Zensorium Tinké	Device for fingertip measurements.	Measures stress, tracks activity, monitors heart rate and provides advanced sleep measurement of to deliver a holistic assessment of your health and reduce stress.	For sale, health claims not validated

Home based sampling of alternative matrices

Moving trial assessments towards the subjects' homes may hamper the ability for frequent blood sampling, which is common practice in clinical trials. However, this also leads to opportunities of frequent, long-term sampling of alternative biological samples for biomarker analysis. Since most patients are not proficient in obtaining venous samples of themselves, investigators will have to use an alternative sampling matrix. For example, saliva and dried blood spot assays may be suitable for biomarker research and for pharmacokinetic analysis when validated correctly[83–85].

Integration of value-based endpoints in the clinical trial of the future

How would the clinical trial of the future look? An imaginary example from the field of asthma. Let's assume a new compound with a novel mechanism of action. Traditionally, such a compound would be given to healthy subjects without asthma to study pharmacokinetics and general tolerability, then in a dose ranging study to subjects with mild asthma and after several years to a larger or more serious group of patients. Patients will be enrolled after an on-site information visit and separate screening visit, and will thereafter be studied at two weekly intervals but eventually even less, which results in a low resolution of measurements. The measurements will be performed by trained study nurses and drug administration will only be performed by a trained physician. Visits will include a general questionnaire on (side) effects, pulmonary function tests, multiple ECGS and blood levels of various exploratory biomarkers. A large trial with will be conducted with a sample size based on hard endpoints such as pulmonary exacerbations. The size of the trial will result in many participating research centers. Since asthma control is quite good in Western Europe, where exacerbations are increasingly rare compared to low- and middle-income countries, the trials will also take place in countries where health care and research practice is less advanced. Trial monitors will travel often to all centers to try to control this widely divergent group of investigators and cultures. The process will span years and cost close to a billion dollars, if successful.

In a home-based version of these trials the patients will be informed of the study by their physician or social media. They will download an app on their smartphone with detailed but concise information about entry criteria and study assessments. Patients will have the opportunity to call or chat with a study physician before deciding to give electronic consent to enroll in the study. The subject will then come to a clinical research unit once for a screening and baseline visit and for in-person training regarding study assessments, administration of medication and a limited amount of tests to be performed by study staff. Subjects will then leave for their homes equipped with a smartwatch that measures movement, heart rate and sleep quality and ECG. They will also have an app on their smartphone that allows regular data collection and instruct the patients to measure blood pressure, weight and pulmonary function with a small wireless device. The endpoints of the study will not only focus on pulmonary function or exacerbations, but also on measures that indicate value for patients, such as the (daytime or nocturnal) duration a subject spends coughing, the exertional capacity and the perceived dyspnea a subject has. They will administer the medication at home and will use the camera on their phone to provide evidence of adherence. Because the relation between salivary concentrations and blood concentrations has been studied earlier, patients will also collect saliva at preset moments or generate a dried blood spot for pharmacokinetic, biomarker and safety analysis, which they will then store in their freezer at home. The data will come in in an automatically monitored platform in the cloud and are therefore monitored on the fly rather than at preset moments. Any deviations will result in a notification for the study team and will result in a call and subsequent visit of a help-team. Patients communicate through their phone app with the study physician for any advice and can record events using their phone app. Finally, the devices and stored samples will be collected by a courier at the end of the study period.

This represents our vision regarding the value-based clinical trials of the future. This trial is far from science fiction. We believe all the techniques described could be incorporated in trials today. When executed, this trial is extremely data-rich, and participants visit a research site only once or twice. The value-based endpoints will measure outcomes patients, and ultimately investigators, care about. Outcomes will be measured with high frequency, at home, with the use of innovative technologies and incorporation of EPROS. The coordination of the value-based trial will take place at limited amount of research centers and will be designed with clear predetermined questions in mind. The improved efficiency will lower trial costs, while improving health care by reporting the value of health care interventions, both positive and negative. This is opposed to the event-based trial, which generally focuses on value for a small proportion of patients, although a combination of value-based endpoints and hard endpoints in clinical trials may emerge as the best of both worlds.

An alternative view of future trials could focus on new, innovative general trial design. In our opinion, there is little room for opportunity there. After all, the various forms of bias present in the first clinical trials, before blinding and randomization were commonplace, are still able to negatively impact the reliability of outcomes. However, the value-based clinical trials of the future may generate high quality evidence in the form of real-life data. This could lead to value-based, precision endpoints capturing the effect of a health care intervention so extremely well, that the amount of bias removed by the addition of blinding would be considered negligible. After all, even the father of the clinical trial, Austin Bradford Hill, deemed the use of blinding unnecessary in the streptomycin trial for pulmonary tuberculosis. His endpoint was deemed strong enough to overcome the risk of bias. This may eventually be possible with the use of value-based endpoints in the clinical trial of the future as well.

REFERENCES

1 Hay M, Thomas DW, Craighead JL, Economides C, Rosenthal J. Clinical development success rates for investigational drugs. Nat Biotechnol. 2014;32(1):40–51.
2 Wong CH, Siah KW, Lo AW. Estimation of clinical trial success rates and related parameters. Biostatistics. 2018;1–14.
3 The Bible–Book of Daniel, Chapter 1, Verse 1–17.
4 Donaldson I. Ambroise Paré's accounts of new methods for treating gunshot wounds and burns. J R Soc Med. 2015;108(11):457–61.
5 Lind J. A treatise of the scurvy. In three parts. Containing an inquiry into the nature, causes and cure, of that disease. Together with a critical and chronological view of what has been published on the subject. Edinburgh A Kincaid A Donaldson, 1753.
6 MRC Patulin Clinical Trials Committee. Clinical trials of patulin in the common cold. Lancet 1944;ii:373–5.
7 Raistrick H, Scadding JG, Tytler WH, Wilson GS, Hart PDA. Streptomycin treatment of pulmonary tuberculosis. Br Med J. 1948;Oct 30:770–82.
8 Gehan EA, Freireich EJ. Non-randomized controls in cancer clinical trials. N Engl J Med. 1974;290(4):198–203.
9 Doll R. Sir Austin Bradford Hill and the progress of medical science. BMJ. 2009;305(6868):1521–6.
10 ClinicalTrials.gov, U.S. National Library of Medicine [cited 2019 Mar 12]. Available from: https://clinicaltrials.gov/
11 Martin L, Hutchens M, Hawkins C, Radnov A. How much do clinical trials cost? Nat Rev Drug Discov. 2017;16(6):381–2.
12 Pallmann P, Bedding AW, Jaki T, Villar SS, Weir CJ, Wheeler GM, *et al.* Adaptive designs in clinical trials: why use them, and how to run and report them. BMC Med. 2018;16(1):1–15.
13 Thorlund K, Haggstrom J, Park JJ, Mills EJ. Key design considerations for adaptive clinical trials: A primer for clinicians. BMJ. 2018;360(fig 1).
14 Chow S. Adaptive Clinical Trial Design. Annu Rev Med 2014 65405–15.
15 Simon R. Critical Review of Umbrella, Basket, and Platform Designs for Oncology Clinical Trials. Clin Pharmacol Ther. 2017;102(6):934–41.
16 Montgomery AA, Peters TJ, Little P. Design, analysis and presentation of factorial randomised controlled trials. BMC Med Res Methodol. 2003;3:1–5.
17 Li T, Yu T, Hawkins BS, Dickersin K. Design, analysis, and reporting of crossover trials for inclusion in a meta-analysis. PLOS One. 2015;10(8):1–12.
18 Baer L, Ivanova A. When should the sequential parallel comparison design be used in clinical trials? Clin Investig (Lond). 2013;3(9):823–33.
19 Cohen AF, van Gerven JMA, Burggraaf J, Moerland M, Groeneveld GJ. The Use of Biomarkers in Human Pharmacology (Phase I) Studies. Annu Rev Pharmacol Toxicol. 2014;55(1):55–74.
20 Feltner DE, Morgan P, Drummond KS, Wegner CD, Arrowsmith J, Street SDA, *et al.* Can the flow of medicines be improved? Fundamental pharmacokinetic and pharmacological principles toward improving Phase II survival. Drug Discov Today [Internet]. 2011;17(9–10):419–24. Available from: http://dx.doi.org/10.1016/j.drudis.2011.12.020
21 Perry S, Thamer M. Medical Innovation and the Critical Role of Health Technology Assessment. JAMA [Internet]. 1999 Nov 17;282(19):1869–72. Available from: https://doi.org/10.1001/JAMA.282.19.1869
22 Jönsson B. Bringing in health technology assessment and cost-effectiveness considerations at an early stage of drug development. Mol Oncol [Internet]. 2015;9(5):1025–33. Available from: http://www.embase.com/search/results?subaction=viewrecord&from=export&id=L604093588%0Ahttp://dx.doi.org/10.1016/j.molonc.2014.10.009
23 Group A. Effects of Aspirin for Primary Prevention in Persons with Diabetes Mellitus. N Engl J Med. 2018;379(16):1529–39.
24 Warren JB. Translating the Dose Response into Risk and Benefit. Br J Clin Pharmacol [Internet]. 2019;bcp.13949. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/bcp.13949
25 Porter M. What is Value in Healthcare. NEJM. 2010;2477–81.
26 Porter ME, Kramer MR. Strategy and Society: The Link Between Competitive Advantage and Corporate Social Responsibility. Harv Bus Rev [Internet]. 2006; Available from: http://link.springer.com/10.1007/s40134-013-0040-x
27 Porter M, Hills G, Pfitzer M, Patscheke S, Hawkins E. Measuring shared value: How to unlock value by linking social and business results. FSG.org. 2012;1–24.
28 Scott A, Liu M, Yong J. Financial Incentives to Encourage Value-Based Health Care. Med Care Res Rev. 2018;75(1):3–32.
29 Izmailova ES, Wagner JA, Perakslis ED. Wearable Devices in Clinical Trials: Hype and Hypothesis. Clin Pharmacol Ther. 2018;104(1):42–52.
30 Smilowitz NR, Mahajan AM, Roe MT, Hellkamp AS, Chiswell K, Gulati M, *et al.* Mortality of Myocardial Infarction by Sex, Age, and Obstructive Coronary Artery Disease Status in the ACTION Registry-GWTG (Acute Coronary Treatment and Intervention Outcomes Network Registry–Get with the Guidelines). Circ Cardiovasc Qual Outcomes. 2017;10(12):1–8.
31 Lipska KJ, Yao X, Herrin J, McCoy RG, Ross JS, Steinman MA, *et al.* Trends in drug utilization, glycemic control, and rates of severe hypoglycemia, 2006–2013. Diabetes Care. 2017;40(4):468–75.
32 van Diemen MPJ, Akram N, Webb A, Groeneveld GJ. Validation of a pharmacological model for mitochondrial dysfunction in healthy subjects using simvastatin: A randomized placebo-controlled proof-of-pharmacology study. Eur J Pharmacol [Internet]. 2017;815(March):290–7. Available from: http://dx.doi.org/10.1016/j.ejphar.2017.09.031
33 Scott RS, Lintott CJ WM. Simvastatin and side effects. N Z Med J 1991 Nov 27;104(924)493–5.
34 Frank M. Sacks, Marc A. Pfeffer, Lemuel A. Moye EB. Effect of Pravastatin on Coronary Events After Myocardial Infarction in Patients With Average Cholesterol Levels. N Engl J Med. 1996;335(14):1001–9.
35 Parker BA, Capizzi JA, Grimaldi AS, Clarkson PM, Cole SM, Keadle J, *et al.* Effect of statins on skeletal muscle function. Circulation. 2013;127(1):96–103.
36 Noyes AM, Thompson PD. The effects of statins on exercise and physical activity. J Clin Lipidol. 2017;11(5):1134–44.
37 Fried EI. The 52 symptoms of major depression: Lack of content overlap among seven common depression scales. J Affect Disord [Internet]. 2017;208(September 2016):191–7. Available from: http://dx.doi.org/10.1016/j.jad.2016.10.019
38 Zung WWK. A Self-Rating Depression Scale. Arch Gen Psychiatry 1965;12(1)63–70 doi101001/archpsyc1965012031006500 8.
39 Hashim NW, Wilkes M, Salomon R, Meggs J, France DJ. Evaluation of Voice Acoustics as Predictors of Clinical Depression Scores. J Voice [Internet]. 2017;31(2):256.e1–256.e6. Available from: http://dx.doi.org/10.1016/j.jvoice.2016.06.006
40 Mohr D, Zhang M, Schueller SM. Personal Sensing: Understanding Mental Health Using Ubiquitous Sensors and Machine Learning. Annu Rev Clin Psychol. 2017;
41 Joseph PD, Craig JC, Caldwell PHY. Clinical trials in children. Br J Clin Pharmacol. 2015;79(3):357–69.
42 Chaput J, Gray CE, Poitras VJ, Carson V, Gruber R, Olds T, *et al.* Sleep and Health Indicators in School-Aged Children and Youth 1. 2016;282(June).
43 LeBlanc J and. Systematic review of the health benefits of physical activity and fitness in school-aged children and youth. Int J Behav Nutr Phys Act. 2010;7:40.
44 Bush A. Montelukast in paediatric asthma: Where we are now and what still needs to be done? Paediatr Respir Rev [Internet]. 2015;16(2):97–100. Available from: http://dx.doi.org/10.1016/j.prrv.2014.10.007
45 Broughton S, Hussein HR, Bossley CJ, Ruiz G, Gupta A, Brathwaite N. A meta-analysis of montelukast for recurrent wheeze in preschool children. Eur J Pediatr. 2017;176(7):963–9.
46 Kamišalić A, Fister I, Turkanović M, Karakatič S. Sensors and functionalities of non-invasive wrist-wearable devices: A review. Sensors (Switzerland). 2018;18(6).
47 Lu TC, Fu C-M, Ma M, Fang CC, Turner AM. Healthcare Applications of Smart Watches. Appl Clin Inform [Internet]. 2016;7(3):850–69. Available from: http://www.schattauer.de/index.php?id=1214&doi=10.4338/ACI-2016-03-R-0042
48 Colbert E, Zhou W, Dunn J, Rego S, Schüssler-Fiorenza Rose SM, McLaughlin T, *et al.* Digital Health: Tracking Physiomes and Activity Using Wearable Biosensors Reveals Useful Health-Related Information. PLOS Biol. 2017;15(1):e2001402.
49 Dunn J, Runge R, Snyder M. Wearables and the medical revolution. Per Med. 2018;15(5):429–48.
50 Pubmed-NCBI-Search for "Wearable" [Internet]. [cited 2019 Mar 13]. Available from: https://www.ncbi.nlm.nih.gov/pubmed/?term=wearable
51 Perry B, Herrington W, Goldsack JC *et al.* Use of Mobile Devices to Measure Outcomes in Clinical Research, 2010–2016: A Systematic Literature Review. Digit Biomark 2018;2(1)11–30 Publ 2018 Jan 31 [Internet]. Available from: doi:10.1159/000486347
52 Casiglia E, Kawecka-Jaszcz K, Dolan E, Malyutina S, Li Y, Ohkubo T, *et al.* The Cardiovascular Risk of White-Coat Hypertension. J Am Coll Cardiol. 2016;68(19):2033–43.
53 Wang R, Blackburn G, Desai M, Phelan D, Gillinov L, Houghtaling P, *et al.* Accuracy of wrist-worn heart rate monitors. JAMA Cardiol. 2017;2(1):104–6.
54 Macinnes M, Martin N, Fulton H, McLeod KA. Comparison of a smartphone-based ECG recording system with a standard cardiac event monitor in the investigation of palpitations in children. Arch Dis Child. 2019;104(1):43–7.
55 Ramos Hernández C, Núñez Fernández M, Pallares Sanmartín A, Mouronte Roibas C, Cerdeira Domínguez L, Botana Rial MI, *et al.* Validation of the portable Air-Smart Spirometer. PLOS One [Internet]. 2018;13(2):e0192789. Available from: https://dx.plos.org/10.1371/journal.pone.0192789
56 Pion-Massicotte J, Godbout R, Savard P, Roy JF. Development and validation of an algorithm for the study of sleep using a biometric shirt in young healthy adults. J Sleep Res. 2018;(December 2017).
57 Fokkert MJ, Van Dijk PR, Edens MA, Abbes S, De Jong D, Slingerland RJ, *et al.* Performance of the freestyle libre flash glucose monitoring system in patients with type 1 and 2 diabetes mellitus. BMJ Open Diabetes Res Care. 2017;5(1):1–8.
58 Story A, Aldridge RW, Smith CM, Garber E, Hall J, Ferenando G, *et al.* Smartphone-enabled video-observed versus directly observed treatment for tuberculosis: a multicentre, analyst-blinded, randomised, controlled superiority trial. Lancet [Internet]. 2019;393(10177):1216–24. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0140673618329933
59 Pratap A, Doerr M, Suver C, Wilbanks J, Bot BM, Klein A, *et al.* The mPower study, Parkinson disease mobile data collected using ResearchKit. Sci Data. 2016;3:160011.
60 Bakos Y, Marotta-Wurgler F, Trossen DR. Does Anyone Read the Fine Print? Testing a Law and Economics Approach to Standard Form Contracts. Work Pap 09–04, NET Institute, Revis Aug 2009.
61 Wilbanks J. Design issues in e-consent. J Law, Med Ethics. 2018;46(1):110–8.
62 Manta CJ, Ortiz J, Moulton BW, Sonnad SS. From the Patient Perspective, Consent Forms Fall Short of Providing Information to Guide Decision Making. J Patient Saf. 2016;00(00):1–6.

63 Stone AA, Shiffman S, Schwartz JE, Broderick JE, Hufford MR. Information in practice Patient non-compliance with paper diaries. Br J Med. 2002;324(73):1193-4.
64 Coughlin SS. Recall bias in epidemiologic studies. J Clin Epidemiol. 1990;43(1):87-91.
65 Coons SJ, Eremenco S, Lundy JJ, O'Donohoe P, O'Gorman H, Malizia W. Capturing Patient-Reported Outcome (PRO) Data Electronically: The Past, Present, and Promise of ePRO Measurement in Clinical Trials. Patient [Internet]. 2015;8(4):301-9. Available from: http://dx.doi.org/10.1007/s40271-014-0090-z
66 Khan NA, Spencer HJ, Abda E, Aggarwal A, Alten R, Ancuta C, *et al.* Determinants of discordance in patients' and physicians' rating of rheumatoid arthritis disease activity. Arthritis Care Res. 2012;64(2):206-14.
67 Rebhan M, Murphy SA, Röst H, Spang R, Schuppert A, Fröhlich H, *et al.* From hype to reality: data science enabling personalized medicine. BMC Med. 2018;16(1):1-15.
68 Bardakjian T, Gonzalez-Alegre P. Towards precision medicine. Handb Clin Neurol. 2018;147:93-102.
69 Cummings J, Ward TH, Dive C. Fit-for-purpose biomarker method validation in anticancer drug development. Drug Discov Today [Internet]. 2010;15(19-20):816-25. Available from: http://dx.doi.org/10.1016/j.drudis.2010.07.006
70 Tuijn S, Janssens F, Robben P, Van Den Bergh H. Reducing interrater variability and improving health care: A meta-analytical review. J Eval Clin Pract. 2012;18(4):887-95.
71 K'Watch Glucose [Internet]. [cited 2019 Apr 8]. Available from: https://www.pkvitality.com/ktrack-glucose/
72 Cyrcadia iTBraTM [Internet]. [cited 2019 Apr 8]. Available from: http://cyrcadiahealth.com/
73 Automated Device for Asthma Monitoring and Management (ADAMM) [Internet]. [cited 2019 Apr 8]. Available from: http://healthcareoriginals.com/
74 Lee J, Bedra M, Finkelstein J. A critical review of consumer health devices for stress self-management. Stud Health Technol Inform. 2014;202(April):221-4.
75 Bodytrak-Smart Biometric Sensor Technology [Internet]. [cited 2019 Apr 8]. Available from: http://www.bodytrak.co/
76 Samsung S Skin Analyzes and Improves Your Skin [Internet]. [cited 2019 Apr 8]. Available from: https://www.medgadget.com/2017/01/samsung-s-skin-analyzes-improves-skin.html
77 AmpStrip cancelled: Fitness plans end as Indiegogo refunds begin [Internet]. [cited 2019 Apr 8]. Available from: https://www.wareable.com/fitness-trackers/ampstrip-cancelled-fitness-plans-end-as-indiegogo-refunds-begin-1816
78 Motio HW by Neogia and Kyomed [Internet]. [cited 2019 Apr 8]. Available from: https://neogia.xyz/press/
79 Bonafide CP, Localio AR, Ferro DF, Orenstein EW, Jamison DT, Lavanchy C, *et al.* Accuracy of Pulse Oximetry-Based Home Baby Monitors. JAMA [Internet]. 2018 Aug 21;320(7):717-9. Available from: https://dx.doi.org/10.1001/JAMA.2018.9018
80 Bonafide CP, Jamison DT, Foglia EE. The emerging market of smartphone-integrated infant physiologicmonitors. JAMA - J Am Med Assoc. 2017;317(4):353-4.
81 UrgoTech. UrgoNight Brain Training for Sleep [Internet]. Available from: https://urgonight.com/
82 Wislowska M, Gnjezda M-T, Griessenberger H, Hoedlmoser K, Heib DPJ, Schabus M. Better than sham? A double-blind placebo-controlled neurofeedback study in primary insomnia. Brain. 2017;140(4):1041-52.
83 Rittau AM, McLachlan AJ. Investigating paracetamol pharmacokinetics using venous and capillary blood and saliva sampling. J Pharm Pharmacol. 2012;64(5):705-11.
84 Bista SR, Haywood A, Norris R, Good P, Tapuni A, Lobb M, *et al.* Saliva versus Plasma for Pharmacokinetic and Pharmacodynamic Studies of Fentanyl in Patients with Cancer. Clin Ther [Internet]. 2015;37(11):2468-75. Available from: http://dx.doi.org/10.1016/j.clinthera.2015.11.003
85 Jager NGL, Rosing H, Schellens JHM, Beijnen JH. Procedures and practices for the validation of bioanalytical methods using dried blood spots: A review. Bioanalysis. 2014;6(18):2481-514.

CHAPTER 2

Development of novel, value-based, digital endpoints for clinical trials - A structured approach towards fit-for-purpose validation

Pharmacol Rev. 2020 Oct;72(4):899-909. doi:10.1124/pr.120.000028. pmid:32958524

MD Kruizinga,[1,2,3] FE Stuurman,[1,3] V Exadaktylos,[1] RJ Doll,[1] DT Stephenson,[4] GJ Groeneveld,[1,3] GJA Driessen,[2] AF Cohen[1,3]

1 Centre for Human Drug Research, Leiden, the Netherlands
2 Juliana Children's Hospital, Haga Teaching Hospital, The Hague, the Netherlands
3 Leiden University Medical Center, Leiden, the Netherlands
4 Critical Path for Parkinson's Consortium, Critical Path Institute, Tucson, Arizona, USA

Abstract

Novel digital endpoints gathered via wearables, small devices or algorithms hold great promise for clinical trials. However, implementation has been slow due to a lack of guidelines regarding the validation process of these new measurements. In this paper, we propose a pragmatic approach towards selection and fit-for-purpose validation of digital endpoints. Measurements should be value-based, meaning the measurements should directly measure or be associated with meaningful outcomes for patients. Devices should be assessed regarding technological validity. Most importantly, a rigorous clinical validation process should appraise the tolerability, difference between patients and controls, repeatability, detection of clinical events and correlation with traditional endpoints. When technically and clinically fit-for-purpose, case-building in interventional clinical trials starts to generate evidence regarding the response to new or existing health care interventions. This process may lead to the digital endpoint replacing traditional endpoints such as clinical rating scales or questionnaires in clinical trials. We recommend initiating more data sharing collaborations to prevent unnecessary duplication of research and integration of value-based measurements in clinical care to enhance acceptance by health care professionals. Finally, we invite researchers and regulators to adopt this approach in order to ensure a timely implementation of digital measurements and value-based thinking in clinical trial design and health care.

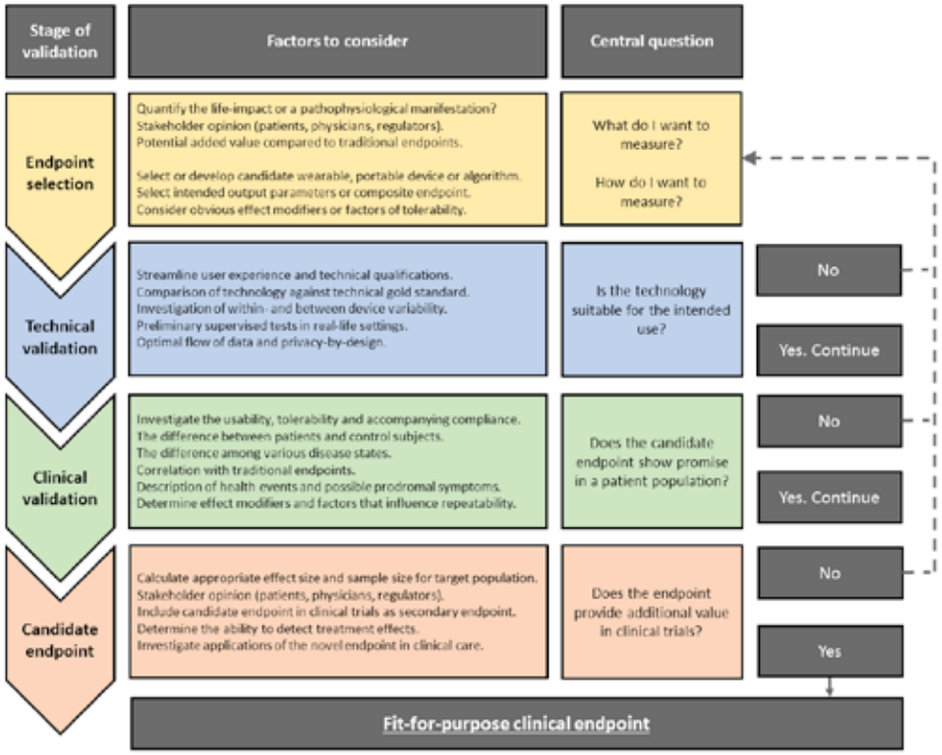
Introduction

Traditional clinical endpoints, such as mortality or clinically validated rating scales, have limitations despite their accepted status as the gold standard in clinical trial design. The worldwide improvement in standards of care means that 'hard' endpoints such as mortality are increasingly rare and necessitate oversized and overly expensive clinical trials[1]. Traditional measurements such as the 6-minute walk test or a single pulmonary function test capture no more than a snapshot of the burden of disease, obtained in clinics rather than the real world[2], while they are only loosely related to health-related quality of life[3]. Health care is moving away from these traditional metrics with the implementation of value-based health care[4]. However, the implementation of a value-based approach in clinical trials has lagged[1]. The use of digital and wearable technology can help in this endeavor, as it could ensure the burden of disease is measured in a more realistic manner: objectively, more frequently, at home, and on an individual level[5].

Despite this potential, the exact value and measurement properties of many digital measurements in the context of monitoring disease is unclear[6]. Full integration as primary or secondary endpoint in clinical trials would imply the inferential value of these measurements is sufficient to change clinical care or lead to the registration of new medicines, which, bar some exceptions[7], is not the case at this moment. Therefore, novel value-based and digital endpoints must be validated before they can be accepted by clinicians or regulators[8]. While a framework for the qualification process of novel endpoints has been proposed, digital endpoints may require a more focused and pragmatic approach[9,10]. Validation steps for digital endpoints must include technical validation, which focuses on the properties of the measurements of devices or software, and clinical validation, focused on the value of the measurements when used as endpoint for patients. While the FDA and EMA both recognize the need for validation and have released guidance[11-13], there is a lack of clarity regarding the exact criteria a digital endpoint should fulfill. Therefore, the question remains when a novel measurement is fit-for-purpose as an endpoint in clinical trials.

In this paper, we propose a pragmatic stepwise approach towards technical and clinical validation of digital endpoints (*Figure 1*), comparable to the fit-for-purpose validation employed for traditional biomarkers[14,15]. We relate each step to example cases in the fields of asthma, cystic fibrosis, pediatrics and orthopedics.

Figure 1. Structured approach to develop a candidate endpoint into a fit-for-purpose endpoint



Potential advantages of value-based digital endpoints

The inherent characteristics of digital endpoints can add value to clinical trials in numerous ways[1]. They allow measurements to be conducted completely at home, increasing participation rates and enabling trials to be conducted in vulnerable populations with chronic diseases such as the elderly, psychiatric patients, and children. These patient groups have traditionally been neglected in clinical research due to a lack of mobility, additional ethical barriers, and low recruitment rates. An added advantage is the ability to measure effects of an intervention in the natural environment of patients, resulting in increased ecological or "real world" validity. The objective nature of measurements can lead to a higher sensitivity and objectivity compared to clinical rating scales.

Wearable technology also offers high frequency and situation-relevant measurements, moving away from the artificially contrived intervals used in clinical trials. The move towards the home is in line with a trend in health care, which is also increasingly delivered outside hospitals. The lack of direct supervision on all assessments generates a potential problem in clinical trials, where complete control of participants via standardized measurements and environments is the standard. However, the inclusion of these measurements to clinic-based trials and data collection adds a new dimension of real-world data to traditional trial designs. If novel endpoints prove to be of near-identical of even superior value, they may eventually lead to reduced visits to the clinic and improved efficiency.

Selection of candidate digital endpoints

Choosing the right digital biomarker for evaluation in a specific clinical condition is challenging. There is an increasing amount of potential digital biomarkers and a large array of start-up companies vying for investors and patient users. In this paper, we assume that devices used for the monitoring, diagnosis or prognosis of disease adhere to the medical device regulations, which is the responsibility of the manufacturer[16]. The regulations contain subtle differences between countries and stratify devices in several classes based on the context and risk associated with their use[17,18]. While not a device per se, health-related software is regarded as a medical device and must comply with applicable regulations.

Table 1 lists several candidate digital biomarkers, of which many have already been investigated in early feasibility studies[19] or industry-sponsored clinical trials[20]. A good candidate endpoint should be accurate, reliable and value-based and therefore directly measure meaningful outcomes for the individual subject (e.g. sleep, activity, pain, ability to move, gait) or be associated with important clinical outcomes (e.g. occurrence of mortality, morbidity, complications). Furthermore, the assessment or device must be usable, tolerable and suitable for the intended users. Conceptually, there must be clear benefit of using the digital measurement over existing methods. The candidate endpoint should also have a plausible relationship with the studied disease or general quality of life. These factors together imply that a close collaboration with patients and patient advocacy groups is vital during the selection and validation process.

Table 1- Potential devices and candidate digital endpoints for use in clinical trials outside of clinical units

Device / sensor	Candidate endpoint	Patient domain
Accelerometer	Physical activity Steps Gait patterns Tremor analysis	Mobility Symptom severity
Blood pressure meter	Blood pressure	Cardiovascular health
Camera	Dermatological assessments Treatment adherence	Dermatological health Compliance
Dynamometer	Muscle strength	Musculoskeletal health
ECG	Event detection	Cardiovascular health
Glucose monitor	Glucose	Diabetic control
Oximeter	Oxygen saturation	Pulmonary Healthy
GPS	GPS mobility Location type	Mobility Social behavior
Light sensor	Light intensity	Environment
Microphone	Event detection Voice analysis	Clinical events Mood
PPG	Heart rate	Cardiovascular health
Smartphone	App use Phone use (calls, sms) Patient reported outcomes	Social behavior Symptom severity
Spirometer	Pulmonary function	Pulmonary health
Thermometer	Temperature	Infection control Thermoregulation
Touch screen	Response time Speed of typing Custom tests	Dexterity Coordination Various

Abbreviations: PPG: photoplethysmography; ECG: electrocardiogram; GPS: Global positioning system

If the goal, at this point, is to obtain regulatory endorsement of the novel digital endpoint, regulators advise early engagement to identify and define the 'Concept of Interest' that underpins the digital endpoint from a regulatory point of view. This engagement also serves to identify appropriate regulatory interaction channels going forward and to discuss how a clinically meaningful change can be defined and investigated[21]. Even when a device to measure digital biomarkers has been selected, choosing the right way to display and analyze a measurement is difficult. Physical activity currently has at least four separate units: step count, duration of moderate-to-vigorous physical activity, accelerometer counts per minute and average gait speed. In the case of multiple promising and related measurements, machine learning or other artificial intelligence techniques can be used to choose and combine several metrics in an algorithm to produce a composite score comprising of several novel and traditional endpoints[22].

Technical validation

Before using the novel measurement in patients, a robust assessment of the usability, reliability and reproducibility of the technology and flow of data should be performed.

Minimal technological standards

The reliability and consistency of devices should be assessed in the form of inter- and intra-device variability. The flow of data should be automated, requiring as little manual input as possible in order to reduce data (entry) errors. The flow should be consistent and allow for the subjects' privacy by design, for example via encrypted transmission of data[23]. Furthermore, the data flow must be part of the validation and comply with the necessary FDA regulations regarding audit trails and the storage and processing of source data[11]. In the case of technological gold standards that can be used as a reference, a head-to-head comparison should be conducted in a standardized setting to determine the bias and limits of agreement, specificity and sensitivity, depending on the type of measurement.

Furthermore, an analysis should be conducted with non-patient test subjects to ensure that a novel measurement truly captures the behavior, symptom, or activity it attempts to quantify in various real-life situations. For example, a smartphone accelerometer is capable of counting steps taken per day, but some may leave their phones on a desk or in a locker, bag or jacket when going for a walk, which leads to low accelerometer counts with little information of the underlying reason. While it is not possible to simulate all situations that might occur in daily life, a supervised test of limited duration may result in the detection of easily addressable confounders. In this case, the exact clinical relevance will not yet be determined, and the test subject merely functions as a free-living data generator in a closely observed setting. In this phase of validation, it is also advised to consider the amount of training and instruction that will be necessary to ensure measurements are conducted correctly by patients.

Not all these steps are feasible in all cases. For example, when there is no obvious technical gold standard or when the measurement is a completely new concept, a head-to-head comparison is impossible. In this case, technical validation is necessarily more limited and clinical validation gains a larger role.

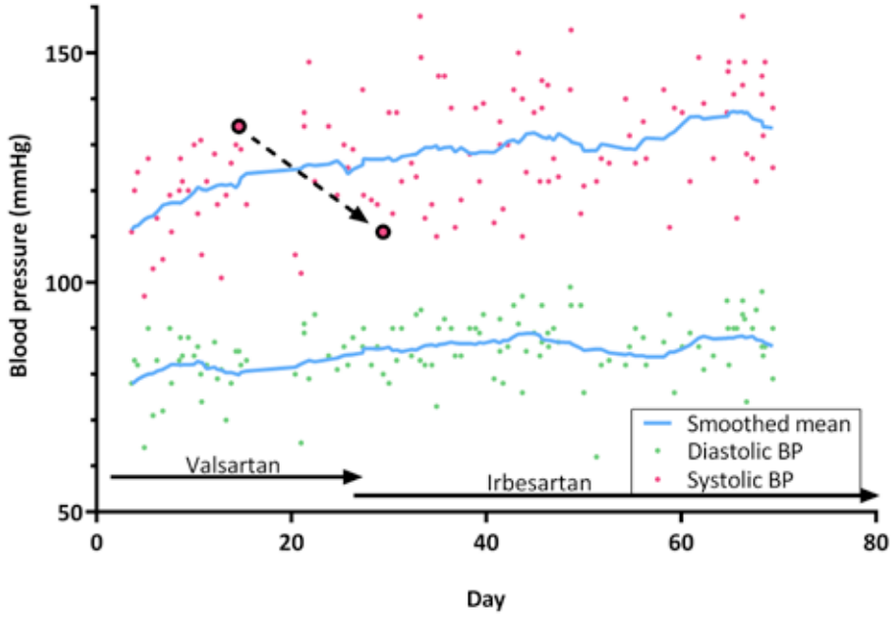
Handling of discordance

Technological validation of novel measurements is vital for the validation process, and if there is a near perfect agreement with an unchallenged technical gold standard, the validation process may end there. However, often there will be some 'technical noise' or measurement bias associated with the miniaturization process, potentially undermining the accuracy of the home-based measurements. Although medical grade devices will likely have less technical noise compared to consumer devices, they are often extremely expensive and unwieldy, which limits widespread implementation and causes a reduced compliance compared to more user-friendly (consumer) devices[24]. Furthermore, technical noise or bias is not necessarily disqualifying, and random deviations from the gold standard would not significantly alter treatment effect estimates in a clinical trial[25]. The major advantage of home-based monitoring is that the resolution of data can be very high and will be likely to outperform traditional endpoints despite suboptimal technological validity. To illustrate this further, *Figure 2* shows serial home-based blood pressure measurements of a patient with a history of hypertension and whose anti-hypertensive was switched by the pharmacy. The graph shows a gradual but clinically significant increase of the systolic blood pressure over time. However, when the patient would have been limited to a small amount of clinic based measurements (e.g. at the time points indicated by the black arrow), one could easily have concluded that blood pressure was much lower on Irbesartan compared to when that patient was on Valsartan. This example shows that, especially for measurements with a high intra-individual variability, increasing the measurement frequency leads to more valid conclusions on an individual level. Even a random measurement error will still lead to a better clinical overview compared to 6-weekly, or even more infrequent, measurements, with the caveat that the error must be significantly lower than the intra-individual variability.

Pitfalls

Investigators should also not be too focused on theoretical measurement errors inherent to home-monitoring. For example, a recent FDA advice requested that activities that may resemble 'steps', such as repetitive movements of the arm must be discriminated from actual steps. Furthermore, a plan had to be in place to make certain that devices are not used by anyone else[26]. While these requests can certainly be relevant on a technical level, there may be many reasons why this may be irrelevant in the context of a clinical trial. For example, if the conditions when this inaccuracy might occur are very rare, or if the impact of the perceived inaccuracy when measuring an association with the severity of a clinical condition is negligible. In such cases, additional requirements such as these might inhibit implementation unnecessarily. Regulators should be wary for a slippery slope: the unsupervised nature of home-based measurements introduces many uncertainties, and new uncertainties could be identified during every evaluation. During the qualification process of the 6-minute walk test, we imagine there was no request for a plan to ensure a subject does not walk slower on purpose. In practice, most of these randomly generated data quality issues will be mitigated by the application of randomization and blinding[25], which will remain the standard in pivotal clinical trials. When regulators focus on a clinical validation process that is more rigorous than for traditional rating scales, any inaccuracies and uncertainties that are found can be appraised in the perspective of the clinical condition.

Figure 2. Increased frequency can overcome variability. Individual patient who measured blood pressure in a home-setting for a period of 70 days. The patient was switched to a different anti-hypertensive drug by his pharmacy and, on average, measures slightly higher systolic blood pressures over time. When monitored during regular outpatient clinic visits, for example at the timepoints indicated by the black arrow, this effect could have been completely missed and a completely opposite conclusion could have been drawn.



Clinical validation

A rigorous clinical evaluation of the candidate endpoint must be performed to determine the potential clinical value. We propose five criteria that should be assessed during this process (*Figure 1*). Most of the characteristics can be assessed in observational studies, and some criteria may not be applicable to some measurements. This applies to measurements that are well known in clinic-based trials and are merely miniaturized or streamlined for use in a home-setting. For example, an absolute difference between patients and healthy controls is less important for a wearable device for glucose monitoring[27] than for a novel algorithm quantifying negative symptoms in schizophrenia[28]. Once the glucose monitoring device has been shown to adequately measure glucose in various situations in a home-setting, it would already be close to fit-for-purpose for clinical trials. It would be unnecessarily burdensome to also associate use of the device with traditional indicators of disease, such as a decrease in the incidence cardiovascular complications. Furthermore, multiple criteria can be investigated in a single observational study. It is therefore important to critically appraise the novel endpoint prior to initiation of the clinical validation process in order to define the most important questions and investigate these early and extensively. This question-based approach has been described for prototypical drug development and can be adapted to digital endpoint development as well[15].

Tolerability and usability for patients

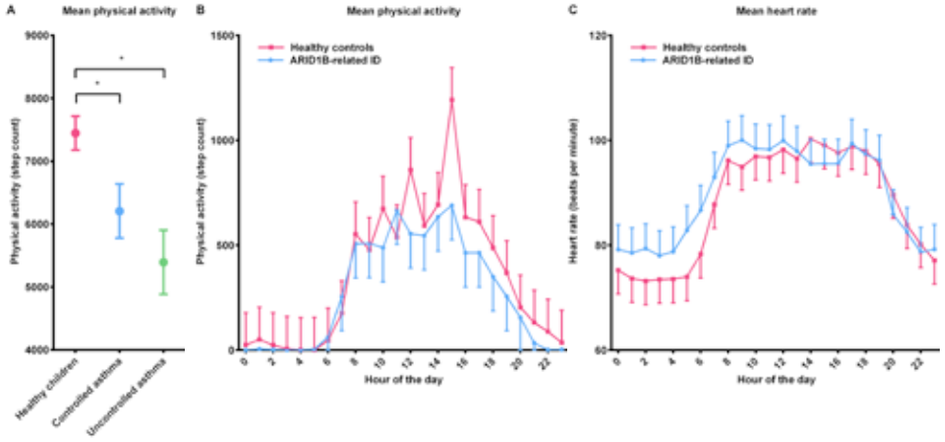
First, assessments should be tolerable for the target population. In general, that means the assessment should be minimally invasive and require as little manual input as possible. Since the use of digital endpoints means that clinical trials will be increasingly decentralized, the user experience of participants is vital to optimize adherence and retention in trials. Technology may allow for a longer follow-up period with a lower burden for subjects, but only when the study assessments are tolerable to conduct for extended periods of time. Tolerability is also important when investigating vulnerable populations, such as children, the elderly and patients that are otherwise impaired. While problems may appear trivial at first, small technical or usability issues, such as decreased smartphone battery life, may lead to low compliance, workarounds by users or even dropouts. Researchers must adapt to population-specific needs in an early stage or conclude that the included measurement is unsuitable. An example is the use of a smartwatch in children. In a recent observational study including 391 pediatric subjects, we observed children aged 6-16 were enthusiastic and demonstrated a drop-out rate of 1.4%. On the other hand, children aged 2-5 were less amused: 17% did not complete the study period (unpublished data). While the user experience should be optimized to increase compliance, low compliance rates in decentralized studies may still lead to a high number of observations that can provide valuable results. A study by Lipsmeier *et al.*, which investigated novel smartphone-based tests in 44 Parkinson's disease patients for a duration of 6 months, demonstrated that patients exhibited an average compliance of only 61%[29]. However, this resulted in a dataset consisting of 5,135 test outcomes, and appeared to allow for the detection of subtle symptomatology unlikely to result in a change in traditional symptom questionnaire scores[29]. Still, compliance with the use of digital devices has been a challenge and there is growing recognition of the need to improve alignment with patients at all stages of development[30,31].

Difference between patients and controls

An important validation criterion is the difference between patient groups and control groups, which should be assessed vigorously. The magnitude of the difference and the accompanying variability can be used to determine what improvement could be considered clinically relevant for new measurements. The data can also aid in the prospective calculation of the sample size needed to detect an appropriate treatment effect. Further development of novel measurements seems futile when there is no detectable or clinically significant difference, since an effective treatment would have to result in the patient group outperforming the control group for the particular measurement. An example from the field of pediatric asthma; *Figure 3A* shows the difference in physical activity between children with (un)controlled asthma and healthy children. The candidate endpoint appears to be able to differentiate both between healthy children and asthmatics, but also between the two disease states. Increasing the resolution of data can provide interesting insights regarding the time of day responsible for group differences (*Figure 3B* and *Figure 3C*), in this example between healthy subjects and subjects with ARID1B-related intellectual disability[32]. In the case of completely new measurements or algorithms, we believe reference values should be obtained for the target populations with special interest towards the influence of age, gender, lifestyle choices and socio-economic status, as these factors are undoubtedly influential in home-based measurement outcomes. When

appraising the difference between patients and controls and estimating relevant treatment effects, a distinction can be made between (partially) reversible and invariably progressive diseases. In the case of progressive disease, the expected gain from treatment may be a decreased speed of deterioration, and not an absolute improvement towards reference values.

Figure 3. Difference between patients and healthy controls. A: Average physical activity (95% ci) per day of healthy children, children with controlled asthma and children with uncontrolled asthma. B. Average physical activity per hour of the day of healthy children, children with controlled asthma and children with uncontrolled asthma. The estimated averages are corrected for age and sex in a mixed model analysis of variance.



* p < 0.05. B: Mean (95% CI) physical activity per hour of the day of children with at-rich interactive domain-containing protein 1B (ARID1B)-related intellectual disability (id) and healthy age-matched controls. C: Mean (95% CI) heart rate per hour of the day of children with ARID1B related intellectual disability and healthy age-matched controls.

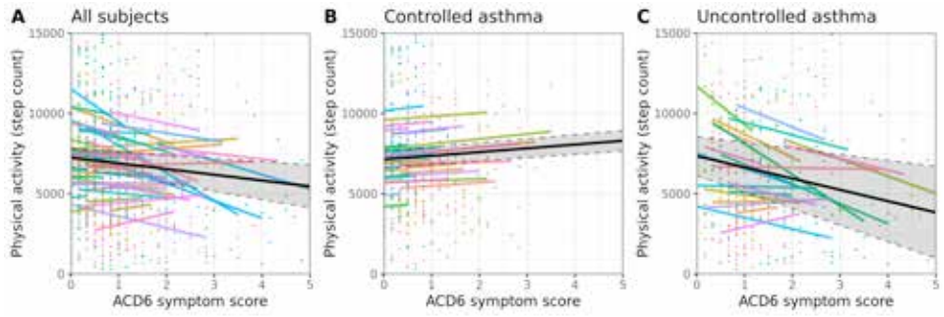
Repeatability and variability

Measurements should be stable over time in the absence of an intervention or change in disease-activity and a change of the outcome variable should be either due to improvement or exacerbation of the disease. While this may seem unrealistic for home-based and continuous measurements, it implies a need to consider the impact of real-world variability induced by factors such as season, weather and location[33], and the impact of baseline factors, such as age and social circumstances. Additionally, concomitant drug use can be a confounding factor in free-living conditions. For example, an improvement of osteoarthritis symptoms may not lead to detectable behavioral changes in the individual patient detected by digital endpoints. When the improvement leads to a reduction in the use of opiates, the improvement is certainly clinically meaningful. A better understanding of the effects of real-world variability also allows for a better quantification of treatment effects in specific individuals, one of the hallmarks of value-based health care. Furthermore, in diseases that are progressive by nature, natural history studies regarding the novel measurement can help to quantify the rate of progression and estimate relevant treatment effects in this regard[34].

Correlation with existing disease metrics

The next step is to correlate the novel endpoint with traditional endpoints, ideally the gold standard. However, perfect correlations will never be achieved considering the nature of both digital and traditional endpoints. Investigators therefore must critically appraise the data in order to determine whether a suboptimal correlation is due to limitations of the novel endpoint, limitations of the 'gold' standard, or because both quantify different aspects of the disease. An uncomplicated example would be to correlate number of steps taken per day versus the 6-minute walk test in patients with COPD[35]. Here, both endpoints are conceptually the same, but measured and expressed differently. Interpretation becomes more challenging when correlating GPS mobility with schizophrenia symptom burden[28], or asthma control diary scores with steps taken per day in pediatric asthma (*Figure 4*)[36]. In that case, there appears to be no correlation between daily symptoms and daily physical activity for subjects with controlled asthma (*Figure 4B*). This may be because for these patients, asthma symptomology is too limited to interfere with daily life. However, when looking at subjects with uncontrolled asthma (*Figure 4C*), the burden of symptoms is higher and appears to significantly impact physical activity as a result. This may lead to the conclusion that physical activity has added value for monitoring of symptoms of children with uncontrolled asthma only. Interpretation should be done carefully, while also accounting for the analyses performed during the other phases of validation. At this point, the relationship of the novel endpoint with general and disease-related quality of life can also be investigated to assess whether the novel endpoint captures a disease-state that is meaningful for patients.

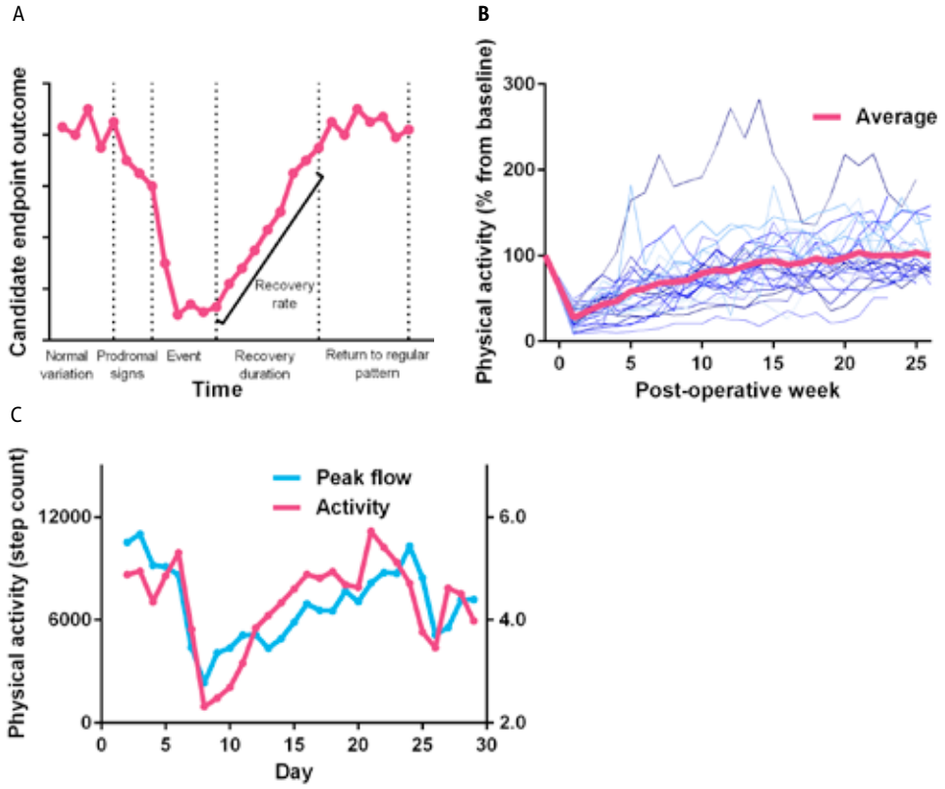
Figure 4. Modelled relationship of asthma symptoms with physical activity. Relationship of asthma control diary 6-questions (acd6) score with physical activity. A mixed model analysis of variance was developed to estimate the relationship with subject as random factor, with a random slope and random intercept. Each dot represents a single day, each color represents a subject. The black line represents the average slope and intercept (95% ci). A: analysis including all subjects with asthma. B: subgroup analysis of subjects with controlled asthma; C: subgroup analysis of subjects with uncontrolled asthma.



Responsive to change in disease state

The final step before declaring a candidate endpoint fit-for-purpose is to investigate whether the endpoint will respond to changes in burden of disease. One method is to investigate the effect of specific health events with expected negative effects, such as a pulmonary exacerbation or sickle cell crisis, or events with expected positive effects, such as a surgical interventions. Several features can be extracted from events, such as prodromal symptoms, the slope or rate of recovery, or the time needed to return to baseline values (*Figure 5A*). An example, *Figure 5B* shows physical activity patterns of elderly patients with osteoarthritis before and after knee replacement surgery. In this graph, a sharp decline in physical activity is visible postoperatively, while the recovery van be visualized over the course of several months. In the future, it may be possible to identify good responders to treatment as the subjects who recover above baseline physical activity, although there are other variables the could reflect improvement, such as the earlier mentioned concomitant medication. Furthermore, *Figure 5C* displays physical activity and pulmonary function of a single subject with cystic fibrosis undergoing a moderate pulmonary exacerbation[36]. The clinical event and recovery period is clearly identifiable and several proposed features in *Figure 5A* can be extracted. Another method is to investigate the effects of known effective treatments on the novel endpoint. However, this final step in the validation process is more easily performed in conditions with an approved disease-modifying therapy or disease with known risk of rapid exacerbations, as opposed to many rare or slowly progressive diseases with no proven treatment. In such cases, validation of novel endpoints could be performed in other, possibly similar, conditions, before turning towards the disease of primary interest.

Figure 5. Response of physical activity to health events. A. Fictional data of a digital endpoint measured during a clinical event. The features that could be extracted from such data are listed in on the X-axis. Additionally, the slope of the line of recovery represents the recovery rate. B. Non-fictional data of the physical activity change from baseline after knee surgery. Individual lines represent individual patients and the pink line represents the average. C. Data of an individual pulmonary event of a pediatric cystic fibrosis patient. The running 2-day average of physical activity (pink) and peak flow (blue) are displayed.



Case-building

The ultimate test for novel endpoints is the interventional clinical trial, which allows the detection of beneficial or deleterious effects of novel or existing treatments. Although it may be tempting to immediately include a novel measurement in trials, it is important to systematically complete the validation process in order to reliably assess the usability, potential value and, most importantly, the expected effect size necessary for clinical benefit. When a novel measurement or technology is assessed positively on all criteria, they are fit-for-purpose as clinical endpoint in future interventional clinical trials. Then, case-building starts by including the biomarker as exploratory or secondary endpoint in relevant trials and eventually, when the novel endpoint has proven superior value compared to traditional endpoints regarding study compliance, response to treatment, or distinguishing capability, as primary endpoint in the clinical trial of the future.

Discussion

Regulatory engagement

The current state of regulatory guidance should not deter investigators to include digital measurements in clinical trials. Although guidelines are lacking, the FDA and EMA have provided strong indications that it supports the use of wearable-biosensor-and other real-world data in regulatory decision making[21,37]. Every single state-of-the-art biomarker was exploratory at one time. For example, human immunodeficiency virus (HIV) viral load is the unchallenged gold standard to quantify disease-activity in 2020, but completely exploratory 25 years prior[38]. While it is not compulsory to only use qualified clinical outcome assessments in clinical trials, early engagement with regulators may streamline the process to qualify novel digital endpoints. Until more experience with digital endpoints has been obtained by both regulators and researchers, the qualification process will be unpredictable. Pioneering work in engagement with EMA has been performed in this regard with the qualification of the stride velocity 95TH centile as secondary endpoint in Duchenne muscular dystrophy[7,39]. The EMA qualification opinion describes an iterative process with multiple discussion sessions and a public consultation. Each session provided additional points needing clarification, and many of the requested clarifications feature in this manuscript. Early adoption of the proposed framework may streamline future iterative discussions with regulators.

Table 2 - Checklist to use during the planning and execution of the validation process of novel digital endpoints for clinical trials

Step	Question	Applicable (Y/N)
Endpoint and device selection	What aspect of the disease must be measured?	
	How is this aspect directly relevant to patients?	
	What is the optimal device, wearable or algorithm to measure this aspect?	
	What are the output parameters of this device?	
	How can the measurement be expressed as a single outcome measure?	
	Is there a difference expected between patients and controls?	
Technical validation	Are the technical qualifications on par with researcher requirements?	
	Can the user experience be streamlined?	
	What kind of user training would optimize uptake and retention?	
	How does the measurement compare to the technical gold standard?	
	What is the intra-device and inter-device variability?	
	What is the sensitivity and specificity?	
	Does the measurement capture the aimed behavior or symptom?	
	Is the privacy of subjects guaranteed?	
	Is the dataflow stable and compliant with regulatory requirements?	
Clinical validation	Is the device tolerable and usable for the target population?	
	What is the difference between patients and control subjects?	
	What is the difference between the several states of disease?	
	What influences the day-to-day variability within subjects?	
	Can the endpoint detect and describe clinical events or interventions?	
	Is the endpoint correlated to traditional endpoints?	
Application & Case building	What phase of clinical research is the device most suitable for?	
	Integration in ongoing or upcoming trials as exploratory endpoint?	
	Can the collected data be extrapolated to other populations?	
	Can anonymized data of control subjects be shared with other parties?	

The lack of complete control over subjects is a major disadvantage of digital endpoints. Real-world data collection in free-living conditions will invariably add a factor of uncertainty and, although difficult to quantify, this factor may remain a recurring theme during regulatory appraisal. However, there is currently no data regarding the consequences of this loss of control in terms of bias in estimated treatment effects in clinical trials. It is up to investigators to provide enough data to prove the added value of digital endpoints for clinical trials and the case-building process should start now. *Table 2* outlines the various steps presented in this paper and can be used to prepare and plan the validation process, and to select the various assessments that are applicable to the candidate endpoint.

Recommendations

There are precompetitive collaborative efforts by various stakeholders to support the developmental process of digital endpoints[40]. Examples are the Critical Path Institute's consortia and the Clinical Trials Transformation Initiative (CTTI), however, more could be done to stimulate and incentivize implementation and standardize reporting[41–44]. International adoption of a single device or algorithm is unrealistic, but there is an established difference between manufacturers in, for example, smartwatches. Furthermore, all algorithms are invariably dependent on the original dataset from which they are derived. Nevertheless, data from multiple studies with different devices may eventually be combined for analysis by regulatory agencies or in the context of meta-analysis. For those cases, head-to-head comparisons can be conducted to demonstrate equivalency or to develop conversion factors enabling the comparison of the results of different studies[45]. Regulatory Agencies have recommended the creation of normative databases for device platforms, but this has been adopted only on rare occasions[7].

Continuing collaboration by academia and industry aimed towards data sharing is crucial to avoid unnecessary duplication[46]. For example, by sharing reference values of novel measurements generated by healthy participants. Adoption of universal data sharing at the level of consent of subjects could accelerate the move forward[47] Socially aware wearable companies may also aid in this purpose via data sharing, and could be more open towards researchers regarding their proprietary algorithms and raw data for the common interest of improving health care. In the case of devices subject to firmware updates, care must be taken that changes do not impact the validity of the measurements.

Furthermore, results should be shared and discussed with the same patients and patient advocacy groups that were consulted during initial candidate endpoint selection in order to ascertain that the novel measurements truly captures aspects of the disease important to patients.

Potential for clinical care

To further stimulate acceptance by both patients and health care providers, suitable digital biomarkers with potential in clinical care should be introduced in the clinic as soon as clinically validated. The high sampling frequency of measurements that are important to patients may allow physicians to obtain a holistic overview of patients' well-being. Digital biomarkers have a wide range of possible applications in clinical care. For example, they can be used to support diagnosis of challenging patients, stratify patients in risk categories, serve as pharmacodynamic response marker, provide monitoring in addition to or in place of traditional visits to the outpatient clinic, and even aid in the prediction of health outcomes after a hospital admission[8,48]. Realizing this potentially disruptive new component in value-based health care necessitates a pro-active approach towards an improved data infrastructure in hospitals, which must be capable of processing algorithms for diagnostic and follow-up purposes[49]. Not all digital endpoints with value in clinical trials will add value in clinical care. The context of participating in a clinical trial with incentives such as access to new treatments or financial reward might be quite different from the general clinical health setting when determining the value of digital endpoints for patients. The usability and tolerability of measurements with potential in clinical care need to be assessed in this light. Ultimately, addition of a novel digital endpoints in standard care requires measurements that are minimally invasive that can reliably detect the individual response to health care interventions and finally, are cost-effective as well. These are demanding requirements which will require a long process of clinical validation beyond the steps described in this review.

Conclusion

The proposed stepwise approach towards technical and clinical validation of novel endpoints in clinical trials is pragmatic and can be applied to most types of digital data. We invite researchers and regulators to endorse and adopt this framework in order to ensure a timely implementation of digital measurements and value-based thinking in clinical trial design and health care.

REFERENCES

1 Kruizinga MD, Stuurman FE, Groeneveld GJ, Cohen AF. The Future of Clinical Trial Design : The Transition from Hard Endpoints to Value-Based Endpoints

2 Steinhubl SR, Mcgovern P, Dylan J, Topol EJ. Perspectives Digital medicine The digitised clinical trial. Lancet. 2016;390(10108):2135.

3 Carranza Rosenzweig JR, Edwards L, Lincourt W, Dorinsky P, ZuWallack RL. The relationship between health-related quality of life, lung function and daily symptoms in patients with persistent asthma. Respir Med. 2004;98(12):1157-65.

4 Porter M. What is Value in Healthcare. NEJM. 2010;2477-81.

5 Boehme P, Hansen A, Roubenoff R, Scheeren J, Herrmann M, Mondritzki T, *et al.* How soon will digital endpoints become a cornerstone for future drug development? Drug Discov Today. 2019;24(1):16-9.

6 Babrak LM, Menetski J, Rebhan M, Nisato G, Zinggeler M, Brasier N, *et al.* Traditional and Digital Biomarkers: Two Worlds Apart? Digit Biomarkers. 2019;3(2):92-102.

7 Haberkamp M, Moseley J, Athanasiou D, de Andres-Trelles F, Elferink A, Rosa MM, *et al.* European regulators' views on a wearable-derived performance measurement of ambulation for Duchenne muscular dystrophy regulatory trials. Neuromuscul Disord. 2019;29(7):514-6.

8 Coravos A, Khozin S, Mandl KD. Developing and adopting safe and effective digital biomarkers to improve patient outcomes. NPJ Digit Med. 2019;2(1):1-5.

9 Leptak C, Menetski JP, Wagner JA, Aubrecht J, Brady L, Brumfield M, *et al.* What evidence do we need for biomarker qualification? Sci Transl Med. 2017;9(417):1-5.

10 Coravos A, Doerr M, Goldsack J, Manta C, Shervey M, Woods B, *et al.* Modernizing and designing evaluation frameworks for connected sensor technologies in medicine. NPJ Digit Med. 2020;3(1):1-10.

11 U.S. Food and Drug Administration. Use of Electronic Records and Electronic Signatures in Clinical Investigations Under 21 CFR Part 11-Questions and Answers: Guidance for Industry. 2017;(June 2017). Available from: https://www.FDA.gov/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/default.htm%0Ahttps://www.FDA.gov/BiologicsBloodVaccines/GuidanceComplianceRegulatoryInformation/Guidances/default.htm

12 EMA. Qualification of novel methodologies for drug development : guidance to applicants. 2012;44(January):1-16.

13 U.S. Department of Health and Human Services Food and Drug Administration, (CBER) C for DE and R (CDER) C for BE and R. Biomarker Qualification: Evidentiary Framework Guidance for Industry and FDA Staff draft guidance. FDA. 2018;(December):1-16.

14 Cummings J, Ward TH, Dive C. Fit-for-purpose biomarker method validation in anticancer drug development. Drug Discov Today. 2010;15(19-20):816-25.

15 Cohen AF, van Gerven JMA, Burggraaf J, Moerland M, Groeneveld GJ. The Use of Biomarkers in Human Pharmacology (Phase I) Studies. Annu Rev Pharmacol Toxicol. 2014;55(1):55-74.

16 Ben-Menahem SM, Nistor-Gallo R, Macia G, von Krogh G, Goldhahn J. How the new European regulation on medical devices will affect innovation. Nat Biomed Eng. 2020;6. doi: 10.1038/s41551-020-0541-x

17 Chen YJ, Chiou CM, Huang YW, Tu PW, Lee YC, Chien CH. A Comparative Study of Medical Device Regulations:: US, Europe, Canada, and Taiwan. Ther Innov Regul Sci. 2018;52(1):62-9.

18 Gordon WJ, Landman A, Zhang H, Bates DW. Beyond validation: getting health apps into clinical practice. NPJ Digit Med. 2020;3(1). doi: 10.1038/s41746-019-0212-z

19 Bakker JP, Goldsack JC, Clarke M, Coravos A, Geoghegan C, Godfrey A, *et al.* A systematic review of feasibility studies promoting the use of mobile technologies in clinical research. NPJ Digit Med. 2019;2(1). doi: 10.1038/s41746-019-0125-x

20 Chasse R. Crowdsourced Library of Digital Endpoints in Industry-Sponsored Studies [Internet]. [cited 2020 Mar 4]. Available from: https://docs.google.com/spreadsheets/d/==/

21 Cerreta F, Ritzhaupt A, Metcalfe T, Askin S, Duarte J, Berntgen M, *et al.* Digital technologies for medicines: shaping a framework for success. Nat Rev Drug Discov. 2020;May. doi: 10.1038/d41573-020-00080-6

22 Zhan A, Mohan S, Tarolli C, Schneider RB, Adams JL, Sharma S, *et al.* Using smartphones and machine learning to quantify Parkinson disease severity the mobile Parkinson disease score. JAMA Neurol. 2018;75(7):876-80.

23 Angeletti F, Chatzigiannakis I, Vitaletti A. Towards an architecture to guarantee both data privacy and utility in the first phases of digital clinical trials. Sensors (Switzerland). 2018;18(12). doi: 10.3390/s18124175

24 Schrack JA, Cooper R, Koster A, Shiroma EJ, Murabito JM, Rejeski WJ, *et al.* Assessing Daily Physical Activity in Older Adults: Unraveling the Complexity of Monitors, Measures, and Methods. Journals Gerontol-Ser A Biol Sci Med Sci. 2016;71(8):1039-48.

25 Buyse M, Squiffiet P, Coart E, Quinaux E, Punt CJA, Saad ED. The impact of data errors on the outcome of randomized clinical trials. Clin Trials. 2017;14(5):499-506.

26 Papadopoulos E, Norman L. Request For Qualification Plan. FDA. 2018

27 Ólafsdóttir AF, Attvall S, Sandgren U, Dahlqvist S, Pivodic A, Skrtic S, *et al.* A Clinical Trial of the Accuracy and Treatment Experience of the Flash Glucose Monitor FreeStyle Libre in Adults with Type 1 Diabetes. Diabetes Technol Ther. 2017;19(3):164-72.

28 Depp CA, Bashem J, Moore RC, Holden JL, Mikhael T, Swendsen J, *et al.* GPS mobility as a digital biomarker of negative symptoms in schizophrenia : a case control study. NPJ Digit Med. (Umr 5287). doi: 10.1038/s41746-019-0182-1

29 Lipsmeier F, Taylor KI, Kilchenmann T, Wolf D, Scotland A, Schjodt-Eriksen J, *et al.* Evaluation of smartphone-based testing to generate exploratory outcome measures in a phase 1 Parkinson's disease clinical trial. Mov Disord. 2018;33(8):1287-97.

30 Pratap A, Neto EC, Snyder P, Stepnowsky C, Elhadad N, Grant D, *et al.* Indicators of retention in remote digital health studies: A cross-study evaluation of 100,000 participants. NPJ Digit Med. 2019;1-10.

31 Bot BM, Suver C, Neto EC, Kellen M, Klein A, Bare C, *et al.* The mPower study, Parkinson disease mobile data collected using ResearchKit. Sci Data. 2016;3:1-9.

32 Kruizinga MD, Zuiker RGJA, Sali E, de Kam ML, Doll RJ, Groeneveld GJ, *et al.* Finding Suitable Clinical Endpoints for a Potential Treatment of a Rare Genetic Disease: the Case of ARID1B. Neurotherapeutics. 2020 doi: 10.1007/s13311-020-00868-9

33 Chan CB, Ryan DA. Assessing the effects of weather conditions on physical activity participation using objective measures. Int J Environ Res Public Health. 2009;6(10):2639-54.

34 Jewell NP. Natural history of diseases: Statistical designs and issues. Clin Pharmacol Ther. 2016;100(4):353-61.

35 Steele BG, Holt L, Belza B, Ferris S, Lakshminaryan S, Buchner DM. Quantitating physical activity in COPD using a triaxial accelerometer. Chest. 2000;117(5):1359-67.

36 Kruizinga M, Heide N van der, Nuijsink M, Stuurman R, Cohen A, Driessen G. Activity and pulmonary function collected via a non invasive platform differentiate healthy and asthmatic children-Selected Abstracts from Pharmacology 2019.
Br J Clin Pharmacol. 2019;n/a(n/a). doi: 10.1111/bcp.14266

37 FDA. Framework for FDA's Real-World Evidence Program. FDA Framew. 2018;(December). Available from: www.FDA.gov

38 Piatak M, Saag MS, Yang LC, Clark SJ, Kappes JC, Luk KC, *et al.* High levels of HIV-1 in plasma during all stages of infection determined by competitive PCR. Science (80-). 1993;259(5102):1749-54.

39 Committee for Medicinal Products for Human Use (CHMP). Qualification opinion on stride velocity 95th centile as a secondary endpoint in Duchenne Muscular Dystrophy measured by a valid and suitable wearable device. 2019. Available from: https://www.ema.europa.eu/en/documents/scientific-guideline/qualification-opinion-stride-velocity-95th-centile-secondary-endpoint-duchenne-muscular-dystrophy_en.pdf

40 Coran P, Goldsack C, Grandinetti A. Advancing the Use of Mobile Technologies in Clinical Trials : Recommendations from the Clinical Trials Transformation Initiative. 2019;27701:145-54.

41 Izmailova ES, Wagner JA, Perakslis ED. Wearable Devices in Clinical Trials: Hype and Hypothesis. Clin Pharmacol Ther. 2018;104(1):42-52.

42 Badawy R, Hameed F, Bataille L, Little MA, Claes K, Saria S, *et al.* Metadata Concepts for Advancing the Use of Digital Health Technologies in Clinical Research. Digit Biomarkers. 2019;3(3):116-32.

43 Byrom B, Rowe DA. Measuring free-living physical activity in COPD patients: Deriving methodology standards for clinical trials through a review of research studies. Contemp Clin Trials. 2016;47:172-84.

44 Arneric SP, Cedarbaum JM, Khozin S, Papapetropoulos S, Hill DL, Ropacki M, *et al.* Biometric monitoring devices for assessing end points in clinical trials: Developing an ecosystem. Nat Rev Drug Discov. 2017;16(10):736.

45 Connell SO, Ólaighin G, Kelly L, Murphy E, Beirne S. These Shoes Are Made for Walking : Sensitivity Performance Evaluation of Commercial Activity Monitors under the Expected Conditions and Circumstances Required to Achieve the International Daily Step Goal of 10,000 Steps. 2016;1-14.

46 Getting real with wearable data. 2019;37(April):41587.

47 Hake AM, Dacks PA, Arnerić SP. Concise informed consent to increase data and biospecimen access may accelerate innovative Alzheimer's disease treatments. Alzheimer's Dement Transl Res Clin Interv. 2017;3(4):536-41.

48 Burnham JP, Lu C, Yaeger LH, Bailey TC, Kollef MH. Using wearable technology to predict health outcomes: A literature review. J Am Med Informatics Assoc. 2018;25(9):1221-7.

49 Panch T, Mattie H, Celi LA. The "inconvenient truth" about AI in healthcare. NPJ Digit Med. 2019;2(1):4-6.

PART II

TECHNICAL VALIDATION OF DIGITAL ENDPOINTS

CHAPTER 3

Development and technical validation of a smartphone-based cry detection algorithm

Front. Pediatr. 9:651356. doi:10.3389/fped.2021.651356

Ahnjili ZhuParris,[1] Matthijs D. Kruizinga,[1,2,3] Max van Gent,[1,2] Eva Dessing,[1,2]
Vasileios Exadaktylos,[1] Robert-Jan Doll,[1] Frederik E. Stuurman,[1,3]
Gertjan A. Driessen,[2,4] Adam F. Cohen[1,3]

1 Centre for Human Drug Research, Leiden, the Netherlands
2 Juliana Children's Hospital, Haga Teaching Hospital, The Hague, the Netherlands
3 Leiden University Medical Center, Leiden, the Netherlands
4 Department of pediatrics, Maastricht University Medical Centre, Maastricht, the Netherlands

Abstract

INTRODUCTION The duration and frequency of crying of an infant can be indicative of its health. Manual tracking and labelling of crying is laborious, subjective, and sometimes inaccurate. The aim of this study was to develop and technically validate a smartphone-based algorithm able to automatically detect crying.

METHODS For the development of the algorithm a training dataset containing 897 5-second clips of crying infants and 1263 clips of non-crying infants and common domestic sounds was assembled from various online sources. OPENSMILE software was used to extract 1591 audio features per audio clip. A random forest classifying algorithm was fitted to identify crying from non-crying in each audio clip. For the validation of the algorithm, an independent dataset consisting of real-life recordings of 15 infants was used. A 29-minute audio clip was analyzed repeatedly and under differing circumstances to determine the intra-and inter-device repeatability and robustness of the algorithm.

RESULTS The algorithm obtained an accuracy of 94% in the training dataset and 99% in the validation dataset. The sensitivity in the validation dataset was 83%, with a specificity of 99% and a positive- and negative predictive value of 75% and 100%, respectively. Reliability of the algorithm appeared to be robust within- and across devices, and the performance was robust to distance from the sound source and barriers between the sound source and the microphone.

CONCLUSION The algorithm was accurate in detecting cry duration and was robust to various changes in ambient settings.

Introduction

Crying is a primary indicator of decreased infant well-being[1]. Besides the normal crying-behaviour that is natural for every infant, a change in cry duration, intensity or pitch can be a symptom of illness[2]. Cry duration has been used as a biomarker for diagnostic and follow-up purposes for a wide range of clinical conditions of infancy, such as gastroesophageal reflux and cow milk allergy[3,4]. However, traditional methods to record cry behaviour, such as parent- or nurse- reported cry duration, are subjective and vulnerable to observer bias[5]. On the other hand, more objective manual annotating of audio recordings is labour intensive and may be subject to privacy-concerns by parents. An objective, automated and unobtrusive method to quantify crying behaviour in an at-home and clinical setting may improve the diagnostic process in excessively crying infants, allow for objective determination of treatment effects by physicians, and enable researchers to include objectively determined cry duration as digital biomarker in clinical trials. Therefore, a classification algorithm is necessary for the automatic recognition of cries in audio files. Given the importance for researchers to study the relationship between an infant's crying patterns and their health, automatic detection and quantification of infant cries from an audio signal is an essential step in remote baby monitoring applications[6].

Automatic cry detection has been reported in the form of remote baby monitors for non-intrusive clinical assessments of infants in hospital settings[6–9], and several researchers have shown that classification of cry- and non-cry- sounds is possible with machine-learning algorithms[10–12]. However, most algorithms lack validation in a completely independent dataset, which is crucial to predict performance in new- and real-world settings, while data regarding intra- and inter- device variability and other factors that may influence repeatability is lacking as well (10,13,14). Finally, algorithms are often developed for use on personal computers or dedicated devices. Usability of an algorithm would be increased if it were available on low-cost consumer-devices such as smartphones, which are readily available in most households and are easy to operate. Furthermore, smartphones have adequate processing power to analyse and transmit data continuously for monitoring in real-time. The aim of this study was to develop and validate a smartphone-based cry-detection algorithm that is accurate, reliable, and robust to changes in ambient conditions.

Materials and methods

Location and ethics

This was a prospective study conducted by the Centre for Human Drug Research (CHDR) and Juliana Children's Hospital. The study protocol was submitted to the medical ethics committee Zuidwest Holland (ID 19-003, Leiden, the Netherlands), who judged the protocol did not fall under the purview of the Dutch Law for Research with Human Subjects (WMO). The study was conducted in compliance with the General data protection regulation (GDPR). The algorithm was developed and reported in accordance with EQUATOR guidelines[15]. A schematic overview of the analysis steps is displayed in *Supplementary Figure S1.*

Algorithm development

TRAINING DATASET A training dataset was obtained from various online sources (*Supplementary Table S2*) and consisted of both crying- and non-crying sounds. Non-crying sounds consisted of common real-life sounds and included talking, breathing, footsteps, cats, sirens, dogs barking, cars honking, snoring, glass breaking, and ringing of church clocks. Furthermore, non-crying infant sounds (hiccoughs, wailing, yelling, babbling, gurgles and squeaking), as well as adult crying sounds, were included in the training dataset. All sounds were played back through a loudspeaker and processed into non-overlapping 5-second epochs on a G5 (Motorola, Chicago, IL, USA) or G6 (Motorola, Chicago, IL, USA) smartphones and. A total of 1591 audio features (*Supplementary Text S3*) were extracted from each 5-second epoch with OPENSMILE (version 2.3.0, audEERING, Gilching, Germany)[16] on the smartphone. Each 5-second epoch was manually annotated as crying or non-crying. A 5-second epoch was selected due to the fact that the median cry duration (without a silent break) in the training dataset was 4 seconds.

ALGORITHM TRAINING To prevent overfitting of the algorithm on non-robust audio features provide by the software, manual feature selection was performed to exclude features that exhibited different distributions when analyzed under different conditions (*Supplementary Text S3*). Feature selection was performed using the audio file generated during the robustness tests. The file was played back through a laptop speaker during differing ambient conditions with (see paragraph Robustness tests in Materials & Methods), a dedicated speaker, and processed to OPENSMILE features with the CHDR MORE® application. Additionally, the raw file was processed using OPENSMILE software on a personal computer. Considering the data was derived from the exact same audio file, the distribution of features should be identical during all conditions (*Supplementary Text S3*). However, this was not the case for all features, particularly those that were derived from the extremes of each feature (e.g. Percentile 1% percentile 99%). Therefore, distribution plots were judged visually by the authors and each feature that demonstrated a clear difference in means or standard deviations across conditions was excluded from the final dataset. After selection, 980 features audio features remained in the dataset. Two discriminative classifiers (Random Forest and Logistic Regression[17–20]) and one generative classifier (Naïve Bayes) were considered for the classification of crying and non-crying sounds. For each classifier, a 5-fold cross-validated grid-search to select the best combination of features and hyper-parameters was performed to minimize the error estimates in the final model. The primary objective of the model was to identify crying and therefore, hyper-parameters that optimized for sensitivity were prioritized. This was followed by 5-fold cross-validation to robustly estimate the model performance and generalization of the model. The classifier with the highest Matthew's Correlation Coefficient (MCC) was chosen as the final model and subjected to algorithm validation.

Algorithm validation

DATA COLLECTION An independent validation dataset was obtained from two sources. First, audio recordings were made in an at-home setting of 4 babies aged 0-6 months using the G5 or G6 smartphones. Second, audio recordings were made with the G5 or G6 smartphones of 11 babies aged 0-6 months admitted to the pediatric ward due to various reasons. Audio recordings were made after obtaining informed consent from both parents and were stripped of medical-and personal information prior to analysis.

PERFORMANCE ANALYSIS Each 5-second epoch in the recordings was annotated as crying-and non-crying by one annotator. In the case of doubt on how to classify an epoch, two additional annotators were included, and a choice was made via blinded majority voting. The developed algorithm was used to classify each epoch, and annotations and classifications were compared to calculate the accuracy, MCC, sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) in the complete dataset and in the hospital-and home datasets separately.

POSTPROCESSING OF CRY EPOCHS INTO NOVEL BIOMARKERS Some infants are reported to cry often, but with short intervals in between. Only counting the number of epochs that contain crying for such infants could result in an underestimation of the burden for infants and parents. As such, the duration of 'cry sequences' (periods during which an infant is crying either continuously or occasionally) is an important additional feature. To calculate this, postprocessing of detected cries was performed to calculate the number and duration of cry sequences as separate candidate biomarkers. A cry sequence was defined by the authors with a start criterion (at least six 5-second epochs containing crying within one minute) and a stop criterion (no crying detected for five minutes). Individual timelines were constructed for true- and predicted cry sequences to determine the reliability of the algorithm for this novel biomarker.

ROBUSTNESS TESTS A series of robustness tests was conducted to ensure that the developed algorithm was robust to varying conditions when used with a smartphone with the final application (CHDR MORE®) installed, which is how the algorithm would be deployed in practice. A 29-minute-long clip containing 16.7 minutes of crying was played from a speaker with a smartphone with the CHDR MORE® application in proximity. This application, developed in-house, has incorporated OPENSMILE technology and is able to extract and transmit audio features. The following conditions were tested during this phase of the study: intra-device variability (n=10), inter-device variability (n=10), distance from audio source (0.5, 1, 2 and 4 meter) and by placing the phone behind several barriers and in the presence of background tv sounds. For intra-device variability, a single phone was used 10 times to determine repeatability within a single device. For inter-device variability, 10 different devices of the same type (G6) were used to determine the repeatability across devices. Because it was not technically possible to pair the application output with the raw audio features of the original recording, cumulative cry count plots were construed for each condition and compared with cumulative cries in the original recording.

Results

Algorithm training

The training set consisted of 897 5-second audio clips, as well as 1263 non-crying 5-second clips. Of the three methods applied to develop the algorithm, the Random Forest method achieved the highest accuracy and MCC with 93.8% and 87.3%, respectively (*Table 1*). The 10 most important audio features for the algorithm were derived from Mel Frequency cepstral coefficients, Mel frequency bands and Voicing Probability. A variable importance plot of the 10 most important features included in the final algorithm is displayed in *Supplementary Figure S4*.

Table 1. Performance of the final algorithm

Training dataset		Validation dataset		
Parameter	Performance (Mean (SD))*	Hospital subjects (n = 11)	Home subjects (n = 4)	All subjects (n= 15)
Accuracy	93.8% (+/-1.1%)	98.5%	99.7%	98.7%
MCC	87.3% (+/- 2.2%)	75.5%	98.6%	78.4%
Sensitivity	93.8% (+/-1.1%)	80.6%	97.5%	83.2%
Specificity	94.8% (+/-1.1%)	99.1%	100%	99.2%
PPV	-	72.2%	100%	75.2%
NPV	-	99.4%	99.6%	99.5%

Abbreviations: MCC: Matthew's Correlation Coefficient, PPV: positive predictive value, NPV: negative predictive value. * Mean (SD) performance of 5-fold cross validation

Algorithm validation

The 15 infants (mean age: 2 months (SD 1.9)) created a total of 150 minutes (1,805 5-second epochs) of crying and 4372 minutes (52,464 5-second epochs) of non-crying. The median cry duration of the infants recorded at home was shorter (1.4 minutes, IQR 0.58 - 2.6) compared to children recorded during their admission to the hospital (5.8 minutes, IQR 2.2-16.7). Performance of the algorithm in the independent validation dataset is displayed in *Table 1*. Overall accuracy was 98.7%, but sensitivity was lower (83.2%) compared to the performance in the training dataset. Due to the relatively low crying incidence compared to non-crying incidence, the specificity of 99.2% led to a PPV of 75.2%. *Supplementary Figure S5* displays individual timelines for each infant, displaying the epochs where crying- and misclassifications were present. After post-processing of cry epochs into cry sequences, the median number of cry sequences per infant in the validation dataset was 3 (IQR 1-3), for a total of 39 cry sequences. The median difference between true and predicted cry sequences was 1 (IQR 0.25-1). Furthermore, the median difference between true and predicted cry sequences duration was 6 minutes (IQR 2-15 minutes, *Table 2*). Individual timelines and concordance between true and predicted cry sequences are displayed in *Figure 1*.

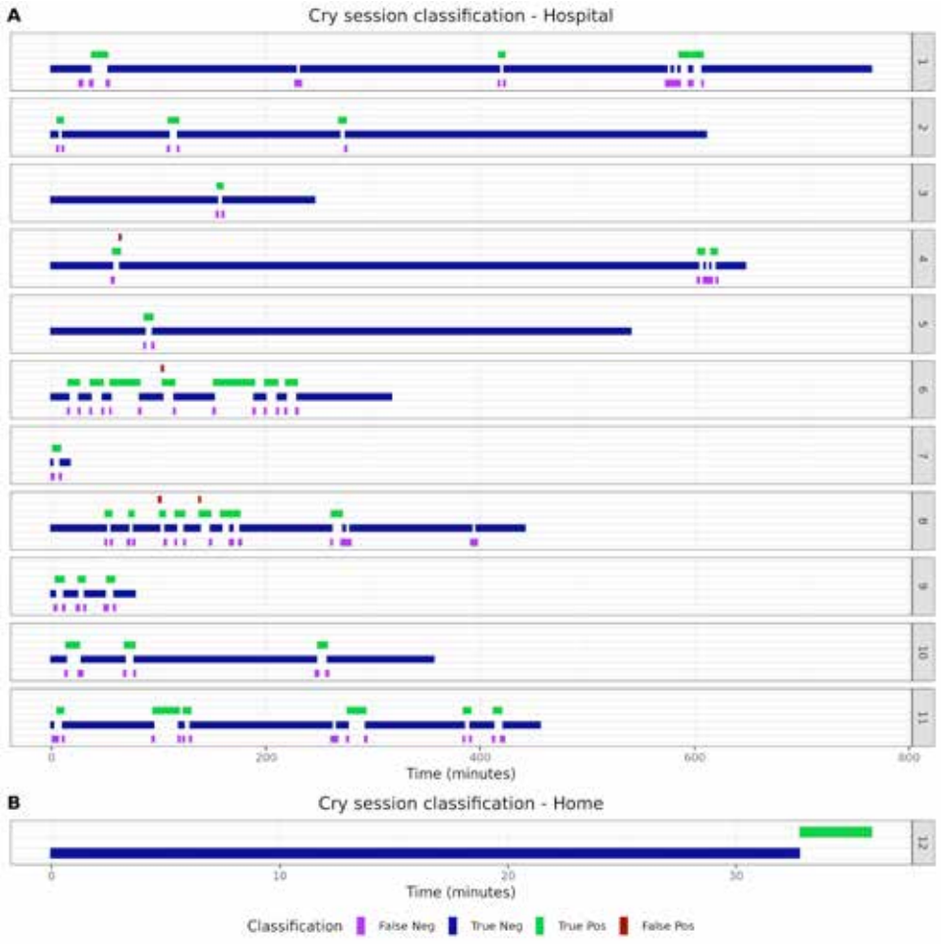
Table 2. Individual algorithm performance

Characteristics		Cry epochs						Cry sessions			
Subject	Duration (min)	Annotated count (n)	Algorithm count (n)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	Annotated cry sequence count (n)	Algorithm cry sequence count (n)	Annotated cry sequence duration	Algorithm cry sequence duration
HOSPITAL DATASET											
1	764	145	120	80%	99.5%	66.2%	99.7%	3	5	37	59
2	610	65	43	90.7%	99.6%	60%	99.9%	3	3	19	21
3	245	12	11	90.9%	99.9%	83.3%	99.9%	1	1	5	6
4	648	52	20	80%	99.5%	30.7%	99.5%	3	3	17	25
5	540	17	12	91.7%	99.9%	64.7%	99.9%	1	1	7	8
6	317	721	711	82.3%	95.6%	81.1%	95.9%	7	7	117	122
7	16.5	26	24	87.5%	97.1%	80.7%	98.2%	1	1	6	8
8	441	200	158	66.5%	98.2%	52.5%	98.9%	7	8	55	72
9	77.5	70	80	75%	98.8%	85.7%	97.7%	3	3	18.5	26
10	356	99	79	62%	98.8%	49.5%	99.3%	3	3	22	36
11	452	320	290	87.9%	98.7%	79.7%	99.3%	6	7	64	80
HOME DATASET											
12	36	38	40	95%	100%	100%	99.5%	1	1	2.8	2.4
13	13	7	7	100%	100%	100%	100%	0	0	0	0
14	2	25	25	100%	100%	100%	100%	0	0	0	0
15	1	8	8	100%	100%	100%	100%	0	0	0	0

Algorithm robustness

To ensure the algorithm and smartphone application performs sufficiently for the intended use, multiple tests were conducted to test robustness with the resulting smartphone application. *Figure 2A* shows the estimated repeatability of the algorithm by repeatedly classifying the same recording with the same device. *Figure 2B* shows the cumulative cry count of 8 different devices of the same type, which gives an indication of the repeatability. The distance from the audio source, up to 4 meters, did not appear to impact the accuracy of the algorithm (*Figure 2C*) Finally, blocking the audio signal by placing the phone behind several physical barriers in front of the audio source demonstrated comparable accuracy across conditions (*Figure 2D*). Creating additional background noise generated by a television appeared to slightly decrease the specificity of the algorithm, as the final cry count according to the algorithm was higher compared to the true number of cries in the audio file.

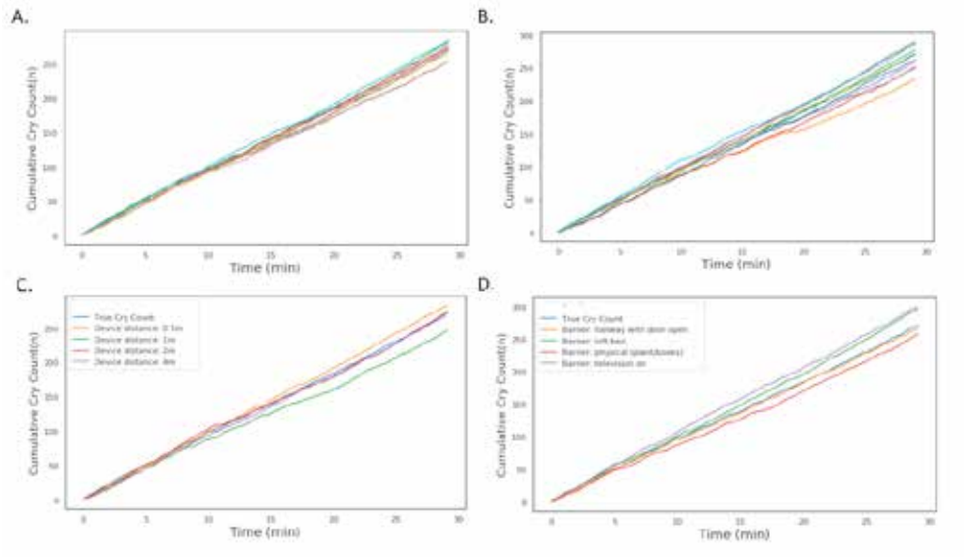
Figure 1. True and predicted cry sequence per infant.



Discussion

This paper describes the development and validation of a smartphone-based cry detection algorithm in infants. A random forest classifier had the highest accuracy in the training dataset and achieved a 98.7% accuracy in an independent validation set. Although the sensitivity of 83.2% was slightly lower compared to the estimated accuracy in the training dataset, the individual classification timelines show that this should not lead to unreliable estimation of cry duration. The fact that most misclassifications occurred directly before

Figure 2. Cumulative cry count during robustness tests. (A) Intra-device repeatability. Each individual line is a different run with the same phone. (B) Inter-device repeatability. Each individual line is a run with a different phone of the same type. (C) Influence of device distance from the audio source. (D) Influence of physical barrier or ambient background noise. In each of the panels, the light-blue line is the reference from the audio file.



or after crying indicates that such misclassifications may be due to cry-like fussing, which are difficult to classify for both the algorithm and the human annotators. Post-processing of the detected cry epochs into cry sequences decreased the mismatch and resulted in excellent performance for each individual infant.

The observed accuracy of the algorithm is comparable to others described in the literature, although there is large variation in reported accuracy. Traditional machine learning classifiers and neural network-based classifiers have been used for infant cry analysis and classification[21]. We found that several studies that explored the use of minimum, maximum, mean, standard deviation and the variance of MFCCS and other audio features to differentiate normal, hypo-acoustic and asphyxia types using the Chillanto database[6]. Support Vector Machines (SVM) are among the most popular infant classification algorithms and routinely outperform neural network classifiers[22,23]. Furthermore, Osmani et al have illustrated that boosted and bagging trees outperform SVM cry classification[24]. Additionally, sensitivities between 35-90% with specificities between 96-98% have been reported using a convoluted neural network approach[10,14]. Ferreti *et al.* and Severini *et al.* also used a neural network approach and achieved a reported precision of 87% and 80%, respectively[11,12]. However, algorithms often lack validation in an independent dataset as, and real-life performance in new and challenging environments will most likely be lower. Our algorithm has several advantages compared to other approaches that have been described in the past. Most importantly, the algorithm was validated on independent and real-life data obtained from two settings where the application could be used in the future. Validation invariably leads to a drop in accuracy compared to the performance of the training data but gives reassurance regarding the generalizability of the algorithm in new settings that were not included during training. Furthermore, the algorithm can be deployed on all Android smartphones and no additional equipment is needed for acquiring the acoustic features. Although it is possible to implement complex deep learning algorithms on portable devices, we demonstrated that a shallow learning algorithm such as a random forest achieves good classifying capability. This means that audio processing and classification can be performed on the device in real-time with the MORE® application, and thus, precludes direct transmission of audio to a central location with inherent preservation of privacy. Finally, the manual feature selection that was performed should lead to further generalizability of the algorithm in new condition, since the observed variability in the excluded audio features would most likely result in a drop in accuracy in challenging acoustic environments. While automated feature selection methods could have been used, automated feature selection requires a static definition of similarity between distributions within features. This is not a straightforward task. Given the nature of the features, we chose to manually exclude features that presented a clearly different distribution from the rest of the features.

All in all, the performance of the algorithm in combination with the mentioned advantages indicate reliability of the algorithm, and may be preferable over manual tracking of cry duration through a diary in several situations. Although the literature regarding sources of inaccuracy in cry monitoring via a diary is sparse, several factors make manual tracking through a diary a subjective assessment[5]. Observer bias can cause parents to overestimate the true duration of crying, and placebo-effects may cause parents to underestimate true cry duration after an intervention[25]. Additionally, parents may underreport nocturnal cry duration when they sleep through short cry sequences during the night. Current tracking of cry duration in clinical settings is performed by nurses, who have other clinical duties as well, possibly making the quality of the cry diary dependent on the number of patients under their care. While the consequences of all of these factors are not

easy to quantify, the combination of these sources of inaccuracy leads to the conclusion that objective and automated cry-monitoring could significantly improve the reliability of objective follow-up of cry duration in both clinical trials and -care. Still, parental report of cry duration and cry behavior will remain an important component of follow-up.

A technical limitation of any Android application, including the MORE® application, is that continuous recording can be interrupted by other smartphone applications apps that also access the microphone, like phone calls. However, using a dedicated smartphone for the purpose of cry monitoring will diminish this limitation. Only Motorola G5/G6 phones were used during each phase of algorithm development and validation. Although performance on other smartphones is uncertain, the approach used in this paper could easily be replicated to adapt the algorithm to other devices and obtain a similar accuracy. In the future, incorporation of covariates such as age, sex or location in the model may improve classifying capability even further, and further stratification could allow to discriminate different types of crying. In this manner cries from asphyxiated infants[26], preterm infants[27], or infants with respiratory distress syndrome could be differentiated from healthy infants[13]. One potential technical limitation of our approach is the use of loudspeakers to create the training dataset. An ideal training dataset would include smartphone-based audio recordings of multiple subjects under different conditions over a long period of time. We found the most appropriate alternative was to re-record open-sourced cry corpus using smartphone. While the playback could have potentially hindered the quality of the openSmile features and thus the classification, it resulted in excellent classification performance of the home and hospital recordings. Hence the impact of the quality of the loudspeaker-based dataset was deemed acceptable. A follow-up study that uses an original smartphone-based cry corpus could potentially improve the accuracy of the classification algorithm. The start- and stop criteria used to determine the beginning and end of a cry sequence are a new proposal that was not previously described in the literature. However, the criteria appear reasonable and individual timeline figures demonstrated that this postprocessing step was able to generate a solid high-level overview of individual cry behaviour. Still, alternative criteria could obtain similar accuracy and may be explored in the future.

The developed algorithm already provides an excellent overview of the cry behaviour of infants and preliminary tests of the robustness of the resulting algorithm show inter- and intradevice repeatability and reliability up to 4 meters from the audio source. The algorithm can replace current methods to track cry behaviour, such as cry diaries, in clinical and at-home settings. However, more research is needed before implementing the cry duration and the amount of cry sequences as digital endpoint in trials. Clinical validation of cry duration and cry sequence count as digital biomarker in a patient population is necessary, and should focus on establishing new normative values for objectively determined cry- and sequence duration and -count, the difference between patients and healthy controls, correlation with disease-severity and sensitivity to change after an intervention[28].

Conclusion

The proposed smartphone-based algorithm is accurate, robust to various conditions and has the potential to improve clinical follow-up of cry behaviour and clinical trials investigating interventions to enhance infant well-being.

SUPPLEMENTARY DATA

Sup. Figure S1	Schematic overview of analysis steps
Sup. Table S2	Audio sources
Sup. Text S3	Audio features and feature selection
Sup. Figure S4	Variable importance
Sup. Figure S5	Figure per baby showing true and predicted crying per epoch

REFERENCES

1 Wolke D, Bilgin A, Samara M. Systematic Review and Meta-Analysis: Fussing and Crying Durations and Prevalence of Colic in Infants. *J Pediatr* (2017) **185**:55-61.e4. doi:10.1016/j.jpeds.2017.02.020

2 Freedman SB, Al-Harthy N, Thull-Freedman J. The crying infant: Diagnostic testing and frequency of serious underlying disease. *Pediatrics* (2009) **123**:841-848. doi:10.1542/peds.2008-0113

3 Moore DJ, Siang-Kuo Tao B, Lines DR, Hirte C, Heddle ML, Davidson GP. Double-blind placebo-controlled trial of omeprazole in irritable infants with gastroesophageal reflux. *J Pediatr* (2003) **143**:219-223. doi:10.1067/S0022-3476(03)00207-5

4 Lucassen PLBJ, Assendelft WJJ, Gubbels JW, Van Eijk JTM, Douwes AC. Infantile colic: Crying time reduction with a whey hydrolysate: A double-blind, randomized, placebo-controlled trial. *Pediatrics* (2000) **106**:1349-1354. doi:10.1542/peds.106.6.1349

5 Barr RG, Kramer MS, Boisjoly C, McVey-White L, Pless IB. Parental diary of infant cry and fuss behaviour. *Arch Dis Child* (1988) **63**:380-387. doi:10.1136/adc.63.4.380

6 Jeyaraman S, Muthusamy H, Khairunizam W, Jeyaraman S, Nadarajaw T, Yaacob S, Nisha S. A review: survey on automatic infant cry analysis and classification. *Health Technol* (*Berl*) (2018)20-29. doi:10.1007/s12553-018-0243-5

7 Saraswathy J, Hariharan M, Yaacob S, Khairunizam W. Automatic classification of infant cry: A review. *2012 Int Conf Biomed Eng ICoBE 2012* (2012)543-548. doi:10.1109/ICoBE.2012.6179077

8 LaGasse LL, Neal AR, Lester BM. Assessment of infant cry: Acoustic cry analysis and parental perception. *Ment Retard Dev Disabil Res Rev* (2005) **11**:83-93. doi:10.1002/mrdd.20050

9 Ntalampiras S. Audio Pattern Recognition of Baby Crying Sound. (2016) doi:10.17743/jaes.2015.0025

10 Lavner Y, Cohen R, Ruinskiy D, Ijzerman H. Baby cry detection in domestic environment using deep learning. *2016* IEEE *Int Conf Sci Electr Eng* ICSEE *2016* (2017) doi:10.1109/ICSEE.2016.7806117

11 Ferretti D, Severini M, Principi E, Cenci A, Squartini S. Infant cry detection in adverse acoustic environments by using deep neural networks. *Eur Signal Process Conf* (2018) **2018**-**Septe**:992-996. doi:10.23919/EUSIPCO.2018.8553135

12 Severini M, Ferretti D, Principi E, Squartini S. Automatic detection of cry sounds in neonatal intensive care units by using deep learning and acoustic scene simulation. IEEE *Access* (2019) **7**:51982-51993. doi:10.1109/ACCESS.2019.2911427

13 Salehian Matikolaie F, Tadj C. On the use of long-term features in a newborn cry diagnostic system. *Biomed Signal Process Control* (2020) **59**:101889. doi:10.1016/j.bspc.2020.101889

14 Choi S, Yun S, Ahn B. Implementation of automated baby monitoring: CC BeBe. *Sustain* (2020) **12**: doi:10.3390/su12062513

15 Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, Shilton A, Yearwood J, Dimitrova N, Ho TB, *et al.* Guidelines for developing and reporting machine learning predictive models in biomedical research: A multidisciplinary view. *J Med Internet Res* (2016) **18**:1-10. doi:10.2196/JMIR.5870

16 Eyben F, Schuller B. openSMILE:). ACM *sigmultimedia Rec* (2015) **6**:4-13. doi:10.1145/2729095.2729097

17 Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning Data Mining, Inference, and Prediction (12th printing). (2009)745. doi:10.1007/978-0-387-84858-7

18 Pranckevičius T, Marcinkevičius V. Comparison of Naive Bayes, Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression Classifiers for Text Reviews Classification. *Balt J Mod Comput* (2017) **5**: doi:10.22364/bjmc.2017.5.2.05

19 Muchlinski D, Siroky D, He J, Kocher M. Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data. *Polit Anal* (2016) **24**:87-103. doi:10.1093/pan/mpv024

20 Czepiel SA. Maximum Likelihood Czepiel, S. A. (2012). Maximum Likelihood Estimation of Logistic Regression Models: Theory and Implementation. Class Notes,1-23. Retrieved from papers3://publication/uuid/4E1E1B7E-9CAC-4570-8949-E96B51D9C91DEstimation of Logistic Regre. *CI Notes* (2012)1-23.

21 Ji C, Mudiyanselage TB, Gao Y, Pan Y. A review of infant cry analysis and classification. *J* audio speech *Music* PROC (2021) doi:https://doi.org/10.1186/s13636-021-00197-5

22 Joshi G, Dandvate C, Tiwari H, Mundhare A. Prediction of probability of crying of a child and system formation for cry detection and financial viability of the system. *Proc-2017 Int Conf Vision, Image Signal Process ICVISP 2017* (2017) **2017**-**November**:134-141. doi:10.1109/ICVISP.2017.33

23 Felipe GZ, Aguiat RL, Costa YMG, Silla CN, Brahnam S, Nanni L, McMurtrey S. Identification of Infants' Cry Motivation Using Spectrograms. *Int Conf Syst Signals, Image Process* (2019) **2019**-**June**:181-186. doi:10.1109/IWSSIP.2019.8787318

24 Osmani A, Hamidi M, Chibani A. Machine learning approach for infant cry interpretation. *Proc-Int Conf Tools with Artif Intell* ICTAI (2018) **2017**-**November**:182-186. doi:10.1109/ICTAI.2017.00038

25 Berseth CL, Johnston WH, Stolz SI, Harris CL, Mitmesser SH. Clinical Response to 2 Commonly Used Switch Formulas Occurs within 1 Day. *Clin Pediatr* (*Phila*) (2009) **48**:58-65.

26 Ji C, Xiao X, Basodi S, Pan Y. Deep Learning for Asphyxiated Infant Cry Classification Based on Acoustic Features and Weighted Prosodic Features. in *2019 International Conference on Internet of Things* (*iThings) and* IEEE *Green Computing and Communications (GreenCom) and* IEEE *Cyber, Physical and Social Computing (CPSCom) and* IEEE *Smart Data* (*SmartData*),1233-1240. doi:10.1109/iThings/GreenCom/CPSCom/SmartData.2019.00206

27 Orlandi S, Reyes Garcia CA, Bandini A, Donzelli G, Manfredi C. Application of Pattern Recognition Techniques to the Classification of Full-Term and Preterm Infant Cry. *J Voice* (2016) **30**:656-663. doi:10.1016/j.jvoice.2015.08.007

28 Kruizinga MD, Stuurman FE, Exadaktylos V, Doll RJ, Stephenson DT, Groeneveld GJ, Driessen GJA, Cohen AF. Development of Novel, Value-Based, Digital Endpoints for Clinical Trials: A Structured Approach Toward Fit-for-Purpose Validation. *Pharmacol Rev* (2020) **72 (4)**:899-909. doi:10.1124/pharmrev.120.000028

CHAPTER 4

Development and technical validation of a smartphone-based pediatric cough detection algorithm

Pediatr Pulmonol. 2021 Dec 29. doi: 10.1002/ppul.25801

MD Kruizinga,[1,2,3*] A Zhuparris,[1*] E Dessing,[1,2] FJ Krol,[1,3] AJ Sprij,[2] RJ Dolll,[1] FE Stuurman,[1] V Exadaktylos,[1] GJA Driessen,[2,4] AF Cohen[1,3]

**Both authors contributed equally*

1 Centre for Human Drug Research, Leiden, the Netherlands
2 Juliana Children's Hospital, Haga Teaching Hospital, The Hague, the Netherlands
3 Leiden University Medical Center, Leiden, the Netherlands
4 Department of pediatrics, Maastricht University Medical Centre, Maastricht, the Netherlands

Abstract

INTRODUCTION: Coughing is a common symptom in pediatric lung disease and cough frequency has been shown to be correlated to disease activity in several conditions. Automated cough detection could provide a non-invasive digital biomarker for pediatric clinical trials or care. The aim of this study was to develop a smartphone-based algorithm that objectively and automatically counts cough sounds of children.

METHODS The training set was composed of 3,228 pediatric cough sounds and 480,780 non-cough sounds from various publicly available sources and continuous sound recordings of 7 patients admitted due to respiratory disease. A Gradient Boost Classifier was fitted on the training data, which was subsequently validated on recordings from 14 additional patients aged 0-14 admitted to the pediatric ward due to respiratory disease. The robustness of the algorithm was investigated by repeatedly classifying a recording with the smartphone-based algorithm during various conditions.

RESULTS: The final algorithm obtained an accuracy of 99.7%, sensitivity of 47.6%, specificity of 99.96%, positive predictive value of 82.2% and negative predictive value 99.8% in the validation dataset. The correlation coefficient between manual-and automated cough counts in the validation dataset was 0.97 ($p < 0.001$). The intra-and interdevice reliability of the algorithm was adequate, and the algorithm performed best at an unobstructed distance of 0.5m-1m from the audio source.

CONCLUSION: This novel smartphone-based pediatric cough detection application can be used for longitudinal follow-up in clinical care or as digital endpoint in clinical trials.

Introduction

Coughing is a physiological mechanism of the respiratory system to clear excessive secretions. It can be caused by various acute and chronic diseases, such as viral upper respiratory tract infections, bacterial infections, asthma, protracted bacterial bronchitis or tic cough, and is a common reason for parents to seek medical consultation for their children[1,2].

Several studies have shown that cough severity is correlated with disease activity in asthma and other pulmonary diseases[3-6], making cough frequency an attractive candidate biomarker for respiratory disease severity. Although coughing is traditionally quantified via self- or parent-report in the form of questionnaires, technological advances allow for more sophisticated (semi-)automatic cough monitoring methods. Indeed, several commercial and academic entities have endeavored to develop cough detection algorithms, with varying success[7]. The most notable and reliable examples are the Leicester Cough Monitor and the VitaloJak[8,9], which record sounds with a dedicated body-contact device and microphone, and subsequently use semi-automated counting methods. Several completely automated cough counting algorithms have been developed, mostly for an adult population, but none have proceeded towards widespread availability[7].

A notable disadvantage of body-contact devices is that they are inconvenient in the field of pediatrics, especially in infants and toddlers. Additionally, pediatric cough sounds exhibit more variability across different ages due to the developing respiratory- and vocal system, which can make robust detection more challenging[10]. An ideal algorithm would require no manual input, be able to monitor from a distance, and be operational on low-cost consumer devices that are readily available, such as smartphones. To date, no such algorithm has been developed in the field of pediatrics. This study aimed to develop an algorithm that objectively and automatically counts cough sounds in children based on audio features collected via a smartphone application.

Materials and Methods

Ethics and logistics

This study was conducted at the Centre for Human Drug Research (CHDR, Leiden, the Netherlands) and the Haga Teaching Hospital, Juliana Children's Hospital (the Hague, the Netherlands). Institutional review board approval was obtained (registration number

T19-080), and the study was conducted in compliance with the general data protection regulation (GDPR). The algorithm was developed as part of the CHDR MORE® system, a remote monitoring clinical trial platform. Reporting was performed in accordance with EQUATOR guidelines[11].

Data collection

A comprehensive training dataset was obtained from multiple sources. First, audio was extracted from 91 publicly available videos on YouTube that contained coughing children with an estimated age between 0-16 years old. Furthermore, 334 non-coughing audio clips were gathered from YouTube, GitHub, and the British Broadcasting Corporation (BBC) sound library. The non-coughing set contained various sounds that were expected to occur in real-life settings, such as talking, breathing, footsteps, cats, sirens, dogs barking, cars honking, snoring, glass breaking, and church clocks. Additionally, 21 children aged 0-16 and admitted due to pulmonary disease were recruited on the ward of Juliana Children's Hospital. Data of the first 7 children were used to supplement the training dataset, with a maximum of the first 150 coughs per child to avoid overrepresentation of a single subject. Remaining cough sounds of the 7 children were discarded. Data from the other 14 subjects were used as validation dataset. All audio clips were manually annotated by an investigator using Audition software (Adobe, San Jose, CA, USA). No filter was applied to remove 'silent' sections of the recording to ensure that the estimated accuracy reflects real-life conditions. As a result, the proportion of cough sounds in the validation dataset was 0.7%. The composition of the final training-and validation dataset are displayed in *Table 1*.

Audio feature extraction and selection

Audio feature were extracted from all audio clips using the OPENSMILE software (version 2.3.0, audEERING, Gilching, Germany)[12]. The software converted all audio clips into 1582 features per epoch. Epoch length was fixed at 0.5 seconds since the average cough duration in the training dataset was 0.3 seconds. The extracted features included several audio domains, such as Mel-frequency cepstral coefficients (MFCCS) and fundamental frequencies (F0) (*Supplementary Text S1*). Using manual inspection, the most robust features across multiple conditions were selected (*Supplementary Text S2*) and only these features were included in the final dataset used for algorithm development.

Table 1. Composition of training-and validation datasets

	Training dataset				Validation dataset
	YouTube (91 clips)	Various sources (334 clips)	Hospital (7 children)	Total	Hospital (14 children)
Cough sounds (n)	2,229		999	3,228	4,123
Non-cough sounds (n)	9,702	39,456	431,622	480,780	100,522
Total (n)	11,931	39,456	432,621	484,008	104,645
Cough proportion (%)*	18.5%	0%	0.2%	0.7%	0.4%
Mean cough duration (s)	0.3	-	0.3	0.3	0.3

* Proportion of 0.5 second epochs that contain cough sounds.

Algorithm development and validation

Two discriminative decision-tree based classifiers were considered for the model: Random Forest and Gradient Boost Classifier. Five-fold cross-validation was used to select the optimal features and hyperparameters for the model. The optimal classifier was selected based on the highest overall Matthew's correlation coefficient (MCC). The selected model was then used to classify all 0.5-second epochs in the validation dataset. The sensitivity, specificity, MCC, positive predictive value (PPV) and negative predictive value (NPV) were calculated for the complete validation dataset and per subject.

Initial robustness tests

Limited robustness tests were conducted to ensure the algorithm performs comparably across a range of different conditions when applied as a smartphone application. First, a 27-minute-long audio-clip was generated which included coughing- and household sounds, as well as sections with silence. The clip was subsequently played repeatedly from a speaker, while a G6 smartphone (Motorola, Chicago, IL, USA) with the CHDR MORE® application was placed in proximity. The application has incorporated OPENSMILE software and is able to calculate and transmit the generated audio features. The following conditions were tested: first, the intra-device variability was tested by repeating the assessment 7 times with the same device; second, the inter-device variability was tested by repeating the assessment 4 times with different devices of the same type;

third, the effect of device distance (0.5m, 1m and 4m) from the audio source was assessed and finally, accuracy was assessed when a small (plant and book) or large (loft bed) barrier was placed in front of the audio source and when television sounds were played in the background. Because the 0.5-second epochs from the original file and the output of the MOORE® application could not be paired, cumulative cough count plots were generated and compared across conditions.

Results

Algorithm training

The training set consisted of 3,424 0.5-second cough epochs of various sources, as well as 431,622 0.5-second non-cough epochs. The final algorithm, fitted through a Gradient Boost Classifier, achieved an accuracy of 99.6%, MCC of 73.7%, sensitivity of 99.6% and specificity of 99.9% in the training set (*Table 2*). The most important audio features the algorithm relied on were derived from the mel frequency and loudness categories (*Supplementary Figure S3*).

Table 2. Performance of the final algorithm

Parameter	Training dataset	Validation dataset
	Mean (SD) performance*	Overall performance
Accuracy	99.61% (+/- 0.13%)	99.74%
MCC	73.67% (+/- 0.16%)	62.40%
Sensitivity	99.62% (+/- 0.13%)	47.56%
Specificity	99.89% (+/- 0.09%)	99.96%
PPV	99.65% (+/- 0.08%)	82.16%
NPV	99.82% (+/- 0.02%)	99.78%

Abbreviations: MCC: Matthew's Correlation Coefficient, PPV: positive predictive value, NPV: negative predictive value. * Mean (SD) performance of 5-fold cross-validation

Algorithm validation

For validation, fourteen patients with respiratory disease aged 0-14 were recorded during a hospital admission. The median recording duration was 632 (IQR 477-775) minutes. In total, 4,123 0.5-second epochs contained coughing. The median cough count per subject was 150 (IQR 38-446). *Table 2* displays the overall accuracy of the algorithm in the validation dataset. Overall sensitivity was 47.6% and specificity was 99.96%. Due to the relatively low frequency of cough counts in the dataset, the NPV and PPV in these real-world settings was 99.78% and 82.2%, respectively. The performance of the algorithm differed between subjects. Individual patient characteristics and classification accuracies are displayed in *Table 3*. The correlation coefficient between manual cough count and automated cough count was 0.97 ($p < 0.001$, *Figure 1*).

Figure 1. Correlation manual-and automatic cough count in validation dataset. Pearson correlation between manually counted coughs and automatically detected coughs. Each dot represents an individual subject in the validation dataset.

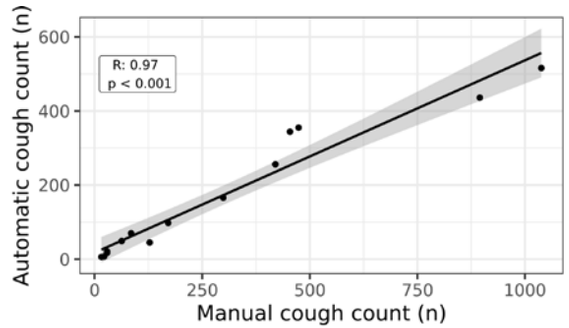


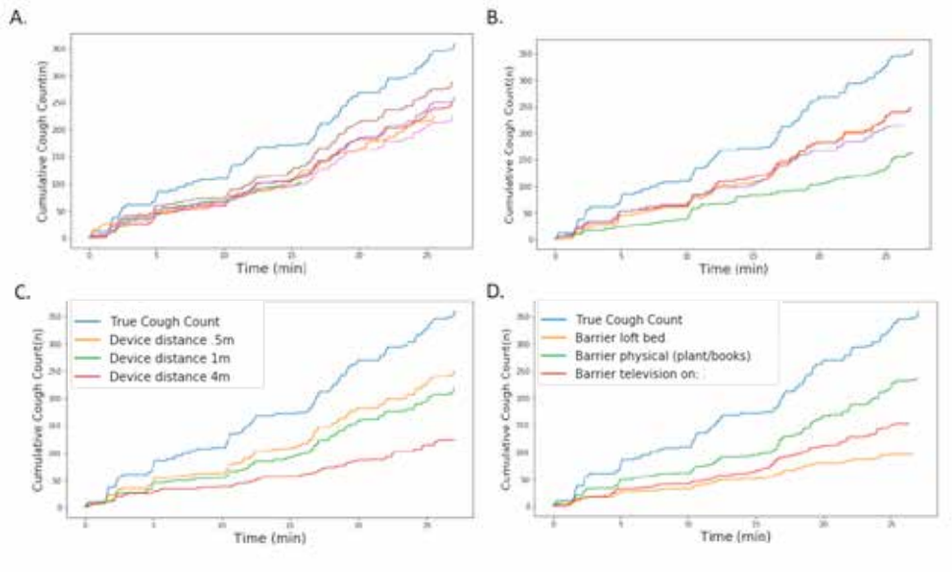
Table 3. Performance of the final algorithm in Individual subjects

Subject (#)	Age	Diagnosis	Recording duration (min)	Manual Count (n)	Algorithm count (n)	Sens.	Spec.	MCC
1	14y	Pneumonia	4	22	7	32%	100%	55%
2	4y	Wheezing	717	63	49	73%	100%	73%
3	5y	Pneumonia	237	29	21	72%	100%	85%
4	1.5y	Pneumonia	609	16	6	19%	100%	31%
5	6w	Bronchiolitis	727	85	70	58%	100%	63%
6	3y	Pneumonia	792	454	344	69%	100%	79%
7	9w	Bronchiolitis	967	895	436	34%	100%	69%
8	4y	Pneumonia/ Wheezing	497	29	17	52%	100%	88%
9	11y	Asthma	598	171	98	56%	100%	73%
10	5w	Bronchiolitis	873	1038	516	37%	100%	53%
11	2y	Pneumonia	434	474	355	70%	100%	81%
12	3y	Pneumonia	470	420	256	54%	100%	68%
13	13w	Bronchiolitis	654	128	45	34%	100%	57%
14	4y	Pneumonia	791	299	166	40%	100%	53%

Limited algorithm robustness tests

Repeated (n=7) tests with the same device and show comparable performance during each iteration (*Figure 2A*), while the inter-device variability tests show some variability in cumulative cough count across devices (*Figure 2B*). The effect of the distance of the device to the audio source was assessed (*Figure 2C*) and demonstrated comparable accuracy for 0.5 and 1m distance. The accuracy was lower when the distance of the monitoring device from the audio source was increased. Finally, the effect of a small- and large barrier was investigated, as well as the effect of ambient television sounds playing in the background (*Figure 2D*). During this test, it appeared that a small physical barrier did not impact algorithm performance, but a large physical barrier and background television sounds led to a lower cumulative cough count.

Figure 2. Performance of the algorithm under varying circumstances. A. Intra-device repeatability. Each individual line represents a different session with the same device. B. Inter-device repeatability. Each individual line represents a different session with a different device of the same type. C. Influence of device distance from the audio source. D. Influence of physical barrier or ambient background noise. In each of the panels, the light-blue line is the reference from the audio file.



Discussion

The current manuscript described the development and initial validation of a novel cough detection algorithm in pediatrics. Publicly available audio recordings were combined with real-life recordings to fit an algorithm that had excellent classification capability in the training dataset. In the validation dataset, a sensitivity of 47.6% and specificity of 99.96% was obtained, which resulted in a PPV of 82.2% and an NPV of 99.8% in these real-world conditions. There was a strong correlation between manual cough count and automatic cough count. The accuracy of the algorithm in the validation set was confirmed by several robustness tests, which repeatedly showed a cumulative cough count that was roughly half of the true cough count across various conditions. The algorithm performed best when there was a relatively unobstructed maximum distance of 0.5m-1m from the audio source.

The current sensitivity is suboptimal but does not disqualify the algorithm, and we envision the current algorithm is already suitable for application in several settings. Algorithm-derived cough count could be incorporated as (secondary) digital endpoint in pediatric pulmonary disease trials. For this application, clinical validation of cough count as digital endpoints should be performed first, focusing on demonstrating a difference between patients and healthy children, correlation of the novel endpoint with traditional endpoints or patient reported outcomes, and sensitivity to change in disease activity[15]. In addition to clinical trials, applying this algorithm in clinical care is likely to be much more reliable than patient- or parent recall regarding cough frequency[13,14]. The strong correlation between manually- and automatically-counted coughs means the algorithm can discriminate children that cough excessively from children that do not and can uncover individual trends over time, e.g. to characterize clinical recovery after a hospital admission, or to assess the effect of treatment in excessively coughing patients with persistent bacterial bronchitis. This is further supported by the very high specificity of the algorithm that is maintained in all validation tests. For example, change in nocturnal cough frequency in the case of an asthma exacerbation could be identified reliably with the current algorithm, and subsequent treatment leading to a significant decrease in nocturnal coughing will also be detectable even with the current sensitivity. In the future, algorithm output could be combined with other non-invasive assessments known to be related to pulmonary disease-activity, such as physical activity-, heart rate- and pulmonary function monitoring, as well as electronic patient reported outcome measures. Together, this could provide a holistic overview of multiple aspects of pulmonary disease-severity and quality of life[16].

Multiple research groups have developed cough detection algorithms in recent years. However, only one was developed specifically for a pediatric population[17]. Although this algorithm was not applied in a mobile device. Still, pediatric cough detection is theoretically more challenging due to changing vocal cord acoustics during various stages of development. In adults, the most widely reported cough detection devices are the Leicester cough monitor and the VitaloJak[7]. These methods have been validated in independent datasets and appear both sensitive (91-99%) and specific (99%), but the use of dedicated microphones is less user-friendly in general and the use contact-devices precludes their use in several age categories in pediatrics. Furthermore, the semi-automated counting method used by both devices remains laborious and requires training, which means that widespread use in large-scale clinical trials or in general care is not feasible. Other algorithms that count coughs automatically have reported sensitivities of 78-99% and specificities of 92-99%[7,17-22], but only a few have been applied on a smartphone[20,21]. The one that most resembles the current study is a smartphone-based algorithm developed by Barata et al, who use a convolutional neural network to classify nocturnal sounds in adult asthmatics and obtained a sensitivity of 99.9% with a specificity of 91.5%[20]. In addition, other projects are often based on data obtained in tightly controlled environments and lack validation in independent or clinical datasets[17,21,22], and may show a similar drop in accuracy during validation as was observed for the algorithm developed here. For example, the PulmoTrack® device, designed for automatic clinic-based monitoring, showed a reduced sensitivity of 26% compared to human annotation during validation in a new cohort[23].

A major advantage of the algorithm developed in this study is the conversion of raw audio into audio features on the smartphone before transmission to the study center, which ensures the privacy of participants. The automated classification is another advantage, allowing devices to analyze-and transmit cough counts in real-time. A limitation was the manual feature selection performed, which introduces a potentially subjective factor to the analysis. Furthermore, a laptop speaker was used during the initial robustness tests and using a higher quality speaker may have led to slightly different performance during these tests. However, we believe the device quality is sufficient for the purpose of testing repeatability and investigating the effects of differing conditions. During this study, a single smartphone type was used, and the observed performance may vary when other devices are used[24]. Another potential problem would arise when the sensitivity of the algorithm would be highly dependent on the underlying disease that is studied, although there is no evidence of this in the validation dataset, such factors need to be studied further during clinical validation for which we can supply the algorithm to other interested academic groups. The current algorithm is developed as a one-size-fits-all solution that can classify coughs of all pediatric patient groups and ages and that only used sound features as input variables. Although the current accuracy appears sufficient to include as digital biomarker in the applications mentioned above, the accuracy of future algorithms could improve significantly with the cost of added complexity. First, accuracy could improve by addition of additional covariates such as age, sex, and diagnosis, although this would require some user input before use. Second, the exponential increase in processing power of mobile devices could allow for the development of personalized models in the future, which would both be trained, validated, and deployed on the participants' own smartphones. A personalized classification model that is tuned to the cough characteristics of an individual could potentially be much more accurate, considering the intra-individual variability in cough sounds is assumed to be smaller compared to inter-individual variability. Future studies could also aim to quantify cough intensity, as this characteristic may have greater impact on quality of life than cough frequency[7].

Conclusion

This novel smartphone-based cough detection application is one of the first of its kind and able to count coughs in pediatric patients with a sensitivity of 47%, specificity of 99.96%, PPV of 82% and NPV of 99.8%. Although the observed sensitivity in the intended use must be improved in the future, the current algorithm may be reliable enough for longitudinal monitoring in the context of clinical trials-or care, which will be evaluated during a clinical validation process.

REFERENCES

1 Kantar A. Phenotypic presentation of chronic cough in children. J Thorac Dis. 2017;9(4):907–13.
2 Goldsobel AB, Chipps BE. Cough in the Pediatric Population. J Pediatr. 2010;156(3):352–358.e1.
3 Theodore AC, Tseng CH, Li N, Elashoff RM, Tashkin DP. Correlation of cough with disease activity and treatment with cyclophosphamide in scleroderma interstitial lung disease: Findings from the scleroderma lung study. Chest. 2012;142(3):614–21.
4 Sato R, Handa T, Matsumoto H, Kubo T, Hirai T. Clinical significance of self-reported cough intensity and frequency in patients with interstitial lung disease: A cross-sectional study. BMC Pulm Med. 2019;19(1):1–10.
5 Li AM, Tsang TWT, Chan DFY, Lam HS, So HK, Sung RYT, *et al.* Cough frequency in children with mild asthma correlates with sputum neutrophil count. Thorax. 2006;61(9):747–50.
6 Van Der Giessen L, Loeve M, De Jongste J, Hop W, Tiddens H. Nocturnal cough in children with stable cystic fibrosis. Pediatr Pulmonol. 2009;44(9):859–65.
7 Cho PSP, Birring SS, Fletcher H V., Turner RD. Methods of Cough Assessment. J Allergy Clin Immunol Pract. 2019;7(6):1715–23.
8 Birring SS, Fleming T, Matos S, Raj AA, Evans DH, Pavord ID. The Leicester Cough Monitor: Preliminary validation of an automated cough detection system in chronic cough. Eur Respir J. 2008;31(5):1013–8.
9 McGuinness K, Holt K, Dockry R, Smith J. P159 Validation of the VitaloJAK 24 Hour Ambulatory Cough Monitor. Thorax. 2012;67(Suppl 2):A131–A131.
10 Chang AB. Pediatric cough: children are not miniature adults. Lung. 2010 Jan;188 Suppl:S33–40.
11 Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, *et al.* Guidelines for developing and reporting machine learning predictive models in biomedical research: A multidisciplinary view. J Med Internet Res. 2016;18(12):1–10.
12 Eyben F, Schuller B. openSMILE:). ACM SIGMultimedia Rec. 2015;6(4):4–13.
13 Morey MJ, Cheng AC, McCallum GB, Chang AB. Accuracy of cough reporting by carers of Indigenous children. J Paediatr Child Health. 2013;49(3). doi: 10.1111/jpc.12118
14 Chang AB, Newman RG, Carlin JB, Phelan PD, Robertson CF. Subjective scoring of cough in children: Parent-completed vs child-completed diary cards vs an objective method. Eur Respir J. 1998;11(2):462–6.
15 Kruizinga MD, Stuurman FE, Exadaktylos V, Doll RJ, Stephenson DT, Groeneveld GJ, *et al.* Development of Novel, Value-Based, Digital Endpoints for Clinical Trials: A Structured Approach Toward Fit-for-Purpose Validation. Pharmacol Rev. 2020;72 (4)(October):899–909.
16 Kruizinga MD, Stuurman FE, Groeneveld GJ, Cohen AF. The Future of Clinical Trial Design: The Transition from Hard Endpoints to Value-Based Endpoints. 2019;371–97.
17 Amrulloh YA, Abeyratne UR, Swarnkar V, Triasih R, Setyati A. Automatic cough segmentation from non-contact sound recordings in pediatric wards. Biomed Signal Process Control. 2015;21:126–36.
18 Coyle M, Keenan D, Henderson L, Watkins M, Haumann B, Mayleben D, *et al.* Evaluation of an ambulatory system for the quantification of cough frequency in patients with chronic obstructive pulmonary disease. Cough. 2005;1(1):3–3.
19 Vizel E, Yigla M, Goryachev Y, Dekel E, Felis V, Levi H, *et al.* Validation of an ambulatory cough detection and counting application using voluntary cough under different conditions. Cough. 2010;6(1):1–8.
20 Barata F, Tinschert P, Rassouli F, Steurer-Stey C, Fleisch E, Puhan MA, *et al.* Automatic recognition, segmentation, and sex assignment of nocturnal asthmatic coughs and cough epochs in smartphone audio recordings: Observational field study. J Med Internet Res. 2020;22(7). doi: 10.2196/18082
21 Monge-Alvarez J, Hoyos-Barcelo C, Lesso P, Casaseca-De-La-Higuera P. Robust Detection of Audio-Cough Events Using Local Hu Moments. IEEE J Biomed Heal Informatics. 2019;23(1):184–96.
22 Pramono RXA, Imtiaz SA, Rodriguez-Villegas E. A cough-based algorithm for automatic diagnosis of pertussis. PLOS One. 2016;11(9):1–20.
23 Turner RD, Bothamley GH. How to count coughs? Counting by ear, the effect of visual data and the evaluation of an automated cough monitor. Respir Med. 2014;108(12):1808–15.
24 Barata F, Kipfer K, Weber M, Tinschert P, Fleisch E, Kowatsch T. Towards device-agnostic mobile cough detection with convolutional neural networks. 2019 IEEE Int Conf Healthc Informatics, ICHI 2019. 2019;1–11.

SUPPLEMENTARY DATA

Sup. Text S1	Opensmile Audio features
Sup. Text S2	Opensmile Feature selection
Sup. Figure S2a	Example of distribution plots of each feature used during the feature selection process
Sup. Figure S3	Feature importance plot

CHAPTER 5

Technical validity and usability of a novel smartphone connected spirometry device for pediatric patients with asthma and cystic fibrosis

Pediatr Pulmonol. 2020 Sep;55(9):2463-2470. doi:10.1002/ppul.24932. Epub 2020 Jul 8

MD Kruizinga,[1,2,3,4] E Essers,[1,2] FE Stuurman,[1,4] A Zhuparris,[1] N van Eik,[2] HM Janssens,[3] I Groothuis,[2] AJ Sprij,[2] M Nuijsink,[2] AF Cohen,[1,4] GJA Driessen[2]

**Both authors contributed equally*

1 Centre for Human Drug Research, Leiden, the Netherlands
2 Juliana Children's Hospital, Haga Teaching Hospital, The Hague, the Netherlands
3 Department of Pediatrics, division of Respiratory Medicine and Allergology, Erasmus Medical Centre/Sophia Children's Hospital, University Hospital Rotterdam, the Netherlands
4 Leiden University Medical Center, Leiden, the Netherlands

Abstract

BACKGROUND Diagnosis and follow-up of respiratory diseases traditionally rely on pulmonary function tests (PFT), which are currently performed in hospitals and require trained personnel. Smartphone connected spirometers, like the Air Next spirometer, have been developed to aid in the home-monitoring of patients with pulmonary disease. The aim of this study was to investigate the technical validity and usability of the Air Next spirometer in pediatric patients.

METHODS Device variability was tested with a calibrated syringe. 90 subjects aged 6-16 were included in a prospective cohort study. 58 subjects performed conventional spirometry and subsequent Air Next spirometry. The bias and the limits of agreement between the measurements were calculated. Furthermore, subjects used the device for 28 days at home and completed a subject satisfaction questionnaire at the end of the study period.

RESULTS Inter-device variability was 2.8% and intra-device variability was 0.9%. The average difference between the Air Next and conventional spirometry was 40 mL for FEV1 and 3 mL for FVC. The limits of agreement were -270 mL and +352 mL for FEV1 and -403 mL and +397 mL for FVC. 45% of FEV1 measurements and 41% of FVC measurements at home were acceptable and reproducible according to ATS/ERS criteria. Parents scored difficulty, usefulness and reliability 1.9, 3.5 and 3.8 out of 5, respectively.

CONCLUSION The Air Next device shows validity for the measurement of FEV1 and FVC in a pediatric patient population.

Introduction

Diagnosis and longitudinal follow-up of pulmonary diseases have relied on pulmonary function tests (PFTS) since the nineteenth century.[1] Traditionally conducted in the clinic, spirometry can be a difficult technique, and the accuracy and repeatability depends on many factors such as equipment, patient effort, and supervision and encouragement of a technician. Nevertheless, a single pulmonary function test is no more than a snapshot of disease-activity and is unable to capture the variability of symptoms in chronic pulmonary disease. Longitudinal data on a regular basis regarding pulmonary health could be very valuable for patients, clinicians, and clinical researchers, and this could be obtained by performing PFTS at the patients' home. An increase in readily available objective longitudinal data could be particularly useful in pediatrics, as children often find it difficult to perceive and express the severity of their symptoms.[2,3]

Researchers have investigated the clinical value of home-based measurements of several devices for pediatric asthma and cystic fibrosis (CF). While outcomes were correlated to disease activity,[4] the devices appeared to offer little benefit for clinical practice in terms of reduced admission rates, better disease control, or slower decline in pulmonary function.[5-7] Since then, improvements in technology have allowed for the development of devices for measurement of complete flow-volume curves at relatively low-cost. An example is the Air Next spirometer, a Bluetooth connected device allowing patients to perform spirometry tests with a smartphone. Use of the device has been reported in adult patients, but not yet in the pediatric population.[8,9] Before implementation in pediatric clinical care or clinical trials, a comprehensive technical validation of the device must be performed, consisting of the assessment of intra- and inter-device variability, comparison with conventional spirometry, as well as the assessment of usability for pediatric patients.

The aim of this study is to determine the agreement between the Air Next spirometer and conventional spirometry and to evaluate the usability of the device for children and parents when used at home.

Materials & Methods

Location and ethics

This study was conducted at the Juliana Children's Hospital (Haga teaching hospital, the Hague, the Netherlands) and Sophia Children's Hospital (Erasmus Medical Centre,

Rotterdam, the Netherlands) from November 2018 until January 2020. The study protocol was reviewed and approved by the Medical Ethics Committee ZuidWest Holland (the Hague, the Netherlands) prior to initiation of the study. The study was conducted according to the Dutch Act on Medical Research Involving Human Subjects (WMO) and in compliance with Good Clinical Practice. Written informed consent was obtained from all parents and children aged 12 years and older. Assent was obtained from children aged younger than 12. The trial was registered at the Dutch Trial Registry (NTR, Trial NL7611).

Subjects and study design

This analysis was part of a study investigating a novel home-monitoring platform (CHDR MORE®) in pediatrics. During this study, pediatric patients with controlled asthma (n=30), uncontrolled asthma (n=30) and cystic fibrosis (n=30) were recruited from the outpatient clinic of the hospitals. All children were aged between six and sixteen years. Asthma control was defined using the GINA criteria and Asthma Control Questionnaire (ACQ, cutoff > 1.5 points).[10,11] Children and parents were given a 10-minute training and practice session and were asked to perform once daily pulmonary function tests with the mobile device for a duration of 28 days. When logistically feasible, children visited the hospital to perform a conventional spirometry test at the outpatient clinic at the beginning or end of the study period and perform an Air Next spirometry test during the same visit. The sequence of tests was chosen based on preference for each patient.

Spirometry

Conventional spirometry was performed on a Masterscreen PFT (Vyaire, Mettawa, IL, USA) at the Juliana Children's Hospital and the Sophia Children's Hospital, calibrated according to ATS/ERS guidelines. Home-based spirometry was performed using the Air Next spirometry device (NuvoAir, Stockholm, Sweden). The device employs a turbine mechanism with disposable mouthpieces and cannot be calibrated by the user. The device uses Bluetooth to connect to a smartphone. Motorola G6 (Motorola, Chicago, IL, USA) phones were used during the study. An accompanying application was installed, which uses age, sex and height to calculate reference values according to the GLI-2012 equations[12], and requires Android 5.0 or higher. The application provides the Forced Expiratory Volume in the first second (FEV1), Forced Vital Capacity (FVC), FEV1/FVC ratio and Peak flow (PEF) per maneuver.

Device variability

The Air Next device cannot be manually calibrated. We used a calibrated syringe (Viasys, Conshohocken, PA, USA) with a capacity of 2994 mL to evaluate accuracy and the inter- and intra-device variability. The syringe was used to push the complete capacity through an Air Next device 20 times per device on 20 devices with a single turbine. Additionally, the syringe was used on 25 different turbines with a single Air Next device.

Test procedures

American Thoracic Society (ATS) and European Respiratory Society (ERS) acceptability guidelines were used to judge and grade PFT quality (grade A-F from best to worst).[13] Spirometry maneuvers were acceptable if the start was rapid and without hesitation, the course of the expiratory maneuver was continuous, without any artefacts or evidence of coughing in the first second and if the end of the maneuver did not show early or abrupt interruption. The difference between the best two acceptable FVC and FEV1 should have been less than 150 mL. At least three maneuvers were performed per spirometry session. When it was difficult to obtain reproducible maneuvers during supervised measurements, a maximum of 10 maneuvers per patient was performed and the usable maneuvers were used. For home-use, subjects were instructed to perform three maneuvers per session and were able to perform 2 additional measurements when appropriate (e.g. misstiming of the forced exhalation or application errors). Subjects were not asked to self-grade repeatability during the study period.

End-of-study questionnaire

At the end of the study period, a questionnaire regarding user experience was completed. Parents and participants were asked to give their opinion about the reliability of the device, the difficulty of using the device, and whether they found the use of the device to be useful or tedious on a 5 point Likert scale.

Statistics

Baseline characteristics were summarized. Inter-, intra- and turbine variability were calculated and expressed as a coefficient of variability (CV). Concordance between Air Next spirometry and conventional spirometry was assessed using the methods described by

Bland and Altman.[14] The mean differences between methods and the 95% limits of agreement were calculated for FEV1, FVC, PEF and FEV1/FVC ratio. For FEV1 and FVC, acceptable bias was no more than 100 mL. For PEF and FEV1/FVC ratio, the acceptable average bias was 300 mL/s and 10% respectively.[13,15] Pearson correlation coefficients between the two methods were calculated. Spirometry measurements at home were graded for quality and the number of maneuvers assigned to each grade were summarized descriptively. A mean grade per subject was calculated. The average mean grades of the three study groups were compared via a one-way ANOVA test and pairs were compared with Tukey's range test to adjust for multiple comparisons. Usability was evaluated by analyzing the end-of-study questionnaire completed by subjects and their parents. R version 3.5.1 was used for statistical analysis and visualization. Promasys® software (OmniComm, Ft. Lauderdale, FL, USA) was used for data management.

Results

Baseline characteristics

A total of 90 subjects were included in the main study. The average age was 10 years (range 6-15). Subjects had performed an average of 12 (SD 11) hospital-based pulmonary function tests before the study. Other baseline characteristics are displayed in *Table 1*.

Table 1. Baseline characteristics

	All participants (n = 90)	Comparison participants* (n = 58)
Age (mean (SD)	10.2 (2.7)	10.2 (2.7)
SEX		
Male, n (%)	54 (60)	37 (65)
Female, n (%)	36 (40)	20 (35)
DIAGNOSIS		
Controlled asthma, n (%)	30 (33.3)	27 (47)
Uncontrolled asthma, n (%)	30 (33.3)	23 (40)
Cystic Fibrosis, n (%)	30 (33.3)	7 (12)
Weight (kg), mean (SD)	39.5 (15.9)	40.8 (16.2)
Body mass index (SDS), mean (SD)	0.6 (1.4)	0.8 (1.4)
Height (m), mean (SD)	144.1 (16.6)	144 (15.5)
ETHNICITY		
Caucasian, n (%)	69 (77)	37 (74)
Other, n (%)	21 (23)	15 (26)
Spirometry experience, n (SD)	12.2 (11)	8.4 (8)

* Comparison participants: patients who also performed conventional spirometry before or at the end of the study period.

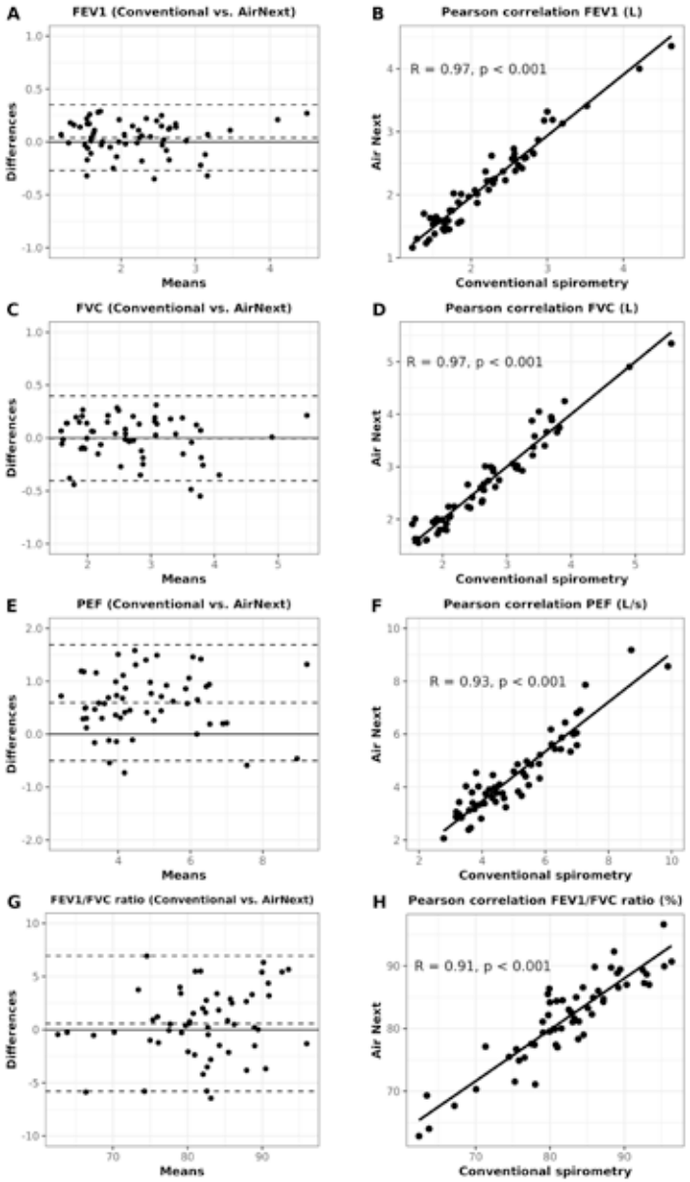
Device variability

Of 400 measurements in 20 devices, the average bias from the calibrated 2994 mL was -40 mL (range -124 - 56 mL). The average intra-device CV was 0.9% (range 0.6-1.2%). Furthermore, the average inter-device CV was 2.8%. Average turbine bias was -70 mL and turbine CV was 1.8%. 4% of measurements with the calibrated syringe exceeded the 3% accuracy threshold advised by ATS standards.

Measurement validity

58 subjects were able to perform hospital and Air Next PFTS subsequently. When comparing output between the two methods, there was one extreme outlier, most likely due to a technical defect resulting in a blockage of the outflow of the Air Next turbine, which was excluded from the statistical analysis. *Figure 1* shows the limits of agreement and correlation between the Air Next and conventional spirometry of the several parameters. For FEV1, the average bias was 40 mL and the 95% limits of agreement were -270 and +352 mL. The Pearson correlation coefficient (R) was 0.97 ($p < 0.001$). The bias of FVC was 3 mL with limits of agreement of -403 mL and +397 mL ($R = 0.97$, $p < 0.001$). Furthermore, the analysis of PEF demonstrated an average difference of 590 mL/s (95% limits of agreement of -500 mL and 1690 mL) and the average difference for the FEV1/FVC ratio was 0.6% (95% limits of agreement of -5.8% and 7.0%). Although the correlation coefficient was lower as compared to FEV1 and FVC, there was still a good correlation between the two methods for both PEF ($R = 0.93$, $p < 0.001$)) and FEV1/FVC ratio ($R = 0.91$, $p < 0.001$). There was no proportional bias for any of the parameters. There was a correlation ($R = -0.33$, $p = 0.01$ for FEV1, $R = -0.26$, $p = 0.05$ for FVC) between the absolute difference in FEV1 and FVC (expressed in % of predicted FEV1 and FVC) and age (*Supplementary Figure S1*), but not between the absolute difference and previous spirometry experience, expressed as the amount of PFTS performed in the past (*Supplementary Figure S2*). There was no statistically significant difference in absolute bias for FEV1 between the three groups ($p = 0.28$, *Supplementary Figure S3*). When the absolute difference between the two methods was expressed as a percentage of the predicted FEV1 and FVC, the mean bias was 6.3% (SD 5%) of predicted FEV1 and 6.7% (SD 5.7%) of predicted FVC. The bias of FEV1 of subjects who performed the comparison at the end of the study period was slightly higher (3% of predicted, $p = 0.009$) compared to subjects who performed the comparison at the beginning of the study period (*Supplementary Figure S4*).

Figure 1. Concordance between Air Next and conventional spirometry. A, C, E, G: Bland-Altman plots displaying the differences between conventional spirometry and Air Next spirometry against the averages of the two techniques for FEV1, FVC, FEV1/FVC ratio and PEF, respectively. Dotted lines reflect the average bias (middle line) and the 95% limits of agreement (outer lines). B, D, F, H: Pearson correlation between the two measurements.



Technique and day-to-day variability

A total of 2047 spirometry measurements were performed with the Air Next device during the study, resulting in an average compliance of 78%. The curves of 1821 sessions were available for analysis. When graded according to the ATS/ERS criteria, 45% of the FEV1 measurements were considered acceptable and reproducible, as well as 41% of the FVC measurements. A significant number of sessions were grade E, meaning they did not produce more than one acceptable maneuver or that the reproducibility was too low. 2% of measurements were neither acceptable nor usable for both FEV1 and FVC. Summarized graded are listed in *Figure 2A* and *Figure 2B*. There was a statistically significant difference in average grade between CF patients and patients with uncontrolled asthma (FEV1, p = 0.02 (*Figure 2C*), FVC, p = 0.03 (*Figure 2D*)). Age and average grade were not correlated (*Supplementary Figure S5*). Day-to-day coefficient of variability (CV) of acceptable trials (grade A-C) was 9.0% (SD 5.7%) for FEV1 and 7.7% (SD 5.4%) for FVC.

Usability

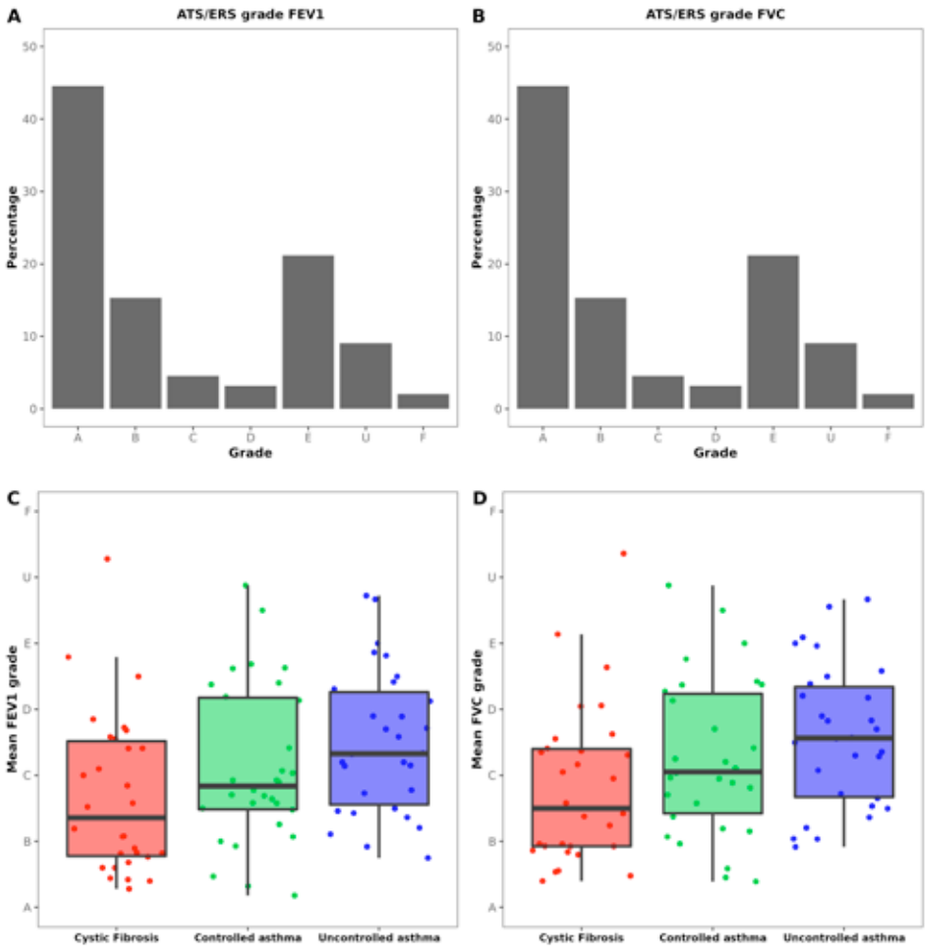
69 (77%) subjects completed the end-of-study questionnaire. In general, parents found the use of the spirometry device to be acceptable. When asked to score their agreement with the statement 'I found the use of the spirometer to be tedious', the average score was 1.8 out of 5 (SD 1.1). Furthermore, parents scored the difficulty 1.9 out of 5 (SD 1.2), usefulness 3.5 out of 5 (SD 0.9) and the perceived reliability 3.3 out of 5 (SD 1.0). Summarized results are displayed in *Supplementary Figure S6*.

Discussion

The current study investigates the technical validity and user experience of the Air Next spirometer for pediatric patients. Air Next spirometer output was compared to the gold standard: conventional spirometry in the clinic. Subjects and their parents also completed a questionnaire regarding the usability of the device.

The inter-device, intra-device and turbine variability were assessed with a calibrated syringe of 2994 mL. All of the measurements were within 125 mL of the reference. Although 125 mL exceeds the 3% accuracy standard advised by the ATS, 96% of measurements fell within the 3% range. The coefficients of variability were all below 3%, which suggests that the repeatability of the device is good.

Figure 2. ERS/ATS grades for measurements performed at home. All spirometry sessions were graded according to ATS/ERS guidelines for FEV1 and FVC separately. Grade A–E represent sessions with acceptable maneuvers but with varying repeatability. Grade U includes session with usable but not with acceptable maneuvers and grade F is reserved for session without acceptable or usable maneuvers. A: proportion of spirometry sessions that were awarded each grade for FEV1. B: proportion of spirometry sessions that were awarded each grade for FVC. C: Boxplot of average FEV1 grade per study group. Dots represent individual averages. There was a statistically significant difference between the CF and uncontrolled asthma group ($p = 0.02$). D: Boxplot of average FVC grade per study group. Dots represent individual averages. There was a statistically significant difference between the CF and uncontrolled asthma group ($p = 0.03$).



Bland-Altman plots displaying the difference between the Air Next measurements and conventional spirometry demonstrated a negligible bias for FEV1, FVC and FEV1/FVC ratio of 40 mL, 3 mL and 0.6% respectively. Furthermore, the 95% limits of agreement for FEV1 and FVC comparable with earlier studies in adults.[8,9] Both FEV1/FVC ratio and PEF showed relatively wide limits of agreement compared to conventional spirometry, while PEF demonstrated bias compared to the gold standard. Interestingly, concordance of PEF was not reported in earlier publications. While the grades of the supervised spirometry sessions with the Air Next were all adequate (A–C) according to ATS/ERS criteria, we suspect the individual differences of FEV1 and FVC measurements, and the consistently lower PEF of the Air Next measurements to be mainly due to differences in technique. Subjects had to coordinate several actions in quick succession: initiating the smartphone application, complete a full forced inspiration, perform a controlled arm movement towards the mouth and finally complete a forced expiration. This is a relatively complex sequence of actions compared to conventional supervised spirometry and could influence the maximum effort given to the forced expiration. The complexity of the sequence of actions may also explain the correlation between absolute difference in FEV1 and age. For most subjects, the spirometry session for comparison was the first time they used the Air Next device. However, more familiarity with the technique did not appear to lead to better concordance, considering the observation that children who performed the comparison at the end of the study period did not exhibit a smaller deviation from conventional spirometry. We hypothesize this may be due a decrease in motivation in children who performed daily PFTS during the preceding 28 days. Another important difference that may explain discordance is that small devices exert low resistance to expiration in comparison to conventional devices, which may affect the way children perform PFTS. While the bias of 0.59 L casts doubt on the absolute accuracy of the device for PEF measurements, the FVC, FEV1/FVC ratio and especially FEV1 are considered to be more important parameters of pulmonary health.[16] Furthermore, the measured PEF may show good correlation with symptom severity in the case of home-monitoring. The limits of agreement for FEV1 and FVC are wider than the bias of the Air Next device determined with the calibrated syringe. This suggests that individual differences between the Air Next and conventional spirometry are the result of bias by both the patient, and the device. A subgroup analysis of 25 children who displayed good technique in the home setting (median grade A–B) showed slightly smaller limits of agreement. (*Supplementary Figure S7*). Limits of agreement of this magnitude are inherent to direct comparisons of spirometers, as demonstrated by the literature on this subject.[17–20]

Still, the relevance of the individual differences of this magnitude is higher in pediatrics, because their smaller expected lung volumes lead to biases that may be clinically relevant.

Subjects used the device at home for 28 consecutive days in the main study. Individual curves were assessed and graded according to the ERS/ATS criteria. The majority of measurements would be considered suitable for further analysis, but 36% of FEV1 measurements and 39% of FVC measurements were graded D, E, U or F, meaning that they were not performed technically adequate.[21] Interestingly, patients with uncontrolled asthma appeared to exhibit worse technique than patients with CF. Several sessions with poor technique could have been the result of dyspnea due to the underlying disease, and the obtained values for FEV1, FVC and PEF could still correlate well with perceived symptoms. However, the difference in technique could also be explained by the fact that children with CF perform a PFT every three months, which results in more familiarity with the technique. Therefore, this observation could also indicate a need for more training sessions, which has been reported to be beneficial for improving inhalation technique.[22] Extensive training could be beneficial for home-based spirometry as well and could be investigated further during a clinical validation study. Although the acceptability criteria that were the cause of a maneuver being unacceptable were not routinely recorded, the unacceptable maneuvers most often did not reach the end of forced expiration criteria. A high back-extrapolation volume was encountered often as well. Both are indicators of insufficient effort during the end and start of the maneuver, respectively.[13]

According to the end-of-study questionnaire, parents and children did not find the measurements to be difficult, although this assessment may change when immediate feedback on the quality of the measurements is provided. During the study, some participants had recurrent Bluetooth connectivity problems, which may be related to the used phone or the particular device that was used. To optimize reliability and usability, more intensive training and strict instructions may be necessary. During this study, participants underwent a 10-minute training, which may not be enough to prevent wrong conduct. Still, issues such as low motivation, technological glitches, or even something as trivial as blocking air inflow with the tongue or air outflow with the hands are difficult to avoid completely without the supervision of a trained technician. This was demonstrated by the extreme outlier excluded in our analysis. Issues such as these may cause false positive or false negative results when used for the remote diagnosis of pulmonary obstruction.

Nevertheless, when correctly performed, the Air Next demonstrates reliability for FEV1 and FVC measurements compared to conventional spirometry and with a good user experience. In clinical care, the device could support home monitoring and provide timely information to patients when to contact a doctor. Furthermore, the device can be used for the purpose of telemedicine, which may be increasingly used during and after the crisis precipitated by the COVID-19 pandemic. Although previous studies have indicated home-based spirometry does not add value to pediatric clinical care, this may change when combined with other assessments, such as a symptom questionnaire[23], a wearable device, or other monitoring techniques.[24] This may help physicians to improve monitoring of pediatric patients, while reducing the burden of disease. In addition, with the increasing popularity of digital endpoints and decentralized clinical trials, the device could play an important role in future clinical trials for pediatric CF, asthma and other pulmonary diseases, which could decrease the burden of clinical trial participation. Finally, the device may be useful for primary care physicians without access to conventional spirometers in low-income countries or rural areas, or at the point of care in patients' homes.

This study has some limitations, one of which is that not all the participants could be included in the validation group. This is mainly due to logistical reasons and the fact that the comparison was part of a secondary analysis of a clinical study. However, there were no large differences in baseline characteristics between the complete cohort and the validation cohort (*Table 1*). The non-randomized order of tests may have influenced the results through spirometry-induced bronchoconstriction.[25] However, we did not diagnose this condition in any of the included subjects. The curves of 226 spirometry sessions were unavailable for review due to application connectivity errors. However, this issue occurred at random and therefore did not impact our overall conclusions. Although we found no correlation between the absolute bias and previous spirometry experience when comparing conventional spirometry to the Air Next, the proportion of highly experienced subjects was low. A higher number of experienced subjects may have resulted in a better correlation. A strength of the study is the inclusion of pediatric patients with controlled asthma, uncontrolled asthma, and CF, giving a representative sample of possible pediatric target populations. The manufacturer has unlocked additional functions of the device since the initiation of this study, allowing for the measurement of the inspiratory measurements FIVC, PIF, MIF and MEF. These functionalities should be independently validated before integration in clinical care or clinical trials. Future clinical validation of home-based measurements with the Air Next will be performed to determine the objectivity and reproducibility of longitudinal unsupervised measurements.

Conclusion

The Air Next spirometer is technically valid for the measurement of FEV1 and FVC in children aged 6 to 16, while PEF measurements show significant bias. The user experience was considered favorable by subjects and their parents. FEV1 and FVC measured at home could add significant value to clinical care and clinical trials, but future studies should determine the clinical value of home-based spirometry measurements for the purpose of monitoring disease-activity or response to treatment, possibly in combination with other home-based measurements.

SUPPLEMENTARY DATA

Sup. Figure S1 Correlation between age and observed absolute difference
Sup. Figure S2 Correlation between spirometry experience and observed absolute difference
Sup. Figure S3 Absolute difference between conventional spirometry and Air Next spirometry by group
Sup. Figure S4 Absolute difference between conventional spirometry and Air Next spirometry - Performed at start of study versus the end of study
Sup. Figure S5 Correlation between average ats/ers grade and age
Sup. Figure S6 End-of-study questionnaire results
Sup. Figure S7 Subgroup analysis of children with median grade B or better

REFERENCES

1 Hutchinson J. Contributions to Vital Statistics, Obtained by Means of a Pneumatic Apparatus for Valuing the Respiratory Powers with Relation to Health. J Stat Soc London 1844;7(3):193.
2 Baker RR, Mishoe SC, Zaitoun FH, Arant CB, Lucas J, Rupp NT. Poor perception of airway obstruction in children with asthma. J Asthma 2000;37(7):613-624.
3 Forno E, Abraham N, Winger DG, Rosas-Salazar C, Kurland G, Weiner DJ. Perception of Pulmonary Function in Children with Asthma and Cystic Fibrosis. Pediatr Allergy, Immunol Pulmonol 2018;31(3):139-145.
4 Brouwer AFJ, Roorda RJ, Brand PLP. Home spirometry and asthma severity in children. Eur Respir J 2006;28(6):1131-1137.
5 Deschildre A, Béghin L, Salleron J, Iliescu C, Thumerelle C, Santos C, Hoorelbeke A, Scalbert M, Pouessele G, Gnansounou M, *et al.* Home telemonitoring (forced expiratory volume in 1 s) in children with severe asthma does not reduce exacerbations. Eur Respir J 2012;39(2):290-296.
6 Shakkottai A, Kaciroti N, Kasmikha L, Nasr SZ. Impact of home spirometry on medication adherence among adolescents with cystic fibrosis. Pediatr Pulmonol 2018;53(4):431-436.
7 Lechtzin N, Mayer-Hamblett N, West NE, Allgood S, Wilhelm E, Khan U, Aitken ML, Ramsey BW, Boyle MP, Mogayzel PJ, *et al.* Home monitoring of patients with cystic fibrosis to identify and treat acute pulmonary exacerbations eICE study results. Am J Respir Crit Care Med 2017;196(9):1144-1151.
8 Ramos Hernández C, Núñez Fernández M, Pallares Sanmartín A, Mouronte Roibas C, Cerdeira Domínguez L, Botana Rial MI, Blanco Cid N, Fernández Villar A. Validation of the portable Air-Smart Spirometer. PLOS One 2018;13(2):e0192789.
9 Plessis E Du, Swart F, Maree D, Van Heerden J, Esterhuizen TM, Irusen EM, Koegelenberg CFN. The utility of hand-held mobile spirometer technology in a resource-constrained setting. South African Med J 2019;109(4):219-222.
10 Juniper EF, Bousquet J, Abetz L, Bateman ED. Identifying "well-controlled" and "not well-controlled" asthma using the Asthma Control Questionnaire. Respir Med 2006;100(4):616-621.
11 Juniper EF, Gruffydd-Jones K, Ward S, Svensson K. Asthma control questionnaire in children: Validation, measurement properties, interpretation. Eur Respir J 2010;36(6):1410-1416.
12 Quanjer PH, Cole TJ, Hall GL, Culver BH. Report of the Global Lung Function Initiative (GLI), ERS Task Force to establish improved Lung Function Reference Values, includng supplement. Eur Respir J 2013;40(6):1324-1343.
13 Graham BL, Steenbruggen I, Barjaktarevic IZ, Cooper BG, Hall GL, Hallstrand TS, Kaminsky DA, McCarthy K, McCormack MC, Miller MR, *et al.* Standardization of spirometry 2019 update an official American Thoracic Society and European Respiratory Society technical statement. Am J Respir Crit Care Med 2019;200(8):E70-E88.
14 Altman DG, Bland JM. Measurement in Medicine: The Analysis of Method Comparison Studies. J R Stat Soc Ser D (The Stat 1983;32(3):307-317.
15 Derom E, Van Weel C, Liistro G, Buffels J, Schermer T, Lammers E, Wouters E, Decramer M. Primary care spirometry. Eur Respir J 2008;31(1):197-203.
16 Tepper RS, Wise RS, Covar R, Irvin CG, Kercsmar CM, Kraft M, Liu MC, O'Connor GT, Peters SP, Sorkness R, *et al.* Asthma outcomes: Pulmonary physiology. J Allergy Clin Immunol 2012;129(3 SUPPL.):S65-S87.
17 Hernández CR, Fernández MN, Sanmartín AP, Roibas CM, Domínguez LC, Rial MIB, Cid NB, Villar AF. Validation of the portable Air-Smart Spirometer. PLOS One 2018;13(2):1-11.
18 Liistro G, Vanwelde C, Vincken W, Vandevoorde J, Verleden G, Buffels J. Technical and functional assessment of 10 office spirometers: A multicenter comparative study. Chest 2006;130(3):657-665.
19 Richter K, Kanniess F, Mark B, Jörres RA, Magnussen H. Assessment of accuracy and applicability of a new electronic peak flow meter and asthma monitor. Eur Respir J 1998;12(2):457-462.
20 Avdimiretz N, Wilson D, Grasemann H. Comparison of a handheld turbine spirometer to conventional spirometry in children with cystic fibrosis. Pediatr Pulmonol 2020;(March):1394-1399.
21 Culver BH, Graham BL, Coates AL, Wanger J, Berry CE, Clarke PK, Hallstrand TS, Hankinson JL, Kaminsky DA, MacIntyre NR, *et al.* Recommendations for a standardized pulmonary function report. An official American Thoracic Society technical statement. Am J Respir Crit Care Med 2017;196(11):1463-1472.
22. Kamps AWA, Brand PLP, Roorda RJ. Determinants of correct inhalation technique in children attending a hospital-based asthma clinic. Acta Paediatr Int J Paediatr 2002;91(2):159-163.
23 Van Den Wijngaart LS, Roukema J, Boehmer ALM, Brouwer ML, Hugen CAC, Niers LEM, Sprij AJ, Rikkers-Mutsaerts ERVM, Rottier BL, Donders ART, *et al.* A virtual asthma clinic for children: Fewer routine outpatient visits, same asthma control. Eur Respir J 2017;50(4):1-10.
24 Kruizinga MD, Stuurman FE, Groeneveld GJ, Cohen AF. The future of clinical trial design: The transition from hard endpoints to value-based endpoints. Handb Exp Pharmacol 2019;260:371-397.
25 Gimeno F, Berg WCHR, Sluiter HJ, Tammeling GJ. Spirometry-Induced Bronchial Obstruction'. Am Rev Respir Dis 1972;105(8):68-74.

PART III

CLINICAL VALIDATION OF DIGITAL ENDPOINTS

CHAPTER 6

Towards remote monitoring in pediatric care and clinical trials - Tolerability, repeatability and reference values of candidate digital endpoints derived from physical activity, heart rate and sleep in healthy children

PLoS One. 2021 Jan 7;16(1):e0244877. doi: 10.1371/journal.pone.0244877

MD Kruizinga[1,2,3], N van der Heide[1,2], A Moll[1,2], A Zhuparris[1], Y Yavuz[1], ML de Kam[1], FE Stuurman[1,3], AF Cohen[1,3], GJA Driessen[2,4]

1 Centre for Human Drug Research, Leiden, the Netherlands
2 Juliana Children's Hospital, Haga Teaching Hospital, The Hague, the Netherlands
3 Leiden University Medical Center, Leiden, the Netherlands
4 Maastricht University Medical Centre, Maastricht, the Netherlands

Abstract

BACKGROUND Digital devices and wearables allow for the measurement of a wide range of health-related parameters in a non-invasive manner, which may be particularly valuable in pediatrics. Incorporation of such parameters in clinical trials or care as digital endpoint could reduce the burden for children and their parents but requires clinical validation in the target population. This study aims to determine the tolerability, repeatability, and reference values of novel digital endpoints in healthy children.

METHODS Apparently healthy children (n=175, 46% male) aged 2-16 were included. Subjects were monitored for 21 days using a home-monitoring platform with several devices (smartwatch, spirometer, thermometer, blood pressure monitor, scales). Endpoints were analyzed with a mixed effects model, assessing variables that explained within- and between-subject variability. Endpoints based on physical activity, heart rate, and sleep-related parameters were included in the analysis. For physical-activity-related endpoints, a sample size needed to detect a 15% increase was calculated.

FINDINGS Median compliance was 94%. Variability in each physical activity-related candidate endpoint was explained by age, sex, watch wear time, rain duration per day, average ambient temperature, and population density of the city of residence. Estimated sample sizes for candidate endpoints ranged from 33-110 per group. Daytime heart rate, nocturnal heart rate and sleep duration decreased as a function of age and were comparable to reference values published in the literature.

CONCLUSIONS Wearable- and portable devices are tolerable for pediatric subjects. The raw data, models and reference values presented here can be used to guide further validation and, in the future, clinical trial designs involving the included measures.

Introduction

Despite several initiatives to stimulate pediatric clinical trial initiation and conduct, the proportion of pediatric trials is 9%[1], even though children are the recipient of 25% of the global disease burden[2]. Recruitment and ethical barriers are often cited as a cause by investigators, while fear for invasive measurements and logistical difficulties, such as school schedules for children and demanding jobs for parents are burdens on the participant's side[3]. However, the improvement of wireless and portable technology may allow decentralized trials by using health devices in a home-setting. The digital endpoints included in such trials could significantly reduce the burden for children and their parents.

Before integration in clinical trials, digital endpoints must be validated fit-for-purpose[4]. During clinical validation, the tolerability and repeatability of the candidate endpoint must be determined.

Furthermore, an important criterion for novel endpoints is a clinically significant difference between patients and healthy controls. In order to determine this in pediatric subjects, data from a large healthy cohort with a wide age range is paramount. While multiple feasibility studies have been conducted with wearables in various age-groups[5–7], no large-scale studies with a wide age-range have been conducted to investigate the measurements in a home-setting. Additionally, there is little pediatric data available regarding the impact of variability that is introduced by performing measurements in free-living conditions. Factors such as school, weather, city of residence and regularly playing sports are likely to influence digitally captured measures like physical activity and heart rate (HR).

It is established that chronically ill children exhibit a lower activity level compared to their healthy peers, and it is plausible that easy to obtain measurements like physical activity, HR, or sleep parameters are correlated to symptom severity in both acutely- and chronically ill children[8]. However, there is a need to translate raw wearable- and sensor data into novel clearly defined endpoints that are validated and sensitive to detect a change in disease severity in children[9]. Validated non-invasive digital endpoints with proven worth for the purpose of monitoring disease-activity could then not only be used in clinical research, but in clinical care as well. For example, home-monitoring of asthma symptoms via an electronic questionnaire has been shown to reduce the need for outpatient clinic visits[10], and including objective measurements with wearable- or portable devices may improve the reliability of home-monitoring even more.

The aim of this study is to investigate the tolerability of a combination of digital health devices via a remote clinical trial regimen, to obtain reference values in healthy children for said devices, to assess conditions that induce variability in free-living conditions and to explore novel features that could be used as candidate digital endpoint in future clinical trials.

Materials & Methods

Location and ethics

This study was conducted at the Juliana Children's Hospital (Haga teaching hospital, the Hague, the Netherlands) and the Centre for Human Drug Research (Leiden, the Netherlands) from November 2018 until March 2020. The study protocol was approved by the Medical Ethics Committee ZuidWest Holland (the Hague, the Netherlands) prior to initiation of the study. The study was conducted in compliance with the Dutch Act on Medical Research Involving Human Subjects (WMO) and Good Clinical Practice. Written informed consent was obtained from all parents and from children aged 12 years and older. Assent was obtained from children aged younger than 12. The trial was registered at the Dutch Trial Registry (NTR, Trial NL7612, registered 18-Mar-2019).

Subjects and study design

Between 10-15 healthy children of each age year between 2 and 16 years old were recruited from the region around the Hague using various recruitment methods (local newspaper advertisement, distributing flyers in the Juliana Children's Hospital and on a local primary school). Exclusion criteria were the presence of (chronic) disease, inability to communicate in Dutch or English and an inability to use the devices. All children were visited at home by a trained investigator for a 30-minute training session and a baseline questionnaire. Children used the devices and completed a daily activity questionnaire for the subsequent 21 days, after which an end-of-study questionnaire was completed, and the devices were retrieved by the investigators. Children received a gift certificate worth €20 for their participation.

Devices

Subjects wore a Steel HR smartwatch (Withings, Issy-les-Molineux, France) during the study period. The watch measures physical activity with a built-in accelerometer. HR was measured every 10 minutes via a photo plethysmography (PPG) sensor on the back of the watch. Furthermore, the watch calculates several sleep-related parameters using the accelerometer and an incorporated temperature sensor. All subjects performed twice weekly temperature measurements with the Withings Thermo device. Subjects ≥ 6 years old performed daily blood pressure (BP) measurements with the Withings BPM, a weekly weight measurement with the Withings Body+ scales and twice-weekly home-based spirometry using the Air Next spirometry device (NuvoAir, Stockholm, Sweden). The Air Next device employs a turbine mechanism with disposable mouthpieces and has been validated for use in children[11]. All devices used Bluetooth to connect to a smartphone (Motorola G6 (Motorola, Chicago, IL, USA)). The smartphone had the Withings Healthmate application, the CHDR MORE® application (used for data collection and aggregation) and an electronic patient reported outcome (EPRO) application pre-installed.

Baseline- and environmental data

Parents of all subjects completed the PedsQL 4.0 questionnaire and provided several baseline variables at the start of the study[12]. The population density of the Children's city of residence was classified using publicly available data of the Dutch Central Office of Statistics. Local weather statistics from a local weather station (Hoek van Holland, the Netherlands) were obtained from the Royal Dutch Meteorological Institute.

Analysis

BASELINE CHARACTERISTICS AND COMPLIANCE Baseline characteristics were summarized. Compliance, an important indicator of the tolerability of the trial regimen, was calculated for each subject individually by dividing the amount of observations in the dataset by the amount of expected observations (calculated per assessment). The same calculation was performed for all assessments together to calculate an overall compliance per subject. For the measurements that were not performed daily, a subject was considered noncompliant when a measurement time point deviated more than 1 day for

spirometry and temperature assessments and more than two days for weight assessments. The median and interquartile range of the compliance within the complete cohort was calculated for each assessment and for the overall compliance. For assessments performed daily, the compliance over time was estimated by calculating the compliance for each individual study day (day 1 through day 21). The proportion of daily watch wear time per day was calculated using aggregated data per hour. An hour with no registered HR and step count data counted as noncompliant ('not worn'), and an hour with either a registered HR or a registered step counted as compliant ('worn'). The proportion of the wear time between 6AM and 10PM was calculated to include as a covariate when analyzing step count data. Data from screening days (n=175) and all days with a watch wear time < 50% between 6AM and 10PM were excluded from analyses regarding daytime measurements (146 study days), while all days with a watch wear time < 50% between 0AM and 5AM (268 study days) were excluded from analyses regarding nocturnal measurements. This 50% threshold has been chosen in earlier studies as well[13,14].

STATISTICAL ANALYSIS OF CANDIDATE ENDPOINTS Candidate endpoints were analyzed with a linear mixed effects model with subject as random factor. Factors that were expected to explain variability were considered as fixed factor or covariate. Spline regression with 2 or 3 degrees of freedom was considered when nonlinear relationships were a possibility. Contribution to model fit was assessed by comparing the Akaike information criterion (AIC) of models and by likelihood ratio tests. A p-value smaller than 0.05 was considered statistically significant. Due to the exploratory nature of the study, no adjustment for multiple comparison was performed. Instead, covariate inclusion was guided by appraising the ΔAIC between candidate models. Interactions between factors or covariates were considered when biologically plausible. Residual plots were inspected for heteroscedasticity and non-normality and logarithmic or square root transformations were considered when assumptions were violated. However, mild non-normality of residuals was accepted due to the large size of the dataset[15]. Marginal effects and the 95% confidence interval of the effect were plotted for each variable in the final model. For selected candidate endpoints, a 90% prediction interval was constructed including random effect variance, in order to display cut-off values for the lowest 5% of measurements. The repeatability of each candidate endpoint was assessed by estimating the intra-subject coefficient of variability (CV). The CV was estimated by taking the square root of the residual variability of each model and dividing it by the estimated population mean.

Physical activity

Several candidate endpoints related to physical activity (measured by step count) were defined prior to analysis. These were divided in measurements per day and measurements per week. For daily measurements, step count per day (Daily PA) and step count during the most active hour per day (Daily PA^{MAX}) were defined. For weekly measurements, average step count per week, 10^{TH} and 90^{TH} centile of step count per day, and the 50^{TH} and 90^{TH} centile of step count per hour (between 6AM-10PM) was chosen. Parameters based on high and low centiles were chosen because peak-, trough-and average activity may exhibit different relationships with disease activity in patients. Factors that were hypothesized to explain variability were included as potential fixed factors or covariates during the analyses. The following parameters were considered: age, sex, body mass index (BMI), quality of life (QOL), rain duration, average temperature, wind speed, proportion of wear time between 6AM-10PM, population density of the subjects' city of residence, day of the week and school-day. A sample size needed to detect a 15% increase with 80% power was calculated for each candidate endpoint. Here, it was assumed that a 15% increase was clinically relevant, and that the calculation was for a hypothetical study with parallel group design and follow-up period of 21 days.

Heart rate

Average HR was summarized as average daytime HR (6AM - 10PM) and average nighttime HR (0AM-5AM). Age and sex were considered as covariate and factor, respectively. Estimated mean daytime HR was compared to the 10^{TH}-90^{TH} centile of reference values obtained from a recent meta-analysis regarding pediatric HR[16]. The relationship between daytime HR and physical activity was explored by including an interaction between step count and age in a separate model.

Accelerometer-derived sleep parameters

Total sleep duration, sleep depth and number of times a subject wakes up (wakeup count) were calculated by the Withings application and the Steel HR. All three were considered as candidate endpoint. Days with a sleep duration shorter than 3 hours and longer than 16 hours were excluded from the analysis as likely inaccurate measurement in this healthy

cohort, considering the limitations of the Steel HR watch and published reference values in pediatrics[17]. Wakeup count was analyzed assuming a negative binomial distribution. Age, sex, school day and average ambient temperature were considered as independent variables in the models.

Spirometry, blood pressure and temperature

All spirometry sessions were graded according to ATS/ERS criteria[18]. Spirometry sessions graded D or worse were excluded from further analyses. FEV1 and FVC (% of predicted) were summarized by age. Temperature and BP measurements were graphically displayed.

Software and data pipeline

Data collected via the Withings devices was automatically transferred to the Withings servers, based on protocols maintained by Withings. A validated data pipeline requested the data from the Withings server and stored it on a Microsoft Azure Datalake (Microsoft, Redmond, WA, USA). From here, PySpark version 2.4.6 was used for data aggregation and tabulation. R version 3.5.1 and the lme4, emmeans, ggeffects and pwr packages were used for statistical analysis.

Results

BASELINE CHARACTERISTICS 175 children were included. Baseline characteristics are displayed in *Table 1*. 45.7% of children was male. 85% of children practiced some type of sports. The average QOL score measured by the PedsQL questionnaire was 90.7 out of 100 (IQR [86-97]).

Compliance and tolerability

The average compliance of each individual measurement is listed in *Table 2*. Median overall compliance was 94% (IQR 87-97%) of expected measurements. Median watch wear time was 23.6 hours per day. Lowest compliance was seen for spirometry and temperature measurements (median 83%). Subjects aged 2 or 3 years exhibited a slightly lower overall compliance (*Supplementary Figure S1*), and 11 subjects (6%) exhibited an overall compliance lower than 70%, including two children (aged 3) who stopped participation due to being uncomfortable wearing the watch continuously. There was no correlation between age and compliance for any of the measurements. Compliance appeared to decrease over time for blood pressure- and questionnaire assessments (*Supplementary Figure S2*).

160 (91%) of all subjects completed the EOS questionnaire. Of these, subjects reported to have spent 8.5 (SD 5.5) minutes per day on study-related assessments. 5% of subjects reported the time spent was too much. In total, 97% of subjects and their parents would participate in similar studies in the future. Other responses in the end-of-study questionnaire regarding tolerability are displayed in *Supplementary Table S3*.

Table 1. Baseline characteristics

	Complete cohort (n = 175)
Age (Mean (SD))	9.1 (4.3)
SEX	
Male	45.7%
Female	54.3%
ETHNICITY	
Caucasian	92%
Other/mixed	8%
Height (cm) (Mean (SD))	138.9 (26.1)
Weight (kg) (Mean (SD))	37.2 (17.5)
BMI SDS (Mean (SD))	0.1 (1.2)
DAYTIME ACTIVITY	
Day care	11%
Primary school	50%
Secondary school	37%
Vocational education	1%
Plays sports (%)	84.6%
POPULATION DENSITY	
< 1000 / km2	1%
1000-1500 / km2	19%
1500-2500 / km2	13%
> 2500 / km2	67%
PedsQL score (Mean (SD))	90.6 (7.3)

Abbreviations: BMI SDS: body mass index standard deviation score

Table 2. Compliance during the study period

Measurement	Median compliance (IQR)
SMARTWATCH	
Step count	100% (100% - 100%)
Heart rate	100% (100% - 100%)
Sleep	95% (85% - 100%)
Wear time per day	23.6h (22.8h - 23.9h)
Questionnaire	90% (81% - 100%)
Temperature	83% (67% - 100%)
Weight*	100% (67% - 100%)
Blood pressure*	95% (85% - 100%)
Spirometry*	83% (67% -100%)
Overall compliance	94% (87% - 97%

* Subjects ≥ 6 years old only

Table 3. Physical activity-related candidate endpoints

Sampling frequency	Candidate endpoint	Associated factors*	Intra-subject CV**	Marginal R^2 / Conditional R^2	ICC	Prediction interval	Sample size needed to detect 15% increase***
Per day	Daily PA	Age, sex, rain duration, temperature, day of the week, population density, watch wear time	18%	0.29 / 0.47	0.25	Figure 1G	n = 37
	Hourly PA^{MAX}		23%	0.18 / 0.30	0.15	Figure 2A	n = 35
Per week	Daily PA^{AVG}	Age, sex, mean rain duration, mean temperature, mean watch wear time, population density	8%	0.50 / 0.78	0.58	Figure 2B	n = 35
	Daily $PA90^{TH}$		9%	0.46 / 0.72	0.47	Figure 2C	n = 33
	Daily $PA10^{TH}$		12%	0.39 / 0.74	0.57	Figure 2D	n = 67
	Hourly $PA90^{TH}$		9%	0.46 / 0.76	0.55	Figure 2E	n = 38
	Hourly $PA50^{TH}$	Age, sex, watch wear time	5%	0.24 / 0.71	0.62	Figure 2F	n = 110

Abbreviations: CV: coefficient of variability, ICC: intraclass correlation coefficient, PA: physical activity. * final model coefficients are displayed in *Supplementary Table S4*. ** Adjusted for associated factors *** Approximate sample size needed per group to be able to detect a 15% increase with 80% power in a hypothetical parallel group study where subjects are monitored for 21 consecutive days and the outcome is adjusted for associated factors.

Physical activity

PHYSICAL ACTIVITY PER DAY 174 subjects provided at least one day of physical activity. Step count per day (Daily PA) was considered as the principal candidate endpoint. The relationship between age and step count was best described as a 3^{RD} order natural spline (*Figure 1A,* ΔAIC 44, $p < 0.001$). On average, the daily step count of male subjects was 1082 steps higher compared to female subjects (95% CI 1609-556, $p < 0.001$, *Figure 1B*), although not for all ages (*Figure 1G*). Rain duration (ΔAIC 38, $p < 0.001$, *Figure 1C*) and ambient temperature (as 3^{RD} order natural spline, ΔAIC 13, $p < 0.001$, *Figure 1D*) were also significantly associated with Daily PA. Furthermore, physical activity was lower on Sundays (*Figure 1E*), and for children with a lower population density of the city of residence compared to children in highly urbanized areas (difference 1199, 95% CI 578-1819, $p < 0.001$, *Figure 1F*). Finally, the average watch wear time between 6AM-10PM was also associated with step count per day (ΔAIC 230, $p < 0.001$). BMI, ethnicity, playing sports and QOL were not significantly associated with step count in this cohort. The model estimated intra-subject CV (corrected for factors rain, temperature, and day of the week) was 18%.

The number of steps taken during the most active hour per day (Hourly PA^{MAX}) was considered as a separate endpoint. The measurement was correlated to Daily PA ($R = 0.8$, $p < 0.001$), and identical variables explained variability. Model-estimated intra-subject CV was 23%, and the marginal R^2 of the multivariate model was 0.18 (*Table 3*). The 90% prediction interval stratified by sex is displayed in *Figure 2A*.

PHYSICAL ACTIVITY PER WEEK Daily step count was averaged per week (Daily PA^{AVG}). Similar factors were associated with this candidate endpoint (*Table 3*). A multivariate model for weekly PA including age, sex, the interaction between age and sex, average temperature, average rain duration, average watch wear time between 6AM-10PM and population density of the subject's place of residence provided the best fit (marginal R^2 of 0.50). The model estimated intra-subject CV was 8%.

Four other endpoints based on physical activity within a week were considered and focused on peak- and trough values: the 10^{TH} and 90^{TH} percentile of daily PA within a week (Daily PA^{10TH} and Daily PA^{90TH}) and the 50^{TH} and 90^{TH} percentile of step count per hour within a week (Hourly PA^{50TH} and Hourly PA^{90TH}). Factors of influence on these endpoints are displayed in *Table 3*. *Figure 2B-2F* display the 90% prediction intervals of the proposed endpoints.

Figure 1. Factors related to daily physical activity and reference values. A. Relationship between mean daily physical activity (95% CI) and age. B. Relationship between daily PA and sex. The interaction between age and sex was excluded from the model for this graph only. C. Relationship between daily rain duration and physical activity. D: estimated mean (95% CI) daily PA with varying ambient temperature. E: estimated mean (95% CI) physical activity for each day of the week. F: estimated mean daily PA for children living in a highly urbanized area (> 2000 people/km²) or a less urbanized area (< 2000 people/km²). G. 90% prediction interval of daily PA stratified by age and sex. Wear time (100%), rain duration (2h) and temperature (11 °C) are held constant. Personalized predictions can be made using the model coefficients in Supplementary Table S4.

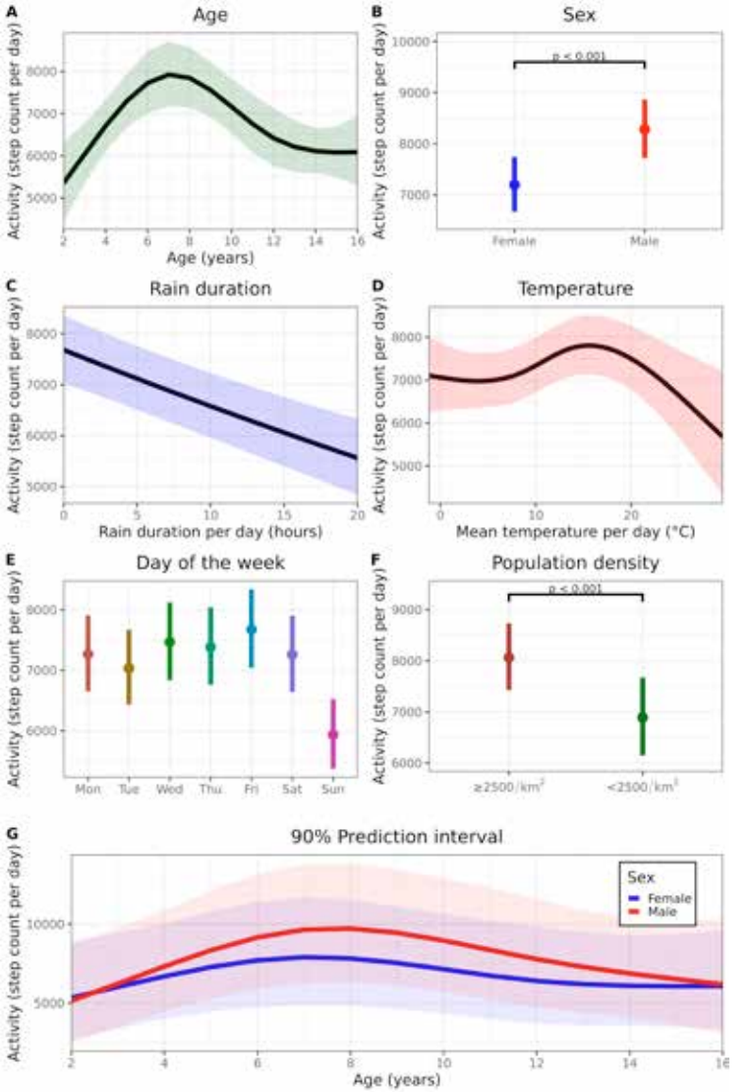
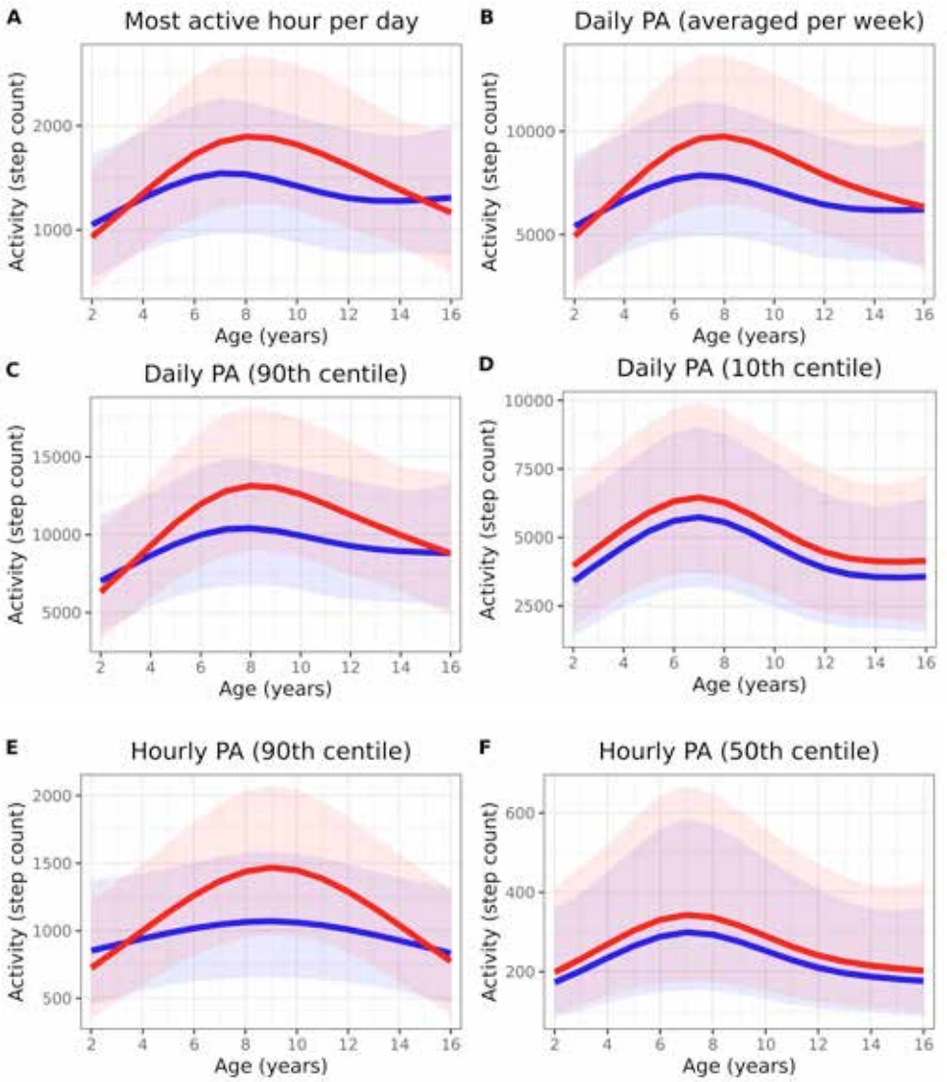


Figure 2. Prediction interval of physical activity candidate endpoints. Prediction interval of several physical-activity related candidate endpoints. Red and blue line represent the estimated mean for male subjects and female subjects, respectively. The shaded areas represent the 90% prediction intervals where watch wear time (100%), rain duration (2h) and temperature (11 °C) are held constant.



The estimated sample size necessary to detect a 15% increase in a parallel study with 21 days of measurements with correction for the factors of influence ranged from 33–110 per group for all candidate endpoints related to physical activity. All candidate endpoint characteristics are listed in *Table 3*. Final model coefficients for physical activity-related endpoints are displayed in *Supplementary Table S4*.

Heart rate

HR data was obtained from 170 subjects. Daytime- and nighttime HR were associated with age (marginal R^2 = 0.62, p < 0.001 for daytime HR and marginal R^2 = 0.50, p < 0.001 for nighttime HR). The relationship was best described with a 3RD degree spline (*Figure 3A*). The population mean and 90% prediction interval were within the reference 10TH–90TH centile of resting HR obtained from the literature[16]. On average, female HR was 2.7 BPM higher compared to male HR (95% CI 1.0–4.4, p = 0.002). Intra-subject CV was 6% (daytime HR) and 8% (nighttime HR). *Figure 3B* shows the average HR during the day for age-groups 2–5, 6–12 and 13–16. There was a statistically significant correlation between HR and step count, although the effect varied by age. Average daytime HR increased by 1.1 BPM (95% CI 0.9–1.3) on average for every 1000 steps taken by 16-year-old children, while increasing 0.5 BPM (95% CI 0.4–0.6) per 1000 steps for 8-year-old children (*Figure 3C*). Model coefficients are listed in *Supplementary Table S5*.

Sleep parameters

Accelerometer-derived nocturnal sleep parameters were obtained from 172 subjects. Sleep duration decreased as a function of age (ΔAIC 65, p < 0.001, *Figure 4A*), and was similar across weekdays, except for weekends of older children (*Supplementary Figure S6*). Average percentage of light sleep was higher as subjects got older (increase of 0.29% per age year (95% CI 0.08 - 0.51, p = 0.006) (*Figure 4B*). On average, the proportion of light sleep of female subjects was 2.5% (95% CI 0.7–4.3, p = 0.008) lower than sleep depth of male subjects. Day of the week was not related to sleep depth. On average, older subjects woke up less often compared to younger subjects, although the correlation was weak (p < 0.001, marginal R^2 0.06, *Figure 4C*). There was no statistically significant difference in wakeup count between male and female subjects (difference -0.07, 95% CI -0.4–0.3, p=0.70), and there was no correlation between ambient temperature and wakeups. Model coefficients are listed in *Supplementary Table S7*.

Figure 3. Heart rate. A: average (90% prediction interval) HR during daytime (blue). Average 95% CI nighttime HR (red). Reference values (10TH–90TH percentile of HR per age year) are displayed in green. B: Average HR (95% CI) per hour of the day. C: estimated relationship between HR and step count for ages 2, 8, 12 and 16. A limited amount of age years is shown for readability.

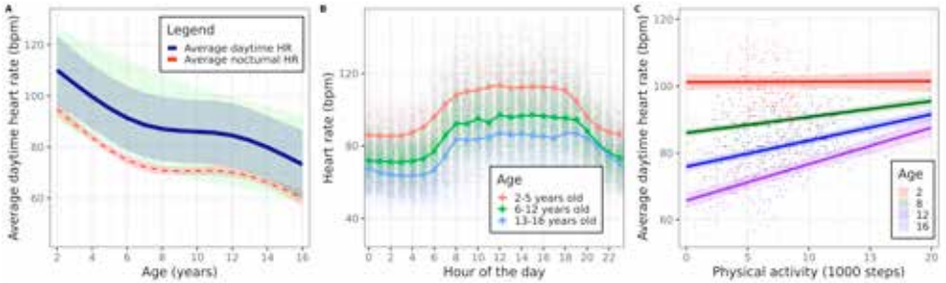
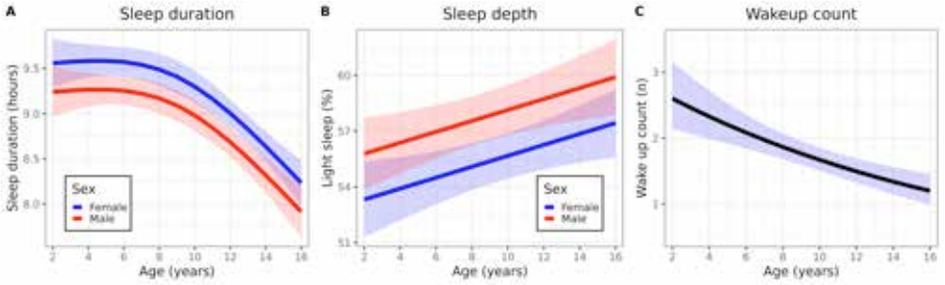


Figure 4. Accelerometer-derived sleep parameters. A: mean (95% CI) sleep duration stratified by sex. B: mean (95% CI) sleep depth stratified by sex. C: mean (95% CI) wakeup count.



Spirometry, blood pressure and temperature

747 spirometry sessions were performed by 126 subjects during the study. 322 sessions (43%) were considered of sufficient quality (ATS grade A-C) for further analysis. Of these, the mean percentage of predicted FEV1 was 94% (95% CI 91–96%) and the mean percentage of predicted FVC was 99% (95% CI 96–102%). Of 2219 BP measurements, mean intra-subject CV was 7% for systolic BP and 10% for diastolic BP. *Supplementary Figure S8* graphically displays BP measurements per age year. Temperature measurements are presented in *Supplementary Figure S9*.

Discussion

There is a need for novel pharmacodynamic clinical endpoints in pediatrics[19]. Continuous monitoring of physical parameters with digital- or wearable devices has potential as

digital endpoint in clinical research and patient care, due to the non-invasive and home-based nature of the measurements. However, extensive clinical validation in the target population is necessary before implementation. Because clear guidelines regarding validation of digital endpoints were lacking, our consortium recently published a pragmatic approach towards fit-for-purpose clinical validation[4]. This study focuses on several crucial steps of this validation process. In the first place, the tolerability of study devices for the target population is important to assess and, if necessary, improve. Secondly, the repeatability of candidate digital endpoints in free-living conditions should be investigated before integration in clinical trials. Furthermore, reference values of candidate digital endpoints in the target population should be obtained and, ideally, datasets should be shared to avoid unnecessary duplication.

Compliance is an important indicator of the tolerability of the clinical trial regimen, and even more so for clinical care, where the follow-up period may be indefinite. In this study, the median compliance was 94%. Based on the literature, a compliance of 70% was determined as a cut-off value to define noncompliant subjects[20,21]. Only 6% of subjects exhibited a compliance lower than this, and the high tolerability was also reflected in the responses in the EOS questionnaire: 97% of subjects and their parents indicated they would participate in similar clinical studies in the future. Smartwatch-related measurements exhibited the highest compliance and may be most suitable to investigate further for their utility in clinical care. The high frequency of smartwatch-related measurements may provide a high-resolution overview of patients' disease state, while objectivity is another advantage compared to traditional questionnaire-based disease scores. Besides subjectivity, questionnaires also suffer from recall bias.

This study investigated the repeatability of candidate endpoints related to physical activity and assessed conditions that explain the observed variability in free-living conditions. Adjusting for such factors in clinical trials will improve statistical power and reduce required sample sizes. For all candidate endpoints, similar covariates and factors were found to explain variability and included age, sex, rain duration, temperature, population density and watch wear time. Some of the reported relationships (age, sex) with physical activity have been reported extensively the past as well[22], including the relationship between population density and step count[23]. This relationship may be explained by children cycling more in less densely populated areas, which is not registered as steps by wrist-worn smartwatches. Interestingly, overall step counts were lower than reported in the literature[22]. This may be partly explained by inter-device variability or by the trend towards a more sedentary lifestyle. BMI, playing sports and QOL (PedsQL questionnaire) were not related to physical activity. However, considering this was a cohort with healthy participants, the range of values of BMI and QOL was small and as such, an association may not have been expected[12].

Candidate endpoints based on physical activity can be divided in three groups: general- (e.g., step count per day), peak- (e.g., hourly PA^{MAX}) and trough physical activity (e.g. weekly PA^{10TH}). We hypothesize that a decrease in peak or trough physical activity may be an early sign of worsening disease activity in chronic disease and therefore be promising for follow-up in clinical care. In the past, research focus has been mainly on moderate-to-vigorous physical activity (MVPA). MVPA cannot easily be derived from step count, although the two are highly correlated[24]. On the other hand, MVPA completely discounts light-intensity activity and is often planned (e.g., gym class, sports). Light-intensity is usually unplanned, more voluntary, and may contain important information relating to disease-activity.

In the end, the optimal endpoint related to activity will be determined by their ability to discriminate healthy from ill children, their association with symptom severity, and the measurement purpose. For example, a high sampling rate of a measurement with a slightly higher intra-subject variability, such as PA^{MAX} will still lead to a precise estimation of the group mean in a clinical trial, while endpoints with lower intra-subject variability, such as weekly PA^{AVG}, will be more suitable for individual follow-up. Clinical validation in pediatric patient groups should be performed to identify the most suitable endpoints for the different aims. Except for endpoints based on trough physical activity, calculated sample sizes needed to detect a 15% increase in candidate endpoints appear feasible.

This study aimed to obtain reference values in healthy children for candidate digital endpoints related to physical activity in the form of 90% prediction intervals. While these graphically presented prediction intervals could be used as a screening tool for pediatric patients, individual predictions that take weather condition, wear time and city of residence into account may be more appropriate and can be calculated using the model coefficients in *Supplementary Table S4*.

HR has been proposed as a candidate biomarker in, among others, pediatric pulmonology, intensive care and psychiatry[25-27], and remote non-invasive HR monitoring could extend this measurement to the home-setting. In the past, Pelizzo *et al.* have shown that HR measured via PPG technology can be reliable in children[28]. Although performed in a surgery setting, direct comparison to ECG-derived HR demonstrated reliability. In this

study, the average heart rate in free-living conditions was compared to known reference values in pediatrics and showed good concordance. Furthermore, a subtle but statistically significant difference in mean HR was found between boys and girls, a finding that has been reported before[29,30]. The lower nocturnal HR compared to daytime HR[31], as well as the correlation that was found between average daytime HR with step count was expected but provides another indication of the validity of PPG measurements in a pediatric population in free-living conditions. The absence of this positive correlation in younger children has been reported before by Herzig *et al.*[32].

Accelerometer-derived sleep duration and depth have been shown to be less accurate when compared to polysomnography, which is the gold standard. However, average total sleep duration in this study appeared to correlate well with published nighttime sleep duration norms[17], and new endpoints for use in free-living conditions should be compared to current standards in the home-setting, such as, comparably reliable, sleep-diaries[33]. Nevertheless, careful interpretation of sleep duration, especially in preschool subjects is necessary. At least three hours of inactivity are needed to register sleep, and daytime naps often do not meet this criterion[34]. Comparison of sleep parameters with pediatric patients with known difficulty sleeping or nocturnal unrest could be helpful to determine the usability of sleep parameters obtained from smartwatches. For example, patients with ARID1B-related intellectual disability showed a higher frequency of wakeups in a recent study[35], and similar differences could be expected in attention deficit hyperactivity disorder, or pediatric asthma.

Temperature, BP, and spirometry measurements are well established measurements with known normative values and are routinely employed in a clinical trial setting. In this study, the measurements were generally within normal range. However, although the used BP monitor has been validated in adults, formal validation has not been performed in children. As a result, clinical interpretation of the measurements should be done with extreme care. However, tolerability and usability of Bluetooth connected devices in a home-setting has not been investigated before and was sufficient in this study, although compliance was slightly lower compared to measurements with a smartwatch.

This study has several limitations. Although the study period of three weeks was long compared to other studies, it was relatively short for the purpose of simulating a possible clinical trial regimen. Compliance to study tasks may decrease with longer follow-up periods. While many variables were collected and related to physical activity, it is possible that a portion of unexplained variability could be accounted for by factors such as socio-economic-status of the parents, which was not registered in this study. Future studies may investigate the influence of variables such as these to increase our understanding of the drivers of physical activity and allow for better isolation of disease- or treatment effects. Conversely, factors that were not registered in this study could change the estimated effects between age and physical activity or heart rate. However, the observed effects of age are plausible and have been reported in the literature in the past. Children were instructed to always wear the watch, including during sports. However, it is possible that subjects took off the watch, for example, during contact sports. This may negatively impact the total step count per day. Although adjustment for watch wear time was performed during the statistical analysis, no information regarding the exact reason of missing data was available and could be either due to not wearing the watch, connectivity errors between the smartwatch and smartphone, or inappropriate handling of the device during the collection or transfer of data. However, this lack of information regarding the cause of missing data is inherent to the field of remote monitoring and finding statistical methods to adjust for this discrepancy is important. During this study, the watch wear time was derived by appraising both the heart rate and step count data during each individual hour. If either was registered, it was determined that the watch was worn during that hour. However, a mismatch was observed between step count and HR data. Some days included more hours with viable HR data than step count data and vice versa. In the case of missing HR data, this discrepancy could be due to an inadequate position of the smartwatch on the wrist of the child, possibly during movement. While this mismatch and the general presence of missing data is suboptimal compared to conventional supervised measurements in the clinic, the fact that the size of the current dataset is much larger compared to conventional trials may outweigh the disadvantages of randomly missing data points. Still, out of the possible 88,200 hours and 3675 days of step count and HR data, between 91% (HR per hour) and 96% (day with wear time > 50%) of observations were included in the final analysis set, which indicates that the impact of missing data on the overall conclusions is likely negligible[36]. The compliance cut-off of 50% wear time per day is lower than employed in other studies in adults and could be revisited in future analyses. It was estimated a priori that children, especially the youngest, would find it difficult to wear the watch 24 hours a day and using this lower cut-off in combination with the statistical adjustment for wear time was expected to lead to valid conclusions. Increasing the threshold to a higher range, for example 70% would lead to exclusion of only a small number (2%) of additional observations. Future clinical validation in patient groups

is necessary to confirm whether the employed analysis methods can detect the effects of treatment and changes in disease-activity despite the potential impact of missing data.

To our knowledge, this study is the first to investigate multiple smartwatch- and digital health-related measurements in a pediatric cohort of this size and age-range. The field of digital endpoints is relatively undeveloped and data sharing may accelerate development and avoid unnecessary duplication. *Supplementary File S10* contains the complete dataset that was used during analysis. A legend of included variables is listed in *Supplementary Table S11*. Future research may focus on tolerability and compliance during a longer follow-up period and further clinical validation of the proposed endpoints in pediatric patient groups.

Conclusion

The investigated home-monitoring platform with a range of wearables and other home-monitoring devices has a good tolerability and led to high compliance. We propose several candidate endpoints related to physical activity that could be used in pediatric trials. Observed variability in endpoints was largely explained by a combination of age, sex, and weather circumstances. In the future, the reference values provided for the candidate endpoints could be used in clinical care and for clinical trial design. Heart rate and sleep data provided by the smartwatch are comparable to pediatric reference values and appear a valid option for pediatric clinical trials in a home-setting. Further clinical validation in pediatric patient groups is currently ongoing and necessary to further define the value of the proposed digital endpoints.

SUPPLEMENTARY DATA

Sup. Figure S1	Median (IQR) compliance to study tasks by age
Sup. Figure S2	Decrease in compliance over time
Sup. Table S3	Tolerability questionnaire
Sup. Table S4	Model Coefficients of physical activity-related candidate endpoints.
Sup. Table S5	Model coefficients of heart rate parameters
Sup. Table S6	Model coefficients of accelerometer-derived sleep parameters
Sup. Figure S7	Sleep duration by day of the week
Sup. Figure S8	Blood pressure measurements per age year
Sup. Figure S9	Temperature measurements per age year
Sup. Data S10	Complete dataset generated during this study
Sup. Table S11	Dataset legend
Sup. File S12	R code used during analysis
Sup. File S13	End-of-study questionnaire

REFERENCES

1 Pasquali SK, Lam WK, Chiswell K, Kemper AR, Li JS. Status of the pediatric clinical trials enterprise: An analysis of the US ClinicalTrials.gov registry. Pediatrics. 2012;130. doi:10.1542/peds.2011-3565

2 Roser M, Ritchie H. Burden of Disease. 2016 p. Published online at OurWorldInData.org. Retrieved.

3 Greenberg RG, Corneli A, Bradley J, Farley J, Jafri HS, Lin L, *et al.* Perceived barriers to pediatrician and family practitioner participation in pediatric clinical trials: Findings from the Clinical Trials Transformation Initiative. Contemp Clin Trials Commun. 2018;9: 7-12. doi:10.1016/j.conctc.2017.11.006

4 Kruizinga MD, Stuurman FE, Exadaktylos V, Doll RJ, Stephenson DT, Groeneveld GJ, *et al.* Development of Novel, Value-Based, Digital Endpoints for Clinical Trials: A Structured Approach Toward Fit-for-Purpose Validation. Pharmacol Rev. 2020;72 (4): 899-909. doi:10.1124/pharmrev.120.000028

5 Ridgers ND, McNarry MA, Mackintosh KA. Feasibility and Effectiveness of Using Wearable Activity Trackers in Youth: A Systematic Review. JMIR mHealth uHealth. 2016;4: e129. doi:10.2196/mhealth.6540

6 Müller J, Hoch AM, Zoller V, Oberhoffer R. Feasibility of physical activity assessment with wearable devices in children aged 4-10 Years-A Pilot study. Front Pediatr. 2018;6: 1-5. doi:10.3389/fped.2018.00005

7 Mackintosh KA, Chappel SE, Salmon J, Timperio A, Ball K, Brown H, *et al.* Parental perspectives of a wearable activity tracker for children younger than 13 years: Acceptability and usability study. JMIR mHealth uHealth. 2019;7: 1-16. doi:10.2196/13858

8 Elmesmari R, Reilly JJ, Martin A, Paton JY. Accelerometer measured levels of moderate-to-vigorous intensity physical activity and sedentary time in children and adolescents with chronic disease: A systematic review and meta-analysis. PLOS One. 2017;12: 1-20. doi:10.1371/journal.pone.0179429

9 Kruizinga MD, Stuurman FE, Groeneveld GJ, Cohen AF. The Future of Clinical Trial Design : The Transition from Hard Endpoints to Value-Based Endpoints.

10 Van Den Wijngaart LS, Roukema J, Boehmer ALM, Brouwer ML, Hugen CAC, Niers LEM, *et al.* A virtual asthma clinic for children: Fewer routine outpatient visits, same asthma control. Eur Respir J. 2017;50: 1-10. doi:10.1183/13993003.00471-2017

11 Kruizinga MD, Essers E, Stuurman FE, Zhuparris A, van Eik N, Janssens HM, *et al.* Technical validity and usability of a novel smartphone-connected spirometry device for pediatric patients with asthma and cystic fibrosis. Pediatr Pulmonol. 2020; 2463-2470. doi:10.1002/ppul.24932

12 Varni JW, Seid M, Kurtin PS. PedsQLTM 4.0: Reliability and Validity of the Pediatric Quality of Life InventoryTM Version 4.0 Generic Core Scales in Healthy and Patient Populations. Med Care. 2001;39. Available: https://journals.lww.com/lww-medicalcare/Fulltext/2001/08000/PedsQL_4_0_Reliability_and_Validity_of_the.6.aspx

13 Evans EW, Abrantes AM, Chen E, Jelalian E. Using Novel Technology within a School-Based Setting to Increase Physical Activity: A Pilot Study in School-Age Children from a Low-Income, Urban Community. Biomed Res Int. 2017;2017. doi:10.1155/2017/4271483

14 Byun W, Lau EY, Brusseau TA. Feasibility and effectiveness of a wearable technology-based physical activity intervention in preschoolers: A pilot study. Int J Environ Res Public Health. 2018;15. doi:10.3390/ijerph15091821

15 Lumley T, Diehr P, Emerson S, Chen L. The importance of the normality assumption in large public health data sets. Annu Rev Public Health. 2002;23: 151-169. doi:10.1146/annurev.publhealth.23.100901.140546

16 Fleming S, Thompson M, Stevens R, Heneghan C, Plüddemann A, MacOnochie I, *et al.* Normal ranges of heart rate and respiratory rate in children from birth to 18 years of age: A systematic review of observational studies. Lancet. 2011;377: 1011-1018. doi:10.1016/S0140-6736(10)62226-X

17 Iglowstein I, Jenni OG, Molinari L, Largo RH. Sleep duration from infancy to adolescence: Reference values and generational trends. Pediatrics. 2003;111: 302-307. doi:10.1542/peds.111.2.302

18 Graham BL, Steenbruggen I, Barjaktarevic IZ, Cooper BG, Hall GL, Hallstrand TS, *et al.* Standardization of spirometry 2019 update an official American Thoracic Society and European Respiratory Society technical statement. Am J Respir Crit Care Med. 2019;200: E70-E88. doi:10.1164/rccm.201908-1590ST

19 Kelly LE, Sinha Y, Barker CIS, Standing JF, Offringa M. Useful pharmacodynamic endpoints in children: Selection, measurement, and next steps. Pediatr Res. 2018;83: 1095-1103. doi:10.1038/pr.2018.38

20 Carter E V., Hickey KT, Pickham DM, Doering L V., Chen B, Harris PRE, *et al.* Feasibility and compliance with daily home electrocardiogram monitoring of the QT interval in heart transplant recipients. Hear Lung J Acute Crit Care. 2012;41: 368-373. doi:10.1016/j.hrtlng.2012.02.012

21 Lipsmeier F, Taylor KI, Kilchenmann T, Wolf D, Scotland A, Schjodt-Eriksen J, *et al.* Evaluation of smartphone-based testing to generate exploratory outcome measures in a phase 1 Parkinson's disease clinical trial. Mov Disord. 2018;33: 1287-1297. doi:10.1002/mds.27376

22 Tudor-Locke C, Craig CL, Beets MW, Belton S, Cardon GM, Duncan S, *et al.* How many steps/day are enough? For children and adolescents. Int J Behav Nutr Phys Act. 2011;8: 78. doi:10.1186/1479-5868-8-78

23 Sallis JF, Cerin E, Conway TL, Adams MA, Frank LD, Pratt M, *et al.* Physical activity in relation to urban environments in 14 cities worldwide: A cross-sectional study. Lancet. 2016;387: 2207-2217. doi:10.1016/S0140-6736(15)01284-2

24 Colley RC, Janssen I, Tremblay MS. Daily step target to measure adherence to physical activity guidelines in children. Med Sci Sports Exerc. 2012;44: 977–982. doi:10.1249/MSS.0b013e31823f23b1
25 Bazelmans T, Jones EJH, Ghods S, Corrigan S, Toth K, Charman T, *et al.* Heart rate mean and variability as a biomarker for phenotypic variation in preschoolers with autism spectrum disorder. Autism Res. 2019;12: 39–52. doi:10.1002/aur.1982
26 Marsillio LE, Manghi T, Carroll MS, Balmert LC, Wainwright MS. Heart rate variability as a marker of recovery from critical illness in children. PLOS One. 2019;14: 1–12. doi:10.1371/journal.pone.0215930
27 Aurora P, Wade A, Whitmore P, Whitehead B. A model for predicting life expectancy of children with cystic fibrosis. Eur Respir J. 2000;16: 1056–1060. doi:10.1034/j.1399-3003.2000.16f06.x
28 Pelizzo G, Guddo A, Puglisi A, De Silvestri A, Comparato C, Valenza M, *et al.* Accuracy of a Wrist-Worn Heart Rate Sensing Device during Elective Pediatric Surgical Procedures. Children. 2018;5: 38. doi:10.3390/children5030038
29 Rijnbeek PR, Witsenburg M, Schrama E, Hess J, Kors JA. New normal limits for the paediatric electrocardiogram. Eur Heart J. 2001;22: 702–711. doi:10.1053/euhj.2000.2399
30 Semizel E, Öztürk B, Bostan OM, Cil E, Ediz B. The effect of age and gender on the electrocardiogram in children. Cardiol Young. 2008;18: 26–40. doi:10.1017/S1047951107001722
31 Hadtstein C, Wühl E, Soergel M, Witte K, Schaefer F, Kirschstein M, *et al.* Normative Values for Circadian and Ultradian Cardiovascular Rhythms in Childhood. Hypertension. 2004;43: 547–554. doi:10.1161/01.HYP.0000116754.15808.d8
32 Herzig D, Eser P, Radtke T, Wenger A, Rusterholz T, Wilhelm M, *et al.* Relation of heart rate and its variability during sleep with age, physical activity, and body composition in young children. Front Physiol. 2017;8: 1–12. doi:10.3389/fphys.2017.00109
33 Nascimento-Ferreira M V., Collese TS, de Moraes ACF, Rendo-Urteaga T, Moreno LA, Carvalho HB. Validity and reliability of sleep time questionnaires in children and adolescents: A systematic review and meta-analysis. Sleep Med Rev. 2016;30: 85–96. doi:10.1016/j.smrv.2015.11.006
34 Lambrechtse P, Ziesenitz VC, Cohen A, van den Anker JN, Bos EJ. How reliable are commercially available trackers in detecting daytime sleep. Br J Clin Pharmacol. 2018;84: 605–606. doi:10.1111/bcp.13475
35 Kruizinga MD, Zuiker RGJA, Sali E, de Kam ML, Doll RJ, Groeneveld GJ, *et al.* Finding Suitable Clinical Endpoints for a Potential Treatment of a Rare Genetic Disease: the Case of ARID1B. Neurotherapeutics. 2020. doi:10.1007/s13311-020-00868-9
36 Bennett DA. How can I deal with missing data in my study? Aust N Z J Public Health. 2001;25: 464–469. doi:10.1111/j.1467-842X.2001.tb00294.x

CHAPTER 7

Clinical validation of digital biomarkers for pediatric patients with asthma and cystic fibrosis – Potential for clinical trials and clinical care

European Respiratory Journal; DOI: 10.1183/13993003.00208-2021

Matthijs D. Kruizinga[1,2,3], Esmée Essers[1,2], Frederik E. Stuurman[1,3], Yalçin Yavuz[1], Marieke L de Kam[1], Ahnjili Zhuparris[1], Hettie M. Janssens[4], Iris Groothuis[2], Arwen J. Sprij[2], Marianne Nuijsink[2], Adam F. Cohen[1,3], Gertjan J. A. Driessen[2,5]

1 Centre for Human Drug Research, Leiden, the Netherlands
2 Juliana Children's Hospital, Haga Teaching Hospital, The Hague, the Netherlands
3 Leiden University Medical Center, Leiden, the Netherlands
4 Division of Respiratory Medicine and Allergology, Department of Pediatrics, Erasmus Medical Centre/Sophia Children's Hospital, University Hospital Rotterdam, Rotterdam, The Netherlands
5 Department of pediatrics, Maastricht University Medical Centre, Maastricht, the Netherlands

Abstract

BACKGROUND Digital biomarkers are a promising novel method to capture clinical data in a home-setting. However, clinical validation prior to implementation is of vital importance. The aim of this study was to clinically validate physical activity, heart rate, sleep and FEV1 as digital biomarkers measured by a smartwatch and portable spirometer in children with asthma and cystic fibrosis (CF).

METHODS This was a prospective cohort study including 60 children with asthma and 30 children with CF (age 6-16). Participants wore a smartwatch, performed daily spirometry at home and completed a daily symptom questionnaire for 28-days. Physical activity, heart rate, sleep and FEV1 were considered candidate digital endpoints. Data from 128 healthy children was used for comparison. Reported outcomes were compliance, difference between patients and controls, correlation with disease-activity and potential to detect clinical events. Analysis was performed with linear mixed effect models.

RESULTS Median compliance was 88%. On average, patients exhibited lower physical activity and FEV1 compared to healthy children, whereas the heart rate of children with asthma was higher compared to healthy children. Days with a higher symptom score were associated with lower physical activity for children with uncontrolled asthma and CF. Furthermore, FEV1 was lower and (nocturnal) heart rate was higher for both patient groups on days with more symptoms. Candidate biomarkers and showed a distinct pattern before-and after a pulmonary exacerbation.

CONCLUSION Portable spirometer- and smartwatch-derived digital biomarkers show promise as candidate endpoints for use in clinical trials or clinical care in pediatric lung disease.

Introduction

Clinical follow-up of pulmonary diseases, such as asthma and cystic fibrosis (CF), traditionally relies on both self- and parent-reported symptoms in the outpatient clinic and pulmonary function tests (PFTs). Even though this is considered adequate for pediatric clinical care, both self- and parent-reported symptoms generally suffer from recall bias and are considered subjective, while clinic-based PFTS in children are sometimes associated with challenges in obtaining acceptable and repeatable measurements as well[1,2]. Additionally, new treatments have also led to a slower decline of pulmonary function in CF patients and increasing numbers of patients have pulmonary function in the normal range while still perceiving a significant symptom load[3]. Similarly, pediatric clinical trials, which are difficult to conduct due to ethical and logistical barriers and low inclusion rates[4], either rely on subjective endpoints or rare 'hard' endpoints, such as hospital admission. Rare endpoints lead to unrealistically large sample sizes and long and costly studies, and although subjective symptom reports can be valuable from an investigational point of view, ideally, they should be collected together with additional biomarkers that give a more objective indication of disease control[5]. In pediatrics, such biomarkers are preferably non-invasive, which are scarce. These limitations lead to gaps in knowledge[6,7], and new, objective and non-invasive biomarkers for pediatric pulmonary disease with high clinical and practical utility are needed for use in clinical trials and-care[6,8].

Non-invasive measurements with digital and portable devices for home-use may provide such new biomarkers. Physical activity (PA) has shown to be related to asthma severity[9], and it is plausible that heart rate (HR) and parameters related to sleep also correlate well with an increase in disease-activity[10]. Such parameters can be easily and objectively obtained by consumer devices like a smartwatch[11]. Several PA- and HR-derived digital biomarkers, such as daily step count, step count taken during most active hour per day, and daytime- or nocturnal HR, have been proposed and evaluated in healthy children, and these candidate digital biomarkers exhibited reasonable intra-subject variability[12]. Furthermore, portable spirometers for measurement of complete flow–volume curves have been developed, which can be used in a home-setting[13,14].

Before these digital biomarkers can be included in clinical trials or clinical care, careful selection, technical validation and, most importantly, a rigorous clinical validation process in the target population is necessary[15,16]. A natural next step in the validation of biomarkers derived from PA, HR and FEV1 is clinical validation. A stepwise approach has been

proposed, which is not necessarily comparable to the traditional validation steps for outcome measures[15,17]. This clinical validation should include determination of the tolerability for patients, day-to-day variability in patients, difference between patients and healthy controls, correlation with traditional methods to measure disease activity and potential to detect clinical events to assess the utility of the novel biomarkers[15]. The aim of this study is to initiate the clinical validation process for biomarkers derived from physical activity, HR and sleep and for forced expiratory volume in 1 second (FEV1) measured by a smartwatch and portable spirometer in children with asthma and CF.

Materials and methods

This study was conducted by Juliana Children's Hospital (Haga teaching hospital, the Hague, the Netherlands), Sophia Children's Hospital (Erasmus Medical Centre, Rotterdam, the Netherlands) and the Centre for Human Drug Research (CHDR, Leiden, the Netherlands) from November 2018 to February 2020. The study protocol was reviewed and approved by the Medical Ethics Committee Zuidwest Holland (The Hague, The Netherlands), and conducted according to the Dutch Act on Medical Research Involving Human Subjects (WMO). Written informed consent was obtained from all parents and children aged 12 years and older. The trial was registered at the Dutch Trial Registry (NTR, Trial NL7611).

Subjects and study design

Pediatric patients aged 6-16 with controlled asthma (n=30), uncontrolled asthma (n=30), and CF (n=30) were recruited from the outpatient clinic of the hospitals. In our centers, the diagnosis of asthma is based on clinical symptoms combined with PFTS[18], while the diagnosis of CF patients was confirmed by genetic tests. Asthma control was defined using the Global Initiative for Asthma criteria and Asthma Control Questionnaire (cutoff > 1.5 points). Children used multiple devices as described below and completed a daily symptom questionnaire, together with their parents, for 28 consecutive days. Subjects with asthma were instructed to complete the asthma control diary 6-questions (ACD6), and subjects with CF were instructed to complete a daily respiratory symptom questionnaire adapted from an existing questionnaire (*Supplementary Text S1*)[19,20]. This respiratory symptom questionnaire is not formally validated for children with CF. After 28 days, an end-of-study questionnaire was completed, and the devices were retrieved by the study team.

Subjects were instructed to wear a Steel HR smartwatch (Withings, Issy-les-Molineux, France) during the study period. The watch measures PA with a built-in accelerometer. HR was measured using a photo plethysmography (PPG) sensor on the back of the watch. Furthermore, the watch calculates several sleep-related parameters using the accelerometer and an incorporated temperature sensor, validity of which has been investigated in similar devices[21]. Technical validity of the Steel HR smartwatch was previously investigated (*Supplementary Text S2*). Subjects were instructed to perform daily home-based spirometry using the Air Next spirometry device (NuvoAir, Stockholm, Sweden). This device is validated for use in children and measures FEV1 as well as forced vital capacity (FVC)[13], and the subjects' age, sex and height were used to calculate FEV1 and FVC expressed as z-score based on GLI-2012 equations[22]. All devices used Bluetooth to connect to a smartphone (Motorola G6 (Motorola, Chicago, IL, USA)), which had the Withings Healthmate, and CHDR MORE® (used for data collection and aggregation) applications pre-installed.

Baseline- and environmental data

Parents were instructed to complete the PedsQL 4.0 questionnaire (score 0-100, higher scores represent better quality of life) at the start of the study[23]. Subjects with asthma and their parent(s) completed the asthma control questionnaire (ACQ, score 0-6, higher scores represent worse asthma control) and pediatric asthma quality of life questionnaire (PAQLQ, score 1-7, higher scores represent better quality of life), while subjects with CF and their parent(s) completed the Cystic Fibrosis Questionnaire (CFQ-R, score per subdomain 0-100, with higher scores representing lower disease burden)[24-26]. Other baseline characteristics were collected from the electronic patient file. Prescribed medication at the time of inclusion was registered. Weather (rain duration, temperature) statistics from a local weather station (Hoek van Holland, the Netherlands) were obtained from the Royal Dutch Meteorological Institute (KNMI) and used as covariate in physical activity analyses.

Candidate endpoints

Several physical activity-related candidate endpoints were defined prior to analyses: step count per day (Daily PA), step count during the most active hour (Daily PA^{MAX}, representing daily peak activity) and weekly summarized average-, 10^{TH} centile- and 90^{TH} centile of

physical activity. The last three represents the average-, peak- and trough physical activity[12]. Nocturnal (average HR between 0-5AM) and daytime (average HR between 6AM-22PM) HR were selected as separate endpoints, as well as FEV1 and FVC. Finally, accelerometer-derived sleep parameters total sleep duration, sleep depth (proportion of light sleep) and wakeup count were also selected.

Analysis set

Out of the total dataset (2520 study days), all days with a watch wear time less than 50% between 6AM-10PM were excluded from the analysis (8%, 197 study days). All spirometry curves were graded manually according to ATS criteria[1]. Spirometry sessions graded A, B or C were eligible for statistical analysis (64% of all spirometry sessions, 1165 observations).

Validation criteria

TOLERABILITY Tolerability was assessed by calculating the compliance during the study and the end-of-study questionnaire outcomes. The median and interquartile range (IQR) of the proportion of expected measurements that were performed was calculated for each individual endpoint, as well as for the total amount study activities. For PA and HR, a watch wear time of 50% was required for that day to be included in statistical analyses[12,27,28]. Prior to study initiation, a subject with an overall compliance across all measurements < 70% was considered non-compliant.

VARIABILITY Intrasubject variability was estimated for each condition and candidate biomarker via mixed effect models. For each condition (asthma, CF, healthy) and candidate biomarker, a separate model was fitted with subject as random intercept. The intra-class correlation coefficient (ICC) was calculated by dividing the random intercept variance by the total variance.

DIFFERENCE PATIENTS-CONTROLS To assess the difference between patients and healthy children, data from 128 apparently healthy children aged 6-16 who participated in a separate, comparable trial in parallel to this study[12]. Healthy children from the same geographical area wore the Steel HR smartwatch for 21 days and performed biweekly PFTS. The difference between patient groups and healthy subjects was calculated with a mixed effects model with condition (healthy, controlled asthma, uncontrolled asthma, or CF) as fixed effect and subject as random effect. Additional adjustments for covariates identified in that trial (watch wear time, age, sex, rain duration, temperature, type of day, urbanization for PA-derived biomarkers, age and sex for HR-derived biomarkers) were made if they improved model fit according to the Akaike- and Bayesian Information Criterion (AIC/BIC)[12,29]. Daytime HR was adjusted for physical activity during that day. No adjustment for multiple comparisons was performed. A sensitivity analysis for the choice of wear time threshold was performed by repeating the analysis with varying thresholds.

CORRELATION WITH EXISTING DISEASE METRICS To evaluate whether a change in traditional endpoint, in this case symptom questionnaire scores, corresponds with a change in novel biomarker outcomes, the relationship between candidate endpoints and a symptom questionnaire was analyzed via mixed effects models. A model was fitted for each candidate endpoint, where ACD6 score (asthma) and respiratory symptom score (CF) were included as fixed effect and a random intercept and slope was fitted for each subject. No adjustment for baseline disease activity was performed[30]. Adjustments for baseline symptom score and covariates identified in the previous study were made if they improved model fit as described in the previous paragraph. The estimated marginal effect and the 95% CI was plotted, and significance of the overall effect was assessed with a type III test of fixed effects.

DESCRIPTION OF HEALTH EVENTS Asthma exacerbations were defined according to the ATS/ERS criteria as worsening of asthma requiring the use of systemic corticosteroids to prevent a serious outcome[31]. Pulmonary CF exacerbations were defined as the need for additional antibiotic treatment as indicated by a recent change in symptoms or decrease in pulmonary function (≥ 10% of predicted FEV1)[32]. In this analysis, the day when corticosteroids or antibiotic treatment was first prescribed, was defined as day 0. Study data from the previous 7 days and the 14 subsequent days were analyzed with a mixed effects model with day as spline covariate and random slope, to allow for nonlinear trajectories[33]. Due to the limited size of the dataset, no adjustments for covariates were made. To assess whether the observed trajectory was not based on random variability, the trajectory over time was also estimated for the group of subjects that did not experience a pulmonary exacerbation during the study.

SOFTWARE AND STATISTICS PySpark version 2.4.6 was used for data aggregation and tabulation. R version 3.5.1 with the lme4, emmeans, rspiro and ggeffects packages was used for statistical analysis. Promasys® software (OmniComm, Ft. Lauderdale, FL, USA) was used for data management. Statistical analysis was performed as described in individual paragraphs above. A p-value < 0.05 was considered statistically significant. Mixed effect model fit was appraised by evaluating the AIC and BIC of each model. Log or square root transformation of the outcome variable was applied during analyses of PA due to heteroscedasticity. A negative binomial distribution was assumed when analyzing wakeup count. A sample size calculation for the difference in PA was performed based on activity data collected in an earlier pilot study[34]. Assuming a significance level of 0.05, a power of 0.8 and the ability to detect a difference between patients and healthy controls of 2750 steps with a standard deviation of 3750 steps in both groups, we calculated a sample size of 30 patients per group.

Results

Baseline characteristics

Baseline characteristics of subjects with controlled asthma (n=30), uncontrolled asthma (n=30) and CF (n=30) were compared with 128 healthy subjects and are displayed in *Table 1*. The mean age of the four groups ranged between 9.7–11.1 years. Subjects with uncontrolled asthma were least likely (67%) to practice any type of sports. Mean quality of life score (PedsQL) was lowest for subjects with uncontrolled asthma (68.7), followed by subjects with cystic fibrosis (79.5), controlled asthma (80.4) and healthy subjects (90.7).

Tolerability

Tolerability was assessed by reviewing the compliance during the study and by a tolerability questionnaire. Median compliance was 88% (IQR[76–95%]) for all subjects (*Table 2*), whereas subjects with uncontrolled asthma had a lower median compliance (79%, IQR[71–95%], *Supplementary Table S1*). Compliance for physical activity and HR was highest for all study groups, followed by sleep, PFT and questionnaire assessments. Children needed a median of 10 minutes per day for study assessments. Eighty-eight percent of respondents of the end-of-study questionnaire reported to be willing to participate in similar studies in the future.

Table 1. Baseline characteristics

	Controlled asthma (n=30)	Uncontrolled asthma (n=30)	Cystic Fibrosis (n=30)	Healthy subjects (n = 128)
Age (mean (SD))	10.5 (2.4)	10.5 (2.9)	9.7 (2.6)	11.1 (3.1)
Sex (% male)	67	67	47	46
Race (% Caucasian)	70	60	100	93
BMI SDS (mean (SD))	0.7 (1.5)	1.2 (1.5)	-0.1 (0.9)	0.3 (1.2)
Plays sports (%)	83	67	80	91
Admissions year prior (mean [range])	0.13 [0–1]	0.37 [0–1]	0.17 [0–1]	-
Atopic asthma (%)	63	83	-	-
Exercise-related symptoms (%)	40	73	-	-
LABA therapy (%)	40	77	17	-
ICS (%)	97	97	17	-
Oral steroids (%)	0	4	-	-
CFTR mutation (%) Class I Class II Class IV	-	-	 3 93 3	-
Pancreatic insufficiency (%)	-	-	93	-
Past pseudomonas infection* (%)	-	-	27	-
PedsQL score (mean (SD))	80.4 (8.6)	68.7 (13.6)	79.5 (11.4)	90.7 (7.4)
ACQ (mean (SD))	0.7 (0.5)	1.9 (0.8)	-	-
PAQLQ (mean (SD))	6.4 (0.4)	5.3 (1.1)	-	-
CFQ respiratory domain (mean (SD))			83.7 (13.9)	
CFQ health perception (mean (SD))	-	-	74.0 (17.1)	-

Abbreviations: SD: standard deviation, BMI: body mass index, SDS: standard deviation score, LABA: long-acting beta agonists, ICS: inhalation corticosteroids, CFTR: Cystic Fibrosis Transmembrane Conductance Regulator, ACQ: asthma control questionnaire, PAQLQ: pediatric asthma quality of life questionnaire, CFQ: cystic fibrosis questionnaire, PedsQL: pediatric quality of life. * Pseudomonas infection was defined as at least one isolate of pseudomonas aeruginosa in sputum in the last 12 months.

Table 2. Median compliance [IQR] during the study period

Assessment	All subjects (n=90)
Step count	100% [100–100]
Heart rate	100% [96–100]
Sleep	85% [74–89]
Pulmonary function test	79% [46–93]
Questionnaire	78% [68–96]
All assessments	88% [76–95]

Variability

ICCS were calculated separately for each group and candidate biomarker and are displayed in *Table 3*. Patient groups exhibited a lower ICC compared to healthy children for PA-related endpoints, HR and sleep while ICC of patient groups was higher for FEV1.

Table 3. Intra-class correlation coefficient (ICC) of candidate biomarkers

Candidate biomarker	Controlled asthma (95% CI)	Uncontrolled asthma (95% CI)	Cystic fibrosis (95% CI)	Healthy 95% CI
PHYSICAL ACTIVITY				
step count per day	0.22 (0.11-0.31)	0.33 (0.21-0.44)	0.16 (0.08-0.24)	0.35 (0.29-0.41)
PA^{MAX}	0.13 (0.06-0.21)	0.21 (0.12-0.31)	0.08 (0.03-0.14)	0.24 (0.19-0.29)
Daytime HR (BPM)	0.48 (0.33-0.60)	0.55 (0.41-0.67)	0.59 (0.42-0.70)	0.65 (0.58-0.70)
Nocturnal HR (BPM)	0.55 (0.40-0.67)	0.50 (0.35-0.61)	0.61 (0.47-0.72)	0.73 (0.66-0.77)
FEV1 (z-score)	0.59 (0.44-0.71)	0.64 (0.48-0.75)	0.63 (0.47-0.74)	0.55 (0.46-0.63)
Sleep duration (hours)	0.26 (0.14-0.37)	0.35 (0.22-0.48)	0.22 (0.12-0.31)	0.31 (0.24-0.36)

Difference patients - controls

Physical activity per day was lower for all three patient groups when compared to healthy children (*Figure 1A*). The largest adjusted difference compared to healthy children was observed for children with uncontrolled asthma (1264 steps, 95% CI 573-1956, p < 0.001), followed by children with CF (847, 95% CI 138-1555, p=0.019) and children with controlled asthma (731, 95% CI 6-1456, p=0.049). PA^{MAX} (*Figure 1B*) of subjects with uncontrolled asthma was lower compared to healthy subjects (adjusted difference 282, 95% CI 134-429, p < 0.001). Average-, peak- and trough physical activity per week showed similar group differences (*Supplementary Figure S1*). Step count per hour of the day (*Figure 1C*) showed that differences in step count between groups were most pronounced during after-school-hours (3PM-7PM). Subsequently, aggregated PA data during after-school-hours was analyzed as exploratory additional biomarker (*Supplementary Figure S2*).

Adjusted average nocturnal HR of subjects with uncontrolled asthma was significantly higher compared to healthy controls and the two other patient groups (*Figure 1D*). Additionally, daytime HR of all patient groups was higher compared to healthy children. Additional adjustment for β-agonist use in patients with asthma showed smaller differences in heart rate between patients and controls (*Supplementary Figure S3*).

CF subjects showed the longest total sleep duration per night (9.1 hours) and slept significantly longer compared to subjects with asthma (*Figure 1E*). There was no statistically significant difference between groups for the parameters sleep depth and wakeup count.

All PFTS performed with adequate technique were included to estimate the difference in pulmonary function between groups. Average FEV1 (expressed as z-score) of patients was lower compared to healthy subjects (*Figure 1F*). There were no differences in FVC between the groups.

Adjusted and unadjusted absolute differences between patients and controls are displayed in *Table 4* for all candidate biomarkers, as well as the standard errors (SE) of the estimate. A sensitivity analysis for the choice of wear time threshold has been included in *Supplementary Figure S4*.

Table 4. Adjusted and unadjusted differences between patient and control groups

Endpoint	Adjustment	Healthy vs. controlled asthma		Healthy vs. uncontrolled asthma		Healthy vs. cystic fibrosis		Adjusted for
		Estimate of the difference (SE)	p-value	Estimate of the difference (SE)	p-value	Estimate of the difference (SE)	p-value	
Daily PA (step count)	Unadjusted	474 (423)	0.26	1097 (406)	0.007	478 (420)	0.25	Wear time, age, rain duration, day type*, sex
	Adjusted	731 (370)	0.048	1264 (353)	< 0.001	847 (361)	0.02	
Daily PA^{MAX} (step count)	Unadjusted	88 (86)	0.31	241 (82)	0.003	30 (87)	0.73	Age, sex, wear time, rain duration, day type*
	Adjusted	150 (79)	0.059	282 (75)	< 0.001	95 (79)	0.23	
Daytime HR (BPM)	Unadjusted	-2.94 (1.45)	0.043	-5.00 (1.45)	< 0.001	-3.76 (1.45)	0.01	Age, sex, physical activity
	Adjusted	-3.24 (1.23)	0.008	-5.71 (1.23)	< 0.001	-2.70 (1.23)	0.03	
Nocturnal HR (BPM)	Unadjusted	-2.97 (1.54)	0.054	-6.82 (1.54)	< 0.001	-2.34 (1.54)	0.13	Age, sex
	Adjusted	-2.92 (1.40)	0.037	-6.77 (1.40)	< 0.001	-0.86 (1.4)	0.54	
Total sleep duration (hr)	Unadjusted	0.14 (0.15)	0.35	0.21 (0.14)	0.14	-0.30 (0.14)	0.04	Age
	Adjusted	0.22 (0.13)	0.08	0.27 (0.12)	0.03	-0.13 (0.13)	0.31	
FEV1 (z-score)	Unadjusted	0.53 (0.24)	0.03	0.8 (0.25)	0.002	0.59 (0.23)	0.01	NA
	Adjusted							
FVC (z-score)	Unadjusted	0.08 (0.26)	0.75	0.32 (0.27)	0.24	0.45 (0.26)	0.08	NA
	Adjusted							

* School day, weekend day or holiday

Figure 1. Difference between patients and control subjects. A: Estimated marginal mean physical activity per day for the four study groups. B: Estimated marginal mean step count during the most active hour on a day. C: Estimated marginal mean physical activity per hour throughout the day. Colors per group are identical as in other panels. D: Estimated marginal mean daytime-and nocturnal heart rate per day. In this estimated average, age is held constant at 12. E: Estimated mean total sleep duration per day, F: Estimated mean FEV1 z-score.

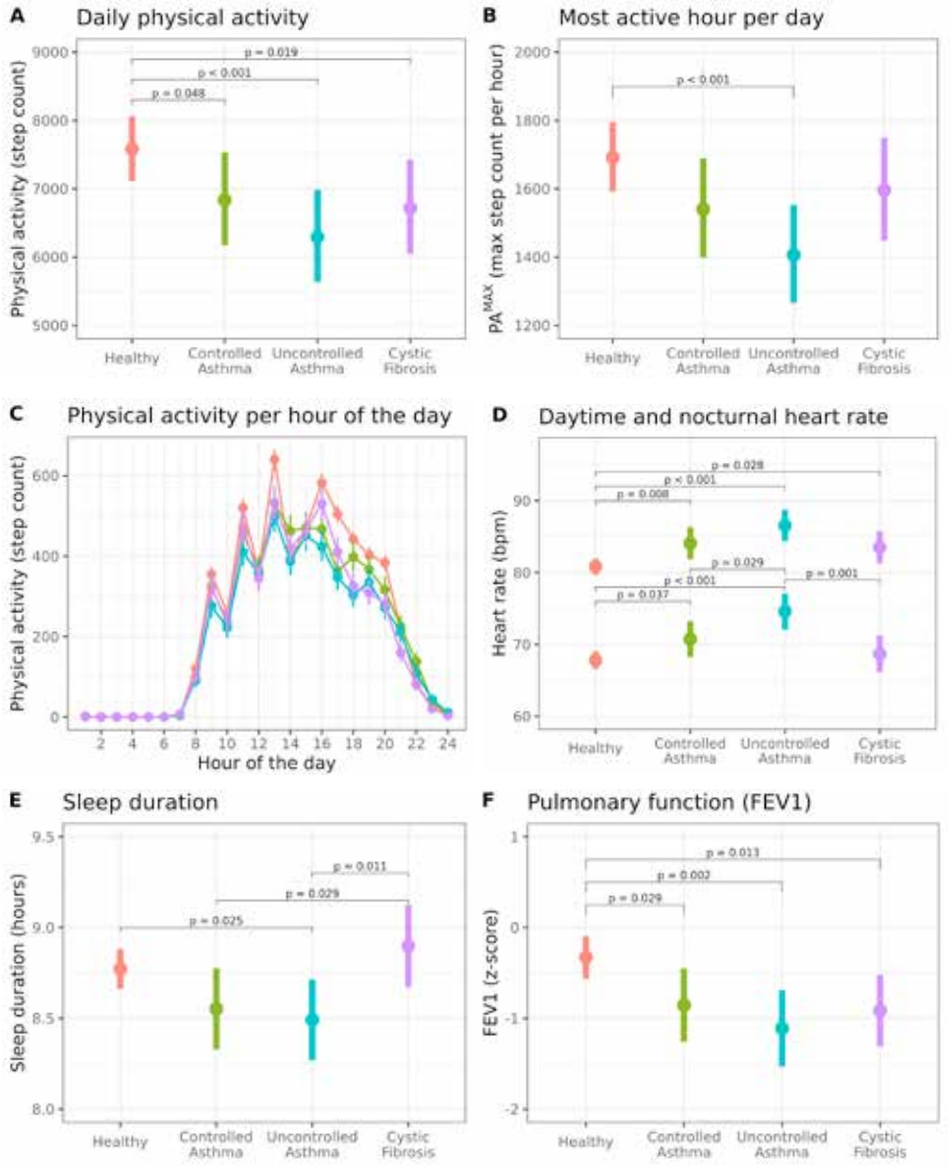


Figure 2. Correlation novel endpoints with traditional endpoints. A-C: Estimated relationship between symptom questionnaire scores and physical activity (step count per day) for subjects with controlled asthma (A), uncontrolled asthma (B) and CF (C). Estimated effects are presented as percentages due to log-transformation of the outcome variable. D-I: Estimated relationship between average daytime HR (D-F) and nocturnal heart rate (G-I) per day and symptom questionnaire scores for subjects with asthma and CF. J-M: Estimated relationship between FEV1 z-score and symptom questionnaire score for subjects with asthma and CF. Bold lines and shaded areas represent the estimated mean and the 95% CI of the relationship. Transparent lines represent individual estimates.

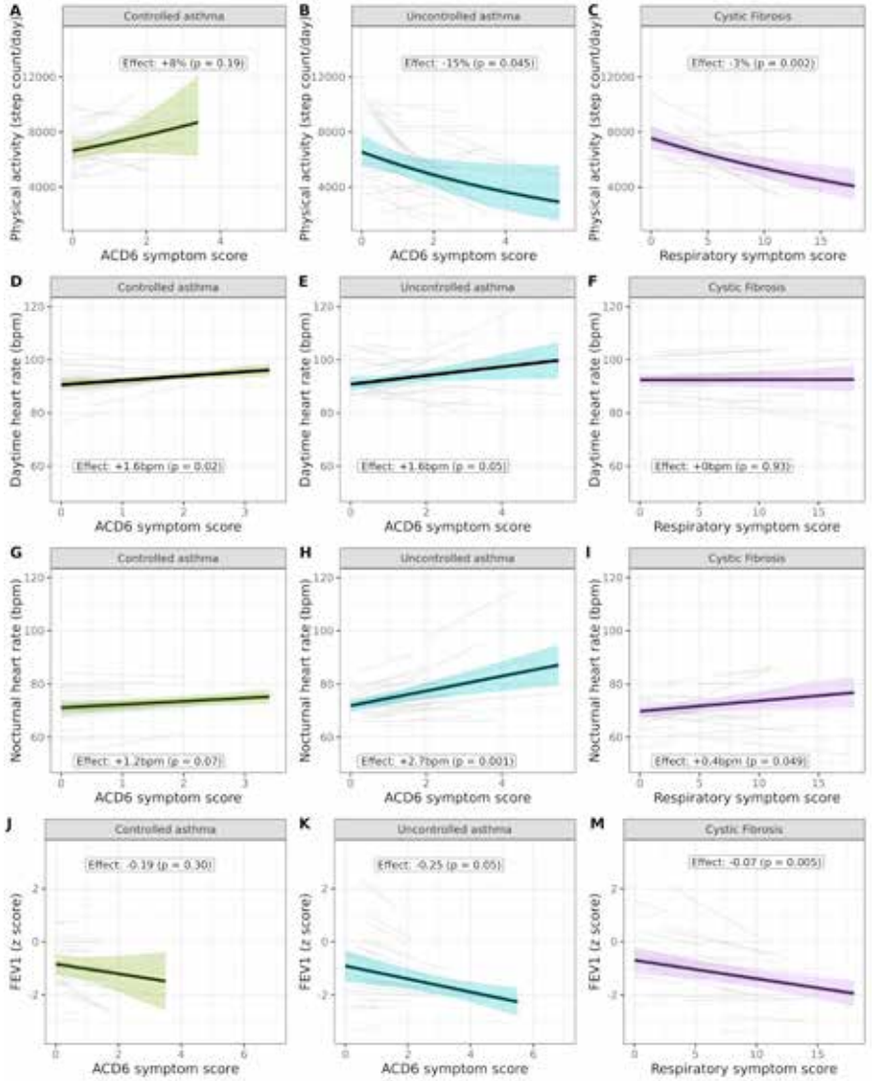
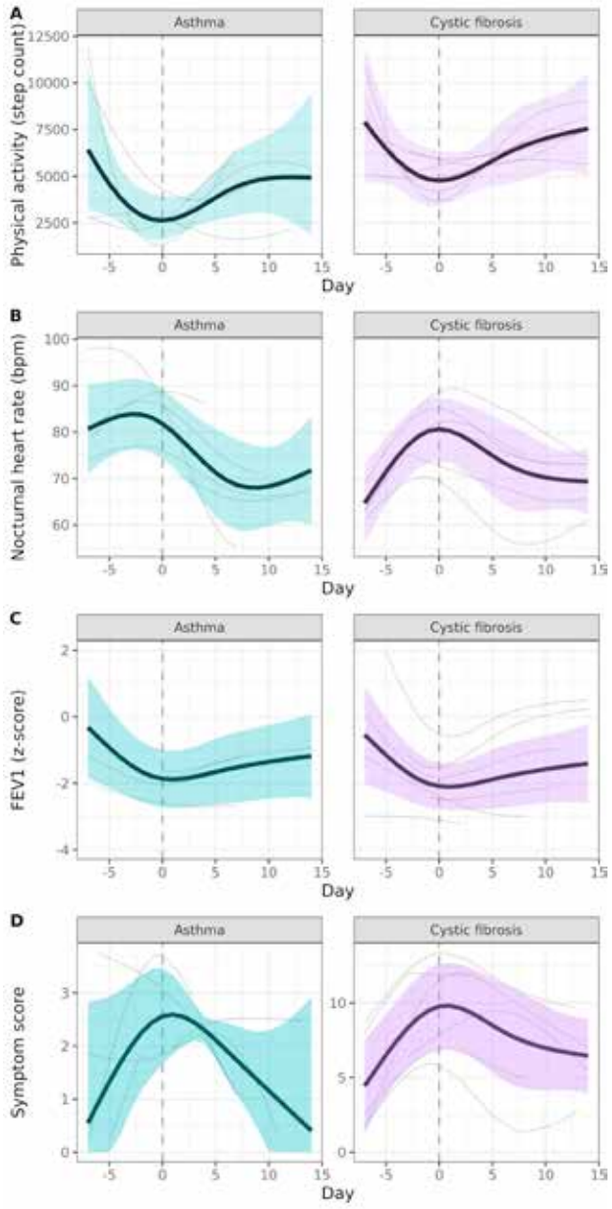


Figure 3. Description of health events. Estimated mean (95% CI) trajectory of physical activity, nocturnal heart rate, FEV1 and symptom score prior-during-and after prescription of rescue therapy (day 0) in the case of exacerbated disease for subjects with asthma (left column) and CF (right column). Bold lines represent the estimated mean. Transparent lines represent individual estimates.



Correlation existing disease metrics

Figure 2 shows the correlation between candidate endpoints and symptom scores. For subjects with uncontrolled asthma, there was a statistically significant relationship between ACD6 score and physical activity per day (15% decrease in step count per point increase in symptom score, 95% CI 0-29% p = 0.045, *Figure 2B*), but not for subjects with controlled asthma (+8% physical activity, 95% CI -5-21%, p=0.19, *Figure 2A*). For subjects with CF, one-point increase in symptom score was associated with a 3% decrease in activity (95% CI 1-5%, p=0.002, *Figure 2C*). Similar effects were found for Daily PA^{MAX} (*Supplementary Figure S5*).

Subjects with controlled asthma had, on average, a daytime HR that was 1.6 BPM higher per point increase in symptom score (95% CI 0.3 - 2.9, p=0.02, *Figure 2D*), and a nocturnal HR that was 1.2. BPM higher (95% CI -0.2 - 2.5, p=0.07) (*Figure 2G*). Daytime HR of subjects with uncontrolled asthma was 1.6 BPM higher per point increase (95% CI 0-3.3, p=0.05, *Figure 2E*), while nocturnal HR was 2.8 BPM higher (95% CI 1.2-4.3, p=0.001, *Figure 2H*). Subjects with CF had a 0.4 BPM higher nocturnal HR per point increase in symptom score (95% CI 0.03-0.75, p=0.049, *Figure 2I*), but no such effect on daytime HR was observed (*Figure 2F*).

Home-measured FEV1 was not correlated to symptom score for subjects with controlled asthma. Uncontrolled asthma subjects had a 0.25 lower FEV1 z-score for each point increase (95% CI 00-0.49, p=0.05 *Figure 2K*), while CF subjects had a 0.07% lower FEV z-score for each point increase (95% CI 0.02-0.12, p=0.005, *Figure 2M*). There was no correlation between FVC and symptom score. PA and FEV1 were correlated for subjects with uncontrolled asthma and CF (*Supplementary Figure S6*).

There was no correlation between ACD6 score and wakeup count, sleep duration or sleep depth. For CF subjects, there was some evidence of an association between wakeup count and respiratory symptom score (RR 1.03, 95% CI 1.00-1.06, p=0.035) but not for sleep duration or sleep depth. Adjustments for baseline disease-activity did not explain additional variance and were not included in the models.

Description of health events

During the study, 5 subjects with asthma and 8 subjects with CF had a case of exacerbated disease and were prescribed systemic corticosteroids and antibiotics, respectively. *Figure 3* displays the estimated mean (95% CI) trajectory of symptom score, physical activity, HR,

and pulmonary function on the 7 days prior to- and the 14 days after the first administration of rescue therapy (day 0). Estimating the same trajectory over time for subjects that did not experience an exacerbation revealed a stable pattern over time (*Supplementary Figure S7*)

Discussion

Innovations in personalized health technology provide a unique opportunity to initiate digital healthcare models and clinical trials that are built around pediatric patients' individual needs[35]. Despite multiple reports on the theoretical promises of wearables and other portable health devices, insufficient research has been performed regarding the clinical application and clinical validation of such measurements[36–38]. This study shows that candidate endpoints physical activity and HR fulfill most of the clinical validation criteria in pediatric patients with asthma and CF[15].

Tolerability and compliance are important predictors of clinical utility[39]. In this study, median overall compliance was 88% for all study assessments, and 100% for HR and physical activity. In addition, subjects found the study enjoyable and 88% of subjects would participate in similar studies. The lower spirometry compliance by subjects with uncontrolled asthma (68%) may be due to the fact that an effort is required for PFTS, leading to lower compliance for children with uncontrolled asthma, who are also generally less adherent to caregiver-instructions regarding their treatment compared to their well-controlled peers[40].

One advantage of monitoring via a wearable device compared to spirometry is the passive nature of data collection. In general, compliance to home-monitoring tasks significantly reduces over time, greatly diminishing the potential benefits[41]. Passive data collection may be less sensitive to this effect. Although spirometry has traditionally been the cornerstone of pulmonary health monitoring, the difficulty of the assessment compared to continuous monitoring by a wearable outside the clinic is a disadvantage in the context of home-monitoring. In this study overall compliance for PFTS was lower, and 36% of PFTS were discarded prior to analysis due to inadequate technique. This significant proportion of missing data impacted the power and generalizability of the analysis, as some subjects were more likely to exhibit bad technique compared to others. Furthermore, this finding raises doubt on the potential of PFTS for home-monitoring purposes. Indeed, previous studies investigating the value of home-based PFTS in pediatrics have reported no- or modest benefits[42–44]. Additionally, the EICE study in adults with CF reported that home-monitoring with PFTS and symptom scores combined with early intervention did not lead to a decrease in decline in pulmonary function compared to standard of care[45]. This study suffered from a similarly low compliance for home-based PFTS, and this indicates that passive monitoring of physical activity may have better value compared to PFT monitoring.

Important validation criteria for digital biomarkers include the difference between patients and controls and a correlation of novel digital endpoints with traditional endpoints, like symptom questionnaires[15]. We found that physical activity was lower in patients compared to controls, which is in agreement with findings that have been reported in the past[9,46], although other studies reported no significant differences in physical activity[47,48]. The differences were especially pronounced between 3PM and 7PM, and future studies may consider using activity during these hours as a separate endpoint (*Supplementary Figure S5*). Furthermore, physical activity was correlated with respiratory symptom scores for both CF and asthma, demonstrating the sensitivity for change in disease-activity. Both the difference in physical activity between asthmatic and healthy children, and the sensitivity of the endpoint to change in disease-activity are supported by Vahlkvist *et al.*, who showed that treatment with inhaled corticosteroids caused a significant increase in physical activity over time for children with recently diagnosed asthma[46]. A limitation of using step count as physical activity endpoint is that it does not capture all types of physical activity, such as cycling or swimming, which may have led to underestimated mean physical activity. The advantages of using a consumer smartwatch are a high compliance and relatively low cost compared to medical-grade devices[49].

For HR, differences between children with asthma and healthy children were observed for nocturnal- and daytime HR, and both were correlated with reported symptom scores. These observations are most likely due to a combination of disease- and pharmacological effects. Children with (uncontrolled) asthma often use (more) β2-agonists, causing elevated HR, which is a positive confounding factor in this analysis and part of the causal pathway between symptoms and heart rate[23,24]. Additional analyses adjusting for this confounder showed lower differences between patients and healthy children (*Supplementary Figure S3*), although the use of a smaller dataset due to questionnaire non-compliance led to increased uncertainty around the estimates. Still, the goal of the current study was to study the association between symptoms and heart rate and to demonstrate that a smartwatch can identify the magnitude of the difference in HR between patients and healthy controls. Considering that there was a difference in HR between healthy

children and patients with asthma and that HR was also responsive to a change in disease activity, we believe that (nocturnal) heart rate is a potential biomarker in real-life settings, irrespective of the underlying physiologic mechanism. Admittedly, digitally monitoring of rescue inhaler use may be a potential biomarker with similar usability[50]. To our knowledge, the application of smartwatch derived heart rate measurements in children with chronic lung disease has not been investigated in the past. In the future, more advanced analyses that integrate heart rate, inhaler use, and physical activity data may be considered to untangle the close relationship of the three variables in patients.

The variability of the investigated candidate biomarkers was assessed previously in healthy children[12], and is an important characteristic for power calculations in future clinical trials planning to utilize the biomarkers as endpoint. We found that ICC was lower for PA- and HR-derived endpoints in patients compared to healthy children, but not for FEV1. We hypothesize the lower ICC for PA and HR, indicating a higher intra-subject variability, is related to fluctuations in disease-activity inherent to the diseases. This is relevant for future clinical trials, since higher intra-subject variability necessitates larger samples sizes to detect clinically significant differences. However, aggregation of daily physical activity in weekly physical activity-related endpoints (*Supplementary Figure S1*) is more stable over time and may be suitable for long-term follow-up studies. A definition of what constitutes a clinically significant improvement is also necessary. Based on the current study, we believe that a change in physical activity or heart rate of 10% could be a reasonable cutoff. However, Future validation studies performed with novel or known (effective) treatments for asthma (e.g., inhalation corticosteroids) and CF should be performed to elucidate the magnitude of improvement in physical activity and other biomarkers that these treatments can elicit.

A final validation criterion is that novel endpoints should be able to discern and describe health events such as pulmonary exacerbations. These events are an important characteristic of severe disease and were defined by the need for rescue therapy during the study. Based on a limited dataset of subjects, physical activity, nocturnal HR and FEV1 appeared to be sensitive to the change in disease activity prior to rescue medication start and showed a distinct recovery curve during the days afterwards. Analysis of the individual trajectories revealed a similar pattern for all subjects throughout the exacerbation period. However, considering the limited sample size and the exploratory nature of this analysis, more research is needed to determine whether prodromal symptoms could provide an early warning sign for subjects and caregivers.

The candidate endpoints included in this study appear to fulfill the predefined clinical validation criteria and may be considered for use in clinical trials- and care. An improvement compared to traditional questionnaire assessments is that the proposed endpoints are objective in nature and less subject to recall bias, and may also assist children who find it difficult to perceive their own asthma-related symptoms[51,52]. Another application in asthma- and CF care that could have value is the prediction of disease control. If a smartphone application with access to the digital measurements predicts an increase in symptoms, it could suggest a specific intervention, which may prevent a pulmonary exacerbation. In the past, researchers have achieved promising results in this respect with asthmatic adults using only peak flow measurements[53], and incorporation of the measurements described in this manuscript may lead to even better predictions. This paper focuses on the necessary preparatory work required, and more longitudinal data of more subjects with more symptom score variability within subjects is needed.

This study has multiple limitations. One of which is that subjects with uncontrolled asthma were included some time after they were seen in the clinic, and an intervention to address the inadequate asthma control may have taken place during that time. Therefore, the true difference between subjects with uncontrolled asthma and subjects with controlled asthma may be more pronounced. Additionally, the smartwatch-derived data obtained is the study was obtained from a single smartwatch model and cannot necessary be extrapolated to smartwatches of other manufacturers. Another limitation is that the daily questionnaire employed for CF patients was not validated in the population. However, no (validated) daily symptom questionnaire was available in pediatric CF patients at the conception of this study.

A major challenge when analyzing datasets is missing data. The mixed effect models use maximum likelihood methods and are robust to randomly missing data, which we believe, based on our exit-interviews with study participants, is the case for the data employed via the smartwatch[54]. However, it is possible that subjects were less likely to perform a spirometry session (with adequate technique) or perform spirometry, on days with a high symptom load. This may have led to an underestimation of the differences between groups. However, the findings in this study may better correspond to the real-world conditions that will apply when the devices will be used in practice.

The sample size for all analyses is limited and future studies should include larger cohorts to increase the generalizability and robustness of the current findings. For example, adjustments for covariates identified in a previous study were performed via mixed

effect models, but only when the previously identified covariates explained additional variance according to prespecified goodness of fit criteria[29]. This was judged to give a good balance between explaining additional variance and risk of overfitting. Future analyses with larger cohorts can adjust for additional covariates with less risk of overfitting.

Strengths of this study include the systematic approach towards clinical validation, which excellently elucidates the characteristics of each individual candidate endpoint. The cohort of healthy children and patients with a wide range of disease-activity is large compared to comparable studies and allowed us to estimate group means representative for the target population. Although the study was not powered to detect many pulmonary exacerbations, the fact that 13 subjects received rescue treatment during the study allowed for a decent description of prodromal indices and recovery after exacerbations. Future research could focus on how to interpret measurements of a single patient in the context of clinical care and on alternative approaches such as fluctuation analyses[55]. Finally, data from a larger cohort with more symptom score variability can give an indication of the predictive capabilities of smartwatch data when monitoring patients with pediatric pulmonary disease.

Conclusion

Remote monitoring with a smartwatch and portable spirometer shows promise as digital biomarkers in pediatric lung disease. Physical activity-, HR-and pulmonary function monitoring is tolerable, can differentiate patients from controls and is correlated to symptom scores.

SUPPLEMENTARY DATA

Sup. Text S1	Respiratory symptom score questionnaire
Sup. Text S2	Technical validation data Steel hr watch
Sup. Table S1	Median compliance [IQR] during the study period for the 4 study groups
Sup. Figure S1	Difference between patients and healthy subjects for physical activity endpoints averaged per week
Sup. Figure S2	Physical activity between 3-7pm as candidate endpoint
Sup. Figure S3	Adjustment for β-agonist use
Sup. Figure S4	Sensitivity analysis
Sup. Figure S5	Correlation symptom score and daily PA^{MAX}
Sup. Figure S6	Correlation between physical activity and pulmonary function
Sup. Figure S7	Trajectory over time of subjects that did not experience pulmonary exacerbation

REFERENCES

1 Graham BL, Steenbruggen I, Barjaktarevic IZ, Cooper BG, Hall GL, Hallstrand TS, Kaminsky DA, McCarthy K, McCormack MC, Miller MR, Oropez CE, Rosenfeld M, Stanojevic S, Swanney MP, Thompson BR. Standardization of spirometry 2019 update an official American Thoracic Society and European Respiratory Society technical statement. *Am. J. Respir. Crit. Care Med.* 2019; 200: E70-E88.

2 Seed L, Wilson D, Coates AL. Children should not be treated like little adults in the PFT lab. *Respir. Care* 2012; 57: 61-74.

3 O'Neill K, Tunney MM, Johnston E, Rowan S, Downey DG, Rendall J, Reid A, Bradbury I, Elborn JS, Bradley JM. Lung Clearance Index in Adults and Children With Cystic Fibrosis. *Chest* [Internet] Elsevier Inc; 2016; 150: 1323-1332Available from: http://dx.doi.org/10.1016/j.chest.2016.06.029.

4 Greenberg RG, Corneli A, Bradley J, Farley J, Jafri HS, Lin L, Nambiar S, Noel GJ, Wheeler C, Tiernan R, Smith PB, Roberts J, Benjamin DK. Perceived barriers to pediatrician and family practitioner participation in pediatric clinical trials: Findings from the Clinical Trials Transformation Initiative. *Contemp. Clin. Trials Commun.* [Internet] Elsevier; 2018; 9: 7-12Available from: https://doi.org/10.1016/j.conctc.2017.11.006.

5 Kruizinga MD, Stuurman FE, Groeneveld GJ, Cohen AF. The Future of Clinical Trial Design : The Transition from Hard Endpoints to Value-Based Endpoints..

6 Szefler SJ, Chmiel JF, Fitzpatrick AM, Giacoia G, Green TP, Jackson DJ, Nielsen HC, Phipatanakul W, Raissy HH. Asthma across the ages: Knowledge gaps in childhood asthma. *J. Allergy Clin. Immunol.* [Internet] Elsevier Ltd; 2014; 133: 3-13Available from: http://dx.doi.org/10.1016/j.jaci.2013.10.018.

7 Joseph PD, Craig JC, Caldwell PHY. Clinical trials in children. *Br. J. Clin. Pharmacol.* 2015; 79: 357-369.

8 Kelly LE, Sinha Y, Barker CIS, Standing JF, Offringa M. Useful pharmacodynamic endpoints in children: Selection, measurement, and next steps. *Pediatr. Res.* [Internet] Nature Publishing Group; 2018; 83: 1095-1103Available from: http://dx.doi.org/10.1038/pr.2018.38.

9 Lang DM, Butz AM, Duggan AK, Serwint JR. Physical activity in urban school-aged children with asthma. *Pediatrics* 2004; 113.

10 Braido F, Baiardini I, Ferrando M, Scichilone N, Santus P, Petrone A, Di Marco F, Corsico AG, Zanforlin A, Milanese M, Steinhilber G, Bonavia M, Pirina P, Micheletto C, D'Amato M, Lacedonia D, Benassi F, Propati A, Ruggeri P, Tursi F, Bocchino ML, Patella V, Canonica GW, Blasi F. The prevalence of sleep impairments and predictors of sleep quality among patients with asthma. *J. Asthma* [Internet] Taylor & Francis; 2020; 0: 1-7Available from: https://doi.org/10.1080/02770903.2019.1711391.

11 Lu TC, Fu C-M, Ma M, Fang CC, Turner AM. Healthcare Applications of Smart Watches. *Appl. Clin. Inform.* [Internet] 2016; 7: 850-869Available from: http://www.schattauer.de/index.php?id=1214&doi=10.4338/ACI-2016-03-R-0042.

12 Kruizinga MD, Heide N van der, Moll A, Zhuparris A, Yavuz Y, Kam ML de, Stuurman FE, Cohen AF, Driessen GJA. Towards remote monitoring in pediatric care and clinical trials-Tolerability, repeatability and reference values of candidate digital endpoints derived from physical activity, heart rate and sleep in healthy children. PLOS *One* United States; 2021; 16: e0244877.

13 Kruizinga MD, Essers E, Stuurman FE, Zhuparris A, van Eik N, Janssens HM, Groothuis I, Sprij AJ, Nuijsink M, Cohen AF, Driessen GJA. Technical validity and usability of a novel smartphone-connected spirometry device for pediatric patients with asthma and cystic fibrosis. *Pediatr. Pulmonol.* 2020; : 2463-2470.

14 Avdimiretz N, Wilson D, Grasemann H. Comparison of a handheld turbine spirometer to conventional spirometry in children with cystic fibrosis. *Pediatr. Pulmonol.* 2020; : 1394-1399.

15 Kruizinga MD, Stuurman FE, Exadaktylos V, Doll RJ, Stephenson DT, Groeneveld GJ, Driessen GJA, Cohen AF. Development of Novel, Value-Based, Digital Endpoints for Clinical Trials: A Structured Approach Toward Fit-for-Purpose Validation. *Pharmacol Rev* 2020; 72 (4): 899-909.

16 Coran P, Goldsack C, Grandinetti A. Advancing the Use of Mobile Technologies in Clinical Trials : Recommendations from the Clinical Trials Transformation Initiative. 2019; 27701: 145-154.

17 Babrak LM, Menetski J, Rebhan M, Nisato G, Zinggeler M, Brasier N, Baerenfaller K, Brenzikofer T, Baltzer L, Vogler C, Gschwind L, Schneider C, Streiff F, Groenen PMA, Miho E. Traditional and Digital Biomarkers: Two Worlds Apart? *Digit. Biomarkers* 2019; 3: 92-102.

18 Kavanagh J, Jackson DJ, Kent BD. Over-and under-diagnosis in asthma. *Breathe* 2019; 15: e20-e27.

19 Juniper EF, Gruffydd-Jones K, Ward S, Svensson K. Asthma control questionnaire in children: Validation, measurement properties, interpretation. *Eur. Respir. J.* 2010; 36: 1410-1416.

20 El Moussaoui R, Opmeer BC, Bossuyt PMM, Speelman P, De Borgie CAJM, Prins JM. Development and validation of a short questionnaire in community acquired pneumonia. *Thorax* 2004; 59: 591-595.

21 Mantua J, Gravel N, Spencer RMC. Reliability of sleep measures from four personal health monitoring devices compared to research-based actigraphy and polysomnography. *Sensors* (*Switzerland*) 2016; 16.

22 Quanjer PH, Stanojevic S, Cole TJ, Baur X, Hall GL, Culver BH, Enright PL, Hankinson JL, Ip MSM, Zheng J, Stocks J, Schindler C. Multi-ethnic reference values for spirometry for the 3-95-yr age range: The global lung function 2012 equations. *Eur. Respir. J.* 2012; 40: 1324-1343.

23 Varni JW, Seid M, Kurtin PS. PedsQLTM 4.0: Reliability and Validity of the Pediatric Quality of Life InventoryTM Version 4.0 Generic Core Scales in Healthy and Patient Populations. *Med. Care* [Internet] 2001; 39Available from: https://journals.lww.com/lww-medicalcare/Fulltext/2001/08000/PedsQL_4_0_Reliability_and_Validity_of_the.6.aspx.

24 Juniper EF, Bousquet J, Abetz L, Bateman ED. Identifying "well-controlled" and "not well-controlled" asthma using the Asthma Control Questionnaire. *Respir. Med.* 2006; 100: 616–621.

25 Klijn PH, van Stel HF, Quittner AL, van der Net J, Doeleman W, van der Schans CP, van der Ent CK. Validation of the Dutch cystic fibrosis questionnaire (CFQ) in adolescents and adults. *J. Cyst. Fibros.* 2004; 3: 29–36.

26 Raat H, Bueving HJ, De Jongste JC, Grol MH, Juniper EF, Van Der Wouden JC. Responsiveness, longitudinal- and cross-sectional construct validity of the Pediatric Asthma Quality of Life Questionnaire (PAQLQ) in Dutch children with asthma. *Qual. Life Res.* 2005; 14: 265–272.

27 Evans EW, Abrantes AM, Chen E, Jelalian E. Using Novel Technology within a School-Based Setting to Increase Physical Activity: A Pilot Study in School-Age Children from a Low-Income, Urban Community. *Biomed Res. Int.* 2017; 2017.

28 Byun W, Lau EY, Brusseau TA. Feasibility and effectiveness of a wearable technology-based physical activity intervention in preschoolers: A pilot study. *Int. J. Environ. Res. Public Health* 2018; 15.

29 Zuur A, Ieno EN, Walker N, Saveliev AA, Smith GM. Mixed effects models and extensions in ecology with R. Springer Science & Business Media; 2009.

30 Schielzeth H, Forstmeier W. Conclusions beyond support: Overconfident estimates in mixed models. *Behav. Ecol.* 2009; 20: 416–420.

31 Anne Fuhlbrigge, David Peden, Andrea J Apter, Homer A Boushey, Carlos A Camargo Jr, James Gern, Peter W Heymann, Fernando D Martinez, David Mauger, William G Teague CB. Asthma Outcomes: Exacerbations. *J Allergy Clin Immunol. 2012 March; 129(3 Suppl) S34–S48. doi10.1016/j.jaci.2011.12.983.*

32 Bhatt JM. Treatment of pulmonary exacerbations in cystic fibrosis. *Eur. Respir. Rev.* 2013; 22: 205–216.

33 Kruizinga MD, Moll A, Zhuparris A, Ziagkos D, Stuurman FE, Nuijsink M, Cohen AF, Driessen GJA. Postdischarge Recovery after Acute Pediatric Lung Disease Can Be Quantified with Digital Biomarkers. *Respiration* 2021;.

34 Lambrechtse P, Ziesenitz VC, Atkinson A, Bos EJ, Welzel T, Gilgen Y, Gürtler N, Heuscher S, Cohen AF, van den Anker JN. Monitoring the recovery time of children after tonsillectomy using commercial activity trackers. *Eur. J. Pediatr.* European Journal of Pediatrics; 2021; 180: 527–533.

35 Greiwe J, Nyenhuis SM. Wearable Technology and How This Can Be Implemented into Clinical Practice. *Curr. Allergy Asthma Rep.* Current Allergy and Asthma Reports; 2020; 20.

36 Bakker JP, Goldsack JC, Clarke M, Coravos A, Geoghegan C, Godfrey A, Heasley MG, Karlin DR, Manta C, Peterson B, Ramirez E, Sheth N, Bruno A, Bullis E, Wareham K, Zimmerman N, Forrest A, Wood WA. A systematic review of feasibility studies promoting the use of mobile technologies in clinical research. *NPJ Digit. Med.* [Internet] Springer US; 2019; 2Available from: http://dx.doi.org/10.1038/s41746-019-0125-x.

37 Dunn J, Runge R, Snyder M. Wearables and the medical revolution. *Per. Med.* 2018; 15: 429–448.

38 Izmailova ES, Wagner JA, Perakslis ED. Wearable Devices in Clinical Trials: Hype and Hypothesis. *Clin. Pharmacol. Ther.* 2018; 104: 42–52.

39 Müller J, Hoch AM, Zoller V, Oberhoffer R. Feasibility of physical activity assessment with wearable devices in children aged 4–10 Years–A Pilot study. *Front. Pediatr.* 2018; 6: 1–5.

40 Klok T, Kaptein AA, Brand PLP. Non-adherence in children with asthma reviewed: The need for improvement of asthma care and medical education. *Pediatr. allergy Immunol. Off. Publ. Eur. Soc. Pediatr. Allergy Immunol.* England; 2015; 26: 197–205.

41 Gupta RS, Fierstein JL, Boon KL, Kanaley MK, Bozen A, Kan K, Vojta D, Warren CM. Sensor-Based Electronic Monitoring for Asthma: A Randomized Controlled Trial. *Pediatrics* United States; 2021; 147.

42 Shakkottai A, Kaciroti N, Kasmikha L, Nasr SZ. Impact of home spirometry on medication adherence among adolescents with cystic fibrosis. *Pediatr. Pulmonol.* 2018; 53: 431–436.

43 Brouwer AFJ, Roorda RJ, Brand PLP. Home spirometry and asthma severity in children. *Eur. Respir. J.* 2006; 28: 1131–1137.

44 Thompson R, Delfino RJ, Tjoa T, Nussbaum E, Cooper D. Evaluation of daily home spirometry for school children with asthma: New insights. *Pediatr. Pulmonol.* 2006; 41: 819–828.

45 Lechtzin N, Mayer-Hamblett N, West NE, Allgood S, Wilhelm E, Khan U, Aitken ML, Ramsey BW, Boyle MP, Mogayzel PJ, Gibson RL, Orenstein D, Milla C, Clancy JP, Antony V, Goss CH. Home monitoring of patients with cystic fibrosis to identify and treat acute pulmonary exacerbations eICE study results. *Am. J. Respir. Crit. Care Med.* 2017; 196: 1144–1151.

46 Vahlkvist S, Inman MD, Pedersen S. Effect of asthma treatment on fitness, daily activity and body composition in children with asthma. *Allergy Eur. J. Allergy Clin. Immunol.* 2010; 65: 1464–1471.

47 Van Gent R, Van Der Ent CK, Van Essen-Zandvliet LEM, Rovers MM, Kimpen JLL, De Meer G, Klijn PHC. No differences in physical activity in (Un)diagnosed asthma and healthy controls. *Pediatr. Pulmonol.* 2007; 42: 1018–1023.

48 Matsunaga NY, Oliveira MS, Morcillo AM, Ribeiro JD, Ribeiro MAGO, Toro AADC. Physical activity and asthma control level in children and adolescents. *Respirology* 2017; 22: 1643–1648.

49 McLellan G, Arthur R, Buchan DS. Wear compliance, sedentary behaviour and activity in free-living children from hip- and wrist-mounted ActiGraph GT3X+accelerometers. *J. Sports Sci.* [Internet] Routledge; 2018; 36: 2424–2430Available from: https://doi.org/10.1080/02640414.2018.1461322.

50 Himes BE, Leszinsky L, Walsh R, Hepner H, Wu AC. Mobile Health and Inhaler-Based Monitoring Devices for Asthma Management. *J. allergy Clin. Immunol. Pract.* 2019; 7: 2535–2543.

51 Forno E, Abraham N, Winger DG, Rosas-Salazar C, Kurland G, Weiner DJ. Perception of Pulmonary Function in Children with Asthma and Cystic Fibrosis. *Pediatr. Allergy, Immunol. Pulmonol.* 2018; 31: 139–145.

52 Baker RR, Mishoe SC, Zaitoun FH, Arant CB, Lucas J, Rupp NT. Poor perception of airway obstruction in children with asthma. *J. Asthma* 2000; 37: 613–624.

53 Thamrin C, Zindel J, Nydegger R, Reddel HK, Chanez P, Wenzel SE, Fitzpatrick S, Watt RA, Suki B, Frey U. Predicting future risk of asthma exacerbations using individual conditional probabilities. *J. Allergy Clin. Immunol.* 2011; 127.

54 Bennett DA. How can I deal with missing data in my study? *Aust. N. Z. J. Public Health* 2001; 25: 464–469.

55 Frey U, Brodbeck T, Majumdar A, Robin Taylor D, Ian Town G, Silverman M, Suki B. Risk of severe asthma episodes predicted from fluctuation analysis of airway function. *Nature* 2005; 438: 667–670.

CHAPTER 8

Post-discharge recovery after acute pediatric lung disease can be quantified with digital biomarkers

Respiration 2021. doi: 10.1159/000516328

Matthijs D Kruizinga,[1,2,3], Allison Moll,[1,2] Ahnjili Zhuparris,[1] Dimitrios Ziagkos,[1]
Frederik E Stuurman,[1,3] Marianne Nuijsink,[2] Adam F Cohen,[1,3] Gertjan JA Driessen,[2,4]

1 Centre for Human Drug Research, Leiden, the Netherlands
2 Juliana Children's Hospital, Haga Teaching Hospital, The Hague, the Netherlands
3 Leiden University Medical Center, Leiden, the Netherlands
4 Maastricht University Medical Centre, Maastricht, the Netherlands

Abstract

BACKGROUND Pediatric patients admitted for acute lung disease are treated and monitored in the hospital, after which full recovery is achieved at home. Many studies report in-hospital recovery, but little is known regarding the time to full recovery after hospital discharge. Technological innovations have led to increased interest in home-monitoring and digital biomarkers. The aim of this study was to describe at-home recovery of three common pediatric respiratory diseases using a questionnaire and wearable device.

METHODS In this study, patients admitted due to pneumonia (n=30), preschool wheezing (n=30) and asthma exacerbation (n=11) were included. Patients were monitored with a smartwatch and a questionnaire during admission, a 14-day recovery period and a 10-day 'healthy' period. Median compliance was calculated, and a mixed effects model was fitted for physical activity and heart rate to describe the recovery period, and the physical activity recovery trajectory was correlated to respiratory symptom scores.

RESULTS Median compliance was 47% (IQR 33-81%) during the entire study period, 68% (IQR 54-91%) during the recovery period and 28% (IQR 0-74%) during the healthy period. Patients with pneumonia reached normal physical activity 12 days post-discharge, while subjects with wheezing and asthma exacerbation reached this level after 5 and 6 days, respectively. Estimated mean physical activity was closely correlated with estimated mean symptom score. Heart rate measured by the smartwatch showed a similar recovery trajectory for subjects with wheezing and asthma, but not for subjects with pneumonia.

CONCLUSIONS The digital biomarkers physical activity and heart rate obtained via smartwatch show promise for quantifying post-discharge recovery in a non-invasive manner, which can be useful in pediatric clinical trials and clinical care.

Introduction

Pediatric patients who are admitted to the pediatric ward for acute lung disease are treated as in-patients until the clinical condition is stable enough for safe discharge. Although the clinical condition is improved at discharge, complete recovery is usually achieved at home.

While much is known about in-hospital recovery, little data is available regarding the time to-full recovery after hospital discharge. Studies researching acute lung disease, such as community-acquired pneumonia (CAP) and preschool wheezing (PW) define recovery time as the duration of hospital stay, but these studies don't describe the at-home recovery period once patients have been discharged.[1–3] One study investigated at-home recovery time in children admitted for asthma exacerbation (AE), but this definition relied on spirometry only, which does not always correspond with symptoms in children.[4,5] As a result, this important clinical disease characteristic does not reach clinical reviews or reference texts, while insight into duration to full recovery could be valuable for patients, their parents, clinicians and clinical researchers.[6,7] It would allow for better insight in expected short-term disease burden, and also for investigating the effect of treatments, such as steroids and antibiotics, beyond the hospital-setting.[8,9]

Current methods to follow subjects in a home-setting, such as questionnaires, have limitations such as recall bias and are inherently subjective.[10] Frequent monitoring at a subject's home by a physician is time consuming and expensive and daily hospital visits place an unwanted burden on both child and parent.[11,12] Technological innovations have led to wearables and other devices that can continuously measure health parameters such as physical activity, heart rate (HR) and sleep pattern. When combined with electronic patient reported outcomes (EPROS), remote monitoring platforms can collect both objective and subjective high resolution data in an at home-setting, which decreases the burden for children significantly, and may even allow for a personalized medicine approach.[13,14]

However, before implementing digital biomarkers in clinical care or clinical trials, extensive fit-for-purpose technical- and clinical validation in the target population is necessary. Technical validation consists of investigating whether a device actually captures that what is claimed on a technical level, whereas clinical validation focuses on the tolerability, difference between patients and controls and correlation with traditional biomarkers [15]. One of the final clinical validation criteria in a recently published validation strategy is the ability to detect clinically meaningful change after a health event such as a hospital admission, and this study aims to address that criterion for several candidate digital biomarkers.[15]

The aim of this pilot study was to investigate the post-admission recovery time of children admitted due to pneumonia, preschool wheezing, and asthma and to evaluate the potential of remote monitoring with digital biomarkers for pediatric clinical trials and -care.

Materials and methods

This study was performed at the Juliana Children's Hospital (The Hague, the Netherlands) in collaboration with the Centre of Human Drug Research (CHDR) in Leiden between November 2018 until March 2020 and was conducted in compliance with Good Clinical Practice and the Dutch Code of Conduct regarding medical research with minors. Signed informed consent was obtained from parents or the legal guardian and of children aged ≥12 years prior to any study-mandated procedure.

Subjects and study design

Patients aged 2-12 admitted to the pediatric ward of the Juliana Children's Hospital due to CAP (n=30), PW (n=30) and AE (n=11) were recruited. Patients with a history of chronic illness other than the studied disease were excluded. Patients were enrolled in the study as quickly as possible after admittance and were monitored during hospital admission, a 14-day recovery period immediately after hospital discharge and another 10-day period 40 days after hospital discharge. The schedule detailing the study periods can be found in *Figure 1*. Subjects admitted due to nonmedical reasons were excluded from participation. There were no formal discharge criteria. In general, patients were discharged when they were no longer oxygen-therapy dependent during the night. In addition, patients with asthma or preschool wheezing were discharged when nebulization with bronchodilators was no longer necessary.

Figure 1. Study schedule

Study assessments

During the study periods all patients were asked to continuously wear a Withings® Steel HR Smartwatch (Withings, Issy-les-Molineux, France), which collects step count, HR, and sleep pattern. Furthermore, subjects were asked to complete a daily questionnaire. An existing respiratory symptom questionnaire was adapted for a pediatric population, and subjects and their parents of the CAP and PW cohorts were asked to complete the questionnaire at the end of each day to determine a respiratory symptom score (*Supplementary Text S1*).[16] in addition, these patients performed daily temperature measurements with the Withings Thermo (Withings, Issy-les-Molineux, France). Patients with AE completed the Modified Asthma Control Diary (ACD)[17] and spirometry measurements with the Air Next spirometer (NuvoAir, Stockholm, Sweden) every day. The spirometer registers the forced vital capacity (FVC), the forced expiratory volume in the 1ST second (FEV1) and the FEV1/FVC ratio.[18] All devices were connected to a G6 smartphone (Motorola, Chicago, IL, USA)) with the HealthMate, Thermo and CHDR MORE® applications pre-installed. The devices used during the study have previously been used in an initial validation study in healthy children.[19] Additionally, the smartphone calendar was filled with a personalized study schedule. At the end of the study, participants were asked to complete a questionnaire regarding the study experience. Baseline and admission characteristics were obtained from patient charts.

Analysis and Statistics

COMPLIANCE AND BASELINE CHARACTERISTICS Compliance was determined by dividing the sum of the completed measurements by the total of the expected measurements for each subject, and the median compliance and interquartile range (IQR) were calculated. Compliance was calculated for the complete study period and for the recovery- and healthy period separately. Descriptive statistics were used to describe the baseline and admission characteristics.

MODELLING ANALYSIS SET The primary endpoint in this study was physical activity (step count). However, individual exploratory plots of physical activity over the entire study period (*Supplementary Figure S2*) showed a large amount of inter-individual variability between subjects. To define a common point of recovery (return to 'healthy'

physical activity levels), subject data was normalized based on data gathered during the 10-day 'healthy' period (day 15-24), since a true baseline period was obviously not possible within the study design. For each individual subject, the mean step count in the healthy period (minimum of 2 days) was used as their "baseline" physical activity (100%). If no data of the healthy period was available, the mean steps of the last 4 days of the recovery period (minimum of 2 days) were used as reference. Thus, only subjects who performed measurements for at least 2 days in the healthy period or at least 2 days in the last four days of the recovery period were included in the analysis set.

SYMPTOM SCORE MODEL To visualize a recovery trajectory, a descriptive linear mixed model was fitted to model respiratory symptom score and ACD6 score for the three groups separately using the restricted maximum likelihood approach (REML). In this analysis, symptom scores reported before 12PM were assumed as data from the previous day. Time was included as a spline covariate with a maximum of 3 degrees of freedom to assess nonlinear recovery. Subject was included as random intercept. During analysis, contribution to model fit was assessed via a likelihood-ratio test and by appraising the Akaike Information Criterion (AIC) and proportion of variance explained (R^2). Model assumptions were checked by inspection of residual plots. Log transformation was performed in the presence of heteroscedasticity. First order autoregression on the time variable was included to account for temporal autocorrelation.

PHYSICAL ACTIVITY MODEL Physical activity was modelled descriptively using similar methods. In the model, study day number was included as a spline covariate with a maximum of 3 degrees of freedom and subject was included as a random intercept. The watch wear time between 6AM and 10PM was included as separate covariate to adjust for partial noncompliance during the day. Estimated mean (95% confidence interval (CI)) physical activity was calculated over time, with wear time of the watch held constant at 100%. Admission duration, oxygen saturation, respiratory rate and HR at admission were included in the models separately as part of an exploratory analysis to assess their effect on recovery.

HEART RATE Average HR during the day (6AM-10PM) and average nocturnal HR (12AM-5AM) were both fitted with a mixed effects model with subject as random intercept and time as spline covariate. Age was included as additional covariate.

RELATIONSHIP BETWEEN RECOVERY TRAJECTORIES To quantify the relationship between the three estimated recovery trajectories, Pearson correlations were performed to quantify the relationship between estimated mean daily physical activity, HR and symptom score.

SOFTWARE PySpark version 2.4.6 was used for data aggregation and tabulation. The statistical analyses were performed using R version 3.6.1 with R-packages nlme, em means and gg effects.

Results

Baseline & admission characteristics

Of the 71 patients included in the study, 20 subjects dropped out shortly after inclusion due to discomfort for the child and were excluded during analysis. This majority of dropped out subjects were 2 (n=5) or 3 (n=6) years old. The remaining study population consisted of 19 pneumonia patients, 21 preschool wheezing patients and 11 asthma patients. Baseline and admission characteristics are shown in *Table 1*. Of the 51 subjects shown in *Table 1*, 39 subjects completed measurements for either at least two days in the healthy period or at least two days in the last four days of the recovery period and thus were suitable for inclusion in the modelling dataset.

Compliance

For the 51 subjects who completed the study, compliance was determined separately for the entire study period, the recovery period and healthy period (*Figure 2*). Median compliance was 47% (IQR 33-81%) during the entire study period, 68% (IQR 54-91%) during the recovery period and 28% (IQR 0-74%) during the healthy period. There was no clear association between age and compliance (*Supplementary Figure S3*). Raw data of the subjects remaining in the final dataset is presented in *Supplementary Figure S4*. Considering the large number of subjects with multiple days of missing data, a mixed effects modelling approach accounting for both missing data from complete study days with a random effect structure and for partial noncompliance during the day by adjusting for wear time was most appropriate.

Table 1. Baseline and admission characteristics.

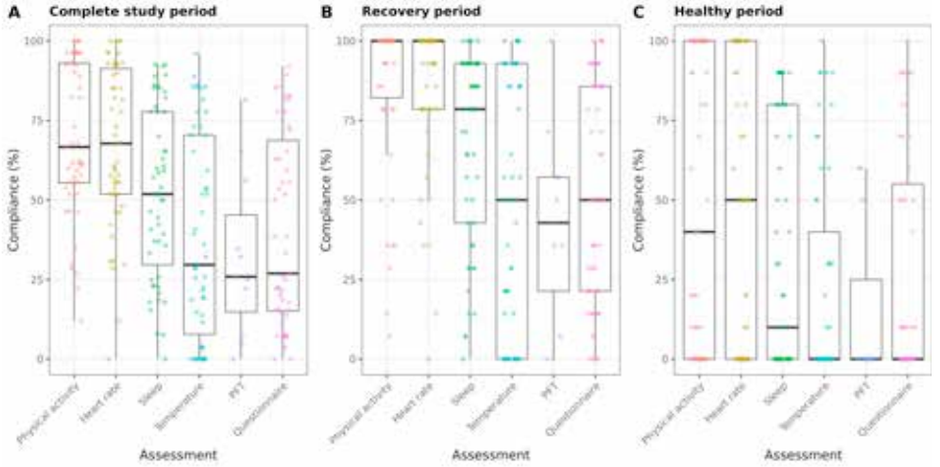
	All subjects n = 51	CAP n = 19	PW n = 21	AE n = 11
Age (mean years ± SD)	4.9 ± 2.9	3.6 ± 2.0	3.7 ± 1.2	9.3 ± 2.2
Gender (Male %)	68.6%	63.2%	76.2%	63.6%
Ethnicity (Caucasian %)	45.1%	52.6%	38.1%	45.5%
Admission duration (mean hrs ± SD)	62.2 ± 47.0	63.6 ± 27.3	48.7 ± 21.6	85.3 ± 84.6
Admission heart rate (mean BPM ± SD)	147.0 ± 23.3	151.6 ± 23.9	152.6 ± 20.1	128.3 ± 18.6
Admission respiratory rate (mean breaths/min)	43.1 ± 15.4	44.8 ± 13.9	45.1 ± 16.8	36.7 ± 14.0
Admission oxygen saturation (mean O2% ± SD)	92.8 ± 3.1	92.2 ± 3.2	92.6 ± 2.9	93.6 ± 3.5
Oxygen therapy (Yes %)	86.3%	94.7%	90.5%	63.6%
Smoking at home (Yes %)	23.5%	26.3%	9.5%	45.5%
Family history of respiratory problems (Yes %)	62.7%	52.6%	61.9%	81.8%
Day care attendance (Yes %)	23.5%	31.6%	28.6%	0%

Abbreviations: CAP: community-acquired pneumonia, PW: preschool wheezing, AE: asthma exacerbation

Symptom score decreases over time after discharge

ACD6- and respiratory symptom score were modelled for each diagnosis. Estimated mean symptom scores are displayed in *Figure 3A-C*. For CAP patients, average respiratory symptom score decreased from 11.2 (95% CI 9.7-12.7) at discharge to 2.8 (95% CI 1.5-4.1) at day 12. PW patients exhibited a mean symptom score of 10.2 (95% CI 8.8-11.6) at discharge, which decreased towards 4.6 (3.2-6.0) at day 5 and plateaued after this timepoint. Finally, subjects with AE used a different questionnaire and exhibited an ACD6 score of 2.5 (95% CI 1.9-3.1) at discharge and reached a plateau after day 6 (mean 0.8, 95% CI 0.1-1.4). Final model coefficients are displayed in *Supplementary Table S5*.

Figure 2. Compliance to study tasks. Median (IQR) compliance for all measurements combined and for individual measurements. Each dot represents an individual subject. Temperature assessments were performed by CAP and PW subjects, while PFTS were performed by AE subjects. A: entire study period, B: 14-day recovery period, C: 10-day healthy period after a 25-day break.

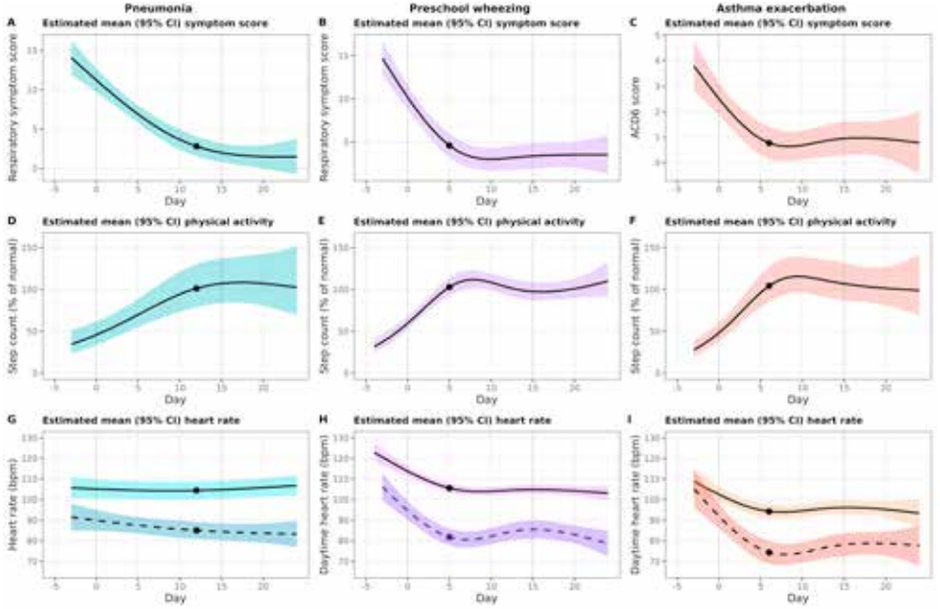


Physical activity displays an inverse pattern compared to symptom score

Estimated mean physical activity for each disease group is displayed in *Figure 3D-F*. At discharge, physical activity of the CAP group was 46% of normal levels, and, on average, patients achieved 100% of normal physical activity levels after 12 days. The PW population had a mean physical activity of around 59% of normal levels at discharge and achieved 100% physical activity earlier compared to CAP patients after 5 days. The AE population had an estimated mean physical activity of around 48% at discharge and reached a mean of 100% after 6 days. Final model coefficients are displayed in *Supplementary Table S5*.

Admission-duration, -O2 saturation, -HR and -respiratory rate were separately introduced to the final models as covariate during exploratory analyses. If these variables influence the average recovery trajectory of patients, model fit should improve significantly. However, only admission duration improved model fit in the case of CAP (δAIC 17, $p < 0.001$) and PW (δAIC 9, $p = 0.002$). Other variables did not improve model fit. Estimated model effects for the CAP and PW group are displayed in *Supplementary Figure S6*.

Figure 3. Estimated recovery trajectory. Estimated mean physical activity over time (A–C), symptom score over time (D–F) and daytime–and nocturnal heart rate (G–I) for pneumonia patients (left column), preschool wheezing patients (central column) and asthma patients (right column). The black lines indicate the estimated population mean; shaded areas represent the estimated 95% confidence intervals of the mean. The black dots indicate at the day where physical activity first reaches 100% of normal levels. For heart rate, darker shaded area's and dotted lines represent nocturnal heart rate and lighter shaded area and solid lines represent daytime heart rate. Estimated heart rate was adjusted for age.



Heart rate decreases in PW and AE

Daytime and nocturnal HR were modelled using the same subjects as for the analysis of physical activity and symptom scores. Mean daytime HR at discharge was 114 BPM for subjects with PW, which decreased to 105 BPM after 5 days. For subjects with asthma, mean HR during the day was 102 BPM at discharge and stabilized at 94 BPM after 6 days, after which HR remained at a stable level until the end of the study. For subjects with pneumonia, no such pattern was observed. Nocturnal HR was modelled separately and displayed similar trends. Estimated mean HR over time is displayed in *Figure 3G–I*. Final model coefficients are displayed in *Supplementary Table S5*.

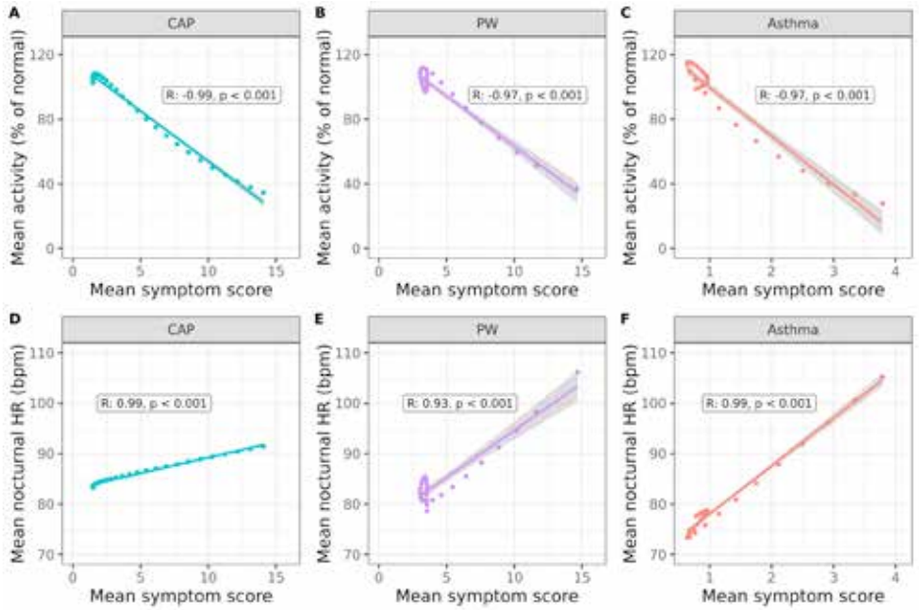
Estimated mean symptom score, physical activity and heart rate are correlated

Correlations between model-predicted mean symptom score, physical activity and nocturnal HR were performed to quantify the relationships between the estimated recovery trajectories (*Figure 4*). Physical activity was inversely related to symptom score and HR was positively correlated with symptom score for all three disease groups

Temperature, sleep, and spirometry

Marginal mean total sleep duration per diagnosis was estimated and is displayed in *Supplementary Figure S7*. Total compliance for PFTS and temperature was considered too low to attempt further analysis. However, individual plots of a highly adherent subject (*Supplementary Figure S8*) showed that PFTS in the home-setting have the potential to differentiate between the acute–and recovered state in the case of good compliance.

Figure 4. Correlation between traditional and novel methods to quantify recovery. Pearson correlations between model-estimated mean symptom score and model-estimated mean physical activity (A–C) and between model-estimated mean symptom score and–nocturnal heart rate (D–F) for each patient group.



Subject satisfaction

Twenty subjects completed the end-of-study questionnaire. The watch scored 2.5 out of 5 on account of being 'painful', but also scored 4.0, 3.5 and 4.1 out of 5 for being 'fun', 'comfortable' and 'easy to wear all day', respectively. Five (25%) parents reported the watch caused some discomfort, itching, irritable skin, or a rash, and 85% of subjects would be willing to participate in a similar study in the future.

Discussion

This study is one of the first to describe the at-home recovery trajectory of some of the most common respiratory diseases in pediatrics. Subjects were monitored using an electronic home-monitoring platform including a questionnaire and smartwatch. Home-monitoring with digital devices is often cited as a promising tool for the future, and the non-invasive nature of the measurements may be particularly useful in the field of pediatrics. In this paper, we use multiple methods to quantify post-discharge recovery.

The first method, a symptom score questionnaire, is well-known and currently the standard in pediatric studies conducted in an at-home setting.[20] On average, children admitted for CAP became symptom-free 5-6 days later compared to PW and AE patients. The duration to recovery for PW and AE corresponds with findings by Bacharier *et al.* and Ahmed *et al.* A study in 522 children with suspected pneumonia found a median recovery time of 7 days and two-thirds of children were completely recovered after 12 days.[21] The difference in recovery time between CAP and PW/AE patients is in line with the view that bacterial infections lead to more severe illness compared to viral infections.[4,22] The second method to quantify recovery in this study was to measure physical activity (step count) with a smartwatch. The recovery trajectory was characterized relative to the normal physical activity levels exhibited during the healthy period or final days of the recovery period. The estimated mean physical activity followed an inverse pattern over time compared to the estimated mean symptom score, and the curve reached a plateau at the same time point compared to modelled symptom score for all three patient populations. Third, we assessed daytime- and nocturnal HR as marker for recovery. Estimated mean HR during recovery showed a sharp reduction from admission until around 6 days after discharge for subjects with PW and AE. This corresponded well with the symptom score- and physical activity recovery trajectories for both disease groups. However, HR remained stable throughout the study period for the CAP group. Although the initial reduction in HR for subjects with PW and AE may be partly due to reduced respiratory distress, we hypothesize the observed HR decrease is also explained by the frequency of β2-agonist administration.[23] Children with PW and AE usually use more bronchodilators at home during the first days after discharge, when compared to their regular treatment regimen.

Interestingly, the three methods estimated highly similar recovery trajectories, as confirmed by the Pearson correlation coefficients. However, each captures a distinctly different health domain in form of parental observations, activity behavior and cardiovascular state. Our data indicates these domains recover at an identical pace, and, although future research should confirm these observations, the three may be used interchangeably for this application. In this regard, an advantage of physical activity monitoring is the clear definition of recovery in the form of return to 100% of normal activity. For HR and symptom scores, such definitions were not used. Interestingly, we found a strong correlation between mean symptom scores and mean nocturnal HR for the CAP group, even though the absolute reduction in HR was much lower compared to the PW and AE groups.

A possible application of remote monitoring is the prediction of the length and trajectory of the recovery for individual patients, based on their diagnosis, disease severity, and baseline characteristics. To evaluate whether this is a valid avenue to pursue, we introduced several admission variables as additional covariate in the physical activity models during exploratory analyses. Only admission duration improved model fit for subjects with CAP and PW. Admission O2 saturation, HR and respiratory rate as covariate did not improve model fit, and the CI of the estimated means were wide, most likely due to the limited size of the dataset, the directions of the estimated effects for, for example, admission O2-saturation are plausible and may indicate that, with more data, it is possible to develop a model that is not only able to describe a study population, but also estimated and individual patients' expected recovery trajectory based on additional clinical characteristics besides diagnosis. Taking this even further, real-time monitoring could detect when children deviate from their expected recovery trajectory and may serve as a warning sign of pending re-admission. To realize this option, a larger and, ideally, more complete dataset is needed to adequately isolate the effects of a large number of covariates on the recovery trajectory simultaneously. Furthermore, a system such a this can only add value to standard-of-care if the data collection and analysis is completely automated, integrated in electronic patient dossiers, and requires very little input from health care providers. If this would be realized, a warning could be sent to caregivers in the case of red flags, prompting re-evaluation and modification of the treatment plans to avoid re-admission of patients.

Our findings also provide perspective for future pediatric clinical trials. Currently, interventional clinical trials in pediatric patients commonly follow up on predefined time points for objective clinic-based measurements, which are limited to a handful of visits, which provide snapshots of a patient's health status[24,25]. Another option is to make extensive use of paper-questionnaires, which are more subjective.[8,26] Remote monitoring platforms could enable researchers to obtain more objective measurements while reducing the necessity of frequent in-hospital follow-up visits. This will automatically lead to a decreased burden for subjects, which may in turn lead to improved recruitment rates, while simultaneously providing a more complete picture of disease activity compared to traditional trial designs.[14]

A major limitation of this study is that significantly less data was collected than originally planned. Of the 71 patients recruited during this study, only 39 subjects were included in the final analysis set. Overall median compliance to study tasks was only 47%, but measurements involving the smartwatch exhibited a higher compliance (67% for physical activity) compared to the more traditional symptom questionnaire (27%), and compliance was higher during the more important recovery period. Several subjects aged 2 and 3 dropped out due to smartwatch discomfort, which could relate to the design of the wearable device, which is marketed to adults. A smaller device capable of collecting the same parameters could improve the adherence to study tasks. For example, via a T-shirt or other smart clothing.[27,28] The 25-day break between study periods could have contributed to the low compliance as well. Another limitation is that the respiratory symptom questionnaire used by patients with CAP and PW is not formally validated for use in pediatrics, although no validated questionnaire was available during the conception of this study.

The observed differences in recovery time between groups could not only be explained by the underlying illness, but also due to the limitations described above. However, we expect that the limitations of this pilot study could be largely negated with an improved study- and wearable design, and we conclude that a smartwatch is a promising tool for remote monitoring of pediatric patients. We were able to determine the length of post-discharge recovery for three common pediatric respiratory diagnoses, which was unclear before this study. A strength of this study is the use of mixed effects models, which can precisely estimate group means in the presence of missing data points. Future studies may replicate the current findings in a new study with improved design and investigate larger sample sizes allowing for inclusion of multiple covariates in models to predict individual recovery trajectories. Furthermore, digital endpoints could be included as secondary outcome in future clinical trials investigating the effect of treatments thought to hasten at-home recovery, such as systemic corticosteroids in the case of PW.

Conclusion

Physical activity and heart rate measured with a smartwatch appears a viable tool for investigating post-admission recovery in children, although the investigated watch was not suitable for children < 4 years old. We believe remote monitoring could significantly benefit observational- and interventional pediatric clinical trials and possibly clinical care.

SUPPLEMENTARY DATA

Sup. Text S1	Symptom score questionnaire
Sup. Figure S2	Individual summary plots of physical activity
Sup. Figure S3	Compliance by age category
Sup. Figure S4	Raw data (mean (SD)) of physical activity, heart rate and symptom score data per group
Sup. Table S5	Model Coefficients
Sup. Figure S6	Influence of admission characteristics
Sup. Figure S7	Recovery trajectory sleep duration over time
Sup. Figure S8	Monitoring device measurements of an ae subject

REFERENCES

1 Williams DJ, Hall M, Shah SS, Parikh K, Tyler A, Neuman MI, Hersh AL, Brogan TV., Blaschke AJ, Grijalva CG. Narrow vs broad-spectrum antimicrobial therapy for children hospitalized with pneumonia. Pediatrics 2013;132(5).

2 Juvén T, Mertsola J, Waris M, Leinonen M, Ruuskanen O. Clinical response to antibiotic therapy for community-acquired pneumonia. Eur J Pediatr 2004;163(3):140-144.

3 Tan TQ, Mason EO, Wald ER, Barson WJ, Schutze GE, Bradley JS, Givner LB, Yogev R, Kim KS, Kaplan SL. Clinical characteristics of children with complicated pneumonia caused by Streptococcus pneumoniae. Pediatrics 2002;110(1 I):1-6.

4 Ahmed S, Jaleel A, Hameed K, Ahmed F, Danish H, Chugtai A, Mustafa S. Serum Vitamin D Concentration in Asthmatic Children and Its Association with Recovery Time from an Asthma Exacerbation. Br J Med Med Res 2015;10(6):1-10.

5 Wildhaber JH, Sznitman J, Harpes P, Straub D, Möller A, Basek P, Sennhauser FH. Correlation of spirometry and symptom scores in childhood asthma and the usefulness of curvature assessment in expiratory flow-volume curves. Respir Care 2007;52(12):1744-1752.

6 Harris M, Clark J, Coote N, Fletcher P, Harnden A, McKean M, Thomson A. British Thoracic Society guidelines for the management of community acquired pneumonia in children: Update 2011. Thorax 2011;66(SUPPL. 2).

7 Ducharme FM, Tse SM, Chauhan B. Diagnosis, management, and prognosis of preschool wheeze. Lancet 2014;383(9928):1593-1604.

8 Mandhane PJ, Paredes Zambrano De Silbernagel P, Nwe Aung Y, Williamson J, Lee BE, Spier S, Noseworthy M, Craig WR, Johnson DW. Treatment of preschool children presenting to the emergency department with wheeze with azithromycin: A placebo-controlled randomized trial. PLOS One 2017;12(8):1-15.

9 Coran P, Goldsack C, Grandinetti A. Advancing the Use of Mobile Technologies in Clinical Trials : Recommendations from the Clinical Trials Transformation Initiative. 2019;27701:145-154.

10 Beigelman A, King TS, Mauger D, Zeiger RS, Strunk RC, Kelly HW, Martinez FD, Lemanske RF, Rivera-Spoljaric K, Jackson DJ, *et al.* Do oral corticosteroids reduce the severity of acute lower respiratory tract illnesses in preschool children with recurrent wheezing? J Allergy Clin Immunol 2013;131(6).

11 Stickland A, Clayton E, Sankey R, Hill CM. A qualitative study of sleep quality in children and their resident parents when in hospital. Arch Dis Child 2016;101(6):546-551.

12 Greenberg RG, Corneli A, Bradley J, Farley J, Jafri HS, Lin L, Nambiar S, Noel GJ, Wheeler C, Tiernan R, *et al.* Perceived barriers to pediatrician and family practitioner participation in pediatric clinical trials: Findings from the Clinical Trials Transformation Initiative. Contemp Clin Trials Commun 2018;9(September 2017):7-12.

13 Izmailova ES, Wagner JA, Perakslis ED. Wearable Devices in Clinical Trials: Hype and Hypothesis. Clin Pharmacol Ther 2018;104(1):42-52.

14 Kruizinga MD, Stuurman FE, Groeneveld GJ, Cohen AF. The Future of Clinical Trial Design : The Transition from Hard Endpoints to Value-Based Endpoints.

15 Kruizinga MD, Stuurman FE, Exadaktylos V, Doll RJ, Stephenson DT, Groeneveld GJ, Driessen GJA, Cohen AF. Development of Novel, Value-Based, Digital Endpoints for Clinical Trials: A Structured Approach Toward Fit-for-Purpose Validation. Pharmacol Rev 2020;72 (4)(October):899-909.

16 El Moussaoui R, Opmeer BC, Bossuyt PMM, Speelman P, De Borgie CAJM, Prins JM. Development and validation of a short questionnaire in community acquired pneumonia. Thorax 2004;59(7):591-595.

17 Juniper EF, Gruffydd-Jones K, Ward S, Svensson K. Asthma control questionnaire in children: Validation, measurement properties, interpretation. Eur Respir J 2010;36(6):1410-1416.

18 Kruizinga MD, Essers E, Stuurman FE, Zhuparris A, van Eik N, Janssens HM, Groothuis I, Sprij AJ, Nuijsink M, Cohen AF, *et al.* Technical validity and usability of a novel smartphone-connected spirometry device for pediatric patients with asthma and cystic fibrosis. Pediatr Pulmonol 2020;(June):2463-2470.

19 Kruizinga MD, Heide N van der, Moll A, Zhuparris A, Yavuz Y, Kam ML de, Stuurman FE, Cohen AF, Driessen GJA. Towards remote monitoring in pediatric care and clinical trials-Tolerability, repeatability and reference values of candidate digital endpoints derived from physical activity, heart rate and sleep in healthy children. PLOS One 2021;16(1):e0244877.

20 Coons SJ, Eremenco S, Lundy JJ, O'Donohoe P, O'Gorman H, Malizia W. Capturing Patient-Reported Outcome (PRO) Data Electronically: The Past, Present, and Promise of ePRO Measurement in Clinical Trials. Patient 2015;8(4):301-309.

21 Örtqvist Å, Hedlund J, Burman LÅ, Elbel E, Höfer MA, Leinonen M, Lindblad I, Sundelöf B, Kalin M. Randomised controlled trial of clinical outcome after chest radiograph in ambulatory acute lower-respiratory infection in children. Lancet 1998;351(9100):404-408.

22 Bacharier LB, Phillips BR, Zeiger RS, Szefler SJ, Martinez FD, Lemanske RF, Sorkness CA, Bloomberg GR, Morgan WJ, Paul IM, *et al.* Episodic use of an inhaled corticosteroid or leukotriene receptor antagonist in preschool children with moderate-to-severe intermittent wheezing. J Allergy Clin Immunol 2008;122(6):1127-1143.

23 Sears MR. Adverse effects of β-agonists. J Allergy Clin Immunol 2002;110(6):S322-S328.

24 Esposito S, Tagliabue C, Picciolli I, Semino M, Sabatini C, Consolo S, Bosis S, Pinzani R, Principi N. Procalcitonin measurements for guiding antibiotic treatment in pediatric pneumonia. Respir Med 2011;105(12):1939-1945.

25 Dimopoulos G, Matthaiou DK, Karageorgopoulos DE, Grammatikos AP, Athanassa Z, Falagas ME. Short-versus Long-Course Antibacterial Therapy for Community-Acquired Pneumonia. 2008;68(13):1841-1854.

26 Papi A, Nicolini G, Baraldi E, Boner AL, Cutrera R, Rossi GA, Fabbri LM. Regular vs prn nebulized treatment in wheeze preschool children. Allergy Eur J Allergy Clin Immunol 2009;64(10):1463-1471.

27 Montoye AHK, Mitrzyk JR, Molesky MJ. Comparative Accuracy of a Wrist-Worn Activity Tracker and a Smart Shirt for Physical Activity Assessment. Meas Phys Educ Exerc Sci 2017;21(4):201-211.

28 Ajami S, Teimouri F. Features and application of wearable biosensors in medical care. J Res Med Sci 2015;20(12):1208-1215.

CHAPTER 9

Objective home-monitoring of physical activity, cardiovascular parameters and sleep in pediatric obesity patients using digital biomarkers

Accepted by Digital Biomarkers

JM Knijff,[1,2] ECAM Houdijk,[2] DCM van der Kaay,[2,6] Y van Berkel,[2,3] L Filippini,[2,3] FE Stuurman,[1,4] AF Cohen,[1,4] GJA Driessen,[2,5] MD Kruizinga[1,2,4]

1 Centre for Human Drug Research, Leiden, the Netherlands
2 Juliana Children's Hospital, Haga Teaching Hospital, The Hague, the Netherlands
3 Department of pediatrics, Haaglanden Medical Centre, the Hague, the Netherlands
4 Leiden University Medical Centre, Leiden, the Netherlands
5 Department of pediatrics, Maastricht University Medical Centre, Maastricht, the Netherlands
6 Sophia Children's Hospital, Erasmus Medical Centre, Rotterdam, the Netherlands

Abstract

BACKGROUND Clinical research and treatment of childhood obesity is challenging, and objective biomarkers obtained in a home-setting are needed. The aim of this study was to determine the potential of novel digital endpoints gathered by a home-monitoring platform in pediatric obesity.

METHODS In this prospective observational study, 28 children with obesity aged 6-16 years were included and monitored for 28 days. Patients wore a smartwatch, which measured physical activity, heart rate and sleep. Furthermore, daily blood pressure measurements were performed. Data from 128 healthy children were utilized for comparison. Differences between patients and controls were assessed via linear mixed effect models.

RESULTS Median compliance for the measurements ranged from 55%-100%. The highest median compliance was observed for the smartwatch-related measurements (81%-100%). Patients had a lower daily and peak physical activity level, a higher daytime and nighttime heart rate, a higher systolic and diastolic blood pressure and a shorter sleep duration compared to controls.

CONCLUSIONS Remote-monitoring via wearables in pediatric obesity has the potential to objectively measure the disease-burden in the home-setting. Future studies are needed to determine the capacity of the novel digital endpoints to detect effect of interventions.

Introduction

Childhood obesity is a chronic disease with an increasing prevalence worldwide.[1] The disease is associated with a wide spectrum of adverse outcomes, including cardiovascular and metabolic diseases, musculoskeletal problems and psychosocial complications.[2] Treatment, follow-up and clinical research in pediatric obesity is challenging, and logistical problems such as travel distance and scheduling conflicts are mentioned as reasons for non-return to a pediatric weight management program.[3] In addition, patient's and parents' recall of physical activity (PA) and food intake is frequently subjective and suboptimal. Reliable blood pressure (BP) measurements are important in the follow-up of pediatric obesity,[4] but BP measurements during outpatient visits can be distorted by white coat hypertension.[5] These examples indicate that there is a need for objective measurements obtained in a home-setting that can monitor disease activity, and this could be provided via remote monitoring with digital biomarkers.[6]

Previous studies already assessed digital monitoring of PA levels of children with obesity and reported that they are less physically active compared to healthy children.[7] However, wearable devices can also capture other digital biomarkers, such as heart rate (HR) and sleep. Capturing PA, HR and sleep simultaneously via wearable technology has not previously been reported in pediatric obesity. Additionally, home-measured BP could provide a better indication of cardiovascular risk status. The combination of these digital biomarkers could provide an objective overview of the child's health and may lead to early detection of complications of childhood obesity. These digital endpoints could be utilized in clinical care and clinical trials to evaluate the effect of lifestyle interventions in the home-setting and may contribute to a reduction of the burden to visit outpatient clinics. Clinical validation of novel digital endpoints must be performed in the target population before integration in clinical care or clinical trials.[8] This process focuses on the tolerability, the ability to detect a significant difference between patients and controls and the correlation with existing disease metrics of candidate endpoints.

The aim of this study was to determine the clinical potential of novel digital endpoints derived from PA, HR, sleep, and BP gathered via a home-monitoring platform in pediatric obesity.

Materials and Methods

Location and ethics

This study was conducted at the Haga Teaching Hospital, Juliana Children's Hospital (The Hague, the Netherlands) and at the Centre for Human Drug Research (Leiden, the Netherlands). The study protocol was approved by the Medical Ethics Committee Zuid West Holland (The Hague, the Netherlands), and was conducted according to the Dutch Act on Medical Research Involving Human Subjects (WMO) and Good Clinical Practice Guideline. Written informed consent was obtained from all parents. Verbal consent was obtained from children aged younger than 12 years and written consent was obtained from children aged 12 years and older. The trial was registered at the Dutch Trial Registry (NTR, Trial NL7611, registered 18-Mar-2019).

Subjects and study design

During this prospective observational case-control study 28 patients, aged between 6-16 years old, were recruited via the outpatient clinic between November 2018 and April 2020. Patients diagnosed with obesity grade 1, 2 or 3 were included.[9] Children diagnosed with a chronic condition, other than obesity, that might impair PA levels were excluded. Data from 128 healthy controls, children aged between 6-16 years, were collected via a separate study.[6] The control group had a similar age and sex distribution as the patient group. Before the start of the study an informative session was planned for education on the study devices. Afterwards, patients were monitored in the home-setting during 28 consecutive days and used a smartphone that connected to other study devices. Patients were expected to wear a Steel HR smartwatch (Withings, Issy-les-Molineux, France) 24 hours per day, which measured PA in step count and several sleep-related parameters via a built-in accelerometer and registered HR through a photoplethysmography (PPG) sensor every 10 minutes. Data were directly uploaded to the server via the CHDR MORE application. Daily BP measurements were performed by a wireless upper arm cuff and oscillometric determination of pressure (Withings BPM) each evening at approximately the same clock time. Patients were instructed not to physically exert themselves just before the measurement. Weekly weight assessments were conducted with Withings Body+Scales. A daily questionnaire was completed on the smartphone regarding the daily screen time of the patient.

Baseline characteristics and environmental data

At the start of the study, baseline characteristics were collected, and the Children's Somatization Inventory (CSI) questionnaire and Pediatric Quality of Life Inventory (PedsQL) 4.0 questionnaire were completed.[10,11] The calculated Body Mass Index Standard Deviation Score (BMI SDS) was adjusted for age and sex at baseline. Data regarding the population density of the child's city of residence were obtained via the Dutch Central Office of Statistics (CBS). Weather data during the study period were collected from the Royal Dutch Meteorological Institute (KNMI) at the weather station located in Hoek van Holland.

Analysis

COMPLIANCE The tolerability of the novel endpoints was assessed by determining the compliance for each measurement type. The compliance was calculated for each participant individually by dividing the amount of completed measurements by the amount of expected measurements. When the weight assessment deviated more than two days from the protocolized time point, this assessment was counted as not completed. The watch wear time between 6AM-10PM was calculated to include as a covariate for the analysis of PA data. Calculation of the watch wear time was based on hourly data of PA and HR. When there was no registration of both HR and PA in a particular hour, it was concluded that the watch was not worn by the subject.

CANDIDATE BIOMARKERS Multiple candidate endpoints based on PA, HR, sleep and BP were defined prior to the analysis. First, proposed PA derived candidate biomarkers consisted of the daily PA, average PA per hour of the day and PA during the most active hour per day (peak physical activity). HR data between 6AM-10PM were summarized as average daytime HR and HR data between 0-5AM were summarized as average nighttime HR. In addition, the average HR per hour of the day was calculated. Moreover, systolic and diastolic BP were considered as candidate biomarkers. Lastly, sleep-related candidate endpoints consisted of the average sleep duration, sleep depth (average proportion light sleep) and the amount of wake-ups per night.

STATISTICAL ANALYSIS OF THE CANDIDATE ENDPOINTS Days with watch wear time <50% between 6AM-10PM were excluded from the analysis. Differences between the two groups were assessed for each candidate endpoint via linear mixed effect models with condition (healthy or obesity) as fixed effect and subject as random effect. Residual plots were inspected, and logarithmic and square root transformations were applied in the case of heteroscedasticity. The following parameters were tested as additional fixed effect in a model when expected to explain additional variance: age, sex, watch wear time, day of the week, type of day (school day/weekend/holiday), population density, rain duration, temperature, sunshine duration, and step count.[6] Polynomial regression with 3 degrees of freedom was utilized when exploratory plots suggested a nonlinear relationship. Inclusion of a covariate or factor and determining the best model fit were based on Akaike's information criterion, Bayesian information criterion and likelihood ratio tests. Interactions between included covariates and factors were tested and considered to be included in the model if the interaction was biologically plausible. Estimated marginal means were calculated for both study groups and plotted with a 95% confidence interval. For all estimated means, fixed effects were held constant to their population average. Models including watch wear time were visualized with a watch wear time of 100%. A p-value smaller than 0.05 was considered statistically significant. Within the patient group, correlations between BMI SDS, quality of life and daily PA were assessed by calculating Pearson's correlation coefficient or Spearman's correlation coefficient.

SOFTWARE Promasys® 7.3 (Anju Software, Fort Lauderdale, Texas, USA) was used for data management of the baseline characteristics. Statistical analysis was performed with R version 3.6.2 with utilization of the lme4, emmeans and ggeffects packages.[12-15]

Results

Baseline characteristics

Baseline characteristics of the children with obesity (n=28) and the healthy children (n=128) are presented in *Table 1.* The mean age was 11 years and 46% of the participants was male in both study groups. The mean BMI SDS of patients was 3.6, versus 0.3 of controls. The average quality of life score measured by the PedsQL questionnaire was 78.6 out of 100 in the patient group versus 90.7 out of 100 in the control group.

Table 1. Baseline characteristics

	Children with obesity (n=28)	Healthy children (n=128)6
Age (years) (Mean (SD))	11.6 (3.1)	11.1 (3.1)
Sex, male (n (%))	13 (46%)	59 (46%)
Weight (kg) (Mean (SD))	77.7 (24.6)	42.7 (15.7)
Height (cm) (Mean (SD))	156.3 (15.5)	151.3 (17.7)
BMI (Mean (SD))	30.9 (5.2)	18.0 (3.1)
BMI SDS (Mean (SD))	3.6 (0.4)	0.3 (1.2)
PedsQL score (Mean (SD))	78.6 (14.2)	90.7 (7.4)
CSI score (Mean(SD))	12.3 (11.4)	-
Obesity Grade (n (%)) Grade 1 Grade 2 Grade 3	 10 (36%) 10 (36%) 8 (29%)	-
Plays sports (n (%))	18 (64%)	117 (91%)
Ethnicity (n (%)) Caucasian Asian/Hindi Other/Mixed	 20 (71%) 3 (11%) 5 (18%)	 122 (95%) 2 (2%) 4 (3%)
Extremely urbanized area (n (%))	25 (89%)	97 (76%)

Compliance

The median compliance of each measurement type is listed in *Table 2*. The median overall compliance was 74% (IQR 55%-85%). The median wear time of the smartwatch was 22.0 hours per day. The lowest median compliance was observed for the daily questionnaire (55%), while the highest median compliance was observed for two smartwatch-related measurements, HR (100%) and step count (100%). Two patients did not complete any BP measurements and two patients did not wear the smartwatch at night.

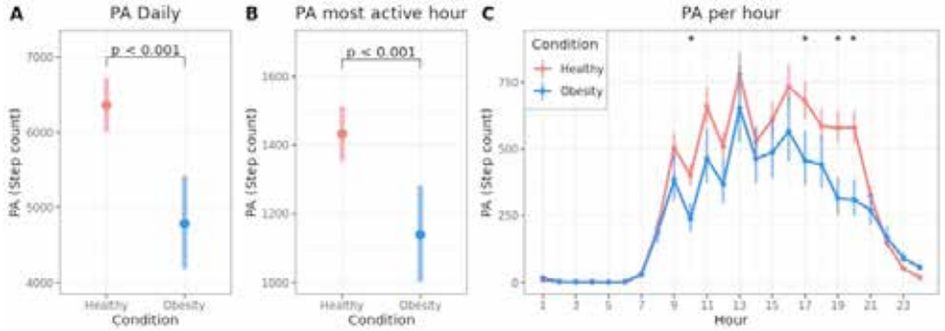
Table 2. Compliance of children with obesity during the study period

Measurement	Median compliance (IQR)
Smartwatch	
Step count	100% (93%-100%)
Heart rate	100% (93%-100%)
Sleep	81% (62%-89%)
Wear time watch per day	22.0h (18.0h-23.3h)
Blood pressure	59% (32%-79%)
Weight	75% (25%-100%)
Daily questionnaire	55% (20%-79%)
Overall compliance	74% (55%-85%)

Difference patients and controls

PHYSICAL ACTIVITY The average daily step count was 4528 for patients versus 6066 for controls (difference 1538 steps, 95% confidence interval (CI) 919–2157, *Figure 1A*). For patients, PA during the most active hour was lower compared to controls with a difference of 289 steps (1099 steps vs 1388 steps, 95% CI 149–428, *Figure 1B*). A separate analysis was performed to calculate the average PA per hour of the day, which was significantly lower for the patient group compared to the control group with a difference of more than 50 steps per hour at 10AM, 5PM, 7PM and 8PM, after controlling for age and sex (*Figure 1C*). The 10TH and 90TH percentile of the daily PA within a week, which were both lower for patients compared to controls. An overview of all analyses with adjusted and unadjusted differences are listed in *Supplementary Table S1.* The relationship between daily PA and daily screen time for both study groups is visualized in *Figure S1*. Daily PA decreased with an increase in screen time for both the patient group and the control group. Patients performed less PA compared with controls with a similar screen time duration.

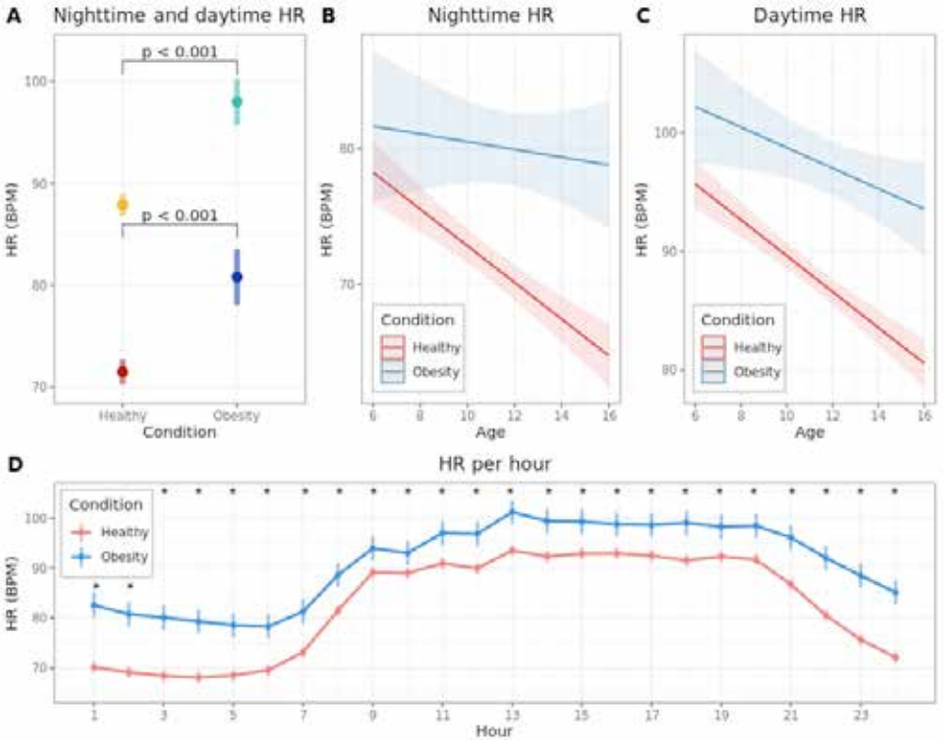
Figure 1. Differences in PA between children with obesity and healthy children. A–B. Estimated marginal mean (95% CI) daily step count (A) and step count during the most active hour per day (B) for children with obesity and healthy children. Age (11 years), rain duration (1.87h), weekday, degree of urbanization and sex were fixed to their population average. Plots are visualized with watch wear time 100% C. Estimated marginal mean (95% CI) PA per hour during the day for children with obesity and healthy children. Age (11 years) and sex are fixed to their population average.



* Indicate hours with a p-value <0.05 for the difference (>50 steps per hour) after Holm's correction for multiple tests.

HEART RATE The average nighttime HR was 81 BPM for patients versus 71 BPM for controls (difference 9.3 BPM, 95% CI 6.3–12.3, *Figure 2A*). In addition, the average daytime HR was also higher for patients compared to controls (98 BPM vs 88 BPM, difference 10.1 BPM, 95% CI 7.6–12.6, *Figure 2A*). The difference in average nighttime HR between patients and controls increased as a function of age by a difference of 1.1 BPM/age year (95% CI 0.1–2.0, *Figure 2B*). This age-related effect was not observed for the daytime HR (*Figure 2C*). A separate analysis per hour showed that patients had a significantly higher average hourly HR compared to controls for every hour of the day, after controlling for age and sex (*Figure 2D*).

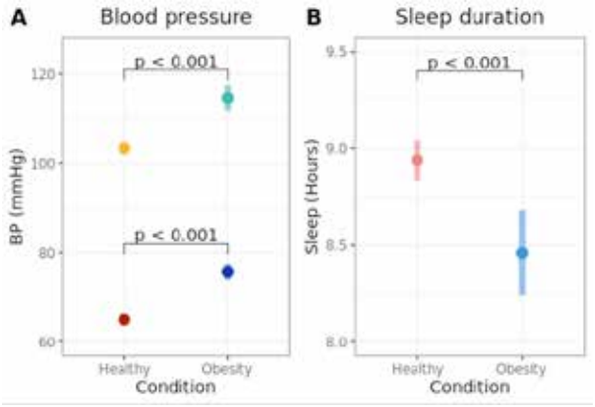
Figure 2. Differences in HR between children with obesity and healthy children. A. Estimated marginal mean (95% CI) nighttime and daytime HR for children with obesity and healthy children. Light colors represent daytime HR, dark colors represent nighttime HR. Age (11 years) and sex for both nighttime and daytime HR and daily step count (7000 steps) for daytime HR only, were fixed to their population average. B. Relationship between age and nighttime HR (difference of 1.1 BPM/age year, 95% CI 0.1–2.0, p = 0.029). C. Relationship between age and daytime HR (difference of 0.6 BPM/age year, 95% CI -0.2 to 1.5, p = 0.119). B–C. Bold lines represent the estimated marginal means, shaded areas indicate the 95% CI of the estimated mean. D. Estimated marginal mean (95% CI) HR per hour of the day for children with obesity and healthy children. Age (11 years) and sex were fixed to their population average.



*Indicate hours with a p-value <0.05 for the difference after Holm's correction for multiple tests.

BLOOD PRESSURE Patients had a higher systolic BP (115 mmHg vs 104 mmHg, difference 11.3 mmHg, 95% CI 8.1–14.5, *Figure 3A*) as well as a higher diastolic BP (76 mmHg vs 65 mmHg, difference 10.7 mmHg, 95% CI 8.7–12.7, *Figure 3A*) compared to controls. The difference in diastolic BP between patients and controls increased significantly as a function of age (difference 0.9 mmHg/age year, 95% CI 0.3–1.5, *Supplementary Figure S2A*). This age-related effect was not observed for the systolic BP (*Supplementary Figure S2B*).

Figure 3. Difference in BP and sleep duration between children with obesity and healthy children.
A. Estimated marginal mean (95% CI) systolic BP and diastolic BP for children with obesity and healthy children. Light colors represent systolic BP, dark colors represent diastolic BP. Age (11 years) was fixed to their population average. B. Estimated marginal mean (95% CI) sleep duration for children with obesity and healthy children. Age (11 years), sex and type of day were fixed to their population average.



SLEEP The adjusted average sleep duration per night was shorter for patients compared to controls (8.5h vs 8.9h, difference 0.5h, 95% CI 0.2–0.7, *Figure 3B*). The average proportion of light sleep per night was 52.6% for the patient group versus 56.9% for the control group (difference 4.3%, 95% CI 1.8–6.8). There was no significant difference found in the average number of wakeups per night between patients and controls.

Correlations in the patient group

No significant correlation was found between daily PA and BMI SDS as well as between daily PA and quality of life in the patient group. Since a period of 28 days was too short to observe evident differences in weight, no further analyses were performed with the weight data obtained by the weekly weight assessments.

Discussion

This study evaluated digital endpoints derived from PA, HR, BP and sleep in pediatric obesity through wearable devices at home. These multiple digital endpoints gathered by a home-monitoring platform show potential for future use in clinical trials and clinical care because of the high tolerability and ability to differentiate patients from controls, which are prerequisites for implementation.[8] Adequate validation must be performed before implementation of the ubiquitous digital devices and applications in clinical research and ultimately in clinical practice. A previously proposed stepwise approach to fit-for-purpose validation of (digital) biomarkers was applied in this study.[8]

One of the most important characteristics of digital biomarkers is the tolerability in the target population.[8] Non-compliance will lead to incomplete datasets and biased results. The smartwatch-related measurements showed the highest tolerability (median compliance 81%–100%). Gathering data via a smartwatch appeared to be superior to gathering data via a questionnaire (median compliance 55%) in children with obesity. Compared to a previous study in healthy children, the overall compliance was lower for children with obesity (94% vs 74%). This difference is predominantly caused by the difference in median compliance for the daily BP measurements (95% for healthy children vs 59% for patients) and daily questionnaire (90% for healthy children vs 55% for patients).[6] Childhood obesity is associated with a less structured home environment.[16] In contrast to the smartwatch related measurements, the daily BP measurements and questionnaire need to be planned and performed actively by the child or parents which might be more difficult in a less structured home environment.

Another important validation criterion for biomarkers is the ability to discriminate healthy children from patients, which was assessed for all candidate biomarkers. The average daily PA was lower for patients compared to controls, which has been cited in numerous previous studies with comparable differences.[7] However, the reported differences in step count throughout the literature vary due to, inter alia, the utilization of a wide selection of pedometers and differences in age of the study populations. There is conflicting evidence whether the awareness of wearing an accelerometer is affecting the level of PA in healthy youth.[17] The majority of studies examining differences in step count between children with obesity and healthy children monitored for a maximum of 8 days.[7] In this study patients were monitored for 28 days. We did not find a decrease in PA levels comparing the first and last week in either study group. As reported by past studies patients

had a lower peak PA compared to controls.[18] We demonstrated a difference in peak PA between the two groups not only by analyzing PA levels during the most active hour per day, but also by calculating the 90TH percentile of the daily PA within a week. The latter endpoint is less variable and could be a useful biomarker for long-term monitoring.[6] Based on the data presented here, interventions focused on PA after school time seem most appropriate, since the biggest differences in PA between the two groups were observed in the after-school period. The combination of PA derived biomarkers provides a wide-ranging and objective overview of the PA level of the patient and can be utilized to promote PA and to provide personal advice.

Multiple candidate biomarkers based on HR were examined in this study. HR was registered through a PPG sensor, which has shown an acceptable validity in adults and has demonstrated to be accurate in measuring HR in children undergoing elective surgery.[19,20] Patients had a higher average nighttime HR compared to controls, with similar absolute differences compared to previous research with other methods of heart rate monitoring.[21] Additionally, children with obesity had a higher daytime HR compared to controls, which also has been reported in the past.[22,23] Analysis of HR per hour clearly displayed the difference in daily HR pattern for patients compared to controls. The higher HR in the patient group can be explained by sympathetic nervous system overactivation,[22,24,25] caused by dysregulation of the release of multiple adipokines (leptin, free fatty acids, TNF-α, IL-6, adiponectin) and baroreflex dysfunction.[26,27] In this study the difference in average nighttime HR between patients and controls increased as a function of age (while no correlation between BMI SDS and age was found). This is a novel observation, possibly explained by the fact that in healthy children a progressive increase in cardiac parasympathetic activity relative to sympathetic activity occurs with an increase in age, while for children with obesity this process is disrupted.[28] This age-related effect was not observed for the average daytime HR, most likely due to the higher proportion of unexplained variability in this data. Weight loss is associated with a decrease in HR, which may suggest that the lower parasympathetic activity is reversible.[29,30] It has been reported that a higher resting HR leads to a higher risk of cardiovascular disease and (non)-cardiovascular death in adults and is associated with dyslipidemia in children.[31–34] Consequently, nighttime HR might be an attractive surrogate biomarker to assess the risk for cardiovascular disease in children with obesity.

Systolic and diastolic BP were also proposed as candidate endpoints, and both were elevated in patients compared to controls. The differences in BP reported here are slightly larger compared to the differences mentioned in previous research but are within the ranges reported in the literature.[4] This relatively large difference between patients and controls could be explained by our study population, which consists of a high percentage of children diagnosed with grade 2 and 3 obesity (65%) compared with other studies. This might have led to a high proportion of patients at risk for cardiovascular problems in our cohort, since an increase in BMI is associated with an increase in BP.[35] Another explanation is the BP cuff used in this study, which was identical for healthy children and children with obesity. This could have resulted in an overestimation of the BP in patients due to a bigger arm circumference, though the observed differences in BP appear too large to be entirely attributed to the utilization of the single sized cuff.[36] Moreover, the Withings device has been validated in accordance with the ESH International Protocol Revision 2010.[37] The pathophysiological mechanism of hypertension in children with obesity is multifactorial and complex. Suggested contributing factors are increased sympathetic nervous system activation, dysfunction of the endocrine system, disturbed sodium homeostasis and vascular damage.[38] The difference in diastolic BP, but not systolic BP, between patients and controls increased as a function of age. Presently, a pathophysiologic explanation for this observation is lacking, and more research regarding hypertension subtypes in children with obesity may elucidate the underlying mechanism.[39] Childhood hypertension has multiple adverse consequences, such as an increased carotid intima-media thickness and left ventricular hypertrophy,[40,41] both precursors to adverse cardiovascular outcomes in adulthood.[42,43] Literature regarding the reversibility of the adverse cardiovascular effects of childhood obesity states that lifestyle interventions improve early markers of atherosclerosis and reduce the BP.[44] Hence, the combination of HR registration and BP measurements appear a valid option to monitor the cardiovascular status of the patient non-invasively.

Multiple sleep parameters were tested as candidate endpoints. Patients had a significantly shorter sleep duration than healthy children, an observation supported by previous studies.[45] Data regarding sleep quality and sleep efficiency in children with obesity compared with healthy children are inconsistent partly due to different measurement methods and definitions.[45] Since sleep parameters were measured via accelerometry, the lower proportion of light sleep for children with obesity compared to the healthy children could be caused by less movement at night due to the habitus of the patients. Also, it must be considered that when interpreting accelerometry-derived sleep measurements, accelerometry has shown to be less accurate compared to polysomnography.[46] On the other hand, sleep registration via accelerometry is less invasive and can be performed multiple

nights in a natural setting in contrast to polysomnography. Furthermore, accelerometry-derived sleep recording has shown to be more reliable than sleep registration via maternal sleep reports and avoids the recall bias related to sleep diaries.[47] In the future, with further improvement of the underlying algorithms, accelerometer-derived parameters might be useful to detect sleep related breathing disorders, such as obstructive sleep apnea, or to monitor children non-invasively after treatment.[48,49]

This study has several limitations. The sample size was limited to 28 patients. Nevertheless, important baseline characteristics (age, sex, and obesity grade), were well distributed in the patient group. The median compliance for the smartwatch-related measurements was high in this study. However, further studies are needed to examine the long-term compliance, which is very important in monitoring treatment effects. If the long-term compliance is sufficient, home-monitoring via wearables could reduce outpatient visits. Moreover, a disadvantage of gathering data in a home-setting is missing data due to non-compliance. Although the amount of missing data was low and therefore unlikely to impact the overall results, watch wear time was included as covariate in the PA models. Finally, when appraising PA-related endpoints, it must be considered that PA has been measured in step count and that activities like cycling and swimming were not registered. Strengths of this study consisted of the utilization of a structured validation process of the candidate endpoints, the inclusion of a large control group, with a similar distribution of age and sex compared to the patient group and the relatively long-term monitoring period of 28 days. Additionally, the linear mixed effect models utilized for the analysis of the candidate endpoints can handle small sample sizes and missing data points. The defined endpoints based on PA, HR, BP and sleep could be utilized to promote and track PA, to assess the risk for cardiovascular disease and to detect sleep-related alterations of childhood obesity.

Conclusion

Remote-monitoring via wearable technology has the potential to objectively measure the disease-burden in the home-setting in pediatric obesity. The digital biomarkers based on PA, HR, BP and sleep have a high tolerability and can demonstrate differences between patients and controls, in line with previous studies gathering the endpoints via conventional clinic-based methods. Future studies are needed to determine the capacity of these novel digital endpoints to detect effect of interventions.

REFERENCES

1 Obesity and overweight. World Health Organization website. https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight. Published April 1, 2020. Accessed January 11, 2021.
2 Kumar S, Kelly AS. Review of Childhood Obesity: From Epidemiology, Etiology, and Comorbidities to Clinical Assessment and Treatment. Mayo Clinic Proceedings. 2017;92(2):251-265.
3 Barlow SE, Ohlemeyer CL. Parent reasons for nonreturn to a pediatric weight management program. Clin Pediatr (Phila). 2006 May;45(4):355-60.
4 Friedemann C, Heneghan C, Mahtani K, Thompson M, Perera R, Ward AM. Cardiovascular disease risk in healthy children and its association with body mass index: Systematic review and meta-analysis. BMJ. 2012;345(7876).
5 Krmar RT. White-coat hypertension from a paediatric perspective. Acta Paediatrica. 2019;108(1):44-49.
6 Kruizinga MD, Heide N van der, Moll A, *et al.* Towards remote monitoring in pediatric care and clinical trials—Tolerability, repeatability and reference values of candidate digital endpoints derived from physical activity, heart rate and sleep in healthy children. PLOS One. 2021;16(1):e0244877.
7 Miguel-Berges ML, Reilly JJ, Aznar LAM, Jiménez-Pavón D. Associations between pedometer-determined physical activity and adiposity in children and adolescents: Systematic review. Clin J Sport Med. 2018;28(1):64-75.
8 Kruizinga MD, Stuurman FE, Exadaktylos V, *et al.* Development of novel, value-based, digital endpoints for clinical trials: A structured approach toward fit-for-purpose validation. Pharmacol Rev. 2020;72(4):899-909.
9 Binsbergen JJ van, Langens FNM, Dapper ALM. Obesity guideline. Dutch College of General Practitioners website. https://richtlijnen.nhg.org/standaarden/obesitas#volledige-tekst-kinderen. Published October, 2010. Updated September 2, 2020. Accessed January 11, 2021.
10 Walker LS, Beck JE, Garber J, Lambert W. Children's somatization inventory: Psychometric properties of the revised form (CSI-24). J Pediatr Psychol. 2009;34(4):430-440.
11 Varni JW, Seid M, Kurtin PS. PedsQLTM 4.0: Reliability and Validity of the Pediatric Quality of Life InventoryTM Version 4.0 Generic Core Scales in Healthy and Patient Populations. Med Care. 2001 Aug;39(8):800-12.
12 R Core Team. R: The R Project for Statistical Computing. https://www.r-project.org/. Published 2019. Accessed January 11, 2021.
13 Bates D, Mächler M, Bolker BM, Walker SC. Fitting linear mixed-effects models using lme4. Journal of Statistical Software. 2015;67(1).
14 Lenth Russell. Emmeans: Estimated Marginal Means, aka Least-Squares Means. https://cran.r-project.org/web/packages/emmeans/index.html. Published 2020. Accessed January 11, 2021.
15 Lüdecke D. ggeffects: Tidy Data Frames of Marginal Effects from Regression Models. J Open Source Softw. 2018;3(26):772.
16 Bates CR, Buscemi J, Nicholson LM, Cory M, Jagpal A, Bohnert AM. Links between the organization of the family home environment and child obesity: a systematic review. Obes Rev. 2018;19(5):716-727.
17 Clemes SA, Biddle SJH. The Use of Pedometers for Monitoring Physical Activity in Children and Adolescents: Measurement Considerations. J Phys Act Health. 2013 Feb;10(2):249-62.
18 Elmesmari R, Martin A, Reilly JJ, Paton JY. Comparison of accelerometer measured levels of physical activity and sedentary time between obese and non-obese children and adolescents: A systematic review. BMC Pediatr. 2018;18(1).
19 British Association of Sport and Exercise Sciences. Validity of Wrist-Worn photoplethysmography devices to measure heart rate: A systematic review and meta-analysis. J Sports Sci. 38(17):2021-2034.
20 Pelizzo G, Guddo A, Puglisi A, *et al.* Accuracy of a Wrist-Worn Heart Rate Sensing Device during Elective Pediatric Surgical Procedures. Children (Basel). 5(3).
21 Archbold KH, Johnson NL, Goodwin JL, Rosen CL, Quan SF. Normative heart rate parameters during sleep for children aged 6 to 11 years. J Clin Sleep Med. 2010 Feb 15;6(1):47-50.
22 Sorof JM, Poffenbarger T, Franco K, Bernard L, Portman RJ. Isolated systolic hypertension, obesity, and hyperkinetic hemodynamic states in children. J Pediatr. 2002;140(6):660-666.
23 Fernandes RA, Freitas IF, Codogno JS, Christofaro DGD, Monteiro HL, Roberto Lopes DMH. Resting heart rate is associated with blood pressure in male children and adolescents. J Pediatr. 2011;158(4):634-637.
24 Rossi RC, Vanderlei LCM, Gonçalves ACCR, *et al.* Impact of obesity on autonomic modulation, heart rate and blood pressure in obese young people. Auton Neruosci. 2015;193:138-141.
25 Rodríguez-Colón SM, Bixler EO, Li X, Vgontzas AN, Liao D. Obesity is associated with impaired cardiac autonomic modulation in children. Int J Pediatr Obes. 2011;6(2):128-134.
26 Smith MM, Minson CT. Obesity and adipokines: Effects on sympathetic overactivity. J Physiol. 2012;590(8):1787-1801.
27 da Silva AA, do Carmo J, Dubinion J, Hall JE. The role of the sympathetic nervous system in obesity-related hypertension. Curr Hypertens Rep. 2009 Jun;11(3):206-11.
28 Eyre ELJ, Duncan MJ, Birch SL, Fisher JP. The influence of age and weight status on cardiac autonomic control in healthy children: A review. Auton Neurosci. 2014;186(C):8-21.
29 Pidlich J, Pfeffel F, Zwiauer K, Schneider B, Schmidinger H. The effect of weight reduction on the surface electrocardiogram: A prospective trial in obese children and adolescents. Int J Obes. 1997;21(11):1018-1023.

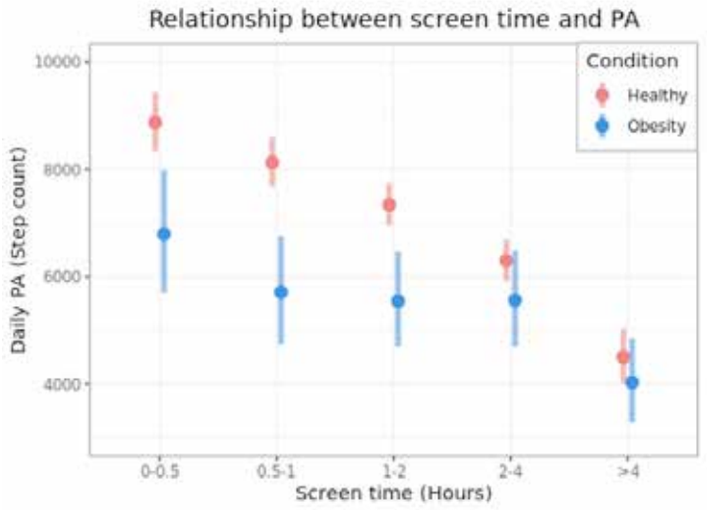
30 Arone LJ, Mackintosh R, Rosenbaum M, Leibel RL, Hirsch J. Autonomic nervous system activity in weight gain and weight loss. Am J Physiol. 1995 Jul;269(1 Pt 2):R222-5.
31 Cooney MT, Vartiainen E, Laakitainen T, Juolevi A, Dudina A, Graham IM. Elevated resting heart rate is an independent risk factor for cardiovascular disease in healthy men and women. Am Heart J. 2010;159(4).
32 Kannel WB, Kannel C, Paffenbarger RS, Cupples LA. Heart rate and cardiovascular mortality: the Framingham Study. Am Heart J. 1987 Jun;113(6):1489-94.
33 Jensen MT, Suadicani P, Hein HO, Gyntelberg F. Elevated resting heart rate, physical fitness and all-cause mortality: A 16-year follow-up in the Copenhagen Male Study. Heart. 2013;99(12):882-887.
34 Freitas Júnior IF, Monteiro PA, Silveira LS, *et al.* Resting heart rate as a predictor of metabolic dysfunctions in obese children and adolescents. BMC Pediatr. 2012;12.
35 Dong J, Guo XL, Lu ZL, *et al.* Prevalence of overweight and obesity and their associations with blood pressure among children and adolescents in Shandong, China. BMC Public Health. 2014;14(1).
36 Whincup PH, Cook DG, Shaper AG. Blood pressure measurement in children: the importance of cuff bladder size. J Hypertens. 1989 Oct;7(10):845-50.
37 Topouchian J, Agnoletti D, Blacher J, *et al.* Validation of four devices: Omron M6 Comfort, Omron HEM-7420, Withings BP-800, and Polygreen KP-7670 for home blood pressure measurement according to the European society of hypertension international protocol. Vasc Health Risk Manag. 2014;10:33-44.
38 Wirix AJG, Kaspers PJ, Nauta J, Chinapaw MJM, Kist-van Holthe JE. Pathophysiology of hypertension in obese children: A systematic review. Obes Rev. 2015;16(10):831-842.
39 Li Y, Haseler E, Chowienczyk P, Sinha MD. Haemodynamics of Hypertension in Children. Curr Hypertens Rep. 2020;22(8).
40 Lande MB, Carson NL, Roy J, Meagher CC. Effects of childhood primary hypertension on carotid intima media thickness: A matched controlled study. Hypertension. 2006;48(1):40-44.
41 Jing L, Nevius CD, Friday CM, *et al.* Ambulatory systolic blood pressure and obesity are independently associated with left ventricular hypertrophic remodeling in children. J Cardiovasc Magn Reson. 2017;19(1).
42 Bots ML, Dijk JM, Oren A, Grobbee DE. Carotid intima-media thickness, arterial stiffness and risk of cardiovascular disease: current evidence. J Hypertens. 2002 Dec;20(12):2317-25.
43 Artham SM, Lavie CJ, Milani R v., Patel DA, Verma A, Ventura HO. Clinical Impact of Left Ventricular Hypertrophy and Implications for Regression. Prog Cardiovasc Dis. 2009;52(2):153-167.
44 Ayer J, Charakida M, Deanfield JE, Celermajer DS. Lifetime risk: Childhood obesity and cardiovascular risk. European Heart Journal. 2015;36(22):1371-1376.
45 Morrissey B, Taveras E, Allender S, Strugnell C. Sleep and obesity among children: A systematic review of multiple sleep dimensions. Pediatr Obes. 2020;15(4).
46 Kolla BP, Mansukhani S, Mansukhani MP. Consumer sleep tracking devices: a review of mechanisms, validity and utility. Expert Rev Med Devices. 2016;13(5):497-506
47 Martinez SM, Greenspan LC, Butte NF, *et al.* Mother-reported sleep, accelerometer-estimated sleep and weight status in Mexican American children: Sleep duration is associated with increased adiposity and risk for overweight/obese status. J Sleep Res. 2014;23(3):328-336.
48 Bixler EO, Vgontzas AN, Lin H-M, *et al.* Sleep disordered breathing in children in a general population sample: prevalence and risk factors. Sleep. 2009 Jun;32(6):731-6.
49 Gruwez A, Bruyneel AV, Bruyneel M. The validity of two commercially-available sleep trackers and actigraphy for assessment of sleep parameters in obstructive sleep apnea patients. PLOS One. 2019;14(1):e0210569.

Sup. Table S1. Unadjusted and adjusted differences between children with obesity and healthy children.

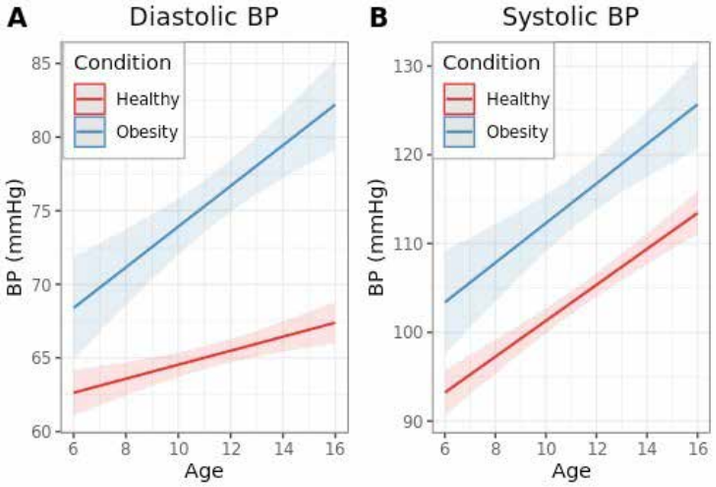
Candidate Endpoints	Unadjusted		Adjusted		Adjusted for
	Difference* (95% CI)	p-value	Difference* (95% CI)	p-value	
Daily step count	2143(1318-2968)	<0.001	1538 (919-2157)	<0.001	Age, sex, wear time watch, weekday, rain duration, degree of urbanization
Step count during most active hour	368(194-542)	<0.001	289 (149-428)	<0.001	Age, sex, wear time watch, weekday, rain duration, degree of urbanization
10TH percentile daily step count	1574(891-2256)	<0.001	948 (325-1571)	0.003	Age, sex, wear time watch, rain duration
90TH percentile daily step count	2878(1880-3877)	<0.001	2257 (1315-3199)	<0.001	Age, sex, wear time watch, rain duration
Nighttime heart rate	-8.7(-12.1 to -5.3)	<0.001	-9.3(-12.3 to -6.3)	<0.001	Age, sex
Daytime heart rate	-7.0(-10.2 to -3.7)	<0.001	-10.1(-12.6 to -7.6)	<0.001	Age, sex, daily step count**
Systolic blood pressure	-11.9(-16.1 to -7.8)	<0.001	-11.3 (-14.5 to -8.1)	<0.001	Age
Diastolic blood pressure	-10.8(-13.0 to -8.7)	<0.001	-10.7(-12.7 to -8.7)	<0.001	Age
Sleep duration	0.5(0.2-0.8)	<0.001	0.5(0.2-0.7)	<0.001	Age, sex, type of day
Sleep depth (% light sleep)	4.2(1.6-6.8)	0.002	4.3(1.8-6.8)	0.001	Age, sex
Number of wake-ups	-0.3(-0.8 to 0.3)	0.326	-0.3(-0.8 to 0.3)	0.326	-

* Differences presented are 'estimated mean healthy controls'-'estimated mean children with obesity'
** 3RD degree polynomial.

Sup. Figure S1. Relationship between screen time and daily step count for children with obesity and healthy children. Relationship between screen time and daily PA for healthy children and children with obesity. Estimated marginal means are plotted (95% CI) for both study groups. Age (11 years) and sex were fixed to their population average.



Sup. Figure S2. Relationship between BP and age for children with obesity and healthy children. A. Relationship between age and diastolic BP (difference of 0.9 mmHg/age year, 95% CI 0.3–1.5, p = 0.005). B. Relationship between age and systolic BP (difference of 0.2 mmHg/age year, 95% CI −0.8 to 1.2, p = 0.693). A–B. Bold lines represent the estimated marginal means, shaded areas indicate the 95% CI of the estimated mean.



CHAPTER 10

Remote monitoring with digital biomarkers in pediatric patients with sickle cell disease: A pilot study

Submitted to Pediatric Blood and Cancer

MD Kruizinga,[1,2,3] A Verbaan,[2] JM Knijff,[1] FE Stuurman,[1,3] FG Hofstede,[2] AF Cohen,[1,3] GJA Driessen[2,4]

1 Centre for Human Drug Research, Leiden, the Netherlands
2 Juliana Children's Hospital, Haga Teaching Hospital, The Hague, the Netherlands
3 Leiden University Medical Centre, Leiden, the Netherlands
4 Department of pediatrics, Haaglanden Medical Centre, the Hague, the Netherlands

Abstract

Digital biomarkers are a novel method to capture clinical data in a home-setting. This study aimed to assess the potential of smartwatch-derived biomarkers in pediatric sickle cell disease (SCD). An observational study was conducted with 15 children with SCD. Participants wore a smartwatch and completed a daily symptom questionnaire for 28-days. Data from 128 healthy children was used for comparison. Median compliance for physical activity monitoring was 96%. On average, physical activity, heart rate and sleep duration per day were different when compared to patient groups. Heart rate was correlated with pain scores. Future studies should investigate clinical applications.

Introduction

Pediatric patients with sickle cell disease (SCD) in high-income countries are a vulnerable population. Although patients visit the outpatient clinic every 3-6 months[1], recall of symptoms between visits is probably unreliable. Patients and caregivers often remember acute and severe pain crises but forget more frequent chronic pains and fatigue, which makes objectifying and managing the disease-burden difficult for clinicians[2]. Furthermore, no-shows are a frequent phenomenon in pediatric SCD outpatient clinics, since not all patients understand the benefits of follow-up when children are feeling well, and some do not have the financial resources to visit the hospital[3,4].

Continuous remote monitoring of patients with mobile platform digital measurements in an at-home setting could provide more objective information while simultaneously decreasing the burden of clinical follow-up[5]. However, objective validated (digital) biomarkers obtained outside of the clinic are not readily available[6]. Digital biomarkers could be captured with low-cost consumer-devices like a smartwatch, which registers physical activity (PA), heart rate (HR) and sleep duration[7]. However, a lack of data regarding expected effect sizes in pediatrics makes evaluating these biomarkers difficult, and it is plausible but uncertain that PA and HR are correlated with disease-activity. Clinical fit-for-purpose validation is necessary, and this consists of an assessment of the tolerability, difference between patients and controls, correlation with traditional endpoints and ability to describe clinical events of novel biomarkers[8]. This study aimed to assess the potential of smartwatch-derived biomarkers in pediatric SCD.

Materials & Methods

Ethics and study design

This study was approved by the Medical Ethics Committee Zuidwest Holland (The Hague, The Netherlands) and conducted by Juliana Children's Hospital and the Centre for Human Drug Research. 15 subjects with SCD were recruited in the outpatient clinic between November-2018 and February-2020. After obtaining informed consent, study participants completed a baseline questionnaire[9], wore Steel HR smartwatch (Withings, Issy-les-Molineux, France) and completed a daily pain score and weekly pain questionnaire[10] for 28 days. Data from healthy subjects were collected in a separate study[11].

Endpoints and statistics

Step count per day, steps taken during the most active hour, mean daytime HR (6AM-10PM), mean nocturnal HR (12AM-5AM) and total sleep duration were considered as candidate endpoints and assessed regarding the aforementioned criteria[8]. First, the differences between SCD patients and healthy controls were estimated for all candidate endpoints via mixed effect models with subject as random intercept and condition as fixed factor. Additional adjustments were performed for age, sex, rain duration- and watch wear time per day where appropriate[11]. Daily pain scores were correlated with PA and HR. Clinical events were described descriptively. Tolerability was assessed by calculating the median compliance of the included subjects.

Results

Baseline characteristics and compliance

Baseline characteristics of SCD patients (n=15) and healthy subjects (n=128) are displayed in *Table 1*. Median compliance for sickle cell patients was 96% and 100% for PA and HR, 70% for sleep duration and 46% for questionnaire assessments.

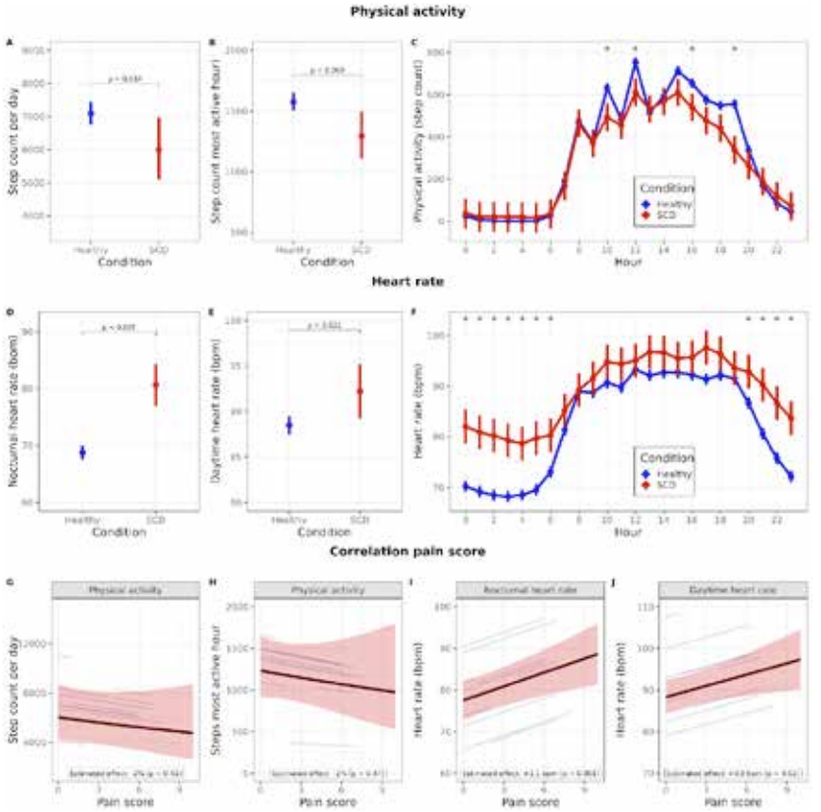
Difference between patients and controls

SCD patients were less physically active and had higher nocturnal- and daytime HR (*Figure 1A-B/D-E*). A separate analysis per hour confirmed the observed differences (*Figure 1C/1F*) and revealed the observed differences in PA were mainly explained by less activity during after-school-hours. On average, children with SCD slept 0.5h (95% confidence interval (CI) 0.1-0.8h, p = 0.008) shorter compared to healthy controls. Adjusted and unadjusted differences are displayed in *Table 2*.

Correlation with daily pain scores

PA-related endpoints were not correlated with daily pain scores (*Figure 1G-H*). However, daytime- and nocturnal HR were related to daily pain scores ('Worst pain while moving', *Figure 1I-J*). For daytime HR, the estimated effect was an increase of 0.9 BPM for each point increase in pain score (95% CI 0.2-1.6, p=0.02), while the estimated effect for nocturnal HR was an increase of 1.1 BPM (95% CI 0.4-1.8, p=0.001).

Figure 1. Differences between SCD patients and healthy controls and correlation with daily pain scores. A-B. Estimated marginal mean (95% CI) step count per day (A) and step count during most active hour per day for healthy children and SCD patients. C. Estimated marginal mean (95% CI) step count per hour of the day for healthy children and SCD patients. D-E Estimated marginal mean (95% CI) nocturnal (D) and daytime (E) HR per day for healthy children and SCD patients. F. Estimated marginal mean (95% CI) HR per hour of the day for healthy children and SCD patients. G-J: Estimated marginal trend (95% CI) for the relationship between pain score while moving and step count per day (G), steps taken during most active hour per day (H), nocturnal HR (I) and daytime HR (J), respectively. The bold line indicates the estimated effect, shaded areas represent the 95% CI of the effect. Individual blue lines represent individual estimates of each subject as random effect. Effects for physical-activity-related measurements are displayed as percentage due to log transformation of the dependent variable. For all graphs, fixed factors (Age and sex for HR and PA, watch wear time and rain duration for PA only) were fixed to their population average.



* Indicates a p-value < 0.05 for the difference after Holm's correction for multiple tests.

Table 1. Baseline characteristics

	SCD patients (n=15)	Healthy subjects (n=128)
Age (mean (SD))	11.6 (3.9)	11.1 (3.1)
SEX		
Male (n (%))	7 (47%)	46%
Female (n (%))	8 (53%)	54%
Weight (kg, mean (SD))	43 (18.8)	43 (16)
Height (m, mean (SD))	1.52 (0.23)	1.51 (0.18)
BMI SDS (mean (SD))	-0.3 (1.2)	0.3 (1.2)
ETHNICITY (N (%))		
Caucasian	0 (0%)	93%
Black	15 (100%)	0 (0%)
Other/mixed	0 (0%)	7%
PedsQL score* (mean (SD))	74.8 (18.1)	90.7 (7.4)
PROMIS Pain questionnaire**	50 (9.3)	-
Nr of painful areas (mean (SD))	1.6 (1.4)	-
Hydroxycarbamide use (n (%)	9 (60%)	-

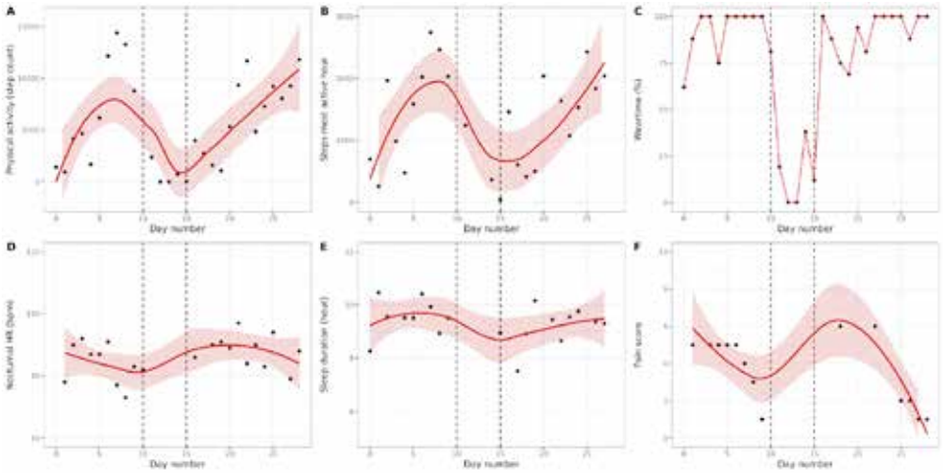
Abbreviations: SD: standard deviation; kg: kilograms; m: meter, BMI: Body mass index; SDS: standard deviation score.
* Higher indicates better QOL, ** PROMIS Pediatric pain questionnaire (scaled), completed weekly per subject. Score was averaged across weeks.

Table 2. Adjusted and unadjusted differences between SCD patients and healthy controls

Candidate biomarker	Unadjusted		Adjusted		Adjusted for
	Difference* (95% ci)	p-value	Difference* (95% ci)	p-value	
Step count per day	1565 (397-2733)	0.009	1094 (82-2105)	0.034	Age, sex**, wear time**, rain duration
Step count during most active hour	314 (78-550)	0.009	218 (71-491)	0.009	Age, sex, rain duration
Daytime HR	-2.7 (-6.9 - 1.5)	0.21	-3.7 (-6.9 - -0.6)	0.02	Age, sex
Nocturnal HR	-11.2 (-15.7 - -6.7)	< 0.001	-11.9 (-15.8 - -8.0)	< 0.001	Age, sex
Sleep duration	0.5 (0.1-0.9)	0.02	0.5 (0.1-0.8)	0.008	Age

* Difference shown are 'estimated mean healthy controls' - 'estimated mean SCD patients'
** interaction between condition and fixed effect included in final model

Figure 2. Description of trajectory before-and after clinical event. Individual data from a subject admitted to the pediatric ward due to a pain crisis. Admission was on day 10 and discharge on day 15. Data shown represents physical activity (step count per day) (A), step count during the most active hour (B), watch wear time per day, expressed as percentage of hours between 6AM and 10PM (C), heart rate at night (D), sleep duration (E) and pain score questionnaire (F) before, during and after hospital admission. Red line represents a loess smoothed regression line. Each dot represents an observation. The subject removed the watch during the hospital admission (C), which must be considered when appraising the other panels.



Clinical events

One subject was admitted to the hospital due to a pain crisis. *Figure 2* shows the individual trajectory of PA, HR, and sleep prior to- and after hospital admission.

Discussion

Novel parametric endpoints that are captured with non-invasive devices and are related to symptom severity or quality of life could improve clinical follow-up and research in pediatric SCD, while simultaneously improving disease-understanding and self-care for patients[5]. In this study, the potential of several smartwatch-derived biomarkers was investigated and appraised regarding predefined criteria. The smartwatch was generally well tolerated, and median compliance for PA and HR measurement was 96% and 100%, respectively. This was much higher than for traditional symptom questionnaire assessments, and indicate that, when clinically validated, continuous monitoring may lead to

superior accuracy and time-resolution of data collection, which may be of particular importance in the case of SCD, where symptomatology over time is variable.

Besides tolerability, a second validation criterion is a difference between patients and controls. Both nocturnal- and daytime HR were higher in SCD patients, despite exhibiting lower physical activity. This is physiologically plausible considered the increased cardiac output commonly reported in SCD,[12] and the observation that heart rate was higher on days with more pain may indicate an increase in sympathetic drive caused by pain is a factor as well[13]. Photoplethysmography (PPG) sensors have been shown to be reliable in a pediatric surgery setting and in a recent study, population averages were comparable to published reference values.[11,14] Although the measurement error inherent to PPG sensors has been reported to be larger for people with a pigmented skin[15], the observed difference appears too large to be attributed only to this.

Patients with SCD were significantly less active compared to a healthy cohort, which has been reported in the past[16]. Further analysis revealed the observed differences were explained by less activity during after-school-hours. This difference could be disease-related and caused by pain and fatigue, but also due to socio-economic factors not captured in the current study, such as the ability to afford sports. We found no correlation between PA and pain scores, possibly due to the low questionnaire compliance. On the other hand, causes and effects of pain in SCD are complex, multifactorial and may not be captured entirely by PA monitoring[17]. Still, descriptive analysis of the period around a sickle cell crisis displayed a clear recovery trajectory of physical activity after the hospital admission. Describing and, ideally, predicting clinical health events is a potentially valuable tool both for clinical trials and clinical care, and future work could focus on prodromal symptoms detectable prior to admission. Additionally, as potential treatments for SCD are developed, digital biomarkers may assist in more reliable evaluation of treatment effects.

The small sample size is a limitation of this study. Small samples are inherent to the field of rare diseases and finding biomarkers that can overcome this limitation is important[18]. Strengths are the systematic fit-for-purpose validation approach followed, the inclusion of patients with a wide age-range, the large cohort of control subjects and the use of mixed effect models able to account for small samples and missing data points. Our study, the first to investigate applications of smartwatch-related biomarkers in pediatric SCD, indicates that PA and HR are different in patients compared to healthy controls and that health events, such as pain crisis, can potentially be detected and monitored. Future studies should focus on longer follow-up period of more patients to determine the potential for clinical trials and -care.

Conclusion

Remote monitoring with a wearable device is feasible in pediatric SCD. There are significant differences in PA and HR between SCD patients and healthy controls, while (nocturnal) HR is correlated to daily pain scores. Future studies are necessary to elucidate the full potential of remote monitoring in SCD.

REFERENCES

1. Yawn BP, Buchanan GR, Afenyi-Annan AN, Ballas SK, Hassell KL, James AH, Jordan L, Lanzkron SM, Lottenberg R, Savage WJ, Tanabe PJ, Ware RE, Murad MH, Goldsmith JC, Ortiz E, Fulwood R, Horton A, John-Sowah J. Management of sickle cell disease: Summary of the 2014 evidence-based report by expert panel members. JAMA - J Am Med Assoc, 2014 312: 1033-1048.
2. Brandow AM, DeBaun MR. Key Components of Pain Management for Children and Adults with Sickle Cell Disease [Internet]. Hematol Oncol Clin North Am, 2018 32: 535-550. Available from: https://doi.org/10.1016/j.hoc.2018.01.014
3. Crosby LE, Modi AC, Lemanek KL, Guilfoyle SM, Kalinyak KA, Mitchell MJ. Perceived barriers to clinic appointments for adolescents with sickle cell disease. J Pediatr Hematol Oncol, 2009 31: 571-576.
4. Cronin RM, Hankins JS, Byrd J, Pernell BM, Kassim A, Adams-Graves P, Thompson AA, Kalinyak K, DeBaun MR, Treadwell M. Modifying factors of the health belief model associated with missed clinic appointments among individuals with sickle cell disease. Hematology, 2018 23: 683-691.
5. Shah N, Jonassaint J, De Castro L. Patients welcome the sickle cell disease mobile application to record symptoms via technology (SMART). Hemoglobin, 2014 38: 99-103.
6. Farrell AT, Panepinto J, Carroll CP, Darbari DS, Desai AA, King AA, Adams RJ, Barber TD, Brandow AM, DeBaun MR, Donahue MJ, Gupta K, Hankins JS, Kameka M, Kirkham FJ, Luksenburg H, Miller S, Oneal PA, Rees DC, Setse R, Sheehan VA, Strouse J, Stucky CL, Werner EM, Wood JC, Zempsky WT. End points for sickle cell disease clinical trials: Patient-reported outcomes, pain, and the brain. Blood Adv, 2019 3: 3982-4001.
7. Lu TC, Fu C-M, Ma M, Fang CC, Turner AM. Healthcare Applications of Smart Watches [Internet]. Appl Clin Inform, 2016 7: 850-869. Available from: http://www.schattauer.de/index.php?id=1214&doi=10.4338/ACI-2016-03-R-0042
8. Kruizinga MD, Stuurman FE, Exadaktylos V, Doll RJ, Stephenson DT, Groeneveld GJ, Driessen GJA, Cohen AF. Development of Novel, Value-Based, Digital Endpoints for Clinical Trials: A Structured Approach Toward Fit-for-Purpose Validation. Pharmacol Rev, 2020 72 (4): 899-909.
9. Varni JW, Seid M, Kurtin PS. PedsQLTM 4.0: Reliability and Validity of the Pediatric Quality of Life InventoryTM Version 4.0 Generic Core Scales in Healthy and Patient Populations [Internet]. Med Care, 2001 39. Available from: https://journals.lww.com/lww-medicalcare/Fulltext/2001/08000/PedsQL_4_0_Reliability_and_Validity_of_the.6.aspx
10. Cunningham NR, Kashikar-Zuck S, Mara C, Goldschneider KR, Revicki DA, Dampier C, Sherry DD, Crosby L, Carle A, Cook KF, Morgan EM. Development and validation of the self-reported PROMIS pediatric pain behavior item bank and short form scale [Internet]. Pain, 2017 158. Available from: https://journals.lww.com/pain/Fulltext/2017/07000/Development_and_validation_of_the_self_reported.18.aspx
11. Kruizinga MD, Heide N van der, Moll A, Zhuparris A, Yavuz Y, Kam ML de, Stuurman FE, Cohen AF, Driessen GJA. Towards remote monitoring in pediatric care and clinical trials-Tolerability, repeatability and reference values of candidate digital endpoints derived from physical activity, heart rate and sleep in healthy children. PLOS One, 2021 16: e0244877.
12. Voskaridou E, Christoulas D, Terpos E. Sickle-cell disease and the heart: Review of the current literature. Br J Haematol, 2012 157: 664-673.
13. Connes P, Coates TD. Autonomic nervous system dysfunction: Implication in sickle cell disease [Internet]. Comptes Rendus - Biol, 2013 336: 142-147. Available from: http://dx.doi.org/10.1016/j.crvi.2012.09.003
14. Pelizzo G, Guddo A, Puglisi A, De Silvestri A, Comparato C, Valenza M, Bordonaro E, Calcaterra V. Accuracy of a Wrist-Worn Heart Rate Sensing Device during Elective Pediatric Surgical Procedures. Children, 2018 5: 38.
15. Shcherbina A, Mikael Mattsson C, Waggott D, Salisbury H, Christle JW, Hastie T, Wheeler MT, Ashley EA. Accuracy in wrist-worn, sensor-based measurements of heart rate and energy expenditure in a diverse cohort. J Pers Med, 2017 7: 1-12.
16. Melo HN, Stoots SJM, Pool MA, Carvalho VO, De Carvalho Aragão ML, Gurgel RQ, Agyemang C, Cipolotti R. Objectively measured physical activity levels and sedentary time in children and adolescents with sickle cell anemia. PLOS One, 2018 13: 1-10.
17. Karlson CW, Baker AM, Bromberg MH, Elkin TD, Majumdar S, Palermo TM. Daily pain, physical activity, and home fluid intake in pediatric sickle cell disease. J Pediatr Psychol, 2017 42: 335-344.
18. Gagne JJ, Thompson L, O'Keefe K, Kesselheim AS. Innovative research methods for studying treatments for rare diseases: Methodological review [Internet]. BMJ, 2014 349: 1-10. Available from: http://dx.doi.org/doi:10.1136/BMJ.G6802

CHAPTER 11

Finding suitable clinical endpoints for a potential treatment of a rare genetic disease: The case of ARID1B

Neurotherapeutics. 2020 Jul;17(3):1300-1310. doi:10.1007/s13311-020-00868-9

MD Kruizinga,[1,2], RGJA Zuiker[1], E Sali[1], ML de Kam[1], RJ Doll[1], GJ Groeneveld[1], GWE Santen[3], AF Cohen[1]

1 Centre for Human Drug Research, Leiden, the Netherlands
2 Juliana Children's Hospital, Haga Teaching Hospital, The Hague, the Netherlands
3 Department of Clinical Genetics, Leiden University Medical Centre, Leiden, the Netherlands.

Abstract

There is a lack of reliable, repeatable, and non-invasive clinical endpoints when investigating treatments for intellectual disability (ID). The aim of this study is to explore a novel approach towards developing new endpoints for neurodevelopmental disorders, in this case for ARID1B-related ID. In this study, twelve subjects with ARID1B-related ID and twelve age-matched controls were included in this observational case-control study. Subjects performed a battery of non-invasive neurobehavioral and neurophysiological assessments on two study days. Test domains included cognition, executive functioning, and eye tracking. Furthermore, several electrophysiological assessments were performed. Subjects wore a smartwatch (Withings® Steel HR) for 6 days. Tests were systematically assessed regarding tolerability, variability, repeatability, difference with control group and correlation with traditional endpoints. Animal fluency, adaptive tracking, body sway and smooth pursuit eye movements were assessed as fit-for-purpose regarding all criteria, while physical activity, heart rate and sleep parameters show promise as well. The event-related potential waveform of the passive oddball and visual evoked potential tasks showed discriminatory ability, but EEG assessments were perceived as extremely burdensome. This approach successfully identified fit-for-purpose candidate endpoints for ARID1B-related ID and possibly for other neurodevelopmental disorders. Next, results could be replicated in different ID populations, and the assessments could be included as exploratory endpoint in interventional trials in ARID1B-related ID.

Introduction

Historically, treatment of intellectual disability (ID) and other neurodevelopmental disorders has focused primarily on the symptoms. Except for a few enzyme deficiency disorders, no treatments of underlying etiology have been incorporated in standard care[1]. Although animal models mimicking clinical ID syndromes have shown promising preclinical data, subsequent trials in humans have failed to show beneficial treatment effects[2,3]. While translation from mice to humans with ID seems unpredictable, it is generally accepted that a lack of reliable clinical endpoints plays a large role in this disparity.[4] Before conducting further interventional trials in subjects and children with ID, new trial designs and especially endpoints are needed[5]. In this study, we investigate the syndrome ARID1B-related ID as a model to illustrate how to develop new biomarkers for a rare neurodevelopmental disorder.

ARID1B-related ID is caused by haploinsufficiency of ARID1B. Pathogenic variants in ARID1B have been identified as a cause of Coffin-Siris syndrome in 2012 for the first time[6,7]. Since then, over 143 patients have been identified and the gene has now been associated with a variable array of phenotypes, ranging from Coffin-Siris syndrome to mild behavioral abnormalities[8]. Most commonly, patients suffer from ID, speech and vision impairment and (partial) agenesis of the corpus callosum, and display distinct facial features[8,9]. Mice with *ARID1B* haploinsufficiency showed similar symptoms and were found to have a reduced number of inhibitory GABA-ergic interneurons, causing a presumed inhibition-excitation imbalance which could be partly reversed by the GABA-A positive allosteric modulator clonazepam[10]. Since clonazepam is considered safe, it is a good candidate drug to investigate in patients. However, without fit-for-purpose endpoints, a trial would be doomed to fail.

Central nervous system (CNS) endpoints in ID trials should be considered fit-for-purpose when they satisfy a number of criteria[11]. In our opinion, they must reflect neurological and functional aspects relevant to the disease and be sensitive to detect pharmacological CNS effects. In the case of ID, endpoints should also be non-invasive and easily conducted. Repeatability should be determined in the targeted population and there should be a clear differentiation between patients and control subjects. Ideally, test results should correlate with existing indicators of disease severity.

The NeuroCart® is a neurological test battery known to be sensitive for the detection of CNS effects of compounds[12]. Using this test battery, non-invasive and data-intensive

studies can be performed to demonstrate specific, time- and dose-dependent, neurophysiological and neuropsychological effects[12]. However, the assessments have not been investigated in patients with ID yet. The aim of this study is to explore the characteristics a battery of non-invasive neurophysiological and neurobehavioral assessments that may be fit-for-purpose as future clinical endpoints for ARID1B-related ID and similar syndromes.

Materials and methods

This study was conducted at the Centre of Human Drug Research (CHDR) in Leiden, the Netherlands from November 2018 until May 2019 and the protocol was reviewed and approved by the Beoordeling Ethiek Biomedisch Onderzoek (BEBO) Foundation Review Board (Assen, the Netherlands). The study was conducted according to the Dutch Act on Medical Research Involving Human Subjects, the Dutch codes of conduct regarding medical research with minors and expression of objection by people with mental disabilities and in compliance with Good Clinical Practice.

Subjects and study design

During this case-control study, twelve patients with pathogenic variants in ARID1B were recruited via the Coffin-Siris expertise centre of the Leiden University Medical Centre. Twelve age-matched healthy controls were also recruited. Age difference between patients and controls was no more than 2 years, except for adult subjects. Subjects who regularly used benzodiazepines were excluded from the study. Tests were assessed for suitability for two age-groups (2-4 and ≥ 5 years old), based on the expected capabilities of the subjects. Tests were performed on two consecutive Saturdays and were repeated 2-4 times during the study. The schedule of assessments is listed in *Supplementary Figure S1*. Study visits lasted 3-5 hours. Baseline characteristics, including the last measured intelligence quotient (IQ) score were obtained from patient charts. Parents completed the Aberrant Behaviour Checklist (ABC) at the start of the study[13].

Selection of candidate endpoints

NeuroCart® tests were selected based on the following criteria:

1. Tests must have demonstrated potential in detecting CNS effects of compounds;
2. Tests must investigate a CNS domain assumed to be affected in patients with ARID1B-related ID;
3. It must be reasonably expected that the tests can be conducted by children and patients with ARID1B-related ID;
4. Ideally, an improvement in test outcomes potentially results in symptom reduction or improvement of quality of life. Selected tests and the accompanying rationale for inclusion are listed in *Table 1*.

Table 1. Rationale for selected tests

	Test	CNS domain	Corresponding ARID1B symptom
Cognition	Animal fluency test	Verbal fluency, semantic memory	Intellectual disability
	VVLT	Memory	Intellectual disability
	Day-Night test	Memory and controlled processing	Impulsiveness and intellectual disability
Eye tracking	Smooth pursuit	Attention and oculomotor function	Expected markers for clonazepam effect
	Saccadic eye movements	Sedation	
Executive functioning	Adaptive tracking	Motor activation and attention	Short attention span
	Finger tapping	Motor activation and fluency	Lethargy and slowness
	Body sway	Balance and attention	Hyperactivity
Electrophysiology	Resting EEG	General CNS activity	
	Passive oddball	Auditory processing	Hypothesized abnormal neuronal organization and general CNS functioning
	Active oddball	Auditory processing	
	VEP	Visual processing	
	ASSR	Auditory processing	
trial@home	Steel HR - Physical activity	General daily activity	Hyperactivity, apathy, and lethargy
	Steel HR - Sleep parameters	Sleep	Insomnia
	Steel HR - Heart rate	Sympathic activation and arousal	Hyperactivity

Abbreviations: VVLT: Visual Verbal Learning Test; EEG: electroencephalography; VEP: Visual Evoked Potential; ASSR: Auditory Steady State response.

Test procedures

COGNITION For the animal fluency test, subjects were asked to verbally produce as many different animals as they could sum up within sixty seconds[14]. Animals that were named twice or more did not count towards the total amount of animals named and neither were infant version of adult animals already named. During the visual verbal learning

test (VVLT), subjects were presented 30 words in three consecutive word trials[15]. Each trial ended with a free recall of the presented words. Thirty minutes after the first trial, subjects were asked to recall the words. Immediately thereafter, subjects underwent memory recognition test, consisting of 15 presented words and 15 'distractors'. The day-night task, a simplified version of the Stroop test suitable for children, was included in the study to assess memory and controlled processing of subjects[16].

EYE TRACKING Recording of eye movements was performed in a quiet room with dimmed illumination. Analysis was conducted with a microcomputer-based system for sampling of eye movements. Disposable electrodes (Ambu Blue Sensor N) were applied on the forehead and beside the lateral canthi of both eyes. Skin resistance was minimalized before measurements. Head movements were restrained using a fixed head support. Subjects were asked to focus on a moving dot displayed on a computer screen. Saccadic eye movements were recorded for stimulus amplitudes of approximately 15 degrees to either side. Fifteen saccades were recorded with interstimulus intervals varying randomly between 3 and 6 seconds. Average values of saccadic peak velocity (degrees/second) of correct saccades were recorded. At least five detected saccades were necessary to include for statistical analysis. For smooth pursuit eye movements, the target moves sinusoidally at frequencies ranging from 0.3 to 1.1 Hz. Four cycles were recorded for each stimulus frequency. The time during which the eyes were in smooth pursuit of the target was calculated and expressed as a percentage of stimulus duration[17].

EXECUTIVE FUNCTIONING ASSESSMENTS The adaptive tracking test is a pursuit-tracking task and was performed as described by Borland and Nicholson[18] using customized equipment and software. The subjects were instructed to keep a dot inside a moving circle by operating a joystick. The speed of the moving circle adapted in response to subject performance. After a run-in period of 30 seconds, the average tracking performance (%) of 3.5 min was used for statistical analysis. The finger tapping test was adapted from the Halstead Reitan Test Battery[19]. Speed of finger tapping was measured for the index finger for the dominant hand; a session contained five performances of 10 seconds. Subjects were instructed to tap a button as quickly as possible with the index finger of the dominant hand. The mean tapping rate was used for statistical analysis. Body sway was conducted by all subjects and assessed using a pot string meter (celesco) based on a Wright ataxiameter, with a string attached to the waist[20]. All body movements over 2 minutes were integrated and expressed as sway in mm. Before starting a measurement, subjects were asked to stand still and comfortable with their hands in a relaxed position. Subjects wore an eye cap to block sight.

ELECTROPHYSIOLOGICAL ASSESSMENTS Complete technical details of measurements and analysis of electrophysiological assessments are listed in *Supplementary Text 1*. To measure general CNS-activity, resting-state EEG with open and blocked eyes was recorded. Spectral analysis of the α1, α2, β1, β2, β3, δ and θ waves was performed to calculate the power of the respective wavebands at FzCz, PzO1, and PzO2. VEPS (visual evoked potentials) were recorded over the scalp overlying the occipital cortex. During the VEP assessment, a pattern reversal paradigm was used with two phase-changing checkerboards (1.0- and 0.25-degree pattern). The oddball paradigm is a neuropsychological test to evoke event related potentials (ERPS). During the passive oddball task, subjects were seated with EEG cap and headphones on and instructed to sit still and relax. Subjects were watching a silent movie while being presented auditory tones as frequent stimuli and infrequent stimuli. For the active oddball task, subjects were to pay attention to the tones and press a button when they heard an infrequent tone. The auditory steady state response (ASSR) is an electrophysiological response to periodic auditory stimulation[21], thought to be generated through entrainment of neuronal populations to periodic stimuli, and reveal the integrity of neuronal networks. During the test, auditory stimuli with a 500ms burst of 1ms monophasic rectangular pulses were presented through headphones. The inter-stimulus interval was 700ms. The ASSR was assessed through spectral modulations and inter trial phase coherence (ITPC).

TRIAL@HOME During the six days between measurements, subjects wore a Steel HR watch (Withings, Issy-les-Moulineaux, France), which is incorporated in the CHDR MORE® trial@home platform and which registered step count, heart rate and several accelerometer-derived sleep parameters. After the final study assessments, parents and children completed a questionnaire regarding the general study experience.

STATISTICS Considering the exploratory nature of this study, no formal power calculation was performed. Statistical analysis was performed with SAS v9.4 (SAS Institute, Cary, NC, USA). The difference between patients and controls was calculated via a repeated measure mixed model analysis of variance with fixed factors group,

measurement and group by measurement and subject as random factor. Based on the model, a minimal detectable effect size (MDES) was calculated for a hypothetical cross-over study with 16 ARID1B subjects. The MDES was expressed as the proportion of the difference between ARID1B-related ID patients and controls to determine whether the effect size is a reasonable goal for future interventional studies. Spearman correlations between mean test outcomes and IQ and ABC subscales were calculated for tests for which a significant difference between patients and controls was demonstrated. Promasys® 7.3 (Omnicomm, Fort Lauderdale, Texas, USA) was used for data management.

CRITERIA FOR CANDIDATE ENDPOINTS Tests where considered fit-for-purpose when fulfilling the following requirements:

1. Tolerable, meaning subjects showed no signs of resistance during the test;
2. Conducted correctly by the study population, with more than 75% of the outcomes suitable for analysis;
3. Stable over time, defined as a coefficient of variability (cv) within the ARID1B group not higher than 50%;
4. Statistically significant difference between healthy and control subjects;
5. Show an mdes which is less than 50% of the difference between ARID1B subjects and control subjects. Ideally, there is an association between test outcome and IQ or relevant ABC subscales.

Results

Baseline characteristics

A total of 20 parents of patients with ARID1B-related ID were approached, of which 12 consented to participate with their child. Twelve healthy age-matched controls were included. Baseline characteristics are displayed in *Table 2*. Subjects with ARID1B-related ID scored highest on the hyperactivity, lethargy and irritability ABC subscales. Parents found the length of study days to be too long (55%), but none indicated they would not participate in a similar study again. The youngest ARID1B subject was 2 years old and found performing tests too difficult, after which the second study day was canceled. There were no adverse events during the conduct of this study. Individual subject characteristics are listed in *Supplementary Table S1*.

Table 2. Baseline characteristics

	ARID1B (n=12)	Controls (n=12)
Age[1] (mean (range))	12.6 [2-31]	11.8 [2-27]
Sex, female (n (%))	9 (75)	12 (100)
Conc. Medication (n (%))	3 (17)	1 (8)
IQ (mean±SD)[2]	74±21	-
Can read age appropriately (n (%))[3]	8 (67)	12 (100)
Can write age appropriately (n (%))[3]	8 (67)	12 (100)
Behavioral problems (n (%))[2]	7 (58)	0 (0)
Speech delay or impairment (n (%))[2]	12 (100)	0 (0)
Vision problems (n (%))[2]	7 (58)	0 (0)
ABC subscale score (mean±SD)		
Irritability	8.3±7.4	-
Lethargy	11.2±17.2	
Stereotypic behavior	2.4±2.1	
Hyperactivity	13.1±10.0	
Inappropriate speech	1.0±1.5	

1 The mean age difference between patients and corresponding controls was 0.75 years.
2 Data obtained from patient charts, when available.
3 Parent-reported.

Candidate endpoints

All conducted tests were assessed according to the specified criteria. Summarized results are displayed in *Table 3*. Individual test performance is displayed in *Supplementary Table S2*.

COGNITION The animal fluency test was successfully conducted by 80% of participants. One nonverbal patient did not complete the test. There was a significant difference between the two study groups and a positive correlation between the number of named animals and IQ (*Figure 1*). Day-night test results did not differ between the two study groups (p = 0.133). The VVLT was considered too difficult and resulted in stress for the first three patients, after which the test was removed from the study.

Table 3. Systematic evaluation of assessments to determine suitability as endpoint in clinical trials

	Test	Tolerable[1]	Conducted correctly[1]	Repeatable (CV[2])	Group difference			MDES cross-over design[3]	Fit-for-purpose
					ARID1B Mean	Control Mean	p-value		
NEUROCART®	Animal Fluency test (n)	Yes	80%	40%	7.8	21.8	0.001	24%	Yes
	VVLT	Yes	33%[4]						No
	Day-Night test (n)	Yes	92%	12%	11.9	15.3	0.13		No
	Smooth pursuit (%)	Yes	100%	34%	14.7	35.9	< 0.001	25%	Yes
	SACCADIC EYE MOVEMENTS								
	Saccadic peak velocity (degr/s)	Yes	40%	3%	515	498	0.64		No
	Saccadic reaction time (ms)			10%	0.28	0.25	0.31		
	Adaptive tracking (%)	Yes	93%	51%	2.47	18.99	< 0.001	8%	Yes
	Finger tapping (n)	Yes	100%	6%	35.31	46.25	0.008	21%	Yes
	Body sway (mm)	Yes	88%	22%	1188	300	0.002	34%	Yes
ELECTROPHYSIOLOGY	Resting EEG	Varying	83%	SUPPLEMENTARY TABLE S1					No
	Passive oddball (MMN latency)	Varying	87%	22-25%[5]	FIGURE 2A-C; SUPPLEMENTARY TABLE S1			112%[6]	No
	Active oddball	Varying	0%						No
	VEP	Varying	69%	11% - 63%[7]	FIGURE 2D; SUPPLEMENTARY TABLE S1			32-73%[6]	No
	ASSR (ITPC)	Varying	74%	232%	0.135	0.186	0.19		No
TRIAL@HOME	Physical activity (step count/day)	Yes	100%	30%	5559	7184	0.06	108%	More research needed
	STEEL HR - SLEEP								
	Sleep duration (h)	Yes	100%	14%	8.7	8.9	0.55	-	
	Light sleep (%)			12%	58%	51%	0.076	108%	
	Times to wake up (n)			51%	3.7	1.7	0.006	101%	
	HEART RATE								
	Daily (BPM)	Yes	100%	-	91	86	< 0.001	-	
	Nocturnal (BPM)			5%	80	73	0.009	55%	

Colors: Green = suitable; Red = unsuitable; Yellow: indeterminate. Abbreviations: CV: coefficient of variability; MDES: minimal detectable effect size; VVLT: visual verbal learning test; EEG: electroencephalography; MMN: mismatch negativity; VEP: visual evoked potential; ASSR: auditory steady state response; ITPC: inter-trial phase coherence; BPM: beats per minute. 1 By ARID1B-related ID subjects. Investigator's assessment after exit interview and end-of-study questionnaire with parents; 2 Coefficient of variability within the group of ARID1B-related ID subjects; 3 Minimal detectable effect size. Expressed as the proportion of the difference between patients and controls that can be detected as improvement in a crossover study with n = 16; 4 Only 1 of the first 3 ARID1B-related ID subjects was able to obtain a valid score, after which the test was removed from the study protocol; 5 Range of CV of the MMN latency at Cz and Fz; 6 Range of MDES calculated only for parameters with a significant difference between ARID1B subjects and controls; 7 Range of CVs of collected parameters.

EXECUTIVE FUNCTIONING AND EYE TRACKING All executive functioning tests (Adaptive tracking, Body sway, Finger tapping) were tolerable and conducted correctly. Notably, finger tapping was the favorite assessment for 85% of subjects and 73% of parents. There was a clear and significant difference between subjects and controls for the three tests, while a correlation was present between the ABC hyperactivity subscale and adaptive tracking results. Patients demonstrated a significantly lower smooth pursuit capability compared to controls and there was a correlation between mean smooth pursuit results and the ABC hyperactivity subscale (*Figure 1*).

ELECTROPHYSIOLOGICAL TESTS All 24 subjects completed at least one resting-state EEG. On average, a slightly higher α2, δ and θ power was detected in the occipital electrodes in ARID1B patients compared to healthy controls. The passive oddball ERP graph is displayed in *Figure 2A-C*. ARID1B patients had a statistically significant difference in mismatch negativity (MMN) latency at Cz compared to controls (183ms vs 141ms, p = 0.014), while MMN latency at Fz and the amplitude were statistically similar. The evoked responses were different (*Figure 2A-B*). The active oddball paradigm was considered too difficult for the first three patients and was subsequently removed from the study. VEP evoked response (*Figure 2D*) demonstrates significantly lower amplitude and longer latency of the P100 peak, as well as a smaller N75-P100 peak-to-peak amplitude. ASSR was performed successfully in 74% of measurements, but there was no significant difference regarding ITPC and evoked power between patients and controls. 50% of patients and controls found setting up the EEG cap to be generally uncomfortable and 42% of parents and 48% of subjects indicated EEG assessments were their least favorite. EEG analysis was suboptimal due to recurrent signal artefacts caused by movements of both ARID1B subjects and the younger control subjects, which led to a low overall signal quality.

TRIAL@HOME The Steel HR watch was tolerated by all subjects. 73% of subjects and 100% of parents indicated 6 days of measurements was enough or short. There was a difference in physical activity of 1625 steps per day between patients and controls, although this did not reach conventional significance (p = 0.06). The sleep duration of both groups was similar (8.7 hours for patients and 8.9 hours for controls), but patients woke up significantly more often (3.7 vs 1.7, p = 0.006). Accordingly, there was a trend towards a lower proportion of light sleep per night and a significantly higher nocturnal heart rate for patients.

Figure 1: Estimated group means and exploratory correlations of NeuroCart® tests. A. Mean outcome of the animal fluency test per subject group and measurement number; B: Linear correlation between historic IQ score and mean animal fluency test score. C: Mean adaptive tracking test outcome per subject group and measurement number; D: Linear correlation between the ABC hyperactivity subscale and mean adaptive tracking test score; E: Mean smooth pursuit eye movement test outcome per subject group and measurement number; F. Linear correlation between the ABC hyperactivity subscale and mean smooth pursuit eye movement test score. The dotted lines demarcate the two study days.

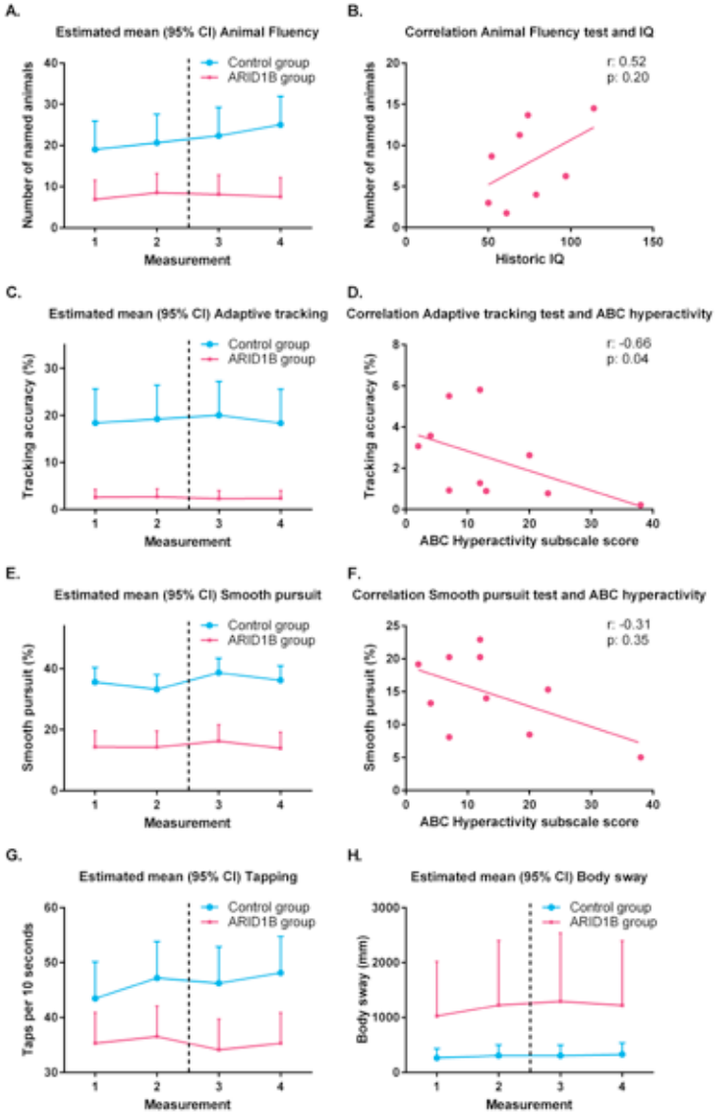
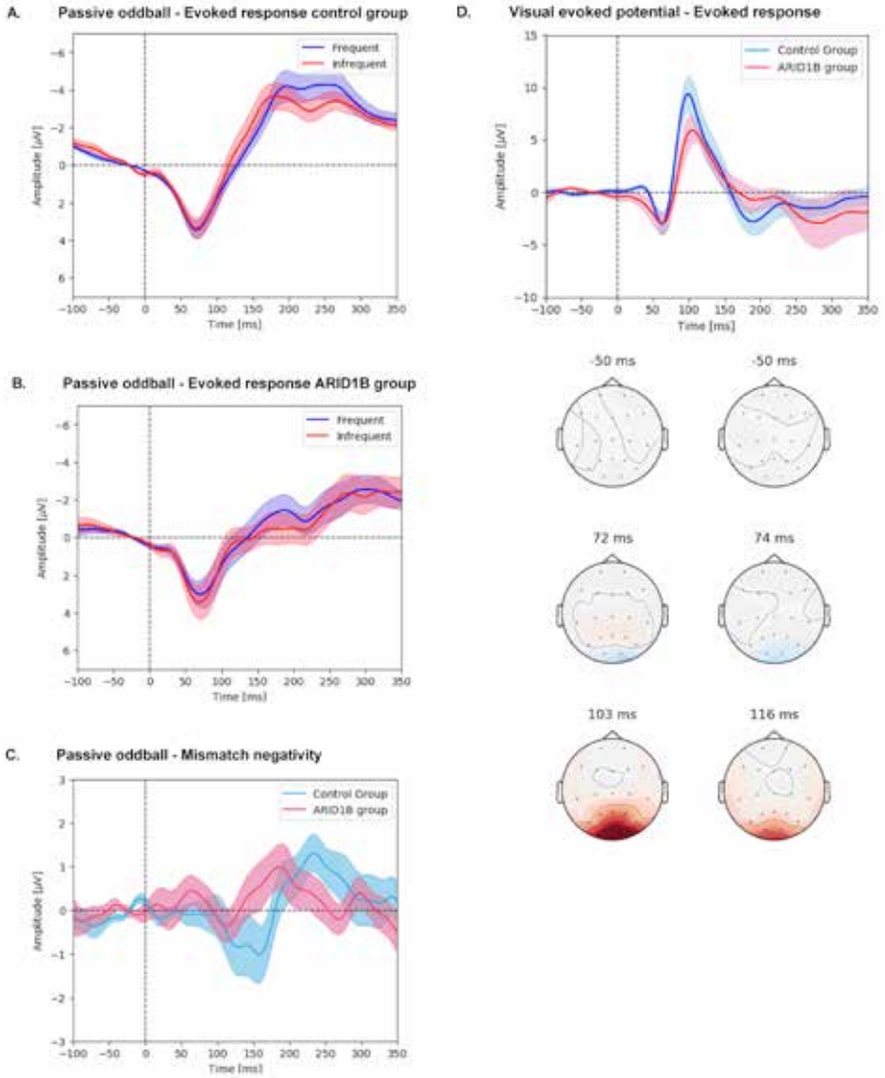


Figure 2. ERPs of patients and controls for passive oddball and VEP assessments. A: Grand mean of the evoked response during the passive oddball task for ARID1B subjects; B: Grand mean of the evoked response during the passive oddball task for control group; C: Mismatch negativity graph; D: Visual evoked potential (VEP) ERP graph after visual stimulation with 1.0-degree phase changing checkerboard, including EEG heat map (Left: control group; Right: ARID1B group). Although subjects were also stimulated with a 0.25 degree checkerboard, the high prevalence of refractive ametropia among ARID1B patients in combination with their disability made it impossible to determine whether all subjects saw the 0.25 degree checkerboard clearly. A statistical summary of passive oddball and VEP analysis are listed in Supplementary Table S1.



Discussion

In this study, twelve patients with ARID1B-related ID and twelve age-matched controls performed a battery of non-invasive neurophysiological and neurobehavioral assessments. All assessments were reviewed for suitability as new clinical endpoint in clinical trials investigating interventions in populations with (ARID1B-related) ID. This study represents the first of its kind, providing an extensive neurobehavioral and neurophysiological phenotype of a population with an ultra-rare condition.

Of the included tests investigating cognition, the animal fluency test was identified as a promising endpoint fulfilling all predefined criteria. Considering the presence of ID in patients, the difference compared to controls was expected. However, the animal fluency test shows stability over time and the absence of a learning effect in the ARID1B group, making it suitable for the assessment of acute and medium-term treatment effects.

Smooth pursuit eye movements fulfilled the criteria for candidate endpoints as well. In contrast, there was no statistical difference in saccadic peak velocity between patients and controls. We hypothesize this is due to the fact saccadic eye movements are a relatively preserved mechanism involving brainstem responses, while smooth pursuit is a function requiring coordination of multiple brain regions, vulnerable to developmental abnormalities[22,23]. This has been previously demonstrated in autism and schizophrenia[24,25]. The difference in the proportion of correctly conducted tests between smooth pursuit and saccadic eye movements was striking. We hypothesize it is more difficult to concentrate on the dot when it randomly changes position during the saccadic eye movement test, as opposed to the continuously moving dot during smooth pursuit. In the future, continuous encouragement during the test may improve the amount of analyzable results.

The three executive functioning tests were found to be suitable candidate endpoints for future studies. This is the first study to investigate these assessments in subjects with ID for this purpose, and overall test results were in line with observed symptoms. For example, patients moved considerably more than controls during the body sway test, reflecting the restlessness that ARID1B subjects exhibit. The patients' slower finger tapping may express the lethargy patients with ARID1B-related ID suffer from. The adaptive tracking test was also positively assessed on all criteria and was correlated with the ABC hyperactivity subscale. The MDES of the ARID1B group in a cross-over study with 16 subjects (1.349), relative to the found difference between the control and ARID1B group (16.52) is 8%, making the adaptive tracking test the most sensitive test within this study to detect potential treatment effects.

We performed a range of electrophysiological assessments investigating general CNS-activity and auditory and visual processing. Interpretation of all electrophysiological assessments was hampered due to a poor signal quality. Final analysis was performed after excluding trials of insufficient quality. For the passive oddball paradigm, ARID1B subjects appeared to exhibit a longer latency of the MMN at Cz, but not the amplitude or the latency at Fz. This may reflect an impaired automatic auditory processing ability also found in one other study in subjects with ID[26]. The general evoked response appeared to be smaller for both frequent and infrequent tones (*Figure 2A-B*), and the grand mean of the MMN (*Figure 2C*) shows two small negative components before and after the average latency. This may be due to unidentified subgroups within ARID1B patients, but displaying grand means of the MMN is not ideal in this study. The MMN matures at increasing age[27]. which leads to varying MMN latencies for the different age groups. In the context of biomarker development, the estimated group means obtained via mixed model analysis may be more suitable. VEP demonstrated a longer latency and lower amplitude of the P100 peak, indicating a slower automatic visual processing ability. The complete electrophysiological substrate of these results goes beyond the scope of this paper. Several studies have shown that pharmacological activity can alter the ERP waveform, making them an interesting, and potentially non-invasive, biomarker for drug effect in neurological disease[28]. While electrophysiological assessments can theoretically be performed by subjects of all ages, assessments were considered quite invasive for ARID1B subjects. The recurrent movement artefacts caused a significantly reduced data quality, and results should be interpreted with care. These findings show the value of our approach: EEG and ERPS could be very useful biomarkers in clinical trials, but recurring tests are unsuitable in this population. One could even argue that incorporation of EEG assessments in future trials would introduce bias, giving only children who are less affected by the disease the chance to repeatedly perform the assessments.

We demonstrated that an unobtrusive smartwatch can be used for home-monitoring of ID patients. Of the collected parameters, notable differences between patients and controls were found in the nocturnal parameters (number of times to wake up, nocturnal heart rate). There was a trend towards a lower physical activity level per day in patients compared to controls, although this was not a significant difference. We expect that it may be possible to detect adverse or unexpected effects of treatments, such as difficulty sleeping or apathy resulting in a decrease in physical activity using the measurements described here. However, other Withings® smartwatch models have shown a lack of

reliability compared to the gold standard regarding measurement of sleep and sleep data should be interpreted with caution[29].

To summarize, we assert that the combination of animal fluency, finger tapping, body sway, adaptive tracking, smooth pursuit eye movements, and possibly home-monitoring with the Steel HR, represents a promising battery of non-invasive tests suitable for interventional studies. In our opinion, this battery of tests is non-invasive and can be conducted correctly by ID patients of 5 years and older. Furthermore, the MDES of tests calculated for a study with a feasible sample size (n = 16) reflect reasonable improvements of less than 50% of the difference between patients and controls.

Except for adaptive tracking, we found no statistically significant correlations with the traditional endpoints IQ and ABC subscales. However, significance was not expected considering this study was not powered adequately for this, and the limitations of endpoints currently used in ID trials. IQ and questionnaires have traditionally been used, but IQ has a high intrasubject variability[30], especially at a young age[31], while interpretations of questionnaires are subjective and invariably suffer from inter-rater bias[4]. Objective and standardized tests with low intra-subject variability are more suitable for early phase drug research in small patient groups. Still, improvement in parent-reported behavior certainly represents value for the individual patients and their parents. A combination of objective tests and subjective parent-reported outcomes in future trials may therefore emerge as the best paradigm.

This study has several limitations. First, the recruitment of patients focused on relatively mentally and physically competent ARID1B subjects thought to be able to tolerate traveling to the research location and being administered the test battery. Therefore, the generalizability of the study results regarding patients with severe ID is unclear. We believe the study subjects represent the population that would participate in any interventional clinical trial as well. Several cut-offs, such as for the CV and MDES, were chosen rather arbitrarily and could be more clearly defined when designing a follow-up study. We used historical IQ as a variable to correlate test outcomes with, introducing another factor of uncertainty. Historical IQ of healthy subjects was not available and could therefore not be compared between the groups. However, none of the included control subjects had learning difficulties and historical IQ was only used in the correlation with cognitive test outcomes. Although correlation of raw cognitive scores with age-adjusted standard scores such as IQ is unconventional, this can't be avoided when no normative values of the included assessments have been determined yet. Age showed some correlation with average test outcome (*Supplementary Figure S2*), but this was expected and does not negatively impact the fit-for-purpose assessment of the most promising tests. A strength of this study is the repeated measurements design, generating robust data about the variability of study assessments. The included battery of tests investigated all functional CNS domains[12,32], resulting in a comprehensive neurophysiological and neurobehavioral phenotype of ARID1B-related ID. While many psychometric properties of the candidate endpoints are unknown in the ID population, most tests have been performed in a pediatric population in the past (unpublished data) and have been extensively investigated in adult neurological disorders[12]. While the included tests have at least a theoretical relationship between disease-severity and test outcome, as outlined in *Table 1*, this relationship must be confirmed in future studies. The included healthy controls allowed us to calculate an MDES relative to the control group, which aids in the interpretation of the effect size. Finally, we have included a relatively large cohort of patients considering the total population of patients with ARID1B-related ID in the Netherlands.

This study shows that our approach towards the identification of fit-for-purpose endpoints in rare neurodevelopmental disorders has been successful in the case of ARID1B-related ID. During the next stage of endpoint development, the identified candidate endpoints could be included as exploratory or secondary endpoint in interventional trials in ARID1B-related ID. Furthermore, since there is a large phenotypic variability within the population, test outcomes could be compared in subgroups throughout the ARID1B spectrum. We expect our results are not specific for ARID1B-related ID. Future studies should also focus on the identified potential endpoints in patients with similar syndromes. Furthermore, prior to conducting trials investigating long-term treatment effects over a timespan of years, studies aiming to uncover natural progression in test outcomes over time in patients should be performed in order to properly isolate long-term treatment effects during analysis[33].

Conclusion

We have identified the animal fluency test, adaptive tracking, smooth pursuit eye movement, finger tapping and body sway as promising endpoints for clinical trials in patients with ARID1B-related ID. More research is needed in the field and physical activity-and sleep monitoring. The results from this study will be used in the preparation of an interventional clinical trial investigating the effects of clonazepam in patients with ARID1B-related ID.

SUPPLEMENTARY DATA

Sup. Figure S1	Schedule of assessments
Sup. Figure S2	Correlation between age and test outcome for the most promising tests
Sup. Text S1	Standard operating procedures for eeg and erp assessments
Sup. Table S1	Individual clinical characteristics
Sup. Table S2	Individual subject characteristics and successfully performed tests

REFERENCES

1 van Karnebeek CDM, Bowden K, Berry-Kravis E. Treatment of Neurogenetic Developmental Conditions: From 2016 into the Future [Internet]. Vol. 65, Pediatric Neurology. Elsevier Inc; 2016. p. 1–13. Available from: http://dx.doi.org/10.1016/j.pediatrneurol.2016.07.010

2 Lee AW, Ventola P, Budimirovic D, Berry-Kravis E, Visootsak J. Clinical development of targeted fragile X syndrome treatments: An industry perspective. Vol. 8, Brain Sciences. 2018. p. 1–15.

3 Berry-Kravis E, Hagerman R, Visootsak J, Budimirovic D, Kaufmann WE, Cherubini M, *et al.* Arbaclofen in fragile X syndrome: Results of phase 3 trials. J Neurodev Disord. 2017;9(1):1–18.

4 Budimirovic DB, Berry-Kravis E, Erickson CA, Hall SS, Hessl D, Reiss AL, *et al.* Updated report on tools to measure outcomes of clinical trials in fragile X syndrome. J Neurodev Disord. 2017;9(1):1–36.

5 Abrahamyan L, Feldman BM, Tomlinson G, Faughnan ME, Johnson SR, Diamond IR, *et al.* Alternative designs for clinical trials in rare diseases. Am J Med Genet Part C Semin Med Genet. 2016;172(4):313–31.

6 Santen GWE, Aten E, Sun Y, Almomani R, Gilissen C, Nielsen M, *et al.* Mutations in SWI/SNF chromatin remodeling complex gene ARID1B cause Coffin-Siris syndrome. Nat Genet [Internet]. 2012;44(4):379–80. Available from: http://dx.doi.org/10.1038/ng.2217

7 Tsurusaki Y, Okamoto N, Ohashi H, Kosho T, Imai Y, Hibi-Ko Y, *et al.* Mutations affecting components of the SWI/SNF complex cause Coffin–Siris syndrome. Nat Genet [Internet]. 2012;44(4):376–8. Available from: http://dx.doi.org/10.1038/ng.2219

8 van der Sluijs E (PJ)., Jansen S, Vergano SA, Adachi-Fukuda M, Alanay Y, AlKindy A, *et al.* The ARID1B spectrum in 143 patients: from nonsyndromic intellectual disability to Coffin–Siris syndrome. Genet Med. 2018;13–24.

9 Santen GWE, Clayton-Smith J, Adachi-Fukuda M, AlKindy A, Baban A, Berry K, *et al.* The ARID1B phenotype: What we have learned so far. Am J Med Genet Part C Semin Med Genet. 2014;166(3):276–89.

10 Jung EM, Moffat JJ, Liu J, Dravid SM, Gurumurthy CB, Kim WY. ARID1B haploinsufficiency disrupts cortical interneuron development and mouse behavior. Nat Neurosci [Internet]. 2017;20(12):1694–707. Available from: http://dx.doi.org/10.1038/s41593-017-0013-0

11 Sahin M, Jones SR, Sweeney JA, Berry-Kravis E, Connors BW, Ewen JB, *et al.* Discovering translational biomarkers in neurodevelopmental disorders. Nat Rev Drug Discov [Internet]. 2018;18(april):7–8. Available from: http://dx.doi.org/10.1038/d41573-018-00010-7

12 Groeneveld GJ, Hay JL, Van Gerven JM. Measuring blood–brain barrier penetration using the NeuroCart, a CNS test battery [Internet]. Vol. 20, Drug Discovery Today: Technologies. The Author(s); 2016. p. 27–34. Available from: http://dx.doi.org/10.1016/j.ddtec.2016.07.004

13 Aman MG, Singh NN, Stewart AW, Field CJ. The aberrant behavior checklist: a behavior rating scale for the assessment of treatment effects. Am J Ment Defic [Internet]. 1985;89(5):485–91. Available from: http://www.ncbi.nlm.nih.gov/pubmed/3993694

14 Tombaugh, T. N., Kozak, J. & Rees L. Normative data stratified by age and education for two measures of verbal fluency : FAS and Animal Naming. Archives of Clinical Neuropsychology, 14, 167–177. 1999;0(2):13–4.

15 van der Elst W, van Boxtel MPJ, van Breukelen GJP, Jolles J. Rey's verbal learning test: Normative data for 1855 healthy participants aged 24–81 years and the influence of age, sex, education, and mode of presentation. J Int Neuropsychol Soc. 2005;11(3):290–302.

16 Gerstadt CL, Hong YJ, Diamond A. The relationship between cognition and action: Performance of 3.5–7 year olds on a Stroop–like day night test. Cognition. 1994;53(1):129–53.

17 Van Steveninck AL, Schoemaker HC, Pieters MSM, Kroon R, Breimer DD, Cohen AF. A comparison of the sensitivities of adaptive tracking, eye movement analysis, and visual analog lines to the effects of incremental doses of temazepam in healthy volunteers. Clin Pharmacol Ther. 1991;50(2):172–80.

18 Borland R, Nicholson A. Visual motor co–ordination and dynamic visual acuity. Br J Clin Pharmacol. 1984;18(1 S):69S–72S.

19 Dikmen SS, Heaton RK, Grant I, Temkin NR. Test-retest reliability and practice effects of expanded Halstead-Reitan Neuropsychological Test Battery. J Int Neuropsychol Soc [Internet]. 1999 May 1 [cited 2019 Sep 29];5(4):346–56. Available from: https://www.cambridge.org/core/product/identifier/S1355617799544056/type/journal_article

20 van Steveninck AL, Gieschke R, Schoemaker HC, Pieters MSM, Kroon JM, Breimer DD, *et al.* Pharmacodynamic interactions of diazepam and intravenous alcohol at pseudo steady state. Psychopharmacology (Berl). 1993;110(4):471–8.

21 Brenner CA, Krishnan GP, Vohs JL, Ahn WY, Hetrick WP, Morzorati SL, *et al.* Steady state responses: Electrophysiological assessment of sensory function in schizophrenia. Schizophr Bull. 2009;35(6):1065–77.

22 Mustari MJ, Ono S, Das VE. Signal processing and distribution in cortical–brainstem pathways for smooth pursuit eye movements. Ann N Y Acad Sci. 2009;1164:147–54.

23 McDowell JE, Dyckman KA, Austin BP, Clementz BA. Neurophysiology and neuroanatomy of reflexive and volitional saccades: Evidence from studies of humans. Brain Cogn [Internet]. 2008;68(3):255–70. Available from: http://dx.doi.org/10.1016/j.bandc.2008.08.016

24 Wilkes BJ, B. Carson T, Patel KP, Lewis MH, White KD. Oculomotor performance in children with high-functioning Autism Spectrum Disorders. Res Dev Disabil [Internet].

2015;38:338–44. Available from: http://dx.doi.org/10.1016/j.ridd.2014.12.022
25 O'Driscoll GA, Callahan BL. Smooth pursuit in schizophrenia: A meta-analytic review of research since 1993. Brain Cogn [Internet]. 2008;68(3):359–70. Available from: http://dx.doi.org/10.1016/j.bandc.2008.08.023
26 Knotht IS, Lippé S. Event-related potential alterations in fragile X syndrome. Vol. 6, Frontiers in Human Neuroscience. 2012. p. 1–17.
27 Shafer VL, Morr ML, Kreuzer JA, Kurtzberg D. Maturation of mismatch negativity in school-age children. Ear Hear. 2000;21(3):242–51.
28 Blokland A, Prickaerts J, Van Duinen M, Sambeth A. The use of EEG parameters as predictors of drug effects on cognition. Eur J Pharmacol [Internet]. 2015;759:163–8. Available from: http://dx.doi.org/10.1016/j.ejphar.2015.03.031
29 Gruwez A, Bruyneel AV, Bruyneel M. The validity of two commercially-available sleep trackers and actigraphy for assessment of sleep parameters in obstructive sleep apnea patients. PLOS One. 2019;14(1):1–11.
30 Krassowski E, Plante E. IQ variability in children with SLI: Implications for use of cognitive referencing in determining SLI. J Commun Disord. 1997;30(1):1–9.
31 Schuerger JM, Witt AC. The Temporal Stability Of Individually Tested Intelligence. J Clin Psychol. 1989;45(2):294–302.
32 De Visser SJ, Van Der Post JP, De Waal PP, Cornet F, Cohen AF, Van Gerven JMA. Biomarkers for the effects of benzodiazepines in healthy volunteers. Br J Clin Pharmacol. 2003;55(1):39–50.
33 Santen GWE, Cohen AF. Rare disease specialists and clinical pharmacologists unite: Increase collection of longitudinal data! Br J Clin Pharmacol. 2019;

PART IV

NON-INVASIVE PHARMACOKINETICS

CHAPTER 12

Theoretical performance of non-linear mixed effect models incorporating saliva as alternative sampling matrix for therapeutic drug monitoring in pediatrics: a simulation study

Ther Drug Monit. 2021 Aug 1;43(4):546-554. doi: 10.1097/ftd.0000000000000904

MD Kruizinga,[1,2,3] FE Stuurman,[1,3] GJA Driessen,[2,4] AF Cohen,[1,3] KR Bergmann,[1] MJ van Esdonk[1]

1 Centre for Human Drug Research, Leiden, the Netherlands
2 Juliana Children's Hospital, Haga Teaching Hospital, The Hague, the Netherlands
3 Leiden University Medical Centre, Leiden, the Netherlands
4 Department of pediatrics, Maastricht University Medical Centre, Maastricht, the Netherlands

Abstract

BACKGROUND Historically, pharmacokinetic studies and therapeutic drug monitoring (TDM) have relied on plasma as sampling matrix. Non-invasive sampling matrices, such as saliva, could reduce the burden for pediatric patients. The variable plasma-saliva relationship can be quantified using population PK models (non-linear mixed effect models (NLMEM)). However, criteria regarding acceptable levels of variability in such models are unclear. This simulation study aimed to propose a saliva TDM evaluation framework and to evaluate model requirements in the context of TDM, using gentamicin and lamotrigine as model compounds.

METHODS Two population pharmacokinetic models for gentamicin in neonates and lamotrigine in pediatrics were extended with a saliva compartment which included a delay constant (K_{SALIVA}), a saliva:plasma ratio and between-subject-variability (BSV) on both parameters. Subjects were simulated with a realistic covariate distribution. Bayesian maximum a posteriori TDM was applied to assess the performance of an increasing number of TDM saliva samples, varying levels of BSV and varying levels of residual variability. Saliva TDM performance was compared to plasma TDM performance. The framework was applied to a known voriconazole saliva model as case study.

RESULTS TDM with saliva resulted in higher target attainment compared to no TDM, and a residual proportional error <25% on saliva observations led to saliva TDM performance comparable to plasma TDM. BSV on K_{SALIVA} did not impact performance, whereas increasing BSV on saliva:plasma ratio >25% for gentamicin and >50% for lamotrigine caused a lower performance. Simulated target attainment for voriconazole saliva TDM was > 90%.

CONCLUSION Saliva as alternative matrix for non-invasive TDM is possible using NLMEM combined with Bayesian optimization. This paper provides a workflow to explore TDM performance for compounds measured in saliva and can be used as evaluation during model building.

Introduction

Historically, pharmacokinetic (PK) studies and therapeutic drug monitoring (TDM) have relied on plasma as the primary sampling matrix.[1] Plasma samples are reliable, reproducible, well-known and easy to obtain from most patients, but pose a hurdle for PK studies and TDM in vulnerable populations such as children and subjects with intellectual disability.[2] Although plasma sampling is possible in these populations when necessary, for example for gentamicin TDM or PK in pivotal studies in rare diseases, the burden is high, which leads to lower recruitment rates.[3] Therefore, alternative non-invasive sampling matrices, such as saliva, would be highly beneficial. Besides the non-invasive nature of saliva as a sampling matrix, advantages are the possibilities of obtaining multiple samples over time and sampling by subjects themselves in a home-setting.

An example of a compound where non-invasive TDM could provide added value is gentamicin. Gentamicin is one of the medications that is most often prescribed to neonates and has a narrow therapeutic range with risks of oto- and nephrotoxicity.[4] Additionally, implementation of non-invasive TDM for several anti-epileptic drugs (AEDS), such as lamotrigine, could have added value as well.[5] Patients with epilepsy often use multiple AEDS, and unpredictable drug-drug interactions between AEDS means TDM can increase the proportion of patients with plasma concentrations within the target range. Currently, TDM is usually performed using commercial software, such as MwPharm or InsightRx, which estimate individual PK parameters using Bayesian methods, incorporating available information about a drugs' population PK, individual patient variables, and measured plasma concentrations.[6]

TDM with saliva samples is relatively straightforward in the case of a low variability and a constant ratio between the plasma and saliva concentrations, for example in the case of morphine or fluconazole.[7] If a non-linear relationship is present due to a delay or a non-linear penetration or when there are multiple sources of inter-individual variability, concentrations in saliva are more difficult to interpret. The development of a non-linear mixed effects population PK model (NLMEM) can solve this problem.[8,9] These models can be used to characterize the distribution kinetics of the drug in plasma and correlate this with the distribution in saliva, and to identify linear or non-linear relationships, including rate constants or interindividual variability on the saliva:plasma ratio.[10] However, there has been sparse published data about the application of this methodology for TDM with alternative sampling matrices.[11,12] Furthermore, not all population PK models are suitable for use

in TDM. For example, if saliva concentrations are associated with multiple and high levels of variability, estimation of individual plasma concentrations based on saliva might not be possible. Therefore, it is important to evaluate the TDM performance of candidate models.

The aim of this study was to propose an evaluation framework, or blueprint, to evaluate the TDM performance of existing population PK models and to evaluate the requirements which a population PK model with an additional saliva compartment must fulfill to achieve adequate performance for the purpose of TDM. To this end, we evaluate two existing literature models describing the pharmacokinetics of gentamicin and lamotrigine in a pediatric population with an additional hypothetical, theoretical, saliva compartment. A combination of different levels of variability in the models was introduced and the performance of saliva TDM was compared to the standard of care using plasma samples. A case study of a recently developed voriconazole population PK model that described the relationship between plasma and saliva concentrations was included to demonstrate the applicability of the framework in practice.

Materials and methods

Plasma population PK models

Gentamicin and lamotrigine were chosen in this study for their frequency of use, known TDM applications and availability of existing population PK models.[13] The gentamicin population PK model from Fuchs *et al.* was used, which is based on data from 1449 neonates.[14] It concerns a 2-compartment model with between-subject variability (BSV) on clearance (CL, 28%) and central distribution volume (VC, 18%). Additionally, the covariates weight, postnatal age, and gestational age (GA) are included on clearance and weight and GA is included on VC. The proportional residual error of the model is 18%, with an additive residual error of 0.1 mg/L. The lamotrigine pediatric population model of He *et al.* is a 1-compartment model with BSV (26%) and comedication-based covariates on CL.[15] The proportional residual error of the model is 21% and no additive residual error was identified. The model was based on steady-state concentrations, and the absorption constant was fixed at 1 h^{-1} by the authors.

R version 4.02[16] and the mrgsolve[17] package were used for simulation. During simulation, ω^2 was defined as ln(BSV^2+1) and σ^2 was defined as the square of the proportional residual error expressed as percentage.

Hypothetical saliva compartment

A hypothetical saliva compartment was included in both models to mimic the distribution from plasma to saliva. Since the expected absolute amount of drug in saliva is too low to influence the plasma PK, no mass transfer or reabsorption from the gastrointestinal tract was included in this model. A schematic representation of the model is displayed in *Figure 1A* and the ordinary differential equation of the saliva compartment is presented below.

Equation 1: $dSaliva/dt = C_{PLASMA} * k_{SALIVA} - Saliva * k_{SALIVA}$

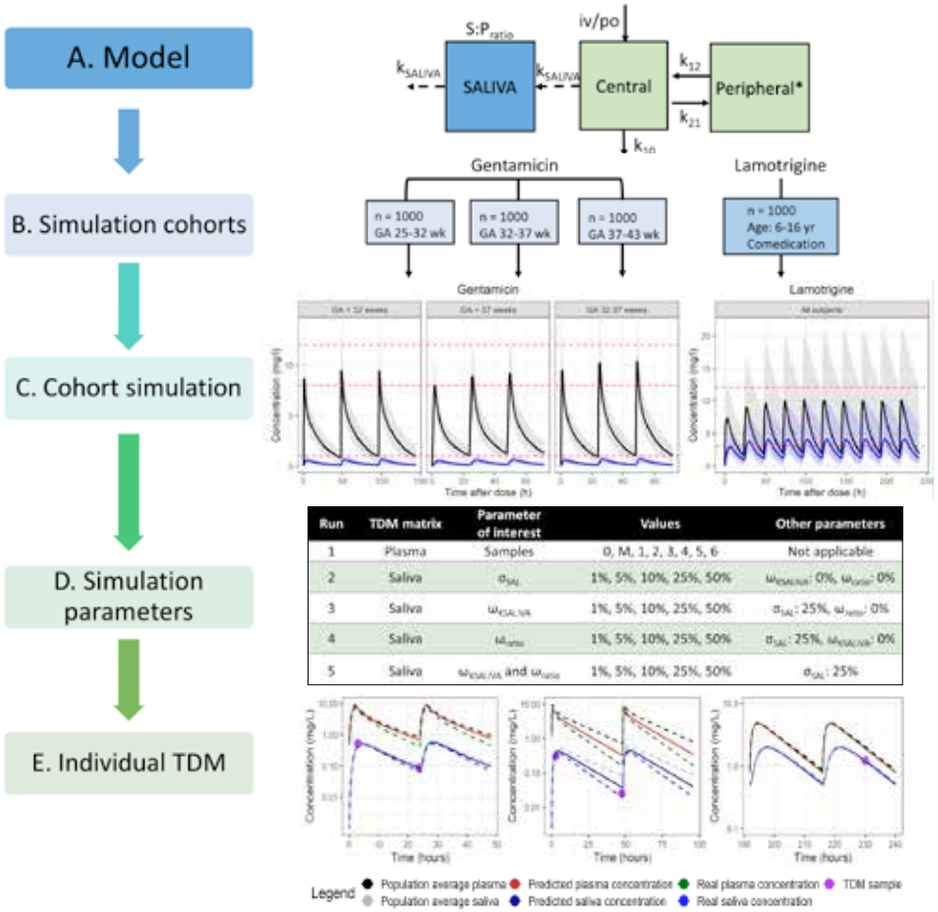
Equation 2: $C_{SALIVA} = (Saliva * saliva{:}plasma\ ratio) * (1+\varepsilon^2)$

To describe the saliva concentrations, five additional parameters were implemented. First, a parameter representing the *saliva:plasma ratio* was included and fixed at 0.1 for gentamicin (based on unpublished clinical trial data) and 0.5 for lamotrigine.[13,18] Second, a constant k_{SALIVA}, representing delayed penetration to the saliva compartment was added and fixed at 0.4h^{-1} in order to mimic a moderate delay. BSV on the ratio (log-normal distributed with mean 0 and variance ω_{RATIO}^2) and the k_{SALIVA} (log-normal distributed with mean 0 and variance $\omega_{KSALIVA}^2$) was included to mimic the inter-individual variability present in the penetration to saliva. A proportional residual error (ε^2, normal distributed with mean 0 and variance σ^2_{SAL}) was included on the saliva observations.

Simulation populations

Four fictional simulation populations were prepared based on normative data for covariates and dosing guidelines for each population (*Figure 1B*). For gentamicin, simulation populations were divided in three groups of 1000 neonates each based on gestational age and corresponding dosing guidelines. Group 1 consisted of neonates with a gestational age (GA) < 32 weeks (dose 5mg/kg/48h iv in 30 minutes), group 2 included subjects with a GA 32-37 weeks (dose 5mg/kg/36h iv) and group 3 included subjects with a GA > 37 weeks (dose 4mg/kg/24h iv). For the lamotrigine simulation, a group consisting of 1000 subjects (dose 10mg/kg, initial maximum of 200mg po) was used with 100 subjects per age year from 6 through 16 years. Comedication was set at valproic acid, carbamazepine, phenobarbital, or none for 250 subjects each. For all cohorts, a uniform distribution of weights was simulated within the 10TH and 90TH percentile of normative data.[19,20]

Figure 1. Visual clarification of the analysis. A. Gentamicin model schematic, green represents the published model, blue represents the hypothetical addition of a saliva compartment with two additional parameters (k_{SALIVA} and S:P ratio). * Gentamicin model only. B. Simulation cohorts were developed with realistic distribution of covariates (GA: Gestational age). C: Simulations performed with model and simulation cohorts. Black line and shaded area represent population median and 80% prediction interval in plasma. Blue line and shaded area's represent population median and 80% prediction interval in saliva. Red lines indicate the target ranges. D. Schematic view of simulation runs and parameter of interest within each simulation run. E. Examples of individual TDM (left: gentamicin, middle: gentamicin, right: lamotrigine) using Bayesian MAP methodology. Purple dots represent simulated TDM samples. Dotted gray and black lines represent the population average concentration-time profile for saliva and plasma, respectively. The dark blue and red lines represent the predicted saliva and plasma concentration-time profile, which is based on the simulated TDM samples. The dotted light blue and green lines represent the 'true' concentration-time profile of these individuals. Ideally, the predicted and 'true' concentration-time profiles overlap completely. Residual or between-subject variability leads to predictions closer to the population average and further from the true concentration-time profile.



TDM sampling schedule and procedure

Population predictions and the 80% prediction interval of plasma- and saliva concentrations of the models are visualized in *Figure 1C*. For the simulation of realistic TDM scenario's, samples were simulated at different timepoints. In the case of one sample, an intermediate (14h post dose) sample was simulated. In the case of two samples, peak (1h post-dose for gentamicin in plasma and 3h post-dose for gentamicin saliva and lamotrigine samples) and trough (0.5h before next dose) samples were simulated after the first dose and compared to an intermediate (14h post-dose) level. Additionally, the combination of the three samples was evaluated. Finally, the effect of additional samples (at 7h, 7h+18h, and at 0.5h+7h+18h, to reach 4, 5 and 6 samples per subject, respectively) was evaluated. For gentamicin, samples were obtained after dose 1. For lamotrigine, samples were obtained after dose 10, to realistically account for the outpatient nature of TDM with anti-epileptics.

Simulation runs

Simulation runs were performed for plasma- and saliva TDM. For plasma TDM, a simulation was performed for each TDM sampling schedule (1-6 samples in total). The outcome of plasma TDM was subsequently compared to the saliva TDM outcomes. For each saliva simulation run, BSV and residual error were varied to simulate different levels of variability. To explore the effect of a single level of variability, residual error was fixed at either 1%, 5%, 10%, 25% or 50%, while fixing BSV on k_{SALIVA} and saliva:plasma ratio at 0%. The BSV on k_{SALIVA} and saliva:plasma ratio or fixed at either 1%, 5%, 10%, 25% or 50% while fixing the other at 0%, and the proportional error at 25%. A residual error of this magnitude has been reported in a recent model incorporating saliva samples.[12] Finally, the combination of BSV on both k_{SALIVA} and saliva:plasma ratio simultaneously was assessed, where both were fixed at either 1%, 5%, 10%, 25% or 50%, with a proportional error of 25%. The various combinations are displayed in *Figure 1D*.

Individual TDM with Bayesian optimization

Bayesian maximum a posteriori (MAP) optimization was used to estimate the most likely CL, VD, k_{SALIVA} and saliva:plasma for each subject based on the obtained plasma- or saliva

samples (*Figure 1E*). Then, based on the estimated CL and VD, the peak and trough concentrations were simulated for each subject, who then entered a basic decision rule optimizing the concentrations by varying the dose and the dosing interval. For gentamicin, the goal was to obtain a peak concentration between 8–12 mg/L and a trough concentration < 1.0 mg/L.[21] Target ranges in the decision rule were deliberately set stricter (peak 9–11 mg/L, trough < 0.8) to account for residual error in the estimations. Optimal plasma concentrations for lamotrigine are a source of controversy.[22,23] For this analysis, trough concentrations between 3–14 mg/L for lamotrigine were targeted, with the decision rule set to optimize between 4–13 mg/L. The optimized dose for gentamicin and lamotrigine was given at dose 3 and 12, respectively, to account for the delay between analyzing the sample and adjusting the dose. For each individual subject, the true and predicted peak- and trough concentrations of gentamicin after the 3RD dose were simulated, as well as the trough concentration of lamotrigine after the 20TH dose to account for the inpatient- and outpatient nature of the TDM process, respectively. This allowed for the assessment of the result of the TDM process for each subject and to assess the accuracy of the model prediction.

Outcome

The proportion of subjects with target attainment after the final dose (plasma concentration within the target levels) was the primary outcome of each simulation run. These outcomes were compared to the proportion of subjects reaching target attainment after following the clinical dosing guidelines during the simulation without dose adjustment, and to the proportion of subjects reaching target attainment after dose adjustment solely based on covariates included in the population PK models, such as weight and comedication.

Case study Voriconazole

To assess whether the simulations with hypothetical saliva parameters are valid, we applied the methodology described above in a case study with the Voriconazole model of Kim *et al.*, which describes the relationship between plasma and saliva based on aggregated data from the literature.[12] The model concerns a one compartment model with a saliva:plasma ratio of 0.5, BSV on CL (36.9%) and a proportional residual error of 27% on saliva concentrations and 24% on plasma concentrations, without any covariates. Thousand simulated subjects received a 4mg/kg twice daily dose. Target attainment was defined as a trough concentration between 1–4 mg/L, and the dose decision rule was programmed to optimize the dose to reach a trough between 1.5 and 3.5 mg/L. The proportion of subjects reaching target attainment was estimated for 0–3 plasma or saliva samples, using the same timepoints as in the analyses above.

Results

Performance of plasma TDM

To be able to compare the performance of saliva TDM to plasma TDM, the first simulation runs were performed with plasma sampling with the models as described in the literature. The proportion of subjects with plasma levels within the target range is displayed in *Figure 2A* (gentamicin) and *Figure 2B* (lamotrigine). For gentamicin, only 41% of subjects achieve plasma concentrations within the target range in our simulation if no TDM was applied and standard dosing guidelines were followed, whereas for lamotrigine this was the case in only 35% of subjects. Optimizing the dose based on the population PK model and covariates weight, age and comedication led to a 72% target attainment for both compounds. Obtaining a single plasma sample 14h post-dose led to 87% and 93% patients achieving sufficient plasma levels for gentamicin and lamotrigine, respectively. This proportion increased slightly for each additional sample taken. These numerical results will be used for comparison of the predictive performance of saliva.

Saliva TDM with increasing proportional residual error (σ_{SAL})

Figure 2C (gentamicin) and *Figure 2D* (lamotrigine) display the proportion of subjects with plasma levels within the target range after saliva TDM with a varying proportional residual error. In this analysis, no BSV on the saliva:plasma ratio and k_{SALIVA} was included. For gentamicin, assuming a TDM regimen with a peak and trough sample, 99% of subjects would reach plasma levels within the target range if σ_{SAL} was 1%, which decreases to 82% when σ_{SAL} is 25% and to 76% if σ_{SAL} is 50%. In the case of a residual error of 25% and 50%, each additional saliva sample taken led to an additional small percentage (0–4%) of subjects reaching plasma levels within the target range. For lamotrigine, similar effects of the

residual error on target attainment were observed, with 98–100% of subjects achieving target attainment with 2 saliva samples and a residual error of 1%–10%. With a 25% and 50% residual error, 91% and 84% achieved plasma concentrations within range.

Figure 2. Proportion of subjects with target attainment with varying residual error (σ2) but without BSV on saliva parameters. A/B: Heat map displaying the proportion of subjects who reach target attainment of gentamicin (A) and lamotrigine (B) after plasma TDM using an increasing number of plasma samples. C/D: Heat map displaying the proportion of subjects who reach target attainment of gentamicin (C) and lamotrigine (D) after saliva TDM using an increasing number of saliva samples and assuming an increasing residual error on each observation sample. No BSV on saliva parameters was incorporated in this analysis. Timepoints where samples were simulated: 0 : no samples, standard dosing according to guidelines; M: no samples, dosing optimized according to population model and individual covariates; 1: sample 14h post-dose; 2: peak sample at 3h post-dose for gentamicin saliva and lamotrigine samples) and trough sample 0.5h before next dose; 3: samples at peak, trough and 14h post-dose; 4: samples at peak, trough, 14h post-dose and 7h post-dose; 5: samples at peak, trough, 14h post-dose, 7h post-dose and 18h post-dose; 6: samples at peak, trough, 14h post-dose, 7h post-dose, 18h post-dose and 0.5h post-dose.

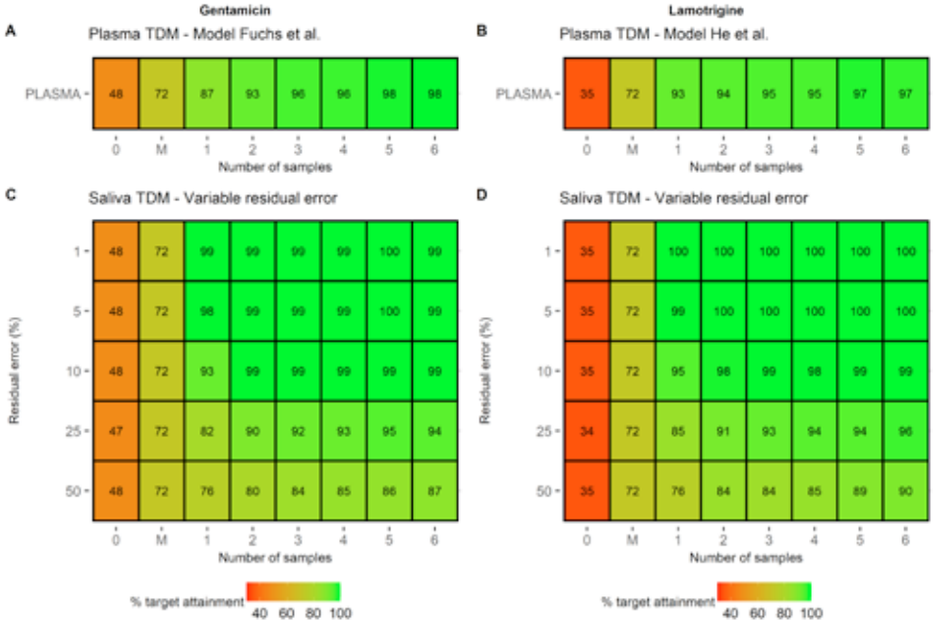
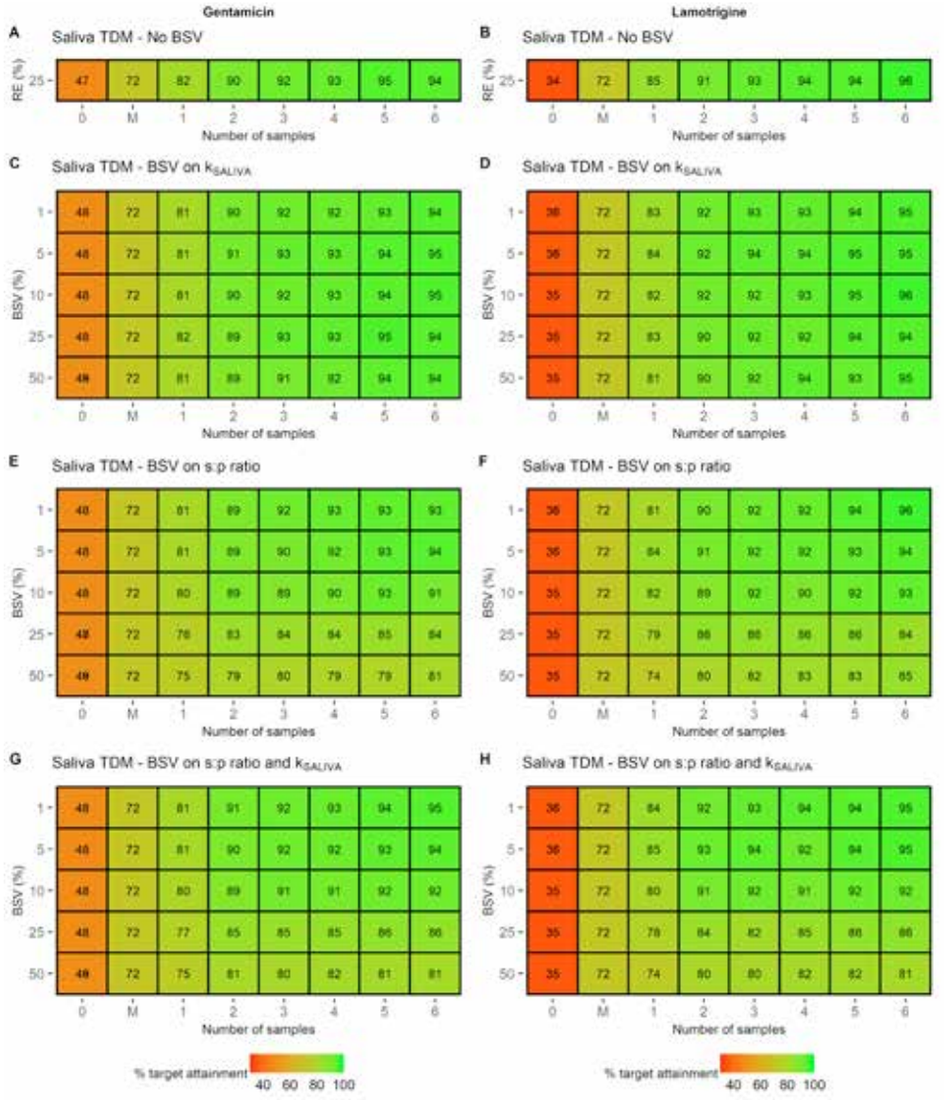


Figure 3. Proportion of subjects with target attainment with varying levels of between-subject variability. A/B. Estimated proportion of subjects within the target concentration range after saliva TDM without BSV. Here, residual error is 25% and no BSV is included on k_{SALIVA} or saliva:plasma ratio. Panel should be used for reference for other panels. C/D: Estimated proportion of subjects within the target concentration when a varying BSV on k_{SALIVA} is incorporated during simulations. E/F: Estimated proportion of subjects within the target concentration when a varying BSV on saliva:plasma ratio is used during simulations. G/H: Estimated proportion of subjects within the target concentration when a varying BSV on both k_{SALIVA} and saliva:plasma ratio is used during simulations (e.g., both 1%, both 5%, etc.).

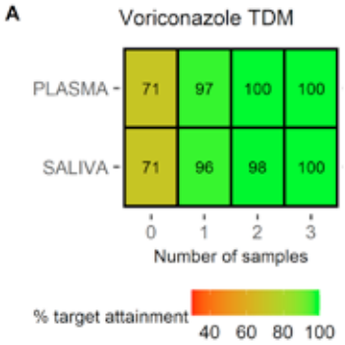


Saliva TDM with increasing BSV on saliva:plasma ratio (ω_{ratio}) and k_{saliva} (ωk_{saliva})

Fitting a model that does not account for BSV in the saliva:plasma ratio and delayed penetration into the saliva compartment may lead to a high amount of unexplained variability, and as a result, lower TDM performance. On the other hand, incorporating BSV, for example on K_{SALIVA}, in a model poses an additional hurdle for the Bayesian optimization to consider, and could result in lower predictive performance of saliva samples to adequately isolate the CL and VD for each subject based on limited samples.

Figure 3 displays the effects of including a varying ω_{RATIO} and $\omega_{KSALIVA}$ on TDM performance. *Figure 3A-B* repeats the outcome of a simulation with a residual error of 25% without additional BSV. *Figure 3C-H* displays the outcome of simulations with varying levels of BSV. Increasing $\omega_{KSALIVA}$ did not lead to a significant reduction of subjects achieving adequate plasma levels for either gentamicin or lamotrigine. Similarly, increasing the BSV on the S:P ratio to up to 10% did not cause a reduction in performance either. However, a ω_{RATIO} of 25% or 50% led to a 3-6% and 5-11% reduction in the proportion subjects achieving target attainment, depending on the number of samples. For lamotrigine, target trough attainment was achieved 2-6% (BSV 25%) and 6-11% (BSV 50%) less in those cases. The combination of both ω_{RATIO} and $\omega_{KSALIVA}$ led to similar performance compared to the simulation run that only included ω_{RATIO} for both gentamicin and lamotrigine.

Fig 4. Case study voriconazole - Proportion of subjects with plasma concentrations within target ranges with varying between-subject variability. Heat map with proportion of subjects with target attainment after TDM with 0-6 saliva- or plasma samples. Simulation based on the voriconazole model of Kim et. al.



Case study with voriconazole saliva model

The estimated proportion of subjects with target attainment for the voriconazole saliva model of Kim *et al.* is displayed in *Figure 4* for several scenarios. Using the 4 mg/kg dosing guidelines, 70% of simulated subjects reached target attainment. Applying TDM with 1 sample led to 97 and 96% of subjects with adequate trough concentrations for plasma and saliva, respectively. Obtaining 3 samples led to 100% target attainment for both plasma and saliva TDM.

Discussion

There have been many studies investigating the penetration of drugs into saliva and using saliva sampling for TDM has been recurrent subject of discussion in the field.[7,13] However, most papers investigating the potential of saliva sampling have focused on a constant ratio between saliva and plasma concentrations, which can vary over time. Non-linear mixed effect models are more flexible, can include covariates and BSV on parameters and, as a result, are better for prediction and estimation of an individual's PK profile. This analysis investigated the effect of several levels of variability structures for two commonly used drugs and this approach can assist pharmacometricians and clinicians when developing novel TDM techniques while assessing the use of saliva in clinical practice. In general, these results show that even with moderate to high levels of variability in the saliva-plasma relationship, TDM with saliva samples is feasible and leads to significantly higher target attainment compared to 'one size fits all' dosing guidelines or even model-based individualized dosing.

When estimating the CL and VD of each individual subject with either saliva or plasma samples, increasing the number of samples caused a small improvement in predictive capability for each additional sample. However, the largest improvement in performance was observed with 2 samples compared to 1 sample. Increasing the proportional error associated with the model increased the uncertainty around each individual estimation. As a result, the difference between the estimated- and true CL and VD was larger for each stepwise increase in the proportional error. If there are variables where incorporating BSV improves model fit, including this in the model will reduce the proportional error, which in turn may lead to increased performance of the model in the context of TDM. Our data demonstrates that including BSV on the delay constant towards the saliva compartment

($\omega_{KSALIVA}$) does not impact TDM performance in the models investigated here, independently of the size of the variability. It was expected this variable would cause uncertainty in the estimation of other parameters, especially VD. We hypothesize the correlation between CL and VD in the gentamicin model of Fuchs *et al.* and the lack of BSV on VD in the lamotrigine model allowed for this stable performance. BSV on the saliva:plasma ratio (ω_{RATIO}) only impacted TDM performance when the variability exceeded 25%.

Of course, the acceptable level of the proportion of subjects reaching adequate plasma levels differs between each compound. Gentamicin has a narrow therapeutic range with potentially debilitating adverse events, which necessitates to aim for accurate TDM, especially when treating for extended periods.[24] On the other hand, lamotrigine has a wider therapeutic range with possible extreme effects of comedication on plasma concentration. In such cases, a larger prediction error could be accepted on the condition that such outliers would be identified reliably, and saliva samples can be used to identify these.

While this paper focuses on saliva, our methods can be applied to other alternative sampling matrices as well, such as dried blood spots, sweat, lacrimal fluid or others.[25–28] The potential applications of alternative sampling matrices are numerous. First, the fact that they are non-invasive allows for widespread application in vulnerable populations that are usually not represented in studies that determine general pharmacokinetic properties, such as children or the mentally impaired. Determining whether these patients obtain adequate plasma concentrations could improve their quality of care. Second, as demonstrated in this paper, increasing the amount of TDM samples obtained from patients leads to better estimations and more patients with plasma concentrations within the target range, and non-invasive sampling matrices make repeated sampling more accessible. Third, non-invasive sampling matrices usually do not require extensive training or supervision to perform. As a result, samples for the pediatric pharmacokinetic clinical trial of the future could be taken in a home-setting, stored in a domestic freezer, and eventually be retrieved by courier. This complements a general trend in clinical trials and -care which moves away from the clinic towards the home.[29]

The purpose of this paper was to propose an evaluation framework and to determine the influence of several model parameters on the performance of TDM with saliva. The saliva:plasma relationship is dependent on several factors, including polarity, molecule size and protein binding capacity.[30] In the end, compounds that reach saliva in too unpredictable or variable ways will be unsuitable for TDM on their own, as estimations based on saliva samples in such models would be driven purely by population effects, such as was observed with high levels of variability of 50% in the current analysis. In NLMEM, this can be quantified by the residual error and BSV. The models that were used during simulation in this study used fictional saliva compartments, and it is unclear whether saliva TDM is viable for the two compounds in the absence of clinical data about the saliva:plasma relationship. The included parameters model a delayed penetration of drug towards the saliva compartment with a given saliva:plasma ratio. This is likely a simplification of the underlying physiology, but there currently is little salivary data available for either gentamicin or lamotrigine. Even so, it is likely that this resembles true physiology more closely compared to the constant saliva:plasma ratio without a delay that is currently employed in most salivary PK studies. In the future, when more data becomes available, modelling may reveal other model structures or differing input and output constants for the saliva compartment. The proposed framework can then be applied on this newly developed model. Although a limitation of the current analysis is that it is not based on real salivary data, it allows to explore the impact of several levels of variability and to determine thresholds of variability that severely impact TDM performance when exceeded. The thresholds of variability around 25% found during simulation appear viable targets during model building. The potential of saliva TDM was confirmed in the simulations of a saliva model of voriconazole12, with target attainment > 90% for all saliva simulation scenarios and highly comparable target attainment compared with plasma TDM. In the future, additional simulations of alternative sampling timepoints or different covariate distributions may lead to different proportions of subjects that achieve target attainment. However, since a large cohort was used that was identical during each simulation run, relative differences across simulations in the proportion of subject with target attainment should remain constant. Future studies should focus on confirming these simulations with real data. Furthermore, advanced modelling techniques such as physiologically based pharmacokinetic modelling (PBPK) of the penetration of the salivary compartment may lead to low levels of variability. Additionally, a hybrid approach combining both saliva- and plasma samples for TDM could also be investigated. For example, a first iteration of TDM could use plasma sampling to determine baseline CL and V, after which outpatient saliva samples could confirm a holding steady-state concentration or provide a warning sign indicative of a need for dose adjustment.

Conclusion

Saliva as alternative matrix for non-invasive TDM in pediatrics may be possible using non-linear mixed effects models combined with Bayesian optimization. Gentamicin or lamotrigine models with low- to moderate levels of variability below 50% on saliva observations achieve TDM performance comparable to TDM with plasma samples according to the simulations presented here. Additionally, this paper provides a workflow to explore the added value of TDM for compounds measured in saliva or other non-invasive sampling matrices.

REFERENCES

1 Soldin OP, Soldin SJ. Review: Therapeutic drug monitoring in pediatrics. Ther Drug Monit 2002;24(1):1–8.
2 Stoltz P, Manworren RCB. Comparison of Children's Venipuncture Fear and Pain: Randomized Controlled Trial of EMLA® and J-Tip Needleless Injection System®. J Pediatr Nurs 2017;37:91–96.
3 Pasquali SK, Lam WK, Chiswell K, Kemper AR, Li JS. Status of the pediatric clinical trials enterprise: An analysis of the US ClinicalTrials.gov registry. Pediatrics 2012;130(5).
4 Simonsen KA, Anderson-Berry AL, Delair SF, Dele Davies H. Early-onset neonatal sepsis. Clin Microbiol Rev 2014;27(1):21–47.
5 Petre M, Strah A. Therapeutic drug monitoring of antiepileptic drugs. Farm Vestn 2015;66(1):35–41.
6 Drennan P, Doogue M, van Hal SJ, Chin P. Bayesian therapeutic drug monitoring software: past, present and future. Int J Pharmacokinet 2018;3(4):109–114.
7 Hutchinson L, Sinclair M, Reid B, Burnett K, Callan B. A descriptive systematic review of salivary therapeutic drug monitoring in neonates and infants. Br J Clin Pharmacol 2018;84(6):1089–1108.
8 Schoemaker RC, Cohen AF. Estimating impossible curves using NONMEM. Br J Clin Pharmacol 1996;42(3):283–290.
9 Jonsson EN, Karlsson MO, Wade JR. Nonlinearity detection: Advantages of nonlinear mixed-effects modeling. AAPS PharmSci 2000;2(3):1–10.
10 Bauer RJ. NONMEM Tutorial Part I: Description of Commands and Options, With Simple Examples of Population Analysis. CPT Pharmacometrics Syst Pharmacol 2019;8(8):525–537.
11 Dobson NR, Liu X, Rhein LM, Darnall RA, Corwin MJ, McEntire BL, Ward RM, James LP, Sherwin CMT, Heeren TC, *et al.* Salivary caffeine concentrations are comparable to plasma concentrations in preterm infants receiving extended caffeine therapy. Br J Clin Pharmacol 2016:754–761.
12 Kim HY, Märtson AG, Dreesen E, Spriet I, Wicha SG, McLachlan AJ, Alffenaar JW. Saliva for Precision Dosing of Antifungal Drugs: Saliva Population PK Model for Voriconazole Based on a Systematic Review. Front Pharmacol 2020;11(June).
13 Patsalos PN, Berry DJ. Therapeutic Drug Monitoring of Antiepileptic Drugs by Use of Saliva. Ther Drug Monit 2013;35(1).
14 Fuchs A, Guidi M, Giannoni E, Werner D, Buclin T, Widmer N, Csajka C. Population pharmacokinetic study of gentamicin in a large cohort of premature and term neonates. Br J Clin Pharmacol 2014;78(5):1090–1101.
15 He DK, Wang L, Lu W, Qin J, Zhang S, Li L, Zhang JM, Bao WQ, Song XQ, Liu HT. Population pharmacokinetics of lamotrigine in Chinese children with epilepsy. Acta Pharmacol Sin 2012;33(11):1417–1423.
16 R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
17 Kyle T. Baron and Marc R. Gastonguay. Simulation from ODE-Based Population PK/PD and Systems Pharmacology Models in R with mrgsolve. J Pharmacokinet Pharmacodyn 2015;42(W-23):S84–S85.
18 Cohen AF, Land GS, Breimer DD, Yuen WC, Winton C, Peck AW. Lamotrigine, a new anticonvulsant: pharmacokinetics in normal humans. Clin Pharmacol Ther 1987;42(5):535–541.
19 Visser GHA, Eilers PHC, Elferink-Stinkens PM, Merkus HMWM, Wit JM. New Dutch reference curves for birthweight by gestational age. Early Hum Dev 2009;85(12):737–744.
20 National Center for Health Statistics, Data Table of Weight-for-age Charts. [accessed 2020 Nov 20]. https://www.cdc.gov/growthcharts/html_charts/wtage.htm#males
21 National Collaborating Centre for Women's and Children's Health (UK). Antibiotics for Early-Onset Neonatal Infection: Antibiotics for the Prevention and Treatment of Early-Onset Neonatal Infection. London: RCOG Press; 2012 Aug. (NICE Clinical Guidelines,.
22 Morris RG, Black AB, Harris AL, Batty AB, Sallustio BC. Lamotrigine and therapeutic drug monitoring: Retrospective survey following the introduction of a routine service. Br J Clin Pharmacol 1998;46(6):547–551.
23 Fröscher W, Keller F, Vogt H, Krämer G. Prospective study on concentration-efficacy and concentration-toxicity: correlations with lamotrigine serum levels. Epileptic Disord 2002;4(1):49–56.
24 Donge T Van, Pfister M, Bielicki J, Csajka C, Rodieux F, Fuchs A. Quantitative Analysis of Gentamicin Exposure in Neonates and Infants Calls into Question Its Current Dosing Recommendations. Antimicrob Agents Chemother 2018;62(4):1–12.
25 Kruizinga MD, Birkhoff WAJ, Esdonk MJ, Klarenbeek NB, Cholewinski T, Nelemans T, Dröge MJ, Cohen AF, Zuiker RGJA. Pharmacokinetics of intravenous and inhaled salbutamol and tobramycin: An exploratory study to investigate the potential of exhaled breath condensate as a matrix for pharmacokinetic analysis. Br J Clin Pharmacol 2020;86(1):175–181.
26 Grumetto L, Cennamo G, Del Prete A, La Rotonda MI, Barbato F. Pharmacokinetics of cetirizine in tear fluid after a single oral dose. Clin Pharmacokinet 2002;41(7):525–531.
27 Marchei E, Farrè M, Pellegrini M, García-Algar Ó, Vall O, Pacifici R, Pichini S. Pharmacokinetics of methylphenidate in oral fluid and sweat of a pediatric subject. Forensic Sci Int 2010;196(1–3):59–63.
28 Jager NGL, Rosing H, Schellens JHM, Beijnen JH. Procedures and practices for the validation of bioanalytical methods using dried blood spots: A review. Bioanalysis 2014;6(18):2481–2514.
29 Kruizinga MD, Stuurman FE, Groeneveld GJ, Cohen AF. The Future of Clinical Trial Design : The Transition from Hard Endpoints to Value-Based Endpoints.
30 Haeckel R. Factors Influencing the Saliva/Plasma Ratio of Drugs. Ann N Y Acad Sci 1993;694(1):128–142.

CHAPTER 13

Population pharmacokinetics of clonazepam in saliva and plasma - steps towards non-invasive pharmacokinetic studies in vulnerable populations

Br J Clin Pharmacol. 2021 Nov 22. doi: 10.1111/bcp.15152.

MD Kruizinga,[1,2,3] RGJA Zuiker,[1] KR Bergmann,[1] AC Egas,[4] AF Cohen,[1,3] GWE Santen,[4] MJ van Esdonk[1]

1 Centre for Human Drug Research, Leiden, the Netherlands
2 Juliana Children's Hospital, Haga Teaching Hospital, The Hague, the Netherlands
3 Leiden University Medical Centre, Leiden, the Netherlands
4 Department of Pharmacy, University Medical Centre Utrecht, Utrecht, the Netherlands
5 Department of Clinical Genetics, Leiden University Medical Centre, Leiden, the Netherlands

Abstract

INTRODUCTION Traditional studies focusing on the relationship between pharmacokinetics (PK) and pharmacodynamics necessitate multiple blood draws which are too invasive for children or other vulnerable populations. A solution is to use non-invasive sampling matrices, such as saliva. The aim of this study was to develop a population PK model describing the relationship between plasma- and saliva clonazepam kinetics and assess whether the model can be used to determine trough plasma concentrations based on saliva samples.

METHODS Twenty healthy subjects, aged 18-30, were recruited and were administered 0.5mg or 1mg of clonazepam solution. Paired plasma- and saliva samples were obtained until 48h post-dose. A population pharmacokinetic model was developed describing the PK of clonazepam in plasma and the relationship between plasma and saliva concentrations. Bayesian maximum a posteriori (MAP) optimization was applied to estimate the predictive accuracy of the model.

RESULTS A two-compartment distribution model best characterized clonazepam plasma kinetics with a mixture component on the absorption rate constants. Oral administration of the clonazepam solution caused contamination of the saliva compartment during the first 4 hours post-dose, after which the concentrations were driven by the plasma concentrations. Simulations demonstrated that the lower and upper limits of agreements between true and predicted plasma concentrations were -28-36% with 1 saliva sample. Increasing the number of saliva samples improved these limits to-18-17%.

CONCLUSION The developed model describes the salivary-and plasma kinetics of clonazepam and can predict steady-state trough plasma concentrations based on saliva concentrations through Bayesian MAP optimization with acceptable accuracy.

Introduction

Traditional studies focusing on pharmacokinetics in plasma necessitate multiple blood samples and therefore may become too invasive for a pediatric population. However, the determination of pharmacokinetic-pharmacodynamic (PKPD) relationships between a drug and an endpoint is an important method for evaluation of drug effects and dose optimization. A possible solution is to determine drug concentrations in sampling matrices that can be collected in a non-invasive manner, such as saliva. Salivary kinetics have been determined in a wide range of compounds, and applications in the field of therapeutic drug monitoring (TDM) have been explored.[1-3] However, analyses have been mainly performed by calculating the plasma concentrations based on a constant ratio over time with the saliva concentrations. Nonlinear mixed effects models have potential advantages, such as accounting for a changing ratio over time and different sources of variability, in the evaluation of the population kinetics in plasma and saliva, and this approach could potentially also be used to improve the prediction of plasma concentrations when only saliva concentrations are available.

Clonazepam is a GABA-A positive allosteric modulator used to treat a range of clinical conditions, such as epilepsy, panic disorder, depression, and bipolar disease.[4,5] Although clonazepam has been prescribed less often in recent years[6], the drug is a candidate for rediscovery in other conditions. For example, pre-clinical evidence shows possible efficacy of clonazepam in ARID1B-related intellectual disability (ID).[7] Conducting clinical trials in a (pediatric) intellectually disabled population is challenging, as trial designs must be unobtrusive to motivate as many patients as possible for participation.[8,9] Inclusion of non-invasive pharmacokinetic assessments may help in this respect.

The aims of this study were to study the population kinetics of clonazepam, to determine the relationship between salivary- and plasma clonazepam concentrations and to investigate the performance of a population PK model describing this relationship to determine trough plasma concentrations based on saliva samples in patients treated with clonazepam.

Materials and Methods

Location and ethics

This study was conducted in preparation of a study researching the therapeutic effects of clonazepam on patients with ARID1B-related ID[7,10] at the Centre for Human Drug Research in Leiden, the Netherlands from June 2020 until July 2020. Ethical approval was obtained from the Beoordeling Ethiek Biomedisch Onderzoek Foundation Review Board (Assen, the Netherlands) prior to initiation of the study. The study was conducted in compliance with the Dutch Act on Medical Research Involving Human Subjects and Good Clinical Practice. Informed consent was obtained prior to study-mandated procedures.

Study design and sample collection

This was an open-label, single dose study in 20 healthy subjects aged 18-30. Exclusion criteria were history of disease- or use of medications that might interfere with saliva production, such as opiates and anticholinergics. Subjects were asked to refrain from alcohol- and caffeine use for 24 hours prior to drug administration until the end of the study, and from any nutrients with CYP-modulating activity for three days prior to drug administration. Subjects were administered a single dose of 0.5 mg (n=10), or 1.0 mg (n=10) clonazepam solution (Rivotril®) dissolved in lemonade, after which paired plasma- and saliva samples were taken at 0.5h, 1h, 2h, 4h, 6h, 8h, 24h and 48h post dose. Subjects thoroughly rinsed their mouth with water 10 minutes prior to saliva sampling. Saliva samples were obtained using the SalivaBio Infant Swab (Salimetrics, Carlsbad, CA, USA) according to the manufacturer's instructions.

Bioanalytical assay

Clonazepam (1 mg/mL, provided by Duchefa Farma (Haarlem, The Netherlands)), clonazepam-d4 (0.1 mg/mL, provided by LGC Standards (Luckenwalde, Germany)), methanol and acetonitril (both provided by Merck BV (Darmstadt, Germany)) were obtained. Assay validation was performed in accordance with European Medicines Agency (EMA) guidelines.[11] From the clonazepam solution, standards were prepared in saliva at the concentrations of 0.1, 0.5, 2.5, 5, 10 and 20 µg/L. The internal standard clonazepam-d4 was diluted with methanol to a final concentration of 30 µg/L. The lower limit of quantification (LOQ), low, medium, and high reference samples were prepared in saliva with a concentration of resp. 0.1, 1, 5 and 15 µg/L. For the measurement of the plasma samples Recipe ClinChek calibrators were used and controls with a linear range of 2 - 72.3 µg/L and the LLQ, LOW, MED and HIGH with a concentration of resp. 2, 5.25, 14.8 and 48.1 µg/L.

Sample preparation of saliva samples was performed by diluting each aliquot of 20 µL saliva with 20 µL internal standard solution in eppendorf cups, which were vortexed for 1 min and centrifuged at 13000 rpm for 5 min. The extract was transferred in vials with insert. For plasma samples, each aliquot of 50 µL plasma was diluted with 50 µL internal standard solution and 150 µl acetonitril in eppendorf cups. The eppendorf cups were vortexed for 1 min and centrifuged at 13000 rpm for 5 min. The extract was transferred in a vial with 150 µl water.

Analysis was performed via Liquid chromatography-mass spectrometry (LC-MS). Extracts (2 µL) were injected onto a Thermo Scientific Hypersil GOLD C18 column, with methanolic mobile phase gradient elution. Clonazepam was detected with a Thermo Scientific TSQ Quantiva triple quadrupole mass spectrometer with positive ionization. Ions monitored in the selected reaction monitoring mode were m/z 316-270 for clonazepam (at 2.97 min) and m/z 320-274 for clonazepam-d4 (at 2.96 min).

Pharmacokinetic modelling

A population PK analysis was performed with a sequential nonlinear mixed effects modelling approach using NONMEM® (version 7.3).[12] Structural plasma model selection was performed by fitting both 1- and 2-compartment models to the plasma concentrations over time. An allometric scaling component (normalized around 70kg) was included on clearance and inter-compartmental clearance, with an exponent of 0.75, and on all volume of distribution parameters, with an exponent of 1, to account for weight-based influences. As only oral data was available, no bioavailability component could be estimated. However, variability on the relative bioavailability (F_{PLASMA}) could still be explored on this parameter. Inter-individual variability following a lognormal distribution on population parameters were selected by a forward inclusion procedure ($p < 0.01$) and the residual error structure was introduced as proportional and checked for appropriateness with goodness-of-fit figures.

The empirical Bayes estimates of the developed plasma model were used for the development of the saliva model. Model structure selection was driven by exploratory figures of

the data and contained a constant plasma:saliva ratio, a non-linear plasma:saliva ratio, and a first-order elimination component to account for the contamination in the saliva compartment immediately after dosing. Inter-individual variability following a lognormal distribution were selected following the same procedure as with the plasma model. Improvements in model fit were judged on a decrease in objective function value (OFV) of 6.64 ($p < 0.01$) after inclusion of 1 parameter, numerical stability as judged by the relative standard errors (RSE) of parameters and shrinkage, and evaluation of goodness-of-fit figures.

After the saliva model was developed, both models were estimated simultaneously, and model predictions were assessed via a prediction-corrected visual predictive check (PCVPC) and the individual model fit over time in both matrices.

Simulations

R version 4.0.2[13] and the mrgsolve package[14] were used to simulate the predictive capability of saliva concentrations in the context of clinical trials. A simulation cohort (n = 2000) with a uniform distribution of age between 6 and 30 and corresponding weights (10TH-90TH centile[15]) was prepared, and twice daily administration of 0.015 mg/kg (max 0.5 mg per dose) was simulated for each subject. As inter-individual variability on the relative bioavailability (F_{PLASMA}) of clonazepam might be lower in the healthy population on which the model was built, this parameter was increased to a coefficient of variation of 50% to allow for the simulation of a wider range of trough concentrations and therewith provide a more conservative simulation. Simulated saliva samples were obtained at 5h, 6h, 8h, 10h and 11.5h post-dose. The accuracy of predicting the trough concentration (C_{TROUGH}) in plasma after dose 1 (based on samples obtained after dose 1) and at steady state (based on samples obtained during steady state) that could be obtained by using 1-5 saliva samples was assessed using Bayesian maximum a posteriori (MAP) optimization and traditional linear regression. An additional scenario (0 samples) without saliva sampling was simulated to establish baseline predictive capability of the model based on population parameters and the weight of a subject, but without any TDM sampling information.

Bayesian maximum a posteriori optimization

The simulated saliva concentrations of each simulated individual were used as input for the Bayesian MAP estimation. During this process, the optimal F_{PLASMA} was estimated for each individual within the constraints provided by the several levels of residual- and between-subject variability in the population PK model, based on the information provided by the saliva samples and covariates.[16] From the individual Bayes estimates, the corresponding plasma C_{TROUGH} after dose 1 (based on samples obtained after dose 1) and at steady-state (after 240 hours of twice daily dosing, based on samples obtained after dose 20) was calculated. As the true simulated Bayes estimates were known for an individual, the estimated C_{TROUGH} were compared with the true plasma C_{TROUGH} to evaluate the predictive performance of this approach. Predictive performance was quantified for each sampling scenario using the root mean squared prediction error (RMSPE), which is a predictive error expressed in the original units (ug/L). Additionally, the average bias and limits of agreement (LOA) of the predictions, expressed as percentage, were calculated according to the methods of Bland and Altman.[17]

Linear regression

Traditionally, plasma:saliva relationships are calculated as a constant ratio or linear regression formula. In order to compare these traditional methods with Bayesian optimization, a linear relationship between the two matrices was estimated via a linear mixed model based on the plasma- and saliva samples obtained during the study. In the model, saliva concentration was considered as fixed effect and subject as random intercept. The derived equation was used to predict plasma concentrations in the simulation cohort after dose 1 based on a single saliva trough samples obtained either at 11.5h post-dose after dose 1, or at 11.5h post-dose at steady state.

Results

Of the 20 subjects included in the study, 9 subjects were male, and the average age was 22 years. Other baseline characteristics are displayed in *Table 1*. Of the 160 saliva samples taken during the study, 154 provided enough volume for analysis. All 160 plasma samples were collected successfully and none of the post-dose saliva or plasma concentrations were below the LOQ. The concentration-time profiles of clonazepam in plasma and saliva are displayed in *Figure 1.* Plasma concentrations showed variability in the C_{MAX} and t_{MAX}, with some subjects immediately reaching the C_{MAX} at the first sample (30 minutes) after dosing. Salivary concentrations were high and could not be correlated with plasma

concentrations directly post-dose, which indicates that clonazepam contamination was present in the saliva. However, the salivary concentration decreased exponentially and appeared correlated with plasma concentrations after 4 hours post-dose.

Figure 1. Individual-and mean (SD) concentration-time profiles of clonazepam in plasma and saliva.
A: Plasma concentration over time for the 0.5mg dose group (left panel) and 1.0mg dose group (right panel).
B: Saliva concentration over time for the 0.5mg dose group (left panel) and 1.0mg dose group (right panel).
Individual concentration-time profiles are displayed as light gray lines. The bold line and dots represent the mean (± SD) concentration on each timepoint. Each gray dot represents a single observation. Each grey line represents a single subject.

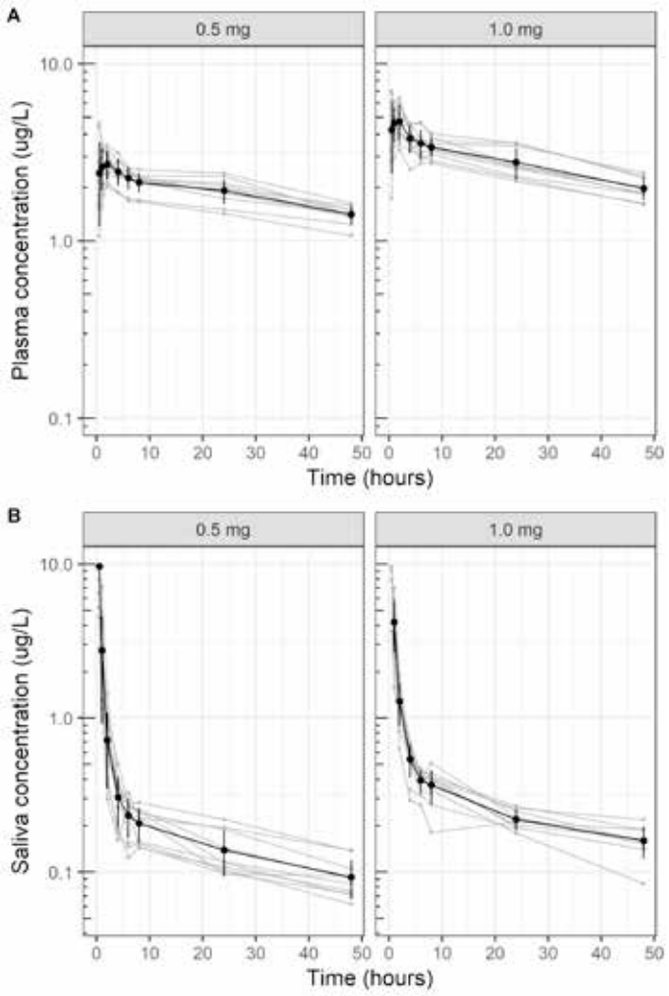


Table 1. Baseline characteristics

Parameter	All subjects (n=20)
Age (years)	22.4 (2.8)
Sex (% male)	45%
Ethnicity (% Caucasian)	100%
Weight (kg)	67.8 (8.3)
Height (cm)	175.1 (7.3)
BMI (kg/m2)	22.2 (2.4)

Data is presented as mean (SD), unless otherwise specified

Population Model

Structural plasma model development resulted in a 2-compartment model which fitted the data best, with a δOFV=-17.34 compared to a 1-compartment model. Inter-individual variability was identified on, in order of inclusion: absorption rate constant (k_A), relative bioavailability, and inter-compartmental clearance. However, the ω^2 on the k_A was high with a value of 0.66 and showed a binominal distribution. This was corrected for by inclusion of a mixture component, in which a fast absorption and a slow absorption population was identified, in which the fast absorption population had a fixed k_A of 100/h. Changing this value to 10/h or 250/h did not change the model fit. This stratification resulted in a significant improvement in model fit and reduced the ω^2 to 0.16, with 75% of subjects in the slow absorption group. No covariates for both subgroups were identified. All parameters were estimated with low RSE's and no changes were made to the proportional residual error structure.

To account for the saliva contamination, a saliva contamination compartment was added, in which a fraction of the full dose (F_{SALIVA}) remained in this compartment. The volume of this compartment was fixed to 1 mL and is represented by the following differential equation:

Equation 1: dxdt Contamination = $-k_{EL}$ * Contamination.

Data exploration showed a nonlinear relationship between the saliva:plasma ratio over the explored concentration range on data > 4 hours post-dose, after which contamination was no contributing factor anymore, in which a steady state ratio was reached at the higher concentrations (*Supplementary Figure S1*). The estimation of a saturable function on the saliva:plasma ratio improved the model fit significantly compared to a constant

saliva:plasma ratio (δOFV = -16.65). As such, the saliva:plasma ratio and saliva concentrations were represented in the model as follows:

Equation 2: Saliva:plasma ratio = $Ratio_{MAX} * C_{PLASMA} / (C_{PLASMA} + Ratio_{KM})$.

Equation 3: $C_{SALIVA} = (Contamination/0.001) + C_{PLASMA} *$ saliva:plasma ratio.

Where 0.001 is the volume of the saliva compartment in liters. Equation 3 therefore accounts for the level of contamination in the initial phase after dosing and for the non-linear saliva:plasma ratio observed in the data.

Inter-individual variability was only identified on the contamination part of the model (F_{SALIVA}, $k_{EL-SALIVA}$) and not on the saliva:plasma ratio. The saliva model gave accurate individual and population model fits over time. The final parameter estimates are displayed in *Table 2* and goodness of fit plots are displayed in *Supplementary Figure S2*. Parameters were estimated with sufficient parameter precision and moderate inter-individual variability and residual error. The plasma residual error was lower than saliva concentrations, indicating that a higher degree of unexplained variability was present in the saliva concentrations over time. The PCVPC show that the model was able to capture the median trend of the data and the level of variability in both matrices correctly.

Simulations

The concentration-time profiles of the simulation cohort are displayed in *Figure 2A*. On average, subjects achieved a median C_{TROUGH} of 2.1 ug/L after dose 1 and 13.7 ug/L at steady state. First, the estimated C_{TROUGH} after dose 1 and at steady state was estimated based on the population PK model parameters and weight of the subject only, which results in a single prediction for each weight, without taking into account any inter-individual variability (equivalent to using '0 samples' for the estimation). This scenario leads to a RMSPE of 1.39 ug/L after dose 1 and 8.8 ug/L at steady state, and an average proportional bias of -3% and -2% (95% LOA -92% - 87%), respectively (*Table 3*), meaning that there was a high level of uncertainty in the predicted individual plasma concentration. In the case of 1 saliva sample, the RMSPE was 3.56 ug/L at steady state, the proportional bias was 4% and the limits of agreement were reduced to -27% - 36% (*Figure 2B*). There was a correlation between true and predicted C_{TROUGH} in this scenario (R 0.93, p < 0.001). Increasing the number of saliva samples improved the accuracy of the prediction, as can be seen by the reduction in the RMSPE and narrowing of the LOA's (*Table 3*). For the simulation scenario with 5 saliva samples, RMSPE was 2.3 ug/L, with a proportional bias of 0% and LOA of -18% - 17%. For all scenario's applying Bayesian optimization, the true and estimated C_{TROUGH} were correlated, with correlation coefficients > 0.93 (p < 0.001, *Supplementary Figure S3 and S4*). There was an eightfold difference in RMSPE between scenario's estimating the C_{TROUGH} after dose 1 compared to scenarios at steady state, which can be explained by the increased concentration after multiple dosing. The proportional bias and LOA's were comparable for the estimation after dose 1 and at steady state.

Figure 2. Population prediction (80% prediction interval) in simulation cohort and visualization of predictive capability based on 1 saliva sample at steady state. A. Median population prediction (solid lines) and 80% prediction interval of the simulation cohort (n = 2000) in plasma (black) and saliva (blue). B. Proportional bias in the prediction (dotted line) and proportional limits of agreement (solid lines) of predicted Ctrough during steady state after Bayesian optimization based on a single saliva sample 11.5h post dose (during steady state). The x-axis displays the mean of the predicted and real Ctrough and the y-axis displayed the proportional difference between the predicted and real Ctrough. C: Pearson correlation between true and predicted plasma Ctrough of the scenario displayed in panel B. Bold black line represents the regression line, thin black line represents the line of unity, and each dot represents a simulated subject. For proportional bias plots and linear correlations of all scenarios, please refer to Supplementary Figures S3 and S4.

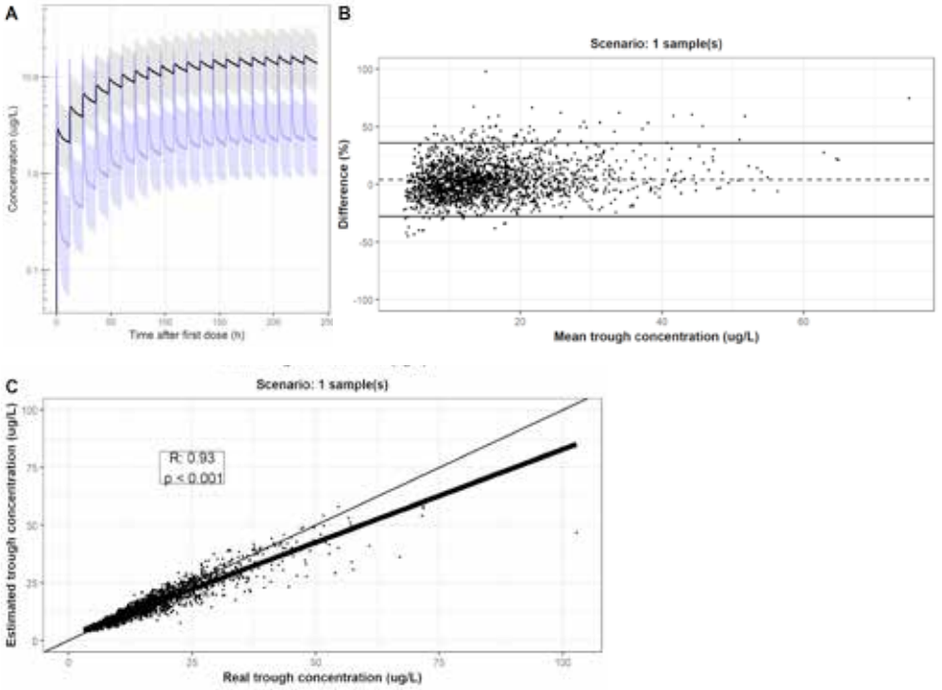


Table 2. Parameter estimates of population pharmacokinetic plasma and saliva model

	Parameter	Estimate [shrinkage %]	SE	RSE (%)
Plasma kinetics	θ k_A - slow group (h-1)	1.106	0.14	12.8
	θ Prob. slow group	0.75	0.10	13.0
	θ VD central (L)	109.5	11.5	10.4
	θ Clearance (L h-1)	2.98	0.16	5.5
	θ VD peripheral (L)	130.6	12.9	9.9
	θ Q (L h-1)	61.37	11.2	18.3
Saliva kinetics	θ F_{SALIVA} (% 1000-1)	0.033	0.005	14.5
	θ K_{EL} saliva (h-1)	1.95	0.12	6.0
	θ $Ratio_{MAX}$	0.195	0.031	16.0
	θ $Ratio_{KM}$ (ug/L)	2.581	0.74	28.5
IIV	ω^2 k_A - slow group	0.16 [2.24%]	0.065	40.8
	ω^2 Q	0.25 [0.33%]	0.12	45.7
	ω^2 F_{PLASMA}	0.026 [3.98%]	0.008	28.9
	ω^2 F_{SALIVA}	0.28 [5.30%]	0.12	42.7
	ω^2 K_{EL} saliva	0.056 [17.2%]	0.025	44.2
Residual error	σ^2 proportional plasma	0.0058 [15.9%]	0.0009	14.9
	σ^2 proportional saliva	0.057 [16.5%]	0.009	15.4

RSE: Relative standard error. IIV: interindividual variability. SE: standard error. / ω^2 and σ^2 are the variances of interindividual variability and residual variability, respectively. / TVCL = θ Clearance * $(WGT/70)^{0.75}$; TVVC = θ VD central * $(WGT/70)^{1}$; TVVP = θVD peripheral * $(WGT/70)^{1}$; TVQ = θ Q * $(WGT/70)^{0.75}$; TVRatio = $Ratio_{MAX}$ * C_{PLASMA} / (C_{PLASMA} + $Ratio_{KM}$)

Table 3. Simulation of predictive capability of using a saliva sample to determine the plasma trough concentration while varying the number of saliva samples used for the analysis.

	C_{TROUGH} after 1st dose				C_{TROUGH} steady state			
Scenario	RMSPE (ug/L)	Bias (%)	LLOA (%)	ULOA (%)	RMSPE (ug/L)	Bias (%)	LLOA (%)	ULOA (%)
Population model	1.39	-2.7	-93	87	8.8	-2.4	-92	87
1 sample, Linear regression[1]	0.57	-9.5	-54	35	4.4	5.2	-40	51
1 sample, Bayesian[1]	0.57	3.7	-28	36	3.6	4.0	-28	36
2 samples, Bayesian[2]	0.44	-0.2	-25	25	2.8	0.1	-25	25
3 samples, Bayesian[3]	0.40	-0.4	-21	21	2.5	-0.2	-21	21
4 samples, Bayesian[4]	0.39	-0.7	-20	19	2.5	-0.4	-20	19
5 samples, Bayesian[5]	0.37	-0.6	-18	17	2.3	-0.3	-18	17

Abbreviations: RMSPE: root mean squared prediction error, LLOA: lower limit of agreement (-2SD), ULOA: upper limit of agreement (+2SD). 1 Sample at 11.5h post-dose. 2 Samples at 5h and 11.5h post-dose. 3 Samples at 5h, 8h, and 11.5h post-dose. 4 Samples at 5h, 6h, 8h, and 10h post-dose. 5 Samples at 5h, 6h, 8h, 10h, and 11.5h post-dose.

Linear regression

The predictive performance of traditional (mixed effects) linear regression was assessed to determine the added value of Bayesian optimization methods. Because of the evident contamination in saliva during the first 4 hours after dosing, the correlation between saliva- and plasma concentrations was not calculated during this time window. The estimated linear regression equation from a linear mixed effects model predicting plasma concentrations from salivary concentrations during this time window was: $C_{PLASMA} = C_{SALIVA} * 5.4172 + 1.1994$ (marginal R^2 0.68). Predicted plasma concentration with the linear regression formula based on the 11.5h post-dose trough sample at steady-state was correlated with the true plasma concentration ($R = 0.91$, $p < 0.001$), but led to a larger RMSPE (4.36 ug/L) compared to Bayesian MAP, with a proportional bias of 4.3% (LOA -41%-50%, *Supplementary Figure S3-4E*).

Discussion

During this study, the pharmacokinetics of saliva- and plasma concentrations of clonazepam was characterized by a nonlinear mixed effects model. Furthermore, the potential use of salivary concentrations for prediction on plasma pharmacokinetics was studied. The developed model was able to predict trough concentrations at steady state with a RMSPE as low as 2.3 ug/L, with an 95% limit of agreement of -18-17%. The model will be used in a future non-invasive study investigating the effects of clonazepam on children with ARID1B-related ID and can be employed in the case of salivary TDM of clonazepam. Additionally, the methodology described in this study can be used to develop and evaluate similar models incorporating both saliva and plasma concentrations to allow for non-invasive pharmacokinetic sampling in future clinical trials in pediatric or other vulnerable populations.

We found high concentrations of clonazepam in saliva samples taken during the first four hours after administration of the oral solution. This indicates significant contamination in the initial saliva samples by clonazepam residue, a finding that has been reported in the past.[18] This precluded the use of rate constants in which there was a slow increase in salivary concentrations driven by plasma pharmacokinetics, because the timepoints earlier than 4h post-dose did not contain any information about the transfer rate from the plasma compartment to the saliva compartment. However, the inclusion of an exponential elimination of the contaminated saliva in combination with a saliva:plasma ratio in the

model led to a good model fit and adequate predictive performance. A consequence of this finding is that saliva samples taken during the first 4 hours do not provide any information regarding the plasma concentrations during that time. However, considering the long half-life and the fact clonazepam therapy is guided via trough concentrations, this has little impact. The estimated model parameters regarding plasma kinetics were comparable to the population model developed by dos Santos *et al.*[19]

Predictive performance was assessed through simulation in a fictional cohort aged 6-30. Simulated salivary samples were obtained that included the residual error component of the model, and the most likely relative bioavailability and the corresponding C_{TROUGH} was estimated via Bayesian MAP optimization. There are several advantages to using Bayesian methodologies for this purpose. First, the prediction error, represented by the RMSPE was lower compared to the prediction error based on linear modelling. Additionally, one can use information obtained from multiple samples to estimate the most likely C_{TROUGH}, reducing the prediction error in the process. We found that, with the current model, predictive performance increases by obtaining additional samples up to 5 samples, and possibly beyond that with even more saliva samples. However, obtaining more than 5 samples in future pediatric PK studies would not be in line with the non-invasive approach taken here. Third, the method allows for convenience sampling at timepoints that are logistically feasible, as long as the chosen timepoints are obtained on timepoints with a valid saliva:plasma correlation, after 4 hours post-dose for the current analysis. Fourth, the optimization process takes residual variability into account, and the prediction shrinks towards the population mean in the case of high residual variability. This prevents that outlier saliva observations are extrapolated to extreme estimated plasma concentrations on which dose adaptions are made. Finally, estimates cannot be outside the constraints provided by the population model, as opposed to linear regression methods that have no such limits. The relevance of several of these advantages are confirmed in our simulations of the predictive capability of Bayesian MAP versus linear regression equations. Predictive capability of the linear regression equation was adequate but inferior to Bayesian MAP based on multiple samples. The simulations confirm that saliva sampling is eminently feasible for monitoring of clonazepam trough concentrations in the context of TDM and for the estimation of individual PK trajectories in the context of clinical trials. Correlations between real and predicted C_{TROUGH} were found, although these showed a slight underprediction at higher concentrations, which can be explained by shrinkage to the population mean during the Bayesian optimization process.

In this study, a saliva:plasma ratio was described which was not constant over time, even after accounting for the initial levels of contamination. The apparent decrease in the ratio driven by decreasing plasma concentrations is a phenomenon that has been reported before[20–22], but the underlying mechanism causing this relationship is unclear. Transport into saliva is partly driven by the free fraction of a drug, but the observed relationship cannot be explained by saturable protein binding. We hypothesize that the observed relationship may be caused by competitive protein binding of clonazepam metabolites, a mechanism previously observed for prednisone[23]. Nevertheless, it remains an important finding in the context of TDM, as this invalidates the use of 'traditional' linear regression equations that do not take this variable saliva:plasma ratio into account. This may be one of the reasons that Bayesian optimization outperformed linear regression, even in the scenario with a single saliva sample.

This study has several limitations. First, this model in this study was based on observations after a single administration of clonazepam. Application of the model for the purpose of TDM or clinical trials will usually occur when subjects have reached steady state, and in that case, it is assumed that the estimates obtained here can be extrapolated to these higher concentrations. As this is the first study systematically exploring the relationship between saliva- and plasma concentrations of clonazepam, this assumption cannot be verified at this time, but can be confirmed in future studies by obtaining paired saliva and plasma samples from subjects who have reached steady state. Using this model in a future pediatric clinical trial is reliant on several other assumptions as well. First, it is assumed that plasma kinetics in children adhere to the allometric scaling employed in this study, which is subject to recurrent discussion.[24,25] However, several studies indicate that this approach is reasonably accurate.[26] The second assumption is that the saliva:plasma ratio is identical in children compared to the young adults included in this study. Little comparative research has been performed on this subject, but Michael *et al.* report highly similar saliva:plasma ratio's in children and adults for voriconazole, and a systematic review regarding saliva:plasma ratio's in infants showed comparability for several compounds.[3,27] Although these assumptions may cause additional variability and less precise results in pediatric patients, we expect the prediction error remains small enough to adequately identify a PKPD relationship based on saliva samples. Furthermore, saliva samples will aid in the identification of ultra-fast metabolizers and subjects that do not adhere to the treatment regimen. The currently developed model can already be applied in adults, where saliva based TDM could be preferable over plasma based TDM.

Future studies may determine whether the underlying assumptions in pediatric populations described above are valid in general and in the case of clonazepam in particular. However, if confirmed, we believe this methodology could be widely implemented to aid clinical trial conduct in pediatrics and apply precision-dosing in pediatric populations.

Conclusion

The developed population pharmacokinetic model describes the salivary- and plasma kinetics of clonazepam well and simulations show that plasma C_{TROUGH} can be predicted from saliva concentrations through Bayesian MAP optimization.

SUPPLEMENTARY DATA

Sup. Figure S1	Nonlinear saliva:plasma ratio
Sup. Figure S2-S5	Goodness of fit plots
Sup. Figure S6	Relative bias, limits of agreement and correlation of predictions in different scenarios after dose 1
Sup. Figure S7	Relative bias, limits of agreement and correlation of predictions in different scenarios at steady state

REFERENCES

1 Patsalos PN, Berry DJ. Therapeutic Drug Monitoring of Antiepileptic Drugs by Use of Saliva. Ther Drug Monit 2013;35(1).
2 Kim HY, Märtson AG, Dreesen E, Spriet I, Wicha SG, McLachlan AJ, Alffenaar JW. Saliva for Precision Dosing of Antifungal Drugs: Saliva Population PK Model for Voriconazole Based on a Systematic Review. Front Pharmacol 2020;11(June).
3 Hutchinson L, Sinclair M, Reid B, Burnett K, Callan B. A descriptive systematic review of salivary therapeutic drug monitoring in neonates and infants. Br J Clin Pharmacol 2018;84(6):1089-1108.
4 Browne TR. Clonazepam. A review of a new anticonvulsant drug. Arch Neurol 1976;33(5):326-332.
5 Cohen LS, Rosenbaum JF. Clonazepam: new uses and potential problems. J Clin Psychiatry 1987;48 Suppl:50-56.
6 Brodie MJ, Chung S, Wade A, Quelen C, Guiraud-Diawara A, François C, Verpillat P, Shen V, Isojarvi J. Clobazam and clonazepam use in epilepsy: Results from a UK database incident user cohort study. Epilepsy Res 2016;123:68-74.
7 Jung EM, Moffat JJ, Liu J, Dravid SM, Gurumurthy CB, Kim WY. ARID1B haploinsufficiency disrupts cortical interneuron development and mouse behavior. Nat Neurosci 2017;20(12):1694-1707.
8 Joseph PD, Craig JC, Caldwell PHY. Clinical trials in children. Br J Clin Pharmacol 2015;79(3):357-369.
9 Gagne JJ, Thompson L, O'Keefe K, Kesselheim AS. Innovative research methods for studying treatments for rare diseases: Methodological review. BMJ 2014;349(November):1-10.
10 Kruizinga MD, Zuiker RGJA, Sali E, de Kam ML, Doll RJ, Groeneveld GJ, Santen GWE, Cohen AF. Finding Suitable Clinical Endpoints for a Potential Treatment of a Rare Genetic Disease: the Case of ARID1B. Neurotherapeutics 2020.
11 Committee for Medicinal Products for Human Use (CHMP). Guideline on bioanalytical method validation. EMEA/CHMP/EWP/192217/2009 Rev 1 Corr 2 2011.
12 Beal SL, Sheiner LB, Boeckmann AJ, *et al.* NONMEM 7.3.0 User Guides. (1989-2013). ICON Dev Solut Hanover, MD.
13 R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
14 Kyle T. Baron and Marc R. Gastonguay. Simulation from ODE-Based Population PK/PD and Systems Pharmacology Models in R with mrgsolve. J Pharmacokinet Pharmacodyn 2015;42(W-23):S84-S85.
15 National Center for Health Statistics, Data Table of Weight-for-age Charts. [accessed 2020 Nov 20]. https://www.cdc.gov/growthcharts/html_charts/wtage.htm#males
16 Kang D, Bae KS, Houk BE, Savic RM, Karlsson MO. Standard error of empirical bayes estimate in NONMEM® VI. Korean J Physiol Pharmacol 2012;16(2):97-106.
17 Altman DG, Bland JM. Measurement in Medicine: The Analysis of Method Comparison Studies. J R Stat Soc Ser D (The Stat 1983;32(3):307-317.
18 Lins RL, Otoul C, De Smedt F, Coupez R, Stockis A. Comparison of plasma and saliva concentrations of levetiracetam following administration orally as a tablet and as a solution in healthy adult volunteers. Int J Clin Pharmacol Ther 2007;45(1):47-54.
19 Santos FM Dos, Gonçalves JCS, Caminha R, Da Silveira GE, Neves CSDM, Gram KRDS, Ferreira CT, Jacqmin P, Noël F. Pharmacokinetic/pharmacodynamic modeling of psychomotor impairment induced by oral clonazepam in healthy volunteers. Ther Drug Monit 2009;31(5):566-574.
20 Haeckel R. Factors Influencing the Saliva/Plasma Ratio of Drugs. Ann N Y Acad Sci 1993;694(1):128-142.
21 Newton R, Broughton LJ, Lind MJ, Morrison PJ, Rogers HJ, Bradbrook ID. Plasma and salivary pharmacokinetics of caffeine in man. Eur J Clin Pharmacol 1981;21(1):45-52.
22 Malone SA, Eadie MJ, Addison RS, Wright AWE, Dickinson RG. Monitoring salivary lamotrigine concentrations. J Clin Neurosci 2006;13(9):902-907.
23 Unadkat JD, Rowland M. Representation and quantitation of the binding interaction between prednisone, prednisolone and corticosteroid binding globulin. J Pharm Pharmacol 1984;36(9):582-585.
24 Mahmood I. Dosing in children: A critical review of the pharmacokinetic allometric scaling and modelling approaches in paediatric drug development and clinical settings. Clin Pharmacokinet 2014;53(4):327-346.
25 Samant TS, Mangal N, Lukacova V, Schmidt S. Quantitative clinical pharmacology for size and age scaling in pediatric drug development: A systematic review. J Clin Pharmacol 2015;55(11):1207-1217.
26 Momper JD, Mulugeta Y, Green DJ, Karesh A, Krudys KM, Sachs HC, Yao LP, Burckart GJ. Adolescent dosing and labeling since the Food and Drug Administration Amendments Act of 2007. JAMA Pediatr 2013;167(10):926-932.
27 Michael C, Bierbach U, Frenzel K, Lange T, Basara N, Niederwieser D, Mauz-Körholz C, Preiss R. Determination of saliva trough levels for monitoring voriconazole therapy in immunocompromised children and adults. Ther Drug Monit 2010;32(2):194-199.

CHAPTER 14

Saliva as sampling matrix for therapeutic drug monitoring of gentamicin in neonates: A prospective population pharmacokinetic- and simulation study

Br J Clin Pharmacol. 2021 Oct 8. doi: 10.1111/bcp.15105

A Samb[1*], MD Kruizinga[2,3,4*], Y Tallahi[1], MJ van Esdonk[2], W van Heel[3], GJA Driessen[3,5], YA Bijleveld[1], FE Stuurman[2,4], AF Cohen[2,4], AH van Kaam[6], TR de Haan[6], RAA Mathôt[1]

**Both authors contributed equally*

1 Department of Pharmacy and Clinical Pharmacology, Amsterdam UMC, Amsterdam, the Netherlands
2 Centre for Human Drug Research, Leiden, the Netherlands
3 Juliana Children's Hospital, Haga Teaching Hospital, The Hague, the Netherlands
4 Leiden University Medical Centre, Leiden, the Netherlands
5 Maastricht University Medical Centre, Maastricht, the Netherlands
6 Department of Neonatology, Emma Children's Hospital Amsterdam UMC, University of Amsterdam, Amsterdam, the Netherlands

Abstract

INTRODUCTION Therapeutic drug monitoring (TDM) of gentamicin in neonates is recommended for safe and effective doses and is currently performed by blood sampling, which is an invasive and painful procedure. In this study, feasibility of a non-invasive gentamicin TDM strategy using saliva was investigated.

METHODS This was a multicenter, observational cohort study including 54 neonates. Any neonate treated with gentamicin was eligible for the study. Up to 8 saliva samples were collected per patient at different time-points. Gentamicin levels in saliva were determined with liquid chromatography coupled with tandem mass-spectrometry. A population PK model was developed using Nonlinear Mixed-Effects Modeling (NONMEM) to describe the relation between gentamicin concentrations in saliva and blood. Simulations were performed to evaluate the efficacy of gentamicin TDM using saliva versus blood.

RESULTS Blood PK was described with an earlier published model. Time profiles of salivary concentrations were quantified using a one-compartment saliva model with first-order input (k^{13} 0.023 h^{-1}) and first-order elimination (k^{30} 0.169 h^{-1}). Inter-individual variability of k^{30} was 38%. Post menstrual age (PMA) correlated negatively with both k^{13} and k^{30}. Simulations demonstrated that TDM with 4 saliva samples was effective in 81% of the simulated cases, versus 94% when performed with 2 blood samples.

CONCLUSION TDM of gentamicin using saliva is feasible, though TDM with 2 blood samples seems to perform better.

Introduction

Neonates admitted to the neonatal intensive care unit (NICU) have a high risk for bacteremia or sepsis due to premature birth, low birth weight and indwelling central venous lines.[1] Intravenous treatment with the aminoglycoside gentamicin provides good gram-negative coverage and is part of the first line antibiotic treatment protocols in many NICU's.

Gentamicin has a narrow therapeutic index, with oto- and nephrotoxicity as its possible concentration-dependent adverse drug events. Neonates are especially vulnerable for adverse events and adequate dosing is complicated by the continuous changes in body composition and clearance caused by a changing kidney function and maturation. Gentamicin concentrations can therefore be unpredictable and therapeutic drug monitoring (TDM) is necessary to ensure adequate plasma concentrations. TDM requires repeated blood sampling, which is invasive, painful and may contribute to clinical anemia or infection.[2] As a result, TDM by blood sampling is complicated in neonates[3], possibly leading to suboptimal individual gentamicin doses and thereby causing a decrease in therapeutic efficacy and/or an increased risk of adverse events.

Therefore, there is a clinical need for non-invasive TDM methods in neonates which would allow for an increased sampling frequency and for safer and more efficacious dosing, while simultaneously decreasing the burden of blood collection. Previous studies have shown that the use of saliva as a matrix for TDM is feasible for several anti-epileptic drugs and caffeine.[4,5] Analyses of salivary gentamicin concentrations and other aminoglycosides during intravenous treatment of children and adults have been published with varying results. Some studies reported a good correlation between gentamicin saliva and blood concentrations, while others reported undetectable aminoglycoside concentrations in saliva.[6-9] So far, no such studies have been performed in a neonatal population.

The aim of this study was to prospectively measure salivary gentamicin concentrations and to compare these to the concentrations in routinely drawn blood samples in neonates.

Materials and methods

Study design

This was a multi-center, prospective, observational pharmacokinetic study conducted in the Emma children's hospital (Amsterdam UMC, Amsterdam, the Netherlands) and the

Juliana children's hospital (Haga Hospital, den Haag, the Netherlands). Gentamicin concentrations were prospectively measured in saliva and compared with blood concentrations, obtained as part of routine TDM. The local ethics committee of the Amsterdam UMC approved this study (number 2018_193). Local feasibility was tested and approved for the Haga hospital. The study was registered in the Dutch Trial Registry (NTR, NL7211).

Subjects

Inclusion of subjects took place between October 8TH 2018 and March 4TH 2020. Any neonate that was treated with gentamicin according to local clinical guidelines was eligible for the study. Patients were included in this study after signed informed consent of both parents was obtained. For the analysis, three distinct subgroups based on gestational age (GA) were pre-specified and treated with intravenous gentamicin according to local dosing protocols: 1) Neonates with GA < 32 weeks (5 mg/kg/48 hours); 2) neonates with GA ≥ 32 weeks - 37 weeks (5 mg/kg/36 hours); and 3) neonates with GA ≥ 37 weeks (4 mg/kg/24 hours at Emma Children's hospital and 5 mg/kg/36 hours at Juliana Children's Hospital). Clinical data were obtained from the digital medical files of the patients (sex, GA, postnatal age (PNA), postmenstrual age (PMA), birth weight (BW), current body weight (WT), perinatal asphyxia, therapeutic hypothermia, and concomitant medication).

Sample sizes could not be accurately calculated, due to absence of data on expected effect-size and variability of estimated saliva PK-parameters. A total of 60 patients (20 patients per group) were scheduled to be enrolled into the study, since 20 patients per subgroup are deemed sufficient for NONMEM analysis as a rule of thumb.[10]

Sample collection

Saliva samples were collected using SalivaBio Infant's Swabs (Salimetrics, Carlsbad, CA, USA). Swabs were placed in the cheek pouch of the neonates for approximately 90 seconds, according to the manufacturer's instructions.[11] After collection, swabs were centrifuged at 4,000 RPM for 5 minutes and extracted saliva was stored at -80° C. Up to 8 saliva samples were collected per patient using an opportunistic sampling schedule. Saliva samples were collected up to 48 hours after the last gentamicin dose. Adsorption of gentamicin to the swab was found to be less than 3.1% at the low concentration level and 8.2% at the high concentration level and therefore below the predetermined acceptable percentage of 15%. Gentamicin concentrations in blood were collected from routine peak and trough TDM measurements (0.5h post infusion (infusion duration 0.5h) and/or between 6-24h post-dose). Additional blood levels were determined in residual material, when available.

Bio-analytical assay

The major components of gentamicin (C1, C1a and C2) were quantified in saliva samples using a previously published LC-MS/MS method.[12] In short, the accuracy and within run precision at the lowest level of quantification (LLOQ) were 118% and 10.2%, respectively. The accuracy and precision were 98.4% and 3.3%, respectively, at the middle level of quantification (MLQ). At the upper limit of quantification (ULOQ), accuracy was 98.7% and precision was 3.2%. Accuracy and precision were within the predetermined acceptable ranges (LLOQ: ±20%, MLQ: ±15%, ULOQ: ±15%). The LLOQ was 0.056 mg/L and minimal sample volume was 10 µl.

Pharmacokinetic analysis

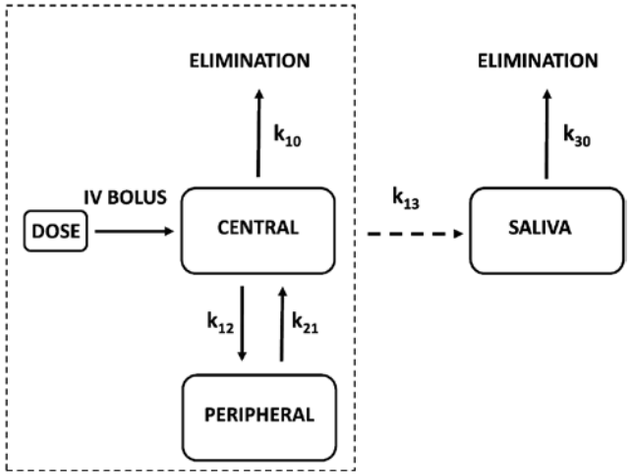
Data handling, data visualization and descriptive statistics were performed using R statistics version 4.0.2. A population PK (POP-PK) model was developed using nonlinear mixed-effects modeling (NONMEM), as implemented in NONMEM version 7.4.0 (ICON Development Solutions, Dublin, Ireland). Gentamicin concentrations in blood and saliva were logarithmically transformed.

An integrated model describing gentamicin in blood and saliva was developed using a stepwise modeling approach. First, blood PK data was described using a previously published model by Fuchs *et al*.[13], fixing the published PK parameters. The control stream for this model was provided by the original authors. This was a 2-compartment model with inter-individual variability (IIV) on clearance (CL) and central volume of distribution (V_c). Model performance was evaluated through the assessment of goodness-of-fit (GOF) plots and visual predictive checks (VPCS).

Following estimation of the individual blood PK parameters, an additional compartment describing the salivary gentamicin concentrations was appended to the model, effectively developing a 3-compartment model. The conceptual model for gentamicin in blood and saliva has been depicted in *Figure 1*. The first-order transport rate from the central

(blood) compartment to the saliva compartment was expressed as k^{13}, whilst the first-order rate of gentamicin elimination from the saliva compartment was expressed as k^{30}. No transport from the saliva compartment to the central and peripheral compartments was modeled, since the oral bioavailability of gentamicin is negligible.[14] Central gentamicin mass decrease due to transport from the central compartment to the saliva compartment was assumed to be negligible as well, as this was expected to be proportionally diminutive compared to the total amount of gentamicin in the central compartment, similar to a hypothetical effect compartment model.[15] Both fixed and random effects of rate constants k^{13} and k^{30} were estimated using the ADVAN6 subroutine in NONMEM. Model parameters were evaluated by assessing changes in the objective function value (OFV) and diagnostic plots. A δOFV of -3.81 corresponds with p = 0.05, which was the significance level for inclusion of any parameter. Gentamicin concentrations in saliva below LOQ were included in the model using the M3-method.[16] First, the structural model was estimated, describing the relations between parameters. Thereafter, the error model was developed, describing the residual error structure in the model. Finally, the covariate model explains part of the variability based on covariates.

Figure 1. Conceptual model for gentamicin PK in blood and saliva. Within dashed lines: Gentamicin in blood. Dose is administered as an iv bolus to the central compartment. k^{12}: Transport rate from central to peripheral compartment. k^{21}: Transport rate from peripheral compartment to central compartment. k^{10}: Elimination rate from the central compartment. Outside dashed lines: gentamicin PK in saliva. k^{13}: Transport rate from central compartment to saliva compartment. The dashed arrow signifies that gentamicin loss from the central compartment is assumed to be negligible. k^{30}: : Elimination rate from saliva.



GA, PNA, PMA, BW, WT, sex, perinatal asphyxia, therapeutic hypothermia, and concomitant drugs were evaluated as covariates for this model. Covariate analysis was performed with stepwise forward inclusion (α=0.05) and backwards elimination (α=0.01). Continuous covariates were included in the model as a power equation function (*Eq.1:* $p = \theta_p * (cov/median)\theta_{cov}$).

Parameter p was calculated from typical parameter θ_p, multiplied with the fractional deviation from the median value of the covariate. The magnitude of the covariate effect was estimated as θ_{cov}. Dichotomous covariates were coded in NONMEM as shown in *Eq.2* (*Eq.2*: $p = \theta_p + cov * \theta_{cov}$).

Dichotomous covariates could take the value of either 0 or 1. Reference parameter value θ_p was estimated, and the parameter difference between covariate parameters was estimated as θ_{cov} to calculate parameter p. Assessments of diagnostic tools, such as GOF plots, parameter residual standard error (RSE), η-shrinkage and ε-shrinkage were used for model evaluation during all steps. Bootstrap analyses (n=1000), as well as the simulation based prediction-corrected VPCS (PCVPC) were employed for assessment of the model robustness and internal validation of the final model.[17]

TDM performance simulation

R version 4.02 and the mrgsolve[18] package were used for Monte Carlo simulations. A simulation cohort (n = 3000) with a uniform distribution of GA and corresponding WT[19] was prepared, and a single administration of 5 mg/kg/48h (GA < 32 weeks), 5 mg/kg/36h (GA ≥ 32-37 weeks) or 4 mg/kg/24h (GA ≥ 37 weeks) was simulated for each subject in accordance with Dutch dosing guidelines.

For blood and saliva TDM, different sampling schedules were simulated with measurements at different time-points after the first dose. First, a schedule with a single intermediate (14h post-dose) sample was simulated and the performance of this schedule in the context of TDM was appraised. Second, a two-sample schedule was evaluated with a peak- (1.5h for blood and 3h for saliva samples) and trough (0.5h before next dose) sample. Next, the combination of peak-, intermediate- and trough samples was evaluated. Finally, schedules were evaluated in which samples were added (at 7 h post-dose; at 7 - 18 h post-dose; at 1 h pre-dose and 7 - 18 h post-dose) were evaluated. Bayesian maximum a posteriori (MAP) optimization was used to estimate the empirical Bayes estimates of the individual CL, VC and k^{30} for each subject based on the simulated samples.[20] Then,

based on the estimated CL and VC, the peak- and trough concentrations were estimated for each subject, who then entered a basic decision rule optimizing the dose to reach a targeted peak concentration between 9-11 mg/L and trough concentration < 0.8 mg/L after the third dose. Target ranges were deliberately set stricter compared to clinical guidelines (peak 8-12 mg/L and trough < 1 mg/L) to account for residual error in the estimations. For each subject, two additional dose intervals of gentamicin were simulated after dose adjustment. Finally, the proportions of subjects with true peak- and trough concentrations within clinical guideline reference ranges (target attainment) after the third dose were calculated. Simulation runs were performed for blood TDM, saliva TDM, model-based dose optimization (using the blood model of Fuchs *et al.*) and 'no TDM' (standard dosing regimen during the complete simulation period). The proportion of subjects with target attainment after each simulated scenario was calculated and compared in order to appraise the added value of saliva and blood TDM.

Table 1. Demographic characteristics of the study population

Demographic	Value
Enrolled patients - n	54
Males - n (%)	31 (57.4)
GA in weeks - median (range)	34.8 (24.3 - 41.7)
< 32 weeks - n (%)	21 (38.9)
32 - 37 weeks - n (%)	13 (24.1)
≥ 37 weeks - n (%)	20 (38.9)
PMA in days - median (range)	244.2 (170.5 - 294.2)
PNA in days - median (range)	1.5 (0.3 - 6.8)
Birth weight in kg - median (range)	2.4 (0.7 - 4.5)
Actual weight in kg - median (range)	2.4 (0.7 - 4.3)
Total saliva samples - n (%)	267 (100)
Analyzed - n (%)	194 (72.7)
Failed - n (%)	73 (27.3)
Analyzed saliva samples per patient - median (range)	3 (1 - 8)
Plasma samples - n	99
Plasma samples per patient - median (range)	2 (1 - 4)
Oro-esophageal congenital anomalies - n	1
Controlled hypothermia - n	3
Perinatal asphyxia - n	3

GA: Gestational age. PNA: Postnatal age. PMA: Postmenstrual age.

Results

Demographic characteristics

Table 1 depicts the demographic characteristics of the included patients. In total 54 of the planned 60 neonates were enrolled in this study due to its early termination during the SARS-COV-2 pandemic, which posed restrictions for clinical research. A total of 267 saliva samples were collected during the study, though 79 (29.6%) saliva samples could not be analyzed because of low sample volume or contamination of the saliva with blood. The demographic characteristics were representative for the population of neonates treated with gentamicin.

Gentamicin pharmacokinetics in blood

Model diagnostic figures indicated that the model provided by Fuchs *et al.* could adequately describe the blood PK data of the study population (*Figure 2*). The model was used to estimate individual blood PK and served as a basis for the construction of the saliva model.

Gentamicin pharmacokinetics in saliva

The salivary PK of gentamicin was described by adding a saliva compartment to the blood model (*Figure 1*). For the structural model, a k^{13} of 0.036 h^{-1} and k^{30} of 0.267 h^{-1} were estimated, as well as IIV on k^{30} of 63.6% (*Table 2*). The estimate of IIV on k^{30} had a η-shrinkage of 17%. Residual error was described with a logarithmic proportional error model, which was estimated as 58.4%. Since 14% of all analyzed saliva samples were found to be below the LLOQ, these measurements were accounted for with the M3 method.[16] Inclusion of additional transit compartments to account for lag in saliva uptake did not improve the model fit; neither did 1^{ST} order transport from the peripheral to the salivary compartment. Though it was also possible to successfully fit a model with estimations for both IIV on k^{30} and k^{13}, η-shrinkage on these parameters was respectively 56% and 34%. These levels of η-shrinkage were unacceptable and therefore that model was rejected.[21]

Stepwise forward inclusion of PMA as a power function covariate on k^{13} led to the largest decrease in OFV (δOFV = -61.33). PMA was also included as a covariate on k^{30} as a

Figure 2. Model diagnostic plots blood PK. A: Population predictions versus observed concentrations in blood. B: Individual predictions versus observed concentrations in blood. C: Population predictions versus conditional weighted residuals (CWRES). D: Time versus CWRES. E: prediction corrected VPC

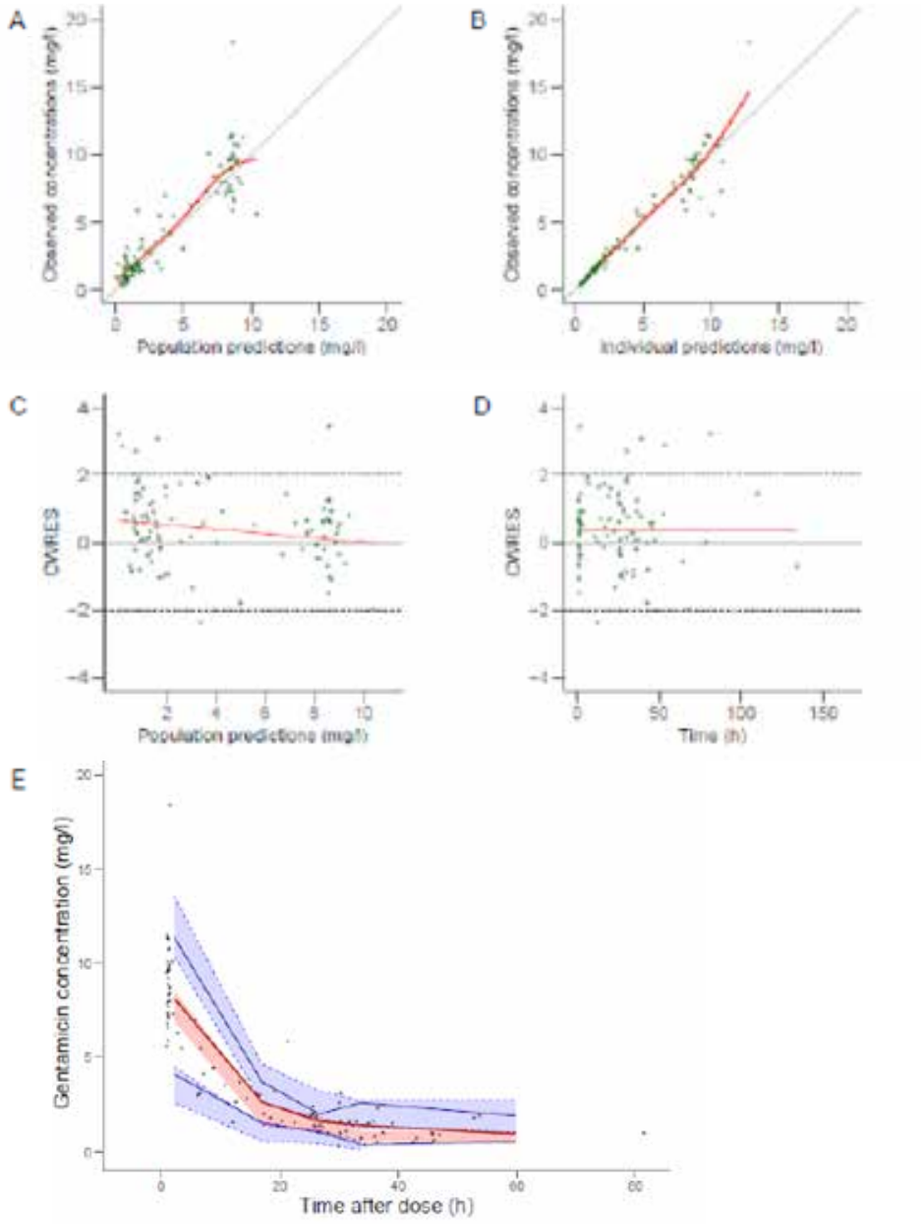


Figure 3. Goodness of fit plots of the final model. A: Population predictions versus observed concentrations in saliva. B: Individual predictions versus observed concentrations. C: Population predictions versus conditional weighted residuals (CWRES). D: Time versus CWRES.

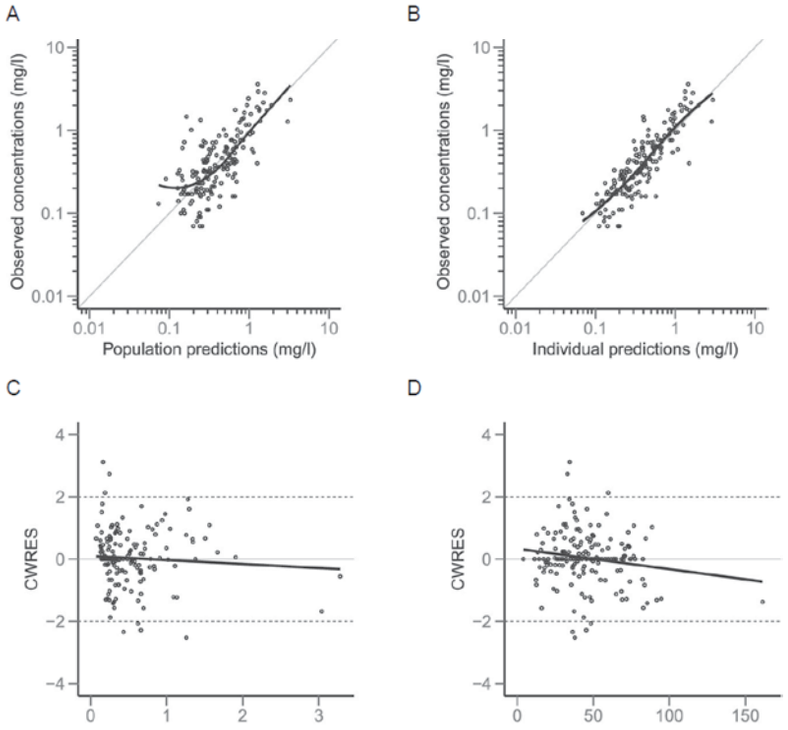
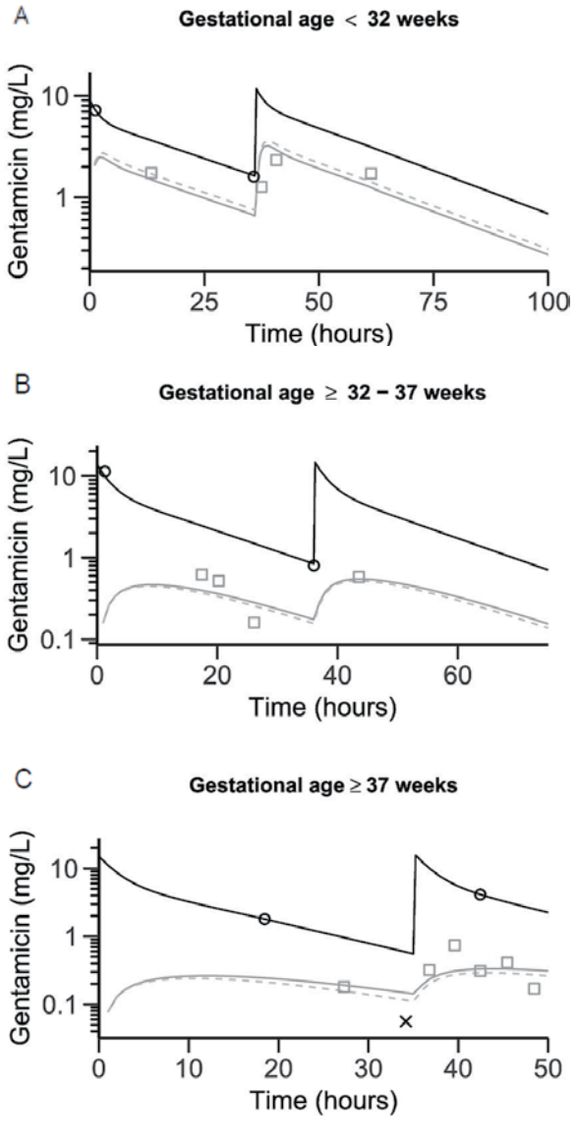


Table 2. Population PK parameters and bootstrap results

Parameter	Structural model		Final model		Bootstrap results		
	OFV = 877.3		OFV = 738.7		(N=1000)		
	Estimate	RSE (%)	Estimate	RSE (%)	Median	2.5th %	97.5th %
θ_{K13} (h^{-1})	0.036	79	0.023	16	0.023	0.016	0.033
θ_{K30} (h^{-1})	0.267	70	0.169	15	0.171	0.123	0.239
θPMA_{K13}	-	-	-8.8	16	-8.7	-11.7	-5.7
θPMA_{K30}	-	-	-5.1	28	-4.9	-8.1	-2.0
σ_{PROP} (%)	58.4	9	49.7	7	49.0	40.8	56.4
IIV_{K30} (%)	63.6	12	38.0	17	37.3	30.5	43.8

θ_{K13}: 1ST order rate constant from central plasma compartment to saliva compartment. θ_{K30}: 1ST order elimination rate constant from saliva compartment. θ_{PMAK13}: Power equation exponent PMA on k_{13}. θ_{PMAK30}: power equation exponent PMA on k_{30}. σ_{PROP}: Proportional error. IIV_{K30}: Inter-individual variability of k_{30}.
$K_{13} = \theta_{K13} * (PMA/244.2)^{\theta PMAK13}$, $K_{30} = \theta_{K30} * (PMA/244.2)^{\theta PMAK30}$

Figure 4. Individual pharmacokinetic profiles of gentamicin in blood and saliva for typical patients of each GA group. A: Individual patient of GA < 32 weeks; B: Individual patient of GA ≥ 32–37 weeks; C: Individual patient of GA ≥ 37 weeks. black circles: observed blood concentrations; gray squares: observed saliva concentrations; solid black line: individual predicted blood concentrations; solid gray line: individual predicted saliva concentrations; dashed gray line: population predicted saliva concentrations; black crosses: observed saliva concentrations < LLOQ.

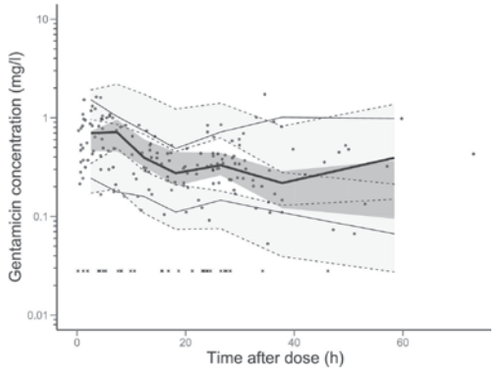


power function (δOFV = -17.25). None of the other tested covariates improved the model. The parameter estimates of the final model are shown in *Table 2*. Final estimates for k^{13} and k^{30} were 0.023 h^{-1} and 0.169 h^{-1}, respectively. IIV of k^{30} was 38% in the final model, whereas proportional residual error was 49.7%. The exponents of PMA as a covariate on k^{13} and k^{30} respectively were -8.8 and -5.1. This describes a negative correlation between PMA and both the transport and elimination rate of gentamicin in saliva, indicating that gentamicin is more readily available in the saliva of patients of low PMA, such as premature neonates. Evaluation of the GOF plots of the final model demonstrated a good description of the observed gentamicin concentrations in saliva (*Figure 3*). For demonstrative purposes, observations and model predictions have been plotted for 1 typical patient per GA group (*Figure 4*).

Bootstrap and internal model validation

The robustness of the final model was evaluated using a bootstrap procedure (n=1000). The median estimates and 95%CI for all parameters are summarized in *Table 2*. In total, 98.3% of the bootstrap runs were successful. For internal validation, a PCVPC (n=1000 samples) of the final model was evaluated (*Figure 5*). The majority of the 10TH, 50TH and 90TH percentiles of the observed values lie within the 95% confidence intervals of the 10TH, 50TH and 90TH percentiles of the simulated values for all bins.

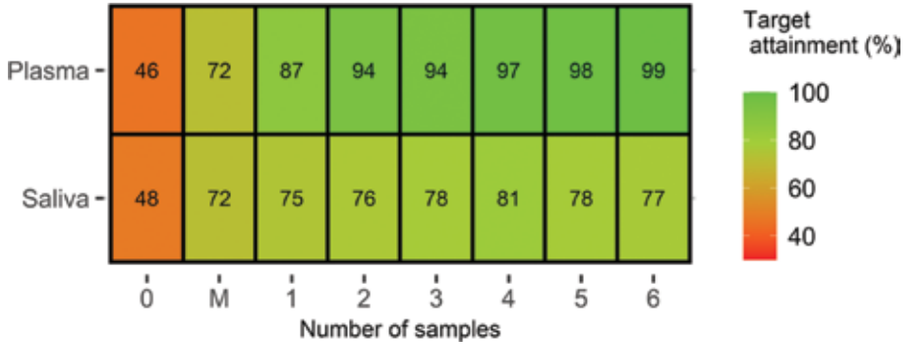
Figure 5. Prediction-corrected visual predictive check of the saliva model. Black circles: Observed gentamicin concentrations; thick black line: median observed concentrations; thin black lines: 80% interval of the observed concentrations; dark gray field: 95% confidence interval of the median prediction; light gray fields with dashed border: 95% confidence intervals of the 10TH and 90TH percentiles of the predictions; red crosses: observations below LLOQ.



Simulations

The simulated proportion of subjects with peak- and trough levels within the target range are displayed in *Figure 6*. Applying TDM using saliva led to a higher percentage of subjects reaching target attainment compared to no TDM (>75% vs 48%, respectively). However, saliva TDM led to a lower percentage of target attainment compared to blood TDM. Obtaining more than four samples for saliva TDM did not result in increased TDM performance. On the contrary, obtaining additional samples at 18h and 1h pre-dose led to a slightly decreased performance (-3% and -4%, respectively) compared to the strategy using four samples.

Figure 6. Heat map displaying the simulated proportion of subjects who reach target attainment of gentamicin after blood- and saliva TDM using an increasing number of samples. Time-points where samples were simulated: 0: standard dosing according to guidelines; M: no samples-dosing optimized according to population model and individual covariates; 1: sample (14h); 2: peak sample (3h for saliva or 1h for blood) and trough sample (0.5h pre-dose); 3: samples at peak, 14h and trough; 4: samples at peak, 7h, 14h and trough; 5: samples at peak, 7h, 14h, 18h and trough; 6: samples at peak, 7h, 14h, 18h, 1h pre-dose and trough.



Discussion

In this study, we have demonstrated the feasibility of monitoring gentamicin concentrations in saliva of neonates. Concentration-time profiles in both blood and saliva were described with an integrated PK model. The potential use of salivary concentrations in the context of TDM was assessed through Monte Carlo simulations. Simulations predicted a target attainment of up to 81% for TDM with 4 saliva samples versus 94% when performed with 2 blood samples.

In the past, several investigators have assessed the use of saliva for TDM of several drugs with varying results.[5-9] Berkovitch *et al.* reported a good correlation between blood and saliva concentrations for a once daily dosing regimen in children.[6] Other investigators reported that aminoglycosides did not penetrate into saliva of children with cystic fibrosis or tuberculosis.[8,9] Work regarding saliva TDM in neonates has covered a wide range of drugs, including caffeine, morphine and antiepileptic drugs.[5] Interestingly, all studies focused on linear correlations. Incorporating saliva concentrations in nonlinear mixed effect models may allow for more flexibility to account for delayed penetration, delayed elimination, and variability in saliva/blood ratio (S/B). To date, this is the first such model to have been developed for gentamicin, and there are only few published models which incorporate this methodology to describe saliva concentrations for other drugs.[22,23]

The model developed during this study was constructed by appending a blood PK model for gentamicin with a saliva compartment. The model by Fuchs *et al.* described the blood PK of the study population.[13] The model of Bijleveld *et al*. did not result in an improved description of the blood PK.[24] This was also true when constructing a new blood PK model with the study data. Gentamicin concentrations in saliva could best be described with drug transport from the central compartment (*Figure 1*). Models incorporating drug transport from the peripheral compartment to saliva were evaluated but did not accurately describe the data. Two separate rate constants were estimated for the saliva model. A 1ST order rate constant k^{13} of 0.023h^{-1} was estimated, whereas an elimination rate k^{30} of 0.169 h^{-1} was estimated. In this case, k^{13} was estimated to be much lower than k^{30}, indicating that transport from the central blood compartment to the saliva compartment is the rate-limiting step determining the concentration-time profile in saliva.[25] When predicting gentamicin concentrations in blood and saliva in typical patients (*Figure 4*), it seems that the S/B ratio stabilizes hours after the last dose is administered. During this phase, the concentration-time curve of saliva is perpendicular to blood, indicating that the salivary gentamicin elimination rate is linear to the blood concentration and therefore is dependent on k^{13}.

Considerable IIV was detected. Part of this was accounted for by including PMA as covariate. It was estimated that IIV on k^{30} was 38% in the final model. Post-menstrual age had a large influence on the salivary PK profile of gentamicin. Inclusion of PMA as a covariate on both k^{30} and k^{13} significantly improved the model. The exponents of the power equation functions were -5.5 and -8.8 for k^{30} and k^{13} respectively, suggesting a very strong age dependency of gentamicin disposition in saliva. With increasing PMA, k^{13} and k^{30}

decrease by a large margin. Indeed, it was observed that salivary gentamicin levels were generally much lower in term neonates, compared to premature neonates. Furthermore, 75% of samples below the LLOQ were from term neonates (PMA>260 days). Though the model did not contain a parameter describing the IIV in k^{13}, inclusion of PMA as a covariate on k^{13} significantly improved model fit, decreased RSE on all parameters and decreased residual error. It was quite notable that gentamicin was more freely distributed in saliva of premature neonates. However, no biological explanation for this phenomenon could be found in literature. Nonetheless, this finding may be indicative that salivary TDM could be more efficacious and possibly more accurate in premature neonates.

TDM performance was assessed through simulation in a fictional cohort of 3,000 neonates with a realistic distribution of covariates.[19] Applying Bayesian MAP during simulation, one can use information obtained from multiple samples to estimate the peak- and trough concentrations, which reduces the prediction error in the process. Additionally, the optimization process takes residual variability into account, and the prediction shrinks towards the population mean in the case of high residual variability. This prevents that outlier saliva observations are extrapolated to extreme estimated blood concentrations on which dose adaptions are made. During the simulations, each virtual subject was subjected to a rigid dose decision rule for dose optimization. In practice, more nuances can be applied. Moreover, results from this simulation study may be quite optimistic, since inter occasion variability (IOV) is not accounted for, such as time-dependent changes in CL. However, the simulations give a crude indication of the expected reliability of TDM with saliva samples versus blood samples, as well as the comparative performance of several sampling schedules.

Simulations indicated that a target attainment of 81% is possible with saliva TDM. Obtaining the 4 saliva samples necessary in this scenario is logistically feasible. However, target attainment following TDM with 2 blood samples was higher (94%). This difference in performance for saliva and blood TDM can be explained by the large difference in residual error between the two matrices. The uncertainty in the Bayesian optimization process introduced by these parameters was too large to achieve adequate precision with additional sampling or different sampling schedules. Moreover, assessed saliva sampling schedules were equal for all dosing regimens, therefore the evaluated additional samples may have had limited value for dosing regimens of 36 or 48 hours. Blood TDM performs better in settings where collection of 2 blood samples is protocol. However, in many clinical settings TDM protocols require a single intermediate concentration sample. In that case, blood TDM has a predicted target attainment of 87% (*Figure 6*). This difference with saliva TDM is substantially smaller. Taken together with the uncertainties of simulations, TDM with 4 saliva samples may be a suitable alternative to blood TDM with a single intermediate concentration sample. Given that gentamicin was more readily available in the saliva of premature neonates and no different sampling strategies were employed based on dose regimen during simulation, the difference in predicted target attainment may not be clinically relevant, especially for premature neonates that could benefit most from a non-invasive TDM method.

This study has several limitations. First, there was a large proportion of saliva samples with insufficient volumes for analysis. This may be due to inadequate sampling technique or insufficient saliva production by subjects, especially with premature neonates. Future studies may employ a different sampling strategy to ensure that an adequate volume of saliva is drawn, such as use of a different swab or cutting the saturated end of the swab.[26,27] Currently no standardized method for the collection of saliva from neonates exists. Nonetheless, many samples were available for model development, thus we do not expect this has influenced the parameter estimates. Second, due to the low volumes of the collected samples, it was not possible to determine pH of the collected samples. Saliva pH has been proposed to influence salivary distribution of drugs.[28] Though little has been published regarding saliva pH of neonates, we expect that fluctuations in saliva pH have little influence on the protonated fraction of gentamicin, since the strongest basic pK_A is 10.18.[29] Regardless, influence of pH on salivary gentamicin concentrations may be assessed, if possible. Finally, assumptions made during simulation, such as the underlying covariate distribution and sampling strategies, have an influence on the proportion of subjects reaching target attainment. However, considering that the goal of the simulation was to compare saliva and blood TDM, the comparative differences found in these simulation scenarios should be independent of these assumptions.

Strengths of this study are the employment of POP-PK, allowing for the description of nonlinear relations between blood and saliva gentamicin concentrations. In addition, a relatively large cohort of neonates of different GA receiving varying dosing regimens originating from both a peripheral pediatric ward and NICU, improved the generalizability of the model. Moreover, use of highly sensitive LC-MS/MS allowed for determination gentamicin concentrations in small sample volumes. The LC-MS/MS method had an LLOQ of 0.056 mg/l, which was substantially lower than earlier publications investigating gentamicin in saliva.[7-9] Population pharmacokinetic modeling allowed for opportunistic

sampling schedules and identification of covariates. The TDM simulations of a wide range of sampling strategies give an adequate overview of the expected performance of saliva TDM in different scenarios.

Conclusion

With this study, we demonstrate that TDM of gentamicin in saliva is feasible. A target attainment of 81% was found based on explorative simulations with 4 saliva samples and performance is close to blood TDM with 1 intermediate sample. In the future, the real-life performance of saliva TDM employing an improved sampling technique should be investigated prospectively in premature neonates, as gentamicin appears more readily in the saliva of premature neonates and these most fragile infants may benefit most from non-invasive TDM.

REFERENCES

1 Simonsen, K. A., Anderson-Berry, A. L., Delair, S. F. & Dele Davies, H. Early-onset neonatal sepsis. *Clin. Microbiol. Rev.* **27**, 21–47 (2014).
2 Widness, J. A. Pathophysiology of anemia during the neonatal period, including anemia of prematurity. *Neoreviews* **9**, (2008).
3 Donge, T. Van *et al.* Quantitative Analysis of Gentamicin Exposure in Neonates and Infants Calls into Question Its Current Dosing Recommendations. *Antimicrob. Agents Chemother.* **62**, 1–12 (2018).
4 Patsalos, P. N. & Berry, D. J. Therapeutic drug monitoring of antiepileptic drugs by use of saliva. *Ther. Drug Monit.* **35**, 4–29 (2013).
5 Hutchinson, L., Sinclair, M., Reid, B., Burnett, K. & Callan, B. A descriptive systematic review of salivary therapeutic drug monitoring in neonates and infants. *Br. J. Clin. Pharmacol.* **84**, 1089–1108 (2018).
6 Berkovitch, M. *et al.* Therapeutic drug monitoring of once daily gentamicin in serum and saliva of children. *Eur. J. Pediatr.* **159**, 697–698 (2000).
7 Madsen, V., Lind, A., Rasmussen, M. & Coulthard, K. Determination of tobramycin in saliva is not suitable for therapeutic drug monitoring of patients with cystic fibrosis. *J. Cyst. Fibros.* **3**, 249–251 (2004).
8 Spencer, H. *et al.* Measurement of tobramycin and gentamicin in saliva is not suitable for therapeutic drug monitoring of patients with cystic fibrosis. *J. Cyst. Fibros.* **4**, 209 (2005).
9 Elsen, S. H. J. van den *et al.* Lack of penetration of amikacin into saliva of tuberculosis patients. *Eur. Respir. J.* **51**, (2018).
10 Vong, C., Bergstrand, M., Nyberg, J. & Karlsson, M. O. Rapid Sample Size Calculations for a Defined Likelihood Ratio Test-Based Power in Mixed-Effects Models. AAPS *J.* **14**, 176–186 (2012).
11 Salimetrics SalivaBio Infant's Swab. (2021).at <https://salimetrics.com/wp-content/uploads/2018/02/infant-swab-saliva-collection-instructions.pdf>
12 Bijleveld, Y. A. *et al.* Altered gentamicin pharmacokinetics in term neonates undergoing controlled hypothermia. *Br. J. Clin. Pharmacol.* **81**, 1067–1077 (2016).
13 Fuchs, A. *et al.* Population pharmacokinetic study of gentamicin in a large cohort of premature and term neonates. *Br. J. Clin. Pharmacol.* **78**, 1090–1101 (2014).
14 Axelrod, H. R. *et al.* Intestinal transport of gentamicin with a novel, glycosteroid drug transport agent. *Pharm. Res.* **15**, 1876–81 (1998).
15 Jacobs, J. R. & Williams, E. A. Algorithm to Control "Effect Compartment" Drug Concentrations in Pharmacokinetic Model-Driven Drug Delivery. IEEE *Trans. Biomed. Eng.* **40**, 993–999 (1993).
16 Beal, S. L. Ways to fit a PK model with some data below the quantification limit. *J. Pharmacokinet. Pharmacodyn.* **28**, 481–504 (2001).
17 Bergstrand, M., Hooker, A. C., Wallin, J. E. & Karlsson, M. O. Prediction-corrected visual predictive checks for diagnosing nonlinear mixed-effects models. AAPS *J.* **13**, 143–151 (2011).
18 Kyle T. Baron and Marc R. Gastonguay Simulation from ODE-Based Population PK/PD and Systems Pharmacology Models in R with mrgsolve. *J Pharmacokinet Pharmacodyn* **42**, S84–S85 (2015).
19 Visser, G. H. A., Eilers, P. H. C., Elferink-Stinkens, P. M., Merkus, H. M. W. M. & Wit, J. M. New Dutch reference curves for birthweight by gestational age. *Early Hum. Dev.* **85**, 737–744 (2009).
20 Kang, D., Bae, K. S., Houk, B. E., Savic, R. M. & Karlsson, M. O. Standard error of empirical bayes estimate in NONMEM® VI. *Korean J. Physiol. Pharmacol.* **16**, 97–106 (2012).
21 Savic, R. M. & Karlsson, M. O. Importance of Shrinkage in Empirical Bayes Estimates for Diagnostics: Problems and Solutions. AAPS *J.* **11**, 558–569 (2009).
22 Dobson, N. R. *et al.* Salivary caffeine concentrations are comparable to blood concentrations in preterm infants receiving extended caffeine therapy. *Br. J. Clin. Pharmacol.* 754–761 (2016).doi:10.1111/bcp.13001
23 Kim, H. Y. *et al.* Saliva for Precision Dosing of Antifungal Drugs: Saliva Population PK Model for Voriconazole Based on a Systematic Review. *Front. Pharmacol.* **11**, (2020).
24 Bijleveld, Y. A., Heuvel, M. E. Van Den, Hodiamont, C. J., Mathôt, R. A. A. & Haan, T. R. De Population pharmacokinetics and dosing considerations for gentamicin in newborns with suspected or proven sepsis caused by gram-negative bacteria. *Antimicrob. Agents Chemother.* **61**, 1–11 (2017).
25 Yáñez, J. A., Remsberg, C. M., Sayre, C. L., Forrest, M. L. & Davies, N. M. Flip-flop pharmacokinetics-delivering a Zreversal of disposition: challenges and opportunities during drug development. *Ther. Deliv.* **2**, 643–672 (2011).
26 Lin, G. C. *et al.* Directed Transport of CRP Across In Vitro Models of the Blood-Saliva Barrier Strengthens the Feasibility of Salivary CRP as Biomarker for Neonatal Sepsis. *Pharmaceutics* **13**, 256 (2021).
27 Gesseck, A. M. *et al.* A Case Study Evaluating the Efficacy of an Ad Hoc Hospital Collection Device for Fentanyl in Infant Oral Fluid. *J. Anal. Toxicol.* **44**, 741–746 (2020).
28 Jusko, W. J. & Milsap, R. L. Pharmacokinetic Principles of Drug Distribution in Saliva. *Ann. N. Y. Acad. Sci.* **694**, 36–47 (1993).
29 Drugbank.ca Gentamicin-DrugBank. (2018).at <http://www.drugbank.ca/drugs/DB00798>

CHAPTER 15

Pharmacokinetics of intravenous and inhaled salbutamol and tobramycin: an exploratory study to investigate the potential of exhaled breath condensate as a matrix for pharmacokinetic analysis

Br J Clin Pharmacol. 2020 Jan;86(1):175-181. doi:10.1111/bcp.14156.

MD Kruizinga,[1] WAJ Birkhoff,[1] MJ van Esdonk,[1,2] NB Klarenbeek,[1] T Cholewinski,[1] T Nelemans,[1] MJ Dröge,[3] AF Cohen,[1] RGJA Zuiker[1]

1 Centre for Human Drug Research, Leiden, the Netherlands
2 Division of Systems Biomedicine and Pharmacology, Leiden Academic Centre for Drug Research, Leiden University, Leiden, the Netherlands
3 Ardena Bioanalytical Laboratory bv, Assen, the Netherlands

Abstract

Concentrations of drugs acting in the lungs are difficult to measure, resulting in relatively unknown local pharmacokinetics. The aim of this study is to assess the potential of exhaled breath condensate (EBC) as a matrix for pharmacokinetic analysis of inhaled and intravenous medication. A 4-way crossover study was conducted in 12 volunteers with tobramycin and salbutamol intravenously and via inhalation. EBC and plasma samples were collected post-dose and analyzed for drug concentrations. Sample dilution, calculated using urea concentrations, was used to estimate the epithelial lining fluid concentration. Salbutamol and tobramycin were largely undetectable in EBC after intravenous administration and were detectable after inhaled administration in all subjects in 50.8% and 51.5% of EBC samples, respectively. Correction of EBC concentrations for sample dilution did not explain the high variability. This high variability of EBC drug concentrations seems to preclude EBC as a matrix for pharmacokinetic analysis of tobramycin and salbutamol.

Introduction

Pharmacological activity is usually derived from the time course of plasma concentrations and pharmacodynamic endpoints for pulmonary drugs, information about local concentrations could potentially improve pharmacokinetic analysis, since pharmacological activity depends on adequate drug levels at the effect site[2]. Lung penetration studies rely on bronchoalveolar lavage (BAL) or sputum induction techniques for the measurement of drugs in the epithelial lining fluid (ELF) of the lung[3,4]. BAL is invasive and potentially risky, which discourages use in vulnerable patients, children, and healthy subjects. Therefore, new sampling matrices are needed. A matrix of interest is exhaled breath condensate (EBC)[5]. EBC is obtained by cooling exhaled breath through contact with a cold condenser. Samples are collected as fluid or frozen material and represent the ELF, diluted in condensed water from inhaled air and some saliva[6–8]. EBC application for biomarker quantification has been limited by poor reproducibility and high variability[9]. Several solutions have been proposed to improve reliability of EBC, such as to control for sample dilution factor, obtained via EBC urea concentrations[10–12]. Drug concentrations have already been measured in exhaled breath following administration of several drugs[13–18]. Nonetheless, no studies have assessed EBC concentrations after inhaled administration or reported proper concentration-time curves of EBC. The aim of this study is to assess the potential of EBC as a matrix for pharmacokinetic analysis of two common respiratory drugs: salbutamol and tobramycin.

Materials and Methods

This study was conducted at the Centre of Human Drug Research (CHDR) in Leiden, the Netherlands from January 2018 until November 2018. The study protocol was reviewed and approved by the Beoordeling Ethiek Biomedisch Onderzoek (BEBO) Foundation Review Board (Assen, the Netherlands) prior to initiation of the study. The study was conducted according to the Dutch Act on Medical Research Involving Human Subjects (WMO) and in compliance with Good Clinical Practice. Twelve healthy, non-smoking male subjects aged 18-65 years and body mass index (BMI) between 18 and 35 kg/m^2 were included in this study. Subjects with a history or evidence of lung disease were excluded from this study.

Study design

This was an open-label, 4-way crossover study. Visits were planned with a washout period of 3-7 days. Salbutamol and tobramycin were determined suitable medication to use in this proof-of-concept study, because they are commonly used, are available for intravenous use and inhalation, and have ample data available regarding plasma pharmacokinetics. Subjects were administered 1 mg/kg tobramycin iv (Obracin diluted in 50-100 mL 0.9% NaCl) over 30 minutes, 250 or 500 µg salbutamol by intravenous infusion (Ventolin Injection 0,5 mg/mL) over 2-8 minutes, 170 mg tobramycin (Vantobra) by inhalation using a Medix AC2000® nebulizer (Clement Clarke International) or 400 µg salbutamol (Ventolin Evohaler, GSK, London, UK) by inhalation with a spacer device (Volumatic). Paired EBC and plasma samples were collected at 2 (salbutamol) or -15 and 5 (tobramycin) minutes pre-dose and at 10-, 20-, 40-, 60-, 80-, 120-, 180-, 240-, 300- and 420-minutes post-dose. Additional plasma and EBC samples were collected from 8 subjects at 600- and 720-minutes post-dose after inhaled administration.

Sample collection and analysis

EBC was collected according to European Respiratory Society (ERS) guidelines during 5 minutes of tidal breathing using the Ecoscreen (Jaeger, Hoechberg, Germany)[19]. EBC samples were stored at -80°C until further analysis. A Liquid chromatography-tandem mass spectrometry (LC-MS/MS) assay was developed by Ardena Bioanalytical Laboratory to quantify tobramycin and salbutamol in plasma and EBC, and urea in EBC (Supplementary Text 1). Urea concentrations were measured as a marker for dilution. At the start of each study day, plasma urea concentration was determined. Dilution factor was calculated (D_{EBC} = $[urea]^{PLASMA}/[urea]_{EBC}$) and was multiplied with EBC tobramycin or salbutamol concentration to get a dilution corrected EBC concentration. The coefficient of variation (CV) was calculated per time point until 180 minutes post-dose and compared to the CV before correction.

Statistics and pharmacokinetic analysis

As this was an exploratory study, no formal power analysis was performed. Pharmacokinetic endpoints were summarized descriptively. When tobramycin or salbutamol was undetectable, 50% of the lowest estimated concentration was imputed for graphical and statistical purposes. Descriptive analysis was performed using SPSS Version 25 (IBM, Armonk, NY). Pharmacokinetic parameters were calculated using R version 3.5.2[20]. The area under the curve (AUC) was calculated as AUC_{0-LAST} for EBC parameters and extrapolated to infinity (AUC_{INF}) for plasma parameters using the terminal elimination rate constant, determined by the log-linear regression of the last observations above the LLOQ. The bioavailability (F) was calculated on an individual level and summarized descriptively. When insufficient data was available for the log-linear regression, the AUC_{INF} was not included in the results and the mean AUC_{INF} was imputed for the calculation of the bioavailability (F). The clearance and volume of distribution were calculated and corrected for by an individual's F. Promasys® software (OmniComm. Ft. Lauderdale, FL, USA) was used for data management.

Results

Subjects, safety and tolerability

Twelve healthy male volunteers were included, and all subjects completed the four study days. Baseline characteristics are shown in *Table 1*. Twelve subjects received 1 mg/kg tobramycin by intravenous infusion, 250 µg (3 subjects) or 500 µg (9 subjects) salbutamol by intravenous infusion, 170 mg tobramycin by inhalation and 400 µg salbutamol by inhalation.

Table 1: Baseline characteristics of participants

Demographic	Value
Male, n (%)	12 (100)
Age (years)	24.6 (6.8)
BMI (kg m^{-2})	22.5 (2.6)
Height (cm)	184 (10.1)
Weight (kg)	75.8 (11.4)
RACE, N (%)	
White	11 (91.7)
Asian	1 (8.3)

Data is presented as mean (standard deviation) unless stated otherwise. All participants completed the four treatment days.
BMI: Body Mass Index.

Table 2: Pharmacokinetic parameters mean (standard deviation) [range]

	Parameter	Tobramycin Intravenous (1 mg/kg)[A]	Tobramycin Inhalation (170 mg)	Salbutamol Intravenous* (500 µg)	Salbutamol inhalation (400 µg)
PLASMA	t_{MAX} (h)	0.09 (0.02) [0.08-0.17]	2.72 (1.9) [1.33-7.0]	0.04 (0.02) [0.03-0.10]	0.31 (0.14) [0.17-0.69]
	C_{MAX} (ng/ml/pg/ml)[B]	5753.3 (1055.2) [4030-7090]	123.2 (55.7) [50.5-247]	17133.3 (4637.1) [10600-248000]	1012.5 (615.1) [486-2600]
	AUC_{INF} (ng*h/mL/pg*h/mL)	14232.8 (1682.8) [11736.2-17036.2]	1090.34 (389.1) [337.7-1715.4]*	19274.4 (4593.0) [13511-26835]	4128.7 (2487.8) [1588.3-10842.9]
	$t_{1/2}$ (hours)	2.23 (0.25) [1.72-2.68]	4.2 (0.63) [3.09-5.08]*	4.43 (0.87) [3.64-5.96]	4.04 (1.15) [2.04-5.89]
	Volume of distribution (L)	17.26 (3.28) [12.23-22.76]	31.1 (4.0) [25.8-36.5]*	168.4 (22.5) [134.12-194.78]	143.4 (57.6) [70.6-262.5]
	Clearance (L/h)	5.39 (0.99) [3.83-6.88]	5.2 (1.0) [3.8-6.9]*	27.23 (6.25) [18.63-37.01]	25.22 (7.75) [8.3-37.01]
	F (%)	-	3.4 (1.1) [1.4-5.5]	-	22.5 (8.5) [10.7-40.6]c
EBC	EBC samples>LLOQ (%)	0	51.5	2.5	50.8
	EBC AUC_{0-LAST} (ng*h/mL/pg*h/mL)	-	8.39 (5.1) [2.39-16.98]	-	260.3 (224.9) [32.2-645.0]
	EBC C_{MAX} (ng/ml/pg/ml)[A]	-	5.23 (4.78) [0.7-15.2]	-	336.25 (468.74) [25.9-1440]
	EBC t_{MAX} (h)	-	2.60 (4.0) [0.17-10.1]	-	1.36 (3.39) [0.12-12.1]
	CV[D] Before correction for D_{EBC}	-	114 (24)	-	124 (28)
	After correction for D_{EBC}	-	129 (32)	-	144 (40)

a Mean dose: 75 mg; b Salbutamol concentrations are reported as pg/ml, tobramycin concentrations are reported as ng/ml; c Calculated on n=9; d Mean CV calculated for all time points until 180 minutes post dose. Abbreviations: AUC: Area Under the Curve, F: bioavailability, D: dilution

Pharmacokinetics

Mean (± SD) plasma and EBC concentration-time curves of all subjects after intravenous and inhaled administration of tobramycin and salbutamol are shown in Figure 1. Pharmacokinetic parameters are summarized in Table 2. Individual concentration-time curves can be found in Supplementary Figure S1 and S2. One subject had an unexplained increase > 2 SD in salbutamol concentration in one plasma sample 2 hours after intravenous infusion, which was excluded during the analysis. After intravenous administration, tobramycin concentrations were undetectable in EBC and a quantifiable concentration of salbutamol was found in only 3 EBC samples (2.5%). However, measurable concentrations of tobramycin and salbutamol after inhaled administration were present in EBC samples of all subjects. Salbutamol was undetectable in 75% of subjects after 120 minutes. The proportion of samples with a quantifiable concentration of salbutamol per time point ranged from 8-92% (Supplementary Figure S3).

Urea and dilution

Urea concentration was measured in all EBC samples with a quantifiable drug concentration. Individual urea concentrations and D_{EBC} estimates are displayed in Supplementary Figure S4. Mean D_{EBC} was 531 (range 32-5719). Concentration-time curves of the estimated ELF drug concentration are displayed in Figure 1. The mean coefficient of variation (CV) per time point after correction for dilution was 144% (SD 40%) for salbutamol, compared to 124% (SD 28%) before correction. Mean CV of tobramycin concentration after correction was 129% (SD 32%), compared to a prior 114% (SD 24%).

Discussion

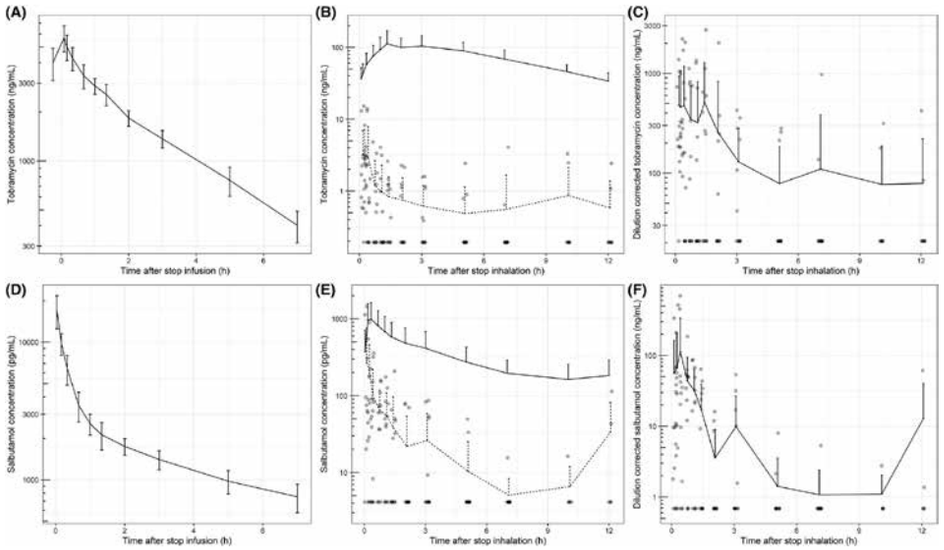
In this study, the potential of EBC as a matrix for pharmacokinetic analysis of drugs acting in the lungs was investigated. A 4-way crossover study was conducted wherein we obtained serial paired plasma and EBC samples after intravenous and inhaled administration of tobramycin and salbutamol.

The plasma pharmacokinetics after administration of intravenous and inhaled salbutamol and tobramycin were determined. Calculated pharmacokinetic parameters of salbutamol in plasma corresponded with results reported in other studies regarding T_{MAX}, C_{MAX} and half-life [21,22]. Plasma C_{MAX} of inhaled tobramycin was lower than expected, which may be a result of the chosen administration technique[23].

We are the first to report serial EBC drug concentrations. We were unable to measure salbutamol and tobramycin in EBC after intravenous administration of salbutamol and tobramycin in 97.5% and 100% of the samples, respectively. Low concentrations were expected in the case of tobramycin, considering the fact that tobramycin is a large (molecular weight 467.52 g/mol), polarized molecule and therefore does not easily pass barriers. This is illustrated by the poor bioavailability of inhaled tobramycin. However, we expected salbutamol to be detectable in more samples, considering salbutamol is a smaller molecule (molecular weight 239.32 g/mol) and the fact that several researchers have

demonstrated the ability to detect intravenously or orally administered drugs with similar properties in exhaled breath or EBC[13–16]. We assume that salbutamol plasma concentrations were insufficient after administration in order to achieve EBC salbutamol concentrations larger than the LLOQ, even with the high dose of 500 µg iv. This hypothesis should be confirmed with a more sensitive assay with a lower LLOQ.

Figure 1. Plasma and EBC concentration–time curves after intravenous administration of tobramycin and salbutamol. A. Concentration–time curve of mean tobramycin concentration in plasma (±SD) after intravenous administration of 1 mg/kg tobramycin (n=12); B. Concentration–time (Mean+SD) curve of plasma (solid line) and EBC (dotted line) tobramycin concentration after inhalation of 170 mg tobramycin; C. Concentration–time curve of mean (SD) EBC tobramycin concentration after inhaled administration of 170 mg tobramycin, corrected for sample dilution; D. Concentration–time curve of mean salbutamol concentration in plasma (±SD) after intravenous administration of 500 µg salbutamol (n=9); E. Concentration–time (Mean+SD) curve of plasma (solid line) and EBC (dotted line) tobramycin concentration after inhalation of 400 µg salbutamol; F. Concentration–time curve of mean (SD) EBC salbutamol concentration after inhaled administration of 400 µg salbutamol, corrected for sample dilution; Sample concentrations < LLOQ were estimated when possible or fixed on 50% of the lowest concentration. Dots represent individual measurements.



Salbutamol and tobramycin were detected in EBC after inhalation of 400µg salbutamol and 170 mg of tobramycin. Mean salbutamol concentrations in EBC after inhaled administration decreased below the LLOQ after approximately two hours, which is about 50% of the amount of time salbutamol is believed to have an effect on pulmonary function[24]. Furthermore, the highest EBC salbutamol concentrations were present during the absorption phase of salbutamol in plasma after inhaled administration. This indicates higher EBC concentrations are detected when there is active exchange of salbutamol between the inhaled air, ELF and plasma. However, variability in EBC was high as shown in the supplementary individual plots.

Mean EBC tobramycin concentration decreased sharply in the first hour, which correlated with the duration of the absorption phase in plasma. After the first hour, tobramycin concentrations remained stable or below the LLOQ,. This may correspond with the relatively long half-life (4.2 hours) of tobramycin in plasma after inhaled administration compared to the half-life after intravenous administration (2.3 hours). Possibly, ELF acts as a tobramycin reservoir enabling flip-flop kinetics, with tobramycin gradually diffusing towards the systemic circulation. This causes a longer elimination phase, while tobramycin also remains detectable in the diluted ELF that is EBC.

After all, EBC reflects a dilution of the ELF[8,19]. Variation in dilution factors of EBC samples has been a frequently hypothesized cause of variability. Urea has been proposed as marker for dilution in EBC[12,25]. We measured urea concentrations in EBC and calculated the estimated dilution. Mean D_{EBC} was slightly lower than reported in other studies[10,25,26] and varied greatly between and within subjects (Supplementary Figure S4), providing an explanation for the observed variability. The CV increased after correction, which did not meet our hypothesis regarding reduced variability.

EBC research in general has been plagued by high variability within and between subjects[9,19,30]. While several authors have reported methods to reduce variability[10–12,25], this too is poorly reproducible and has not resulted in the development of clear guidelines with standardized methods[19]. Consequently, EBC has not yet reached clinical practice. The fact that two commonly used respiratory drugs could not be detected in EBC for the majority of samples post inhalation appears to disqualify EBC as a matrix for pharmacokinetic analysis.

This study has limitations. Variation in inhalation technique could explain the lower than expected plasma concentrations of tobramycin, but also the variability in measured EBC concentrations[31]. In addition, although the EBC device contains a saliva trap, it cannot not be excluded that samples taken directly after administration represent oropharyngeal salbutamol deposition rather than lung pharmacokinetics. Furthermore, almost all sample concentrations were below or on the lower end of the calibration curve. Variability may therefore also result from assay variability as opposed to biological or device variability[7]. Finally, the method to incorporate estimated sample concentrations below the

LLOQ, as well as impute 50% of the lowest concentration when no estimation was possible, is a debated subject[32]. Nevertheless, the use of other methods would not change the main outcome of this study. A main strength of this study is the longitudinal analysis of EBC concentration and EBC dilution. The 4-way crossover design allowed for the calculation of the bioavailability of tobramycin and salbutamol on an individual level. Finally, EBC collection was conducted in a standardized manner in line with ERS guidelines[19].

Conclusion

In conclusion, salbutamol and tobramycin can be quantified in EBC after inhaled administration, especially during the plasma absorption phase, but not after intravenous administration. The high amount of variability of EBC drug concentrations seems to preclude its use for robust pharmacokinetic analysis and as such, we do not recommend its use in this area.

SUPPLEMENTARY DATA

Sup. Text S1	Sample analysis
Sup. Figure S1	Individual concentration-time curves-inhaled tobramycin
Sup. Figure S2	Individual concentration-time curves-inhaled salbutamol
Sup. Figure S3	Proportion of samples above lower limit of detection
Sup. Figure S4	Individual plots concentration-time curve urea and estimated dilution

REFERENCES

1 Lipworth BJ. Pharmacokinetics of inhaled drugs. Br J Clin Pharmacol. 1996;42(6):697-705.
2 Rizk ML, Zou L, Savic RM, Dooley KE. Importance of Drug Pharmacokinetics at the Site of Action. Clin Transl Sci. 2017;10(3):133-42.
3 Chandorkar G, Huntington JA, Gotfried MH, Rodvold KA, Umeh O. Intrapulmonary penetration of ceftolozane/tazobactam and piperacillin/tazobactam in healthy adult subjects. J Antimicrob Chemother. 2012;67(10):2463-9.
4 van Hasselt JG, Rizk ML, Lala M, Chavez-Eng C, Visser SA, Kerbusch T, *et al.* Pooled population pharmacokinetic model of imipenem in plasma and the lung epithelial lining fluid. Br J Clin Pharmacol. 2016;81(6):1113-23.
5 Khoubnasabjafari M, Rahimpour E, Samini M, Jouyban-Gharamaleki V, Chen L, Chen DH, *et al.* A new hypothesis to investigate bioequivalence of pharmaceutical inhalation products. Daru. 2019;27(1):517-24.
6 Sidorenko GI, Zborovskii EI, Levina DI. [Surface-active properties of the exhaled air condensate (a new method of studying lung function)]. Ter Arkh. 1980;52(3):65-8.
7 Davis MD, Montpetit A, Hunt J. Exhaled breath condensate: an overview. Immunol Allergy Clin North Am. 2012;32(3):363-75.
8 Effros RM, Peterson B, Casaburi R, Su J, Dunning M, Torday J, *et al.* Epithelial lining fluid solute concentrations in chronic obstructive lung disease patients and normal subjects. J Appl Physiol (1985). 2005;99(4):1286-92.
9 van Mastrigt E, de Jongste JC, Pijnenburg MW. The analysis of volatile organic compounds in exhaled breath and biomarkers in exhaled breath condensate in children-clinical tools or scientific toys? Clin Exp Allergy. 2015;45(7):1170-88.
10 Effros RM, Biller J, Foss B, Hoagland K, Dunning MB, Castillo D, *et al.* A simple method for estimating respiratory solute dilution in exhaled breath condensates. Am J Respir Crit Care Med. 2003;168(12):1500-5.
11 Reinhold P, Knobloch H. Exhaled breath condensate: lessons learned from veterinary medicine. J Breath Res. 2010;4(1):017001.
12 Esther CR, Jr., Boysen G, Olsen BM, Collins LB, Ghio AJ, Swenberg JW, *et al.* Mass spectrometric analysis of biomarkers and dilution markers in exhaled breath condensate reveals elevated purines in asthma and cystic fibrosis. Am J Physiol Lung Cell Mol Physiol. 2009;296(6):L987-93.
13 Khoubnasabjafari M, Ansarin K, Jouyban-Gharamaleki V, Panahi-Azar V, Hamidi S, Azarmir Z, *et al.* Methadone Concentrations in Exhaled Breath Condensate, Serum and Urine of Patients Under Maintenance Treatment. Iran J Pharm Res. 2017;16(4):1621-30.
14 Beck O, Sandqvist S, Bottcher M, Eriksen P, Franck J, Palmskog G. Study on the sampling of methadone from exhaled breath. J Anal Toxicol. 2011;35(5):257-63.
15 Gamez G, Zhu L, Disko A, Chen H, Azov V, Chingin K, *et al.* Real-time, in vivo monitoring and pharmacokinetics of valproic acid via a novel biomarker in exhaled breath. Chem Commun (Camb). 2011;47(17):4884-6.
16 Perl T, Carstens E, Hirn A, Quintel M, Vautz W, Nolte J, *et al.* Determination of serum propofol concentrations by breath analysis using ion mobility spectrometry. Br J Anaesth. 2009;103(6):822-7.
17 Hamidi S, Khoubnasabjafari M, Ansarin K, Jouyban-Gharamaleki V, Jouyban A. Chiral separation of methadone in exhaled breath condensate using capillary electrophoresis. Anal Methods-Uk. 2017;9(15):2342-50.
18 Jouyban A, Samadi A, Khoubnasabjafari M, Jouyban-Gharamaleki V, Ranjbar F. Amidosulfonic acid-capped silver nanoparticles for the spectrophotometric determination of lamotrigine in exhaled breath condensate. Microchim Acta. 2017;184(8):2991-8.
19 Horvath I, Barnes PJ, Loukides S, Sterk PJ, Hogman M, Olin AC, *et al.* A European Respiratory Society technical standard: exhaled biomarkers in lung disease. Eur Respir J. 2017;49(4).
20 R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013 [Available from: http://www.R-project.org/.
21 Moore A, Riddell K, Joshi S, Chan R, Mehta R. Pharmacokinetics of Salbutamol Delivered from the Unit Dose Dry Powder Inhaler: Comparison with the Metered Dose Inhaler and Diskus Dry Powder Inhaler. J Aerosol Med Pulm Drug Deliv. 2017;30(3):164-72.
22 Morgan DJ, Paull JD, Richmond BH, Wilson-Evered E, Ziccone SP. Pharmacokinetics of intravenous and oral salbutamol and its sulphate conjugate. Br J Clin Pharmacol. 1986;22(5):587-93.
23 Dalhoff A. Pharmacokinetics and pharmacodynamics of aerosolized antibacterial agents in chronically infected cystic fibrosis patients. Clin Microbiol Rev. 2014;27(4):753-82.
24 Filiz A, Ekinci E, Dikensoy O, Bulgur D, Oz M. Comparison of a single dose of aerosol salbutamol and fenoterol/ipratropium on bronchial asthmatic patients. Del Med J. 1994;66(10):549-52.
25 Effros RM, Hoagland KW, Bosbous M, Castillo D, Foss B, Dunning M, *et al.* Dilution of respiratory solutes in exhaled condensates. Am J Respir Crit Care Med. 2002;165(5):663-9.
26 Horvath I, Hunt J, Barnes PJ, Alving K, Antczak A, Baraldi E, *et al.* Exhaled breath condensate: methodological recommendations and unresolved questions. Eur Respir J. 2005;26(3):523-48.
27 McCafferty JB, Bradshaw TA, Tate S, Greening AP, Innes JA. Effects of breathing pattern and inspired air conditions on breath condensate volume, pH, nitrite, and protein concentrations. Thorax. 2004;59(8):694-8.
28 Kullmann T, Barta I, Antus B, Valyon M, Horvath I. Environmental temperature and relative humidity influence exhaled breath condensate pH. Eur Respir J. 2008;31(2):474-5.

29 Michal Gregus PD, Julia Lacna, Frantisek Foret, Petr Kuban. Study of various parameters that influence the content of exhaled breath condensate used in the diagnosis of gastroesophageal reflux disease. Hungarian Journal of Industry and Chemistry. 2018;46(1):29–33.
30 Horvath I. The exhaled biomarker puzzle: bacteria play their card in the exhaled nitric oxide–exhaled breath condensate nitrite game. Thorax. 2005;60(3):179–80.
31 Hindle M, Newton DA, Chrystyn H. Investigations of an optimal inhaler technique with the use of urinary salbutamol excretion as a measure of relative bioavailability to the lung. Thorax. 1993;48(6):607–10.
32 Keizer RJ, Jansen RS, Rosing H, Thijssen B, Beijnen JH, Schellens JHM, *et al.* Incorporation of concentration data below the limit of quantification in population pharmacokinetic analyses. Pharmacol Res Perspe. 2015;3(2).

PART V

DISCUSSION

CHAPTER 16

Trial@home in pediatrics - A framework for remote and non-invasive data collection in pediatric clinical trials

MD Kruizinga, FE Stuurman, GJA Driessen, AF Cohen

The current pediatric clinical trial paradigm

Primary endpoints in pediatric clinical trials are currently very similar to those in adult trials[1], and focus on quantifying or counting hard endpoints like mortality, hospital admissions and length of stay. Additionally, biochemical biomarkers in serum are often measured to assess drug effects on a biochemical level. The occurrence of mortality and hospital admissions is rare thanks to the improvements in clinical care that have occurred in the last century, and adopting these as primary endpoints in clinical trials gives disproportional weight to rare events which most patients will not experience. Conversely, length of stay for many clinical conditions is short, and this duration only captures a small part of the clinical recovery trajectory that patients must undergo.

Besides hard endpoints, clinic-based assessments and composite scores that are related to one or multiple components of the disease clinical assessments are often used as outcome. Although such scores are useful, such scores incorporate a subjective component of the observer in the outcome, while at the same time providing a mere snapshot of disease activity in a clinical setting. As discussed in **Chapter 1**, a shift from hard endpoints towards value-based endpoints is a natural progression of the current clinical trial paradigm.[2] Value-based endpoints represent outcomes that patients care about, are suitable to detect individual outcomes and that are measured frequently in a patients' natural environment. These new objective biomarkers should be able to quantify if a drug or health care intervention *helps* individual patients, as opposed to that the drug *works* on a biochemical or organ system level.

Although one manifestation of value-based thinking in clinical trials is the adoption of patient reported outcome measures (PROS) in the form of questionnaires, PROS may not be the solution in pediatrics. Adequate report of PROS is reliant on adequate disease-understanding and a homogenous perception on what kind of symptoms constitute severe disease. These outcomes are extremely difficult to assess in a reliable way for young patients. Researchers often include the parent in the disease-scoring process, but this creates an additional subjective factor between the actual symptom severity and the PRO. Although PROs can be valuable and demonstrate the capability to characterize the *subjective* burden of disease, inclusion of *objective* monitoring techniques needs to be considered as well.[3]

Implementing more value-based endpoints in clinical trials will likely improve the insights obtained from all trials, but the concept is especially relevant for the pediatric clinical trial. Trials in pediatric populations are difficult to conduct due to several reasons, such as high proportion of non-consent resulting in low recruitment rates. Many reasons for non-participation in a study are unavoidable, such as concerns about a study drug, randomization, or blinding.[4] However, in that case it is only logical that steps must be taken to overcome other barriers as much as possible. One of the major other reasons of parents for non-participation in a clinical trial is that 'study logistics were too complicated or difficult',[5] and a perception by parents that a study may interfere with standard care may be a factor as well.[4] Additionally, other studies cite inconvenience for the child or the parents,[6] a lack of time, and a lack of interest as a frequent reason for nonparticipation.[7] Finally, many researchers report that concerns about the number of blood draws and burdensome logistics in general are a major barrier as well.[8]

Taken together, the factors described above invariably lead to the conclusion that a new perspective on the pediatric clinical trial is necessary, and many of the current barriers and limitations could be surmounted by adopting a remote clinical trial paradigm.

Towards remote data collection in pediatric clinical trials

Although there are numerous locations to monitor children, the most logical location is the home and school for several reasons. In these locations the child is naturally most comfortable, and the environment resembles 'real-world conditions' by definition, which cannot be achieved in hospitals or clinical research units. Since it is evidently unfeasible on financial and logistical grounds to send trained research personnel to the homes of every individual subject frequently, monitoring in the home must be achieved using different means. Technology in the form of digital biomarkers, also termed digital endpoints in the context of clinical trials, is a possible solution.[9]

Before digital biomarkers can be implemented in clinical trials, and eventually in clinical care, they must be rigorously validated to ensure they are fit-for-purpose. In **Chapter 2** we describe a stepwise approach towards endpoint selection, technical validation, and clinical validation of digital endpoints.[10] This approach can be applied to most digital measurements, algorithms, and devices. Technical validation means investigating whether the devices measure that what is claimed and focuses on the variability within- and between devices, the general usability, and the data processing pipeline. **Chapters 3-5**, provide examples of the technical validation of a novel mobile spirometer and smartphone-based

algorithms to detect crying- and coughing in children.[11] Clinical validation means investigating the endpoint in the target population, in this case pediatric patient groups. Important biomarker characteristics that are investigated during this process are the tolerability, repeatability, difference between healthy and ill groups, correlation with traditional endpoints and the description of health events. In **Chapters 6–11**, the clinical validation process of the remote measurement of physical activity, heart rate, forced expiratory volume in 1 second (FEV1), sleep duration, electronic PRO's (EPRO's) and tests incorporated in CHDR's NeuroCart® system is described in several conditions, specifically pediatric asthma, -pneumonia, -preschool wheezing, -obesity, -sickle cell disease, and ARID1B-related intellectual disability. *Table 1* lists the progression of the clinical validation process for various digital endpoints that were investigated in the current work. For several candidate endpoints, all preparatory work has been performed to be able to progress to the final, and arguably most important, step of the validation process: confirming the new candidate endpoint is able to detect the effect of new effective treatments.

Implementation of non-invasive pharmacokinetics can supplement remote clinical trials

Digital endpoints are a novel way to register what is traditionally termed the pharmacodynamics of drugs: the effect of the drug on the patient. However, the other cornerstone of clinical trials, pharmacokinetics (what the patient does to the drug), is traditionally evaluated in hospitals and clinical research units as well. Determination of pharmacokinetics is, among others, necessary in the event of target concentrations, e.g., in the case of antibiotic therapy, and when a specific PK-PD relationship is to be investigated. Traditionally this is done in blood samples, which is not easily transferred to the home environment. Although some advances have been made using dried blood spots, this still requires blood sampling and causes discomfort.[12,13] Developing novel non-invasive methods to obtain pharmacokinetic information may allow pediatric clinical trials to move completely towards the home. To achieve this, the sampling matrix of choice, which is currently plasma, must change to be able to obtain samples in a non-invasive manner at home. One such matrix is saliva[14,15], and in **Chapter 12**, we describe a framework to estimate plasma concentrations in individual patients based on obtained saliva samples with nonlinear mixed effect models and Bayesian maximum a posteriori (MAP) optimization. The framework is applied in practice for clonazepam and gentamicin in **Chapter 13** and **Chapter 14**.

Table 1. Progress of the validation process of (digital) endpoints investigated in the current work

Candidate endpoint	Technical validation	Clinical validation				
		Tolerability	Repeatability and effect modifiers	Difference healthy-ill	Correlation traditional endpoints	Description of health events
Physical activity	Yes[1]	Yes	Yes	Yes	Yes	Yes
Heart rate	Yes[1]	Yes	Yes	Yes	Yes	Yes
Cry (session) duration	Yes	-	-	-	-	-
Cough count	Yes	-	-	-	-	-
Pulmonary function tests	Yes	Yes	Yes	Yes	Yes	Yes
Sleep duration	Yes[1]	Yes	Yes	Varying	No	No
Sleep depth / wakeup count	Yes[1]	Yes	Yes	Varying	No	No
NeuroCart®	Yes[1]	Yes	-	Yes	Yes	-
EPRO's	Yes[1]	Yes	-	Yes	NA	Yes

1 Not investigated in the current work. Legend: blue: validation criterion was met in one or more clinical conditions. Yellow: criterion fulfillment was inconclusive at this point. Red: validation criterion was not met in the current studies. White: not investigated at this point, more research necessary.

In the framework, one could conduct a preparatory study in (young) adult subjects to determine the saliva:plasma correlation for a particular compound, perhaps already during phase I studies. The saliva:plasma relationship is variable across compounds, and therefore difficult to predict due to multiple factors, such as molecule size, polarity, protein binding, and salivary flow.[16] Although the relationship seems stable across several age ranges for multiple compounds, e.g. voriconazole, phenytoin, phenobarbital and lamotrigine,[14,17] extrapolation of the saliva:plasma relationship from adults to pediatrics is a major assumption of the framework, and we believe more confirmatory research is needed before this can be implemented. However, if the assumption is confirmed to be valid, future studies investigating the pharmacokinetics of new compounds in children could be based solely on salivary concentrations with back-calculation of plasma concentration based on the known saliva:plasma relationship determined in adults.

Other matrices that do not involve blood sampling can be implemented as well, given that drug concentrations in the matrix are predictable and related to plasma concentrations.[12,18] In **Chapter 15**, we describe an exploratory study investigating the use of exhaled breath condensate (EBC) in the field of pulmonary pharmacokinetics. However, EBC exhibited extremely high variability for both salbutamol and tobramycin, and the approach did not appear a promising avenue for further research.[19]

Pediatric disease areas that could benefit from digital biomarkers in clinical trials

The current work shows that digital endpoints can be applied to both chronic and acute disease. The largest focus has been on the respiratory diseases in this dissertation, and our results indicate that endpoints such as physical activity and heart rate are ready to proceed towards preliminary implementation in pediatric trials investigating pulmonary diseases. However, remote monitoring can be applied to other chronic diseases, such as obesity and sickle cell disease.

Additionally, there are many other unexplored disease areas that could potentially benefit from adopting the proposed framework. For example, pediatric dermatology trials could benefit from a combination of EPROS and objective digital endpoints, such as photos taken by parents. These photos can be transmitted to researchers for an objective assessment of the condition of the skin. The field of child- and adolescent psychiatry could include objective physical activity-, heart rate- and blood pressure monitoring to assess (adverse) effects of treatment in Attention Deficit Hyperactivity Disorder (ADHD). Trials in pediatric diabetes could incorporate non-invasive digital glucose monitoring devices,[20] combined with physical activity monitoring that may relate to quality-of-life related disease activity, and a heart rate sensor to monitor the cardiovascular consequences of the disease, similarly to what was shown in pediatric obesity in **Chapter 9**. Rheumatology is another example where continuous and objective monitoring could be beneficial beyond the inflammatory markers and patient reported outcomes that trials currently rely on. Finally, trials in rare (genetic) diseases, such as muscular dystrophies or neurological syndromes, will also benefit from adoption of the proposed framework in clinical trials. Patients are even more vulnerable compared to their peers with, and home-based trials would be the least invasive and therefore preferable. Considering the orphan drugs developed for these conditions are extremely high-priced, it is of paramount importance that rare disease clinical trials become more value-based and demonstrate that an expensive treatment translates to real value for patients in addition to, for example, biochemical-or cellular improvement.[21]

In short, there are many possibilities to move clinical trials towards the home and incorporate endpoints that are directly related to the manifestations of the disease that lead to significant burdens in patients. In addition, similar applications of the framework may be possible in numerous other conditions not listed here.

Beyond the clinical trial - Application of digital endpoints and non-invasive TDM in clinical care

The focus of the current work was on the validation of novel methodologies in clinical trials. However, a clinically validated digital endpoint has many potential applications in clinical care as well. Value-based medicine has been gaining popularity since the introduction of the concept by Michael Porter in 2010.[22] In clinical care, like in clinical trials, the goal of implementing more value in clinical care leads to the conclusion that personalized outcome measures are a necessary component of treatment effect evaluation. Objective digital endpoints combined with electronic patient reported outcomes may be suitable for this task.[2]

However, there are other applications of home-monitoring systems in clinical care that can supplement the current telemedicine paradigm. Although more popular in rural countries compared to The Netherlands, telemedicine allows patients to consult their doctors from the comforts of their home via a video- or phone call.[23] This is both more comfortable for the patient and potentially more cost-effective for the health care system, but the drawback is that the physician has no access to information obtained via the physical examination or additional tests. This drawback can be mitigated using digital measurements. For example, imagine the following: the physician has been given access to the 'personal health dashboard' by a patient with, for example, cystic fibrosis with worsening disease-activity. On this dashboard, the physician can see that physical activity, adjusted for weather and other variables, has been decreasing gradually towards levels that are on the low end of reference values, while nocturnal heart rate has been increasing at the same speed. Together with other parameters, such as periodical patient reported outcomes and pulmonary function tests, the holistic overview provided by the dashboard will make clear that a change in treatment is necessary to prevent further worsening of the disease towards a pulmonary exacerbation, which can be initiated promptly.

Looking further ahead, the improvement of artificial intelligence may eventually enable algorithms to analyze multiple sources of data to predict symptom severity during the upcoming days.[24] When the algorithm detects a high probability of worsening disease-activity, an intervention can be suggested to provide timely symptom relief.

Besides home-monitoring of disease-activity, the non-invasive pharmacokinetics based on saliva samples described in this dissertation can be extended to the field of therapeutic drug monitoring (TDM).[25] Although TDM is relatively uncommon in pediatrics, it

enables for precision-dosing in drugs where a clear target range in plasma concentrations is available, for example, in the case of several antimicrobials and anti-epileptic drugs.

Conclusion

The remote pediatric clinical trial paradigm, consisting of digital endpoints and non-invasive pharmacokinetic sampling, has the potential to transform pediatric clinical trials and pediatric clinical care. The process towards implementation is challenging and can only proceed after a rigorous validation process. The current work provides a roadmap towards selection, validation, and implementation of digital endpoints, and describes preliminary steps taken for several candidates. The digital endpoints investigated in this work fulfill several validation criteria in a range of clinical conditions and, combined with non-invasive pharmacokinetics, may move the pediatric clinical trial completely towards the home.

REFERENCES

1 Green DJ, Burnham JM, Schuette P, Liu XI, Maas BM, Yao L, McCune SK, Chen J, van den Anker JN, Burckart GJ. Primary Endpoints in Pediatric Efficacy Trials Submitted to the US FDA. J Clin Pharmacol 2018;58(7):885-890.

2 Kruizinga MD, Stuurman FE, Groeneveld GJ, Cohen AF. The Future of Clinical Trial Design: The Transition from Hard Endpoints to Value-Based Endpoints. 2019:371-397.

3 Kahan BC, Doré CJ, Murphy MF, Jairath V. Bias was reduced in an open-label trial through the removal of subjective elements from the outcome definition. J Clin Epidemiol 2016;77:38-43.

4 Hoberman A, Shaikh N, Bhatnagar S, Haralam MA, Kearney DH, Colborn DK, Kienholz ML, Wang L, Bunker CH, Keren R, *et al.* Factors that influence parental decisions to participate in clinical research: Consenters vs nonconsenters. JAMA Pediatr 2013;167(6):561-566.

5 Greenberg RG, Gamel B, Bloom D, Bradley J, Jafri HS, Hinton D, Nambiar S, Wheeler C, Tiernan R, Smith PB, *et al.* Parents' perceived obstacles to pediatric clinical trial participation: Findings from the clinical trials transformation initiative. Contemp Clin Trials Commun 2018;9(September 2017):33-39.

6 Institute of Medicine (US) Committee on Clinical Research Involving Children; Field MJ, Behrman RE, editors. Ethical Conduct of Clinical Research Involving Children. Washington (DC): National Academies Press (US); 2004. 5, Understanding and Agreeing to Ch.

7 Vermaire JH, Van Loveren C, Poorterman JHG, Hoogstraten J. Non-participation in a randomized controlled trial: The effect on clinical and non-clinical variables. Caries Res 2011;45(3):269-274.

8 Greenberg RG, Corneli A, Bradley J, Farley J, Jafri HS, Lin L, Nambiar S, Noel GJ, Wheeler C, Tiernan R, *et al.* Perceived barriers to pediatrician and family practitioner participation in pediatric clinical trials: Findings from the Clinical Trials Transformation Initiative. Contemp Clin Trials Commun 2018;9(September 2017):7-12.

9 Boehme P, Hansen A, Roubenoff R, Scheeren J, Herrmann M, Mondritzki T, Ehlers J, Truebel H. How soon will digital endpoints become a cornerstone for future drug development? Drug Discov Today 2019;24(1):16-19.

10 Kruizinga MD, Stuurman FE, Exadaktylos V, Doll RJ, Stephenson DT, Groeneveld GJ, Driessen GJA, Cohen AF. Development of Novel, Value-Based, Digital Endpoints for Clinical Trials: A Structured Approach Toward Fit-for-Purpose Validation. Pharmacol Rev 2020;72 (4)(October):899-909.

11 Kruizinga MD, Essers E, Stuurman FE, Zhuparris A, van Eik N, Janssens HM, Groothuis I, Sprij AJ, Nuijsink M, Cohen AF, *et al.* Technical validity and usability of a novel smartphone-connected spirometry device for pediatric patients with asthma and cystic fibrosis. Pediatr Pulmonol 2020;(June):2463-2470.

12 Jager NGL, Rosing H, Schellens JHM, Beijnen JH. Procedures and practices for the validation of bioanalytical methods using dried blood spots: A review. Bioanalysis 2014;6(18):2481-2514.

13 Kloosterboer SM, van Eijk E, van Dijk M, Dieleman GC, Hillegers MHJ, van Gelder T, Koch BC, Dierckx B. Feasibility of Dried Blood Spots in Children with Behavioral Problems. Ther Drug Monit 2020;42(4):648-651.

14 Hutchinson L, Sinclair M, Reid B, Burnett K, Callan B. A descriptive systematic review of salivary therapeutic drug monitoring in neonates and infants. Br J Clin Pharmacol 2018;84(6):1089-1108.

15 Patsalos PN, Berry DJ. Therapeutic Drug Monitoring of Antiepileptic Drugs by Use of Saliva. Ther Drug Monit 2013;35(1).

16 Haeckel R. Factors Influencing the Saliva/Plasma Ratio of Drugs. Ann N Y Acad Sci 1993;694(1):128-142.

17 Michael C, Bierbach U, Frenzel K, Lange T, Basara N, Niederwieser D, Mauz-Körholz C, Preiss R. Determination of saliva trough levels for monitoring voriconazole therapy in immunocompromised children and adults. Ther Drug Monit 2010;32(2):194-199.

18 Grumetto L, Cennamo G, Del Prete A, La Rotonda MI, Barbato F. Pharmacokinetics of cetirizine in tear fluid after a single oral dose. Clin Pharmacokinet 2002;41(7):525-531.

19 Kruizinga MD, Birkhoff WAJ, Esdonk MJ, Klarenbeek NB, Cholewinski T, Nelemans T, Dröge MJ, Cohen AF, Zuiker RGJA. Pharmacokinetics of intravenous and inhaled salbutamol and tobramycin: An exploratory study to investigate the potential of exhaled breath condensate as a matrix for pharmacokinetic analysis. Br J Clin Pharmacol 2020;86(1):175-181.

20 Ólafsdóttir AF, Attvall S, Sandgren U, Dahlqvist S, Pivodic A, Skrtic S, Theodorsson E, Lind M. A Clinical Trial of the Accuracy and Treatment Experience of the Flash Glucose Monitor FreeStyle Libre in Adults with Type 1 Diabetes. Diabetes Technol Ther 2017;19(3):164-172.

21 Cox GF. The art and science of choosing efficacy endpoints for rare disease clinical trials. Am J Med Genet Part A 2018;176(4):759-772.

22 Porter M. What is Value in Healthcare. NEJM 2010:2477-2481.

23 Greiwe J, Nyenhuis SM. Wearable Technology and How This Can Be Implemented into Clinical Practice. Curr Allergy Asthma Rep 2020;20(8).

24 Johnson A, Yang F, Gollarahalli S, Banerjee T, Abrams D, Jonassaint J, Jonassaint C, Shah N. Use of mobile health apps and wearable technology to assess changes and predict pain during treatment of acute pain in sickle cell disease: Feasibility study. JMIR mHealth uHealth 2019;7(12).

25 Soldin OP, Soldin SJ. Review: Therapeutic drug monitoring in pediatrics. Ther Drug Monit 2002;24(1):1-8.

Trial@home bij kinderen - Een raamwerk voor non-invasieve dataverzameling op afstand voor klinische trials binnen de kindergeneeskunde

MD Kruizinga, FE Stuurman, GJA Driessen, AF Cohen

Het huidige paradigma rond klinische trials binnen de pediatrie

De primaire eindpunten die worden gekozen bij klinisch onderzoek met kinderen zijn op dit moment zeer vergelijkbaar met de primaire eindpunten in klinisch onderzoek met volwassenen. De focus ligt op het kwantificeren of tellen van 'harde eindpunten', zoals mortaliteit, ziekenhuisopnames en opnameduur. Daarnaast worden biochemische biomarkers in het bloed gemeten om de medicatie-effecten op biologisch vlak vast te stellen. Gelukkig zijn het optreden van zowel mortaliteit als ziekenhuisopnames dankzij de grote verbeteringen in de geneeskunde steeds zeldzamer geworden bij veel aandoeningen. Door een parameter als dit toch te kiezen als belangrijkste eindpunt in klinisch onderzoek wordt er op een disproportionele manier gewicht gegeven aan deze zeldzame gebeurtenissen die een overgroot gedeelte van de patiënten nooit mee zal maken. Daarnaast bevat een uitkomstmaat als opnameduur slechts een zeer kort onderdeel van het hersteltraject die zich tegenwoordig vooral in de thuissituatie voltrekt. Naast deze 'harde eindpunten' worden vaak klinische beoordelingen of samengestelde scores gebruikt om de ziektelast te kwantificeren. Hoewel zulke scores extreem nuttig kunnen zijn, incorporeren zij automatisch een subjectieve component van de observator in de uitkomst, en zijn zij tegelijkertijd slechts een momentopname van de ziekte-ernst in een klinische omgeving. In **Hoofdstuk 1** wordt een omslag van het paradigma rondom klinisch onderzoek voorgesteld. In plaats van de focus op harde eindpunten te leggen, moet deze verplaatst worden naar 'value-based' eindpunten.

Value-based eindpunten zijn uitkomsten die patiënten belangrijk vinden, die geschikt zijn om individuele uitkomsten te meten en die hoogfrequent geregistreerd kunnen worden in de natuurlijke omgeving van de patiënt. Deze nieuwe objectieve biomarkers moeten kunnen vaststellen of een medicijn of interventie de individuele patiënt *helpt*, in tegenstelling tot het vaststellen of een medicijn *werkt* op biochemisch niveau.

Hoewel het gebruik van vragenlijsten (zogeheten patient reported outcome measures (PROS)) één manifestatie van value-based denken in klinisch onderzoek is, zijn deze mogelijk niet de enige oplossing binnen de kindergeneeskunde. Adequaat invullen van vragenlijsten is afhankelijk van een homogene perceptie van de verschijnselen die behoren bij de ziekte, en dit is moeilijk voor (jonge) kinderen. Onderzoekers vragen daarnaast vaak de ouders om de vragenlijsten in te vullen, maar dit voegt een additionele *subjectieve* factor toe aan het invulproces. Hoewel PROS waardevol kunnen zijn en een uitstekende manier zijn om de *subjectieve* last van de ziekte te kwantificeren, is het noodzakelijk dat *objectieve* manieren om patiënten te volgen ook overwogen worden.

Het implementeren van value-based eindpunten in klinische trials verbeteren mogelijk de inzichten die verzameld worden in alle klinische trials, maar het concept heeft specifieke relevantie binnen de kindergeneeskunde. Klinisch onderzoek bij kinderen is moeilijk vanwege meerdere redenen, zoals het feit dat veel ouders deelname weigeren, waardoor het behalen van adequate groepsgroottes een grote uitdaging is. Veel redenen die ouders opvoeren voor het weigeren van toestemming tot deelname zijn moeilijk te veranderen, zoals zorgen om de studiemedicatie, randomisatie, of blindering. Echter, dit betekent logischerwijs dat alle andere barrières zoveel mogelijk verkleind moeten worden. Deze andere barrières omvatten vaak logistieke problemen, zoals werk van ouders en schoolroosters van kinderen. Daarnaast spelen de perceptie dat klinische studies interfereren met de standaardzorg, een gebrek aan interesse in de studie-uitkomsten en zorgen om de hoeveelheid bloedafnames ook een rol bij non-participatie aan klinisch onderzoek.

Samenvattend leiden de bovenstaande barrières en limitaties voor klinisch onderzoek bij kinderen tot de conclusie dat een nieuw perspectief noodzakelijk is om klinische trials bij kinderen te verbeteren. Een mogelijke oplossing voor veel problemen zou het op afstand uitvoeren van klinisch onderzoek bij kinderen zijn.

De weg naar gegevensverzameling op afstand voor klinisch onderzoek bij kinderen

Uiteraard zijn er ontelbare locaties om gegevens te verzamelen bij kinderen, maar de thuissituatie en de school zijn de meest logische locaties vanwege meerdere redenen. Hier is het kind het meest op zijn/haar gemak, en de 'real-world' omstandigheden waarin metingen in dit geval uitgevoerd worden, kunnen moeilijk geëmuleerd worden in ziekenhuizen of onderzoekscentra. Aangezien het zowel financieel als logistiek niet altijd haalbaar is om getraind personeel regelmatig langs de woning van alle deelnemers te laten gaan, moet de gegevensverzameling plaatsvinden op een andere manier, bijvoorbeeld met behulp van technologie in de vorm van digitale biomarkers, die ook wel digitale eindpunten worden genoemd in de context van klinisch onderzoek.

Digitale biomarkers moeten rigoureus worden gevalideerd voordat zij kunnen worden geïmplementeerd in klinisch onderzoek. In **Hoofdstuk 2** beschrijven wij een stapsgewijze benadering voor de selectie, technische validatie en klinische validatie van digitale

biomarkers. Deze benadering kan worden toegepast op de meeste digitale metingen, algoritmes of apparaten. Tijdens de technische validatie wordt onderzocht of het apparaat datgene meet wat geclaimd of verwacht wordt. Daarnaast wordt de variabiliteit binnen en tussen apparaten onderzocht, alsook het gebruiksgemak en de route die data aflegt voordat deze bij de data analist arriveert. **Hoofdstuk 3, 4 en 5** geven voorbeelden van het technische validatieproces van een nieuwe mobiele spirometer en van twee algoritmes die huilen en hoesten in kinderen kunnen detecteren met behulp van een smartphone. Als een nieuwe biomarker technisch adequaat blijkt, volgt klinische validatie. Dit houdt in dat onderzocht wordt of de biomarker in staat is om ziekteactiviteit te kwantificeren bij patiënten. Belangrijke factoren die tijdens dit proces geëvalueerd worden zijn de tolereerbaarheid, het verschil tussen zieke- en gezonde kinderen, de correlatie tussen de nieuwe meting en traditionele metingen en het vermogen om gebeurtenissen zoals exacerbaties te beschrijven. Ook worden variabelen die de meting kunnen beïnvloeden, zoals weers- en seizoensinvloeden onderzocht. **Hoofdstukken 6 tot en met 11** beschrijven het klinisch validatieproces van fysieke activiteit, hartslagfrequentie, FEV1, slaapduur, elektronische vragenlijsten en tests behorend tot het NeuroCart® systeem bij kinderen met verschillende aandoeningen (astma, pneumonie, bronchiale hyperreactiviteit, obesitas, sikkelcelziekte en ARID1B-gerelateerde verstandelijke beperking). *Tabel 1* laat de voortgang van het klinisch validatieproces van deze metingen zien. Voor meerdere kandidaat eindpunten is het voorbereidende werk verricht dat noodzakelijk is vóórdat het laatste, en belangrijkste, validatiecriterium getoetst wordt: bevestigen dat het nieuwe eindpunt in staat is om de effecten van nieuwe effectieve behandelingen te detecteren.

Non-invasieve farmacokinetische metingen vullen klinisch onderzoek op afstand aan

Digitale eindpunten zijn een nieuwe manier om de farmacodynamiek (het effect van geneesmiddelen op de patiënt) te kwantificeren. Echter, de andere hoeksteen van klinisch onderzoek, de farmacokinetiek (hoe verwerkt het lichaam het geneesmiddel), wordt ook traditioneel in ziekenhuizen of klinische onderzoeksinstellingen uitgevoerd. Het vaststellen van de farmacokinetiek is onder andere noodzakelijk als er sprake is van een specifieke geneesmiddelenconcentratie die bereikt moet worden, bijvoorbeeld bij antibiotica, en wanneer er een bepaalde relatie tussen farmacokinetiek- en dynamiek wordt onderzocht.

Traditioneel wordt farmacokinetisch onderzoek met behulp van bloedmonsters verricht, wat niet eenvoudig is uit te voeren in de thuissituatie. Hoewel 'dried blood spots', bloedmonsters die met behulp van een vingerprik worden verzameld, steeds vaker worden gebruikt, zorgt deze methode nog steeds voor pijn bij kinderen. Nieuwe non-invasieve methoden om farmacokinetische informatie bij kinderen te verzamelen zijn echter belangrijk om klinisch onderzoek bij kinderen volledig in de thuissituatie te kunnen laten plaatsvinden. Om dit te bewerkstelligen moet een nieuwe matrix gebruikt worden, zoals speeksel.

Tabel 1. Voortgang van het validatieproces van (digitale) eindpunten die in dit proefschrift onderzocht zijn

Kandidaat eindpunt	Technische validatie	Clinical validation				
		Tolereerbaarheid	Reproduceerbaarheid en factoren van invloed	Verschil Gezond-ziek	Correlatie traditionele eindpunten	Beschrijving 'health events'
Fysieke activiteit	Ja[1]	Ja	Ja	Ja	Ja	Ja
Hartfrequentie	Ja[1]	Ja	Ja	Ja	Ja	Ja
Huil frequentie	Ja	–	–	–	–	–
Hoest frequentie	Ja	–	–	–	–	–
Longfunctie	Ja	Ja	Ja	Ja	Ja	Ja
Slaapduur	Ja[1]	Ja	Ja	Variabel	Nee	Nee
Slaapdiepte	Ja[1]	Ja	Ja	Variabel	Nee	Nee
NeuroCart®	Ja[1]	Ja	–	Ja	Ja	–
EPROS	Ja[1]	Ja	–	Ja	NA	Ja

1 Niet onderzocht in dit proefschrift. / Legenda: blauw: validatiecriterium is behaald in een of meerdere aandoeningen. Geel: behalen van criterium is nog onduidelijk. Rood: validatiecriterium is niet behaald in de huidige studies. Wit: tot op heden niet onderzocht. Meer onderzoek noodzakelijk.

In **Hoofdstuk 12** beschrijven we een methode om met behulp van non-lineaire mixed effect modellen en Bayesian maximum a posteriori (MAP) optimalisatie plasmaconcentraties in individuele patiënten te schatten aan de hand van speeksel concentraties. Deze methode wordt vervolgens in de praktijk toegepast bij clonazepam (**Hoofdstuk 13**) en gentamicine (**Hoofdstuk 14**). Met deze methode kunnen onderzoekers tijdens een voorbereidende (Fase 1) studie in (jong)volwassen proefpersonen de relatie tussen speeksel en plasmaconcentraties vaststellen. Deze relatie varieert in grote mate tussen verschillende geneesmiddelen door een aantal oorzaken, zoals molecuulgrootte, polariteit, eiwitbinding en speekselflow. Hoewel de relatie stabiel lijkt te zijn gedurende het leven voor verschillende geneesmiddelen, zoals voriconazole, fenobarbital en lamotrigine, is de extrapolatie van de speeksel:plasma relatie van volwassenen naar kinderen een aanname binnen

het raamwerk. Er is meer onderzoek nodig naar deze relatie, maar als deze aanname valide blijkt, kan de farmacokinetiek van nieuwe geneesmiddelen bij kinderen in toekomstige studies volledig plaatsvinden op basis van speekselconcentraties.

Behalve speeksel zijn ook andere matrices die op een non-invasieve manier kunnen worden verkregen mogelijk geschikt bij kinderen, met de voorwaarde dat medicatieconcentraties in deze matrix voorspelbaar en gerelateerd aan plasmaconcentraties zijn. In **Hoofdstuk 15** beschrijven wij een exploratieve studie die het gebruik van *exhaled breath condensate* (EBC) onderzocht om de pulmonale farmacokinetiek van geneesmiddelen te beschrijven. Helaas was de geneesmiddelenconcentratie in EBC extreem variabel, waardoor deze aanpak niet geschikt bleek voor deze toepassing.

Gebieden binnen de kindergeneeskunde waar digitale biomarkers meerwaarde kunnen bieden

Dit proefschrift laat zien dat digitale eindpunten bij zowel acute- als chronische ziekte kunnen worden toegepast. De focus in dit proefschrift lag op de respiratoire aandoeningen, en onze resultaten geven aan dat eindpunten zoals fysieke activiteit en hartslag mogelijk geschikt zijn om te implementeren in klinisch onderzoek bij kinderen met asthma en cystische fibrose. De onderzoeken bij kinderen met obesitas en sikkelcelziekte laten zien dat thuismonitoring ook bij andere chronische aandoeningen gebruikt kunnen worden, maar er zijn daarnaast vele andere gebieden die hier mogelijk baat bij kunnen hebben. Zo kunnen klinische onderzoeken binnen de dermatologie baat hebben bij EPROS in combinatie met digitale eindpunten in de vorm van fotografische beelden van de huid. Klinisch onderzoek binnen de kinderpsychiatrie zou gebruik kunnen maken van objectieve metingen van de fysieke activiteit, hartslag en bloeddruk om de positieve en negatieve effecten van behandeling voor Attention Deficit Hyperactivity Disorder (ADHD) te monitoren. Onderzoek bij kinderen met diabetes met non-invasieve glucosemeters is mogelijk, eventueel in combinatie met andere parameters die gerelateerd zijn aan cardiovasculaire gezondheid en kwaliteit van leven (**Hoofdstuk 9**). De kinderreumatologie is een ander voorbeeld waar objectieve thuismonitoring aanvullend waarde kan hebben, naast de inflammatoire markers en de vragenlijsten waar reumatologie-onderzoek vooralsnog vooral op leunt. Deze opsomming is lang niet compleet: er zijn in vrijwel elke specialisme binnen de kindergeneeskunde mogelijkheden om klinisch onderzoek naar de thuissituatie te verplaatsen door middel van digitale eindpunten.

Applicaties van digitale eindpunten en non-invasieve therapeutic drug monitoring (TDM) in klinische zorg

Hoewel de focus in dit proefschrift ligt op de validatie van nieuwe methodologie in klinisch onderzoek, hebben klinisch gevalideerde digitale eindpunten ook potentiele toepassingen in de klinische zorg. Er is sprake van een toenemende populariteit van 'value-based medicine' sinds de introductie door Michael Porter in 2010. Bij het toepassen van dit concept in de klinische zorg komt echter de realisatie dat voor de implementatie van value-based medicine ook gepersonaliseerde uitkomstmaten van groot belang zijn om behandeleffecten te evalueren. Objectieve digitale eindpunten, gecombineerd met PROS kunnen hiervoor geschikt zijn.

Er zijn echter ook andere toepassingen van thuismonitoringsystemen die de huidige praktijken voor zorg op afstand kunnen verrijken. Hoewel zorg op afstand populairder is in landen die minder stedelijk zijn dan Nederland, biedt het patiënten de kans om hun dokter te spreken vanuit het comfort van hun eigen huis. Dit is mogelijk prettiger voor de patiënt en is mogelijk ook kosteneffectief ten opzichte van poliklinische bezoeken. Een nadeel is dat de arts geen toegang heeft tot informatie die verkregen wordt door middel van lichamelijk en aanvullend onderzoek, en digitale metingen kunnen dit nadeel mogelijk deels mitigeren. Stel u het volgende voor: een arts heeft van een patiënt met, bijvoorbeeld, cystische fibrose met een naderende exacerbatie, toegang gekregen tot een 'persoonlijk gezondheidsdashboard'. Hij ziet dat de fysieke activiteit, gecorrigeerd voor weersinvloeden en andere factoren van invloed, de afgelopen maanden verminderd is en nu aan de ondergrens van normaal zit. Tegelijkertijd is de nachtelijke hartslag de laatste weken gradueel verhoogd. Tegelijkertijd is er een vergelijkbare trend zichtbaar in de wekelijkse symptoomscore en longfunctie tests. Tezamen vormen deze parameters een holistische helikopterview van de patiënt die duidelijk maakt dat een verandering in behandeling noodzakelijk is om verdere achteruitgang tot een fulminante exacerbatie te voorkomen en dit kan direct na het telefonisch consult gestart worden. In de iets verdere toekomst kunnen de verbeteringen in kunstmatige intelligentie de mogelijkheid bieden om verschillende databronnen te analyseren om ziekteactiviteit in de komende dagen te voorspellen. Als het algoritme een hoge kans op verergering van de ziekteactiviteit detecteert, kan automatisch een interventie of consult worden voorgesteld om op tijd in te grijpen voordat een bezoek aan de spoedeisende hulp noodzakelijk is. Non-invasieve farmacokinetische metingen met speeksel kunnen daarnaast ook toegepast worden in de zorg in de vorm

van TDM bij verschillende geneesmiddelen. Hoewel TDM op dit moment slechts bij weinig medicamenten wordt toegepast, kan het zorgen voor precisie-dosering voor medicijnen waar een duidelijke doelconcentratie is, bijvoorbeeld bij antibiotica en anti-epileptica.

Conclusie

Digitale eindpunten en non-invasieve faramcokinetische metingen, waarvan er enkelen zijn beschreven in dit proefschrift, hebben de potentie om zowel klinisch onderzoek als klinische zorg bij kinderen te transformeren. Het implementatieproces is echter zeer uitdagend, en moet alleen voorgezet worden na een rigoureus validatieproces. Dit proefschrift biedt een routekaart voor de selectie, validatie en implementatie van digitale eindpunten, en beschrijft de eerste stappen die gezet zijn voor enkele kandidaat-eindpunten. Deze eindpunten lijken reeds aan verschillende validatiecriteria te voldoen voor meerdere aandoeningen, en als zulke digitale metingen gecombineerd worden met non-invasieve farmacokinetische metingen kan klinisch onderzoek bij kinderen in de toekomst volledig naar de thuissituatie verplaatst worden.

APPENDICES

CURRICULUM VITAE

Matthijs Derk Kruizinga was born on the 18TH of September 1991 in 's-Gravenzande, the Netherlands. After graduating pre-university education in 2010 at the Interconfessionele Scholengemeenschap Westland (ISW), he studied Biomedical Sciences at Leiden University for 2 years, before transferring to Medicine in 2012. During his clinical rotation at the department of pediatrics in Bronovo Hospital, the Hague, he discovered his passion for pediatrics, after which he performed his research internship in pediatrics at the department of pediatric immunology and stem cell transplantation at Leiden University Medical Centre under the supervision of dr. Robbert G.M. Bredius. After graduating as a physician in 2017, he returned to the Bronovo hospital as a resident physician. In 2018, he started as a PhD candidate at the Centre for Human Drug Research and Juliana Children's hospital under supervision of prof. dr. Adam F. Cohen, prof. dr. Gertjan J.A. Driessen, and dr. Rik E. Stuurman, resulting in this thesis. During 2021, Matthijs was employed as a resident physician (ANIOS) at the Juliana Children's Hospital, and will join the pediatric residency program of the Leiden University Medical Centre in January 2022.

LIST OF PUBLICATIONS

Kruizinga MD, Bresters D, Smiers FJ, Lankester AC, Bredius RGM. The use of intravenous pentamidine for the prophylaxis of Pneumocystis pneumonia in pediatric patients. *Pediatr Blood Cancer 2017;64(8).*

Kruizinga MD, Kuijpers TW, Alders M, Kindermann A, Bredius RGM. Immuundysregulatie, polyendocrinopathie, enteropathie, X-gebonden syndroom. Kliniek, diagnostiek en behandeling. *Ned Tijdschr Allerg astma, 2017;1779-84.*

Kruizinga MD, van Tol MJD, Bekker V, Netelenbos T, Smiers FJ, Bresters D, Jansen-Hoogendijk AM, van Ostaijen-ten Dam MM, Kollen WJW, Zwaginga JJ, et al. Risk Factors, Treatment, and Immune Dysregulation in Autoimmune Cytopenia after Allogeneic Hematopoietic StemCell Transplantation in Pediatric Patients. *Biol Blood Marrow Transplant 2018;24(4).*

Kruizinga MD, Stuurman FE, Groeneveld GJ, Cohen AF. The Future of Clinical Trial Design: The Transition from Hard Endpoints to Value-Based Endpoints. *2019:371-397.*

Kruizinga MD, Birkhoff WAJ, Esdonk MJ, Klarenbeek NB, Cholewinski T, Nelemans T, Dröge MJ, Cohen AF, Zuiker RGJA. Pharmacokinetics of intravenous and inhaled salbutamol and tobramycin: An exploratory study to investigate the potential of exhaled breath condensate as a matrix for pharmacokinetic analysis. *Br J Clin Pharmacol 2020;86(1):175-181.*

Kruizinga MD, Zuiker RGJA, Sali E, de Kam ML, Doll RJ, Groeneveld GJ, Santen GWE, Cohen AF. Finding Suitable Clinical Endpoints for a Potential Treatment of a Rare Genetic Disease: the Case of ARID1B. *Neurotherapeutics 2020.*

Kruizinga MD, Essers E, Stuurman FE, Zhuparris A, van Eik N, Janssens HM, Groothuis I, Sprij AJ, Nuijsink M, Cohen AF, et al. Technical validity and usability of a novel smartphone-connected spirometry device for pediatric patients with asthma and cystic fibrosis. *Pediatr Pulmonol 2020;(June):2463-2470.*

Kruizinga MD, Stuurman FE, Exadaktylos V, Doll RJ, Stephenson DT, Groeneveld GJ, Driessen GJA, Cohen AF. Development of Novel, Value-Based, Digital Endpoints for Clinical Trials: A Structured Approach Toward Fit-for-Purpose Validation. *Pharmacol Rev 2020;72 (4) (October):899-909.*

Kruizinga MD, Heide N van der, Moll A, Zhuparris A, Yavuz Y, Kam ML de, Stuurman FE, Cohen AF, Driessen GJA. Towards remote monitoring in pediatric care and clinical trials-Tolerability, repeatability and reference values of candidate digital endpoints derived from physical activity, heart rate and sleep in healthy children. *PLoS One 2021;16(1):e0244877.*

Kruizinga MD, Peeters D, van Veen M, van Houten M, Wieringa J, Noordzij JG, et al. The impact of lockdown on pediatric ED visits and hospital admissions during the COVID19 pandemic: a multicenter analysis and review of the literature. *Eur J Pediatr. 2021 Mar;1-9.*

ZhuParris A, **Kruizinga MD**, Gent M van, Dessing E, Exadaktylos V, Doll RJ, et al. Development and Technical Validation of a Smartphone-Based Cry Detection Algorithm. Front Pediatr. 2021;9:262. *Available from: https://www.frontiersin.org/article/10.3389/fped.2021.651356*

Kruizinga MD, Moll A, Zhuparris A, Ziagkos D, Stuurman FE, Nuijsink M, Cohen AF, Driessen GJA. Postdischarge Recovery after Acute Pediatric Lung Disease Can Be Quantified with Digital Biomarkers. Respiration. *2021 May 18:1-10. doi:10.1159/000516328. Epub ahead of print. PMID:34004601.*

Kruizinga MD, Stuurman FE, Driessen GJA, et al. Theoretical Performance of Nonlinear Mixed-Effect Models Incorporating Saliva as an Alternative Sampling Matrix for Therapeutic Drug Monitoring in Pediatrics: A Simulation Study. *Therapeutic Drug Monitoring. 2021 Aug;43(4):546-554. DOI:10.1097/ftd.0000000000000904. PMID:34250966.*

Samb A, **Kruizinga MD**, Tallahi Y, van Esdonk M, van Heel W, Driessen G, Bijleveld Y, Stuurman R, Cohen A, van Kaam A, de Haan TR, Mathôt R. Saliva as a sampling matrix for therapeutic drug monitoring of gentamicin in neonates: A prospective population pharmacokinetic and simulation study. *Br J Clin Pharmacol. 2021 Oct 8. doi: 10.1111/bcp.15105. Epub ahead of print. PMID:34625981.*

Kruizinga MD, Zuiker RGJA, Bergmann KR, Egas AC, Cohen AF, Santen GWE, van Esdonk MJ. Population pharmacokinetics of clonazepam in saliva and plasma: Steps towards noninvasive pharmacokinetic studies in vulnerable populations. *Br J Clin Pharmacol. 2021 Nov 22. doi: 10.1111/bcp.15152. Epub ahead of print. PMID:34811788.*

Kruizinga MD, Essers E, Stuurman FE, Yavuz Y, de Kam ML, Zhuparris A, Janssens HM, Groothuis I, Sprij AJ, Nuijsink M, Cohen AF, Driessen GJA. Clinical validation of digital biomarkers for pediatric patients with asthma and cystic fibrosis-Potential for clinical trials and clinical care. *Eur Respir J. 2021 Dec 9:2100208. doi:* 10.1183/13993003.00208-2021. Epub ahead of print. *PMID*: 34887326.

Kruizinga MD, Zhuparris A, Dessing E, Krol FJ, Sprij AJ, Doll RJ, Stuurman FE, Exadaktylos V, Driessen GJ, Cohen AF. Development and technical validation of a smartphone-based pediatric cough detection algorithm. *Pediatr Pulmonol. 2021 Dec 29. doi: 10.1002/ppul.25801. Epub ahead of print. PMID:34964557.*

DANKWOORD

Dit proefschrift is tot stand gekomen dankzij de hulp van een enorme hoeveelheid collegae die ik mijn dank verschuldigd ben. In deze sectie heb ik geprobeerd de verschillende bijdragen uiteen te zetten. Hierbij wil ik van tevoren opmerken dat onderstaande uiteenzetting ongetwijfeld niet compleet is, en dat ik ook de bijdragen van niet genoemde personen zeker gewaardeerd heb. Uiteindelijk is research, bij het CHDR nog meer dan op andere plekken, een enorme team-effort waarbij het bijna merkwaardig is dat er maar één persoon de credits krijgt in de vorm van een titel.

Ik wil graag beginnen met het bedanken van mijn promotieteam: Adam Cohen, Gertjan Driessen en Rik Stuurman. Bedankt voor de vrijheid die jullie mij de afgelopen jaren hebben geboden om de projecten op te zetten, uit te voeren, te analyseren en op te schrijven. Een leerzame combinatie die niet iedereen gegeven is, maar die zeker tot een minder bevredigend resultaat zou zijn gekomen zonder jullie bijsturing en begeleiding.

Helaas stonden de promotiereglementen mij niet toe om een 4e en 5e copromotor aan het team toe te voegen, maar als dit wel had gemogen, was er geen twijfel geweest over die dit had moeten zijn. Rob Zuiker, bedankt voor de begeleiding bij het opzetten en uitvoeren van de studies die zich binnen het CHDR afspeelden. Michiel van Esdonk, bedankt voor het inwijden van deze eenvoudige arts in de wereld van non-lineaire populatie pk-modellen. Daaraan gerelateerd verdienen Marieke de Kam en Yalçin Yavuz alle lof voor het (alsmaar) geduldig uitleggen van de statistische methoden die in dit proefschrift zijn gebruikt.

Er zijn daarnaast talloze andere collega's op het CHDR die met hun aanwezigheid of woorden hebben bijgedragen aan het feit dat de afgelopen drie jaar meer dan draaglijk waren. Ik hoop dat jullie mijn aanwezigheid ook konden waarderen, en ik bedank jullie graag persoonlijk op een later (of eerder) moment. Een groep collegae die echter niet onbenoemd kunnen blijven zijn de wetenschappelijk stagiaires: Tessa Nelemans, Tomasz Cholewinski, Esmee Essers, Max van Gent, Nikki van der Heide, Allison Moll, Eva Dessing en Jeanne Knijff. Ik hoop dat ik jullie, naast het bezorgen van een onuitputtelijke lading werk, ook nog iets heb kunnen leren en mogelijk zelfs enthousiast heb kunnen maken voor de wetenschap. Eén ding is zeker, zonder jullie was dit proefschrift niet in dezelfde tijd tot stand gekomen, en aangezien ik de huidige tijdsspanne al ruim voldoende vond, kan ik jullie hier alleen maar hartelijk voor bedanken.

De oplettende lezer heeft gezien dat het overgrote deel van de projecten die beschreven zijn in dit proefschrift zich buiten het CHDR plaatsvonden. Onderzoek zonder de geoliede recruiting-machine van CHDR kan niet zonder enthousiaste collega-artsen, verpleegkundigen en onderzoekers. Binnen het Juliana Kinderziekenhuis zijn dit specifiek Marianne Nuijsink, Iris Groothuis, Arwen Sprij, Danielle van der Kaaij, Mieke Houdijk, Alfred van Meurs, Erika Jongerius, Annemieke Verbaan, de neonatologen, maar ook alle andere ANIOS, AIOS, semi-artsen en coassistenten die in de loop der jaren hebben geholpen bij het benaderen en rekruteren van patiënten. Ik wil Daphne Peeters graag specifiek bedanken voor alle gezelligheid. Een werkplek is niet compleet zonder een buddy, en dankzij deze was de werkplek in het JKZ er één waar de lachspieren in ieder geval goed getraind werden.

Daarnaast heb ik veel geleerd van de samenwerking met Timo de Haan, Ron Mathot, Amadou Samb, Younes Tallahi in het AMC en verliep de samenwerking met Gijs Santen in het LUMC altijd bijzonder prettig. Tot slot hebben Hettie Janssens, Badies Manaï en Els Kooij in het Sophia Kinderziekenhuis mij veel geleerd over hoe soepel een research unit binnen een kliniek kan draaien, wat alleen maar voor meer motivatie heeft gezorgd om deze situatie in de toekomst ook op mijn huidige en toekomstige werkplekken te bewerkstelligen.

Tot slot wil ik op persoonlijke noot nog drie mensen bedanken: Melanie Thomas, Maarten van Tol en Robbert Bredius. Jullie zijn de afgelopen drie jaar niet betrokken geweest bij het onderzoek beschreven in dit proefschrift, maar jullie enthousiasme voor respectievelijk de kindergeneeskunde, het wetenschappelijk onderzoek, en de combinatie van beiden hebben er eigenhandig voor gezorgd dat ik de keuze voor de kindergeneeskunde en een promotie-traject heb gemaakt die uiteindelijk resulteerde in dit proefschrift.