EXPLORING MACHINE LEARNING TECHNIQUES IN THE CONTEXT OF EARLY-STAGE CLINICAL RESEARCH

HEIN VAN DER WALL

EXPLORING MACHINE LEARNING TECHNIQUES IN THE CONTEXT OF EARLY-STAGE CLINICAL RESEARCH

COVER DESIGN Caroline de Lint, Den Haag (caro@delint.nl)
THESIS LAY-OUT Caroline de Lint, Den Haag (caro@delint.nl)
PRINTING Oranje Van Loon Drukkers, Den Haag
© COPYRIGHT H.E.C. van der Wall, 2022
ISBN 978-90-9036804-7

The research described in this thesis was performed at the foundation Centre for Human Drug Research (Leiden, The Netherlands) and the division Drug Discovery and Safety of the Leiden Academic Centre for Drug Research (LACDR), Leiden University (Leiden, The Netherlands). The research was financially supported by the foundation Centre for Human Drug Research, Leiden, the Netherlands.

EXPLORING MACHINE LEARNING TECHNIQUES IN THE CONTEXT OF EARLY-STAGE CLINICAL RESEARCH

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 dinsdag 13 december 2022
klokke 16:15 uur

door

Hein Evert Christiaan van der Wall
geboren te Amsterdam, Nederland
in 1989

PROMOTORES
Prof. Dr. J. Burggraaf
Prof. Dr. G.J.P. van Westen

CO-PROMOTOR
Dr. R.J. Doll

PROMOTIECOMMISSIE
Prof. Dr. H. Irth
Prof. Dr. J.A. Bouwstra
Prof. Dr. C.A.J. Knibbe
Prof. Dr. R. Rissmann
Prof. Dr. T. van Gelder
Dr. R.G.J.A. Zuiker

CHAPTER 1

CHAPTER 2

CHAPTER 3

CHAPTER 4

CHAPTER 5

CHAPTER 6

CHAPTER 7

APPENDICES

CHAPTER 1

INTRODUCTION

BACKGROUND

During the development of new drugs, an increasing amount of data collected from clinical trials have become available.[1] These growing datasets are mainly the result of major advancements in technology in health care. The datasets comprise ,among others, a broad area of diagnostics, such as laboratory data and data obtained via advanced imaging techniques. Also, the information density of standard techniques such as electrocardiographic (ECG) or questionnaires are going beyond the usual read-outs as advanced data analytics allow to obtain more information. Further, the context of the datasets is becoming increasingly important as can be illustrated by the difference in blood pressure measurements as performed in a clinical setting compared to continuous recording of the blood pressure during days or even weeks. It is essential to realize, however, that data by themselves are useless. To be useful, data must be analyzed, interpreted, and acted on.[2]

The goal of focusing on extensive data sets is manyfold. In this thesis the main focus will be on data collected in the context of early phase drug development. The aim is to find adequate biomarkers for prognostication or identification of factors that result in a better understanding of the intended and not-intended effects of pharmacological interventions. The detection of biomarkers by the integration of data from early pharmacological data in human studies could ensure a better informed future clinical drug development.[3] It can also be concluded that this development should be halted at an early stage, which might save a lot of money, as described by Cohen et al.[4]

Although in clinical trials an increasing number of analyses are being performed to find biomarkers, it is not always clear 1) what to do with the collected data and 2) whether all of the large number of analyses are needed to identify the biomarkers. Artificial Intelligence (AI) strategies may assist in answering these questions.

Broadly speaking, AI can be defined as the discipline devoted to the simulation of human cognitive capabilities on the computer. AI is already widely used, and found its way into healthcare and medical fields. For example, AI can improve healthcare efficiency and reduce costs by providing rapid and accurate image interpretation, improving workflow in healthcare delivery, or empowering patients to monitor their health at a new level.[5]

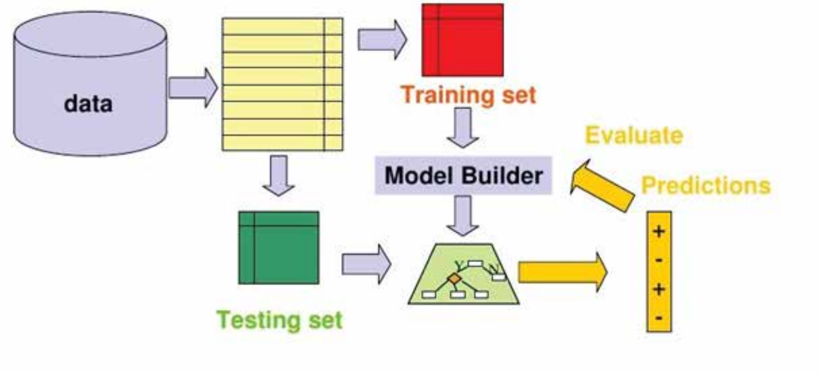
A well-known component of AI is Machine Learning (ML). Generally, machine learning is a type of AI that provides computers with the ability to learn patterns without being explicitly programmed. It focuses on the development of algorithms that can change when exposed to new data. More specifically, machine learning techniques can be used to identify patterns from high-dimensional data and make decisions with minimal human intervention. Additionally, it can be used to obtain insights, predictions, and decisions from vast amounts of data by combining different parameters. For instance, machine learning has already shown its ability to identify key features (markers) and modelling predictive biomarker signature in a variety of medical fields, including oncology[6], neurology[7–9], immunology[10], gastroenterology[11], diabetes[12], and skin diseases.[13,14] Thus, machine learning can play an important role by using data collected in clinical trials to rationally decide on the best application of old and new drugs. The advantages of machine learning techniques over inferential statistical models are to infer relationships between variables for automatic pattern discovery based on multi-dimensional data and the ability to build generalized models.[15]

One of the subsets of machine learning is supervised learning, in which data consisting of a series of measurements are described by a set of features which is used to train an algorithm. Subsequently, this algorithm can be used to make decisions based on concealed data.

In order to evaluate the performance of the algorithm on unseen data, the data set with known output is split into a training and testing/hold out set. The machine learning model is built using the training set, whereupon the outcomes of the testing set are predicted. These predictions are consequently compared with the real outcomes (graphically depicted in Figure 1).

A machine learning model can be a regression or classification algorithm. A regression model is used to predict a continuous value such as price, salary, age, etc. To evaluate this kind of model error metrics such as the r-squared, the mean squared error, the mean absolute error, etc. are calculated. A classification model is used to predict discrete values, such as male or female, normal or abnormal, healthy or unhealthy, etc. Accuracy, specificity, sensitivity, and area under the curve are known error-metrics to evaluate a classification model.

Figure 1 *Process of data splitting to evaluate the performance of the model. The data is split into a training and a testing set. A model is built based on the training set and consequently evaluated using the testing set.*[16]



Machine learning models can also be divided into linear and non-linear models. Linear algorithms such as linear regression and logistic regression assume a linear relationship between features and outcome. In case a linear model is used, it is clear what the role of each feature is in the predictions. When the relation between the features and the outcome is not linear, a non-linear model such as a random forest or neural network should be used. Then the models can become more like a 'black box' and it is not obvious what role each feature plays in making predictions. At that point other ways must be found to explain the prediction.

In this PhD thesis several machine learning techniques applied to data sets derived from early phase clinical research are explored. Machine learning strategies are applied on three types of data:

- Classical data based on electrocardiographic (ECG) measurements such as conduction intervals, electrical-anatomical features of atria and ventricles
- Innovative data based on driving performance tests and driving simulators
- Emerging data based on microbiome data in healthy subjects and patients with skin disease.

OUTLINE OF THIS THESIS

This thesis contains machine learning approaches on a variety of clinical data sets. The classical data consist of electrical signals from the ECG of healthy subjects, the innovative data originate from measurements in a driving simulator, and emerging data are derived from DNA analysis of the microorganisms living on the skin of patients with skin disease.

In **Chapter 1** an introduction to this thesis is given.

In **Chapters 2** and **3**, the application of both classical data analysis and machine learning analysis on the ECG in human subjects are explored.

The ECG is a ubiquitous tool in clinical medicine that has been used for decades since its invention in 1902 by the Dutch physiologist Willem Einthoven from Leiden, who was awarded the Nobel prize in 1924. The ECG is a low-cost, rapid and simple test that is available even in the most resource-scarce settings.[17] In the classical setting the ECG is used to detect abnormal electrical cardiac signals in patients with myocardial infarction, arrhythmias, cardiomyopathy, and other cardiac disorders. The ECG has proven its important and significant value in daily clinical practice. Simple combinations of ECG abnormalities can be recognized and interpreted. However, more complex combinations of ECG deviations are difficult to translate towards a cardiac disease. Machine learning may provide identification of patterns in multiple abnormalities in individual patients, given proper training in a large data set. In this data set the link between specific complex patterns of ECG abnormalities and a certain clinical diagnosis is made. Subsequently, this relationship can be applied on the ECG of a new patient. Compared with the classical ECG analysis, machine learning may thus provide additional ways to identify changes in complex ECG abnormalities in individual patients. The application of machine learning in ECG analyses has just been started recently.[5] Machine learning and other advanced AI methods, such as deep-learning convolutional neural networks, have enabled rapid, human-like interpretation of the ECG. Signals and patterns largely unrecognizable to human interpreters can be detected by multilayer AI networks with precision, making the ECG a powerful, non-invasive biomarker, even more than in the classical setting.[17]

In **Chapter 2** a typical classical analysis is presented, by studying how many ECGs are needed to perform an adequate QT interval analysis. Undesirable side effects of several drugs are the unwanted occurrence of cardiac arrhythmias and subsequent sudden cardiac death, as it can cause prolongation of the QT interval[18] A thorough QT (TQT) study is specifically designed to evaluate the potential prolongation of the QT interval by a novel compound.[19] Although many of these studies have been performed since the introduction of the guideline,[20] the correct performance and the scientific value of a TQT study are still under debate. A TQT study exposes many healthy volunteers or patients to the novel compound and the costs are relatively high.[20–22]

In current practice, several elements to measure a QT prolonging effect of a specific compound are not underpinned by peer-reviewed scientific data. This includes the number of ECG replicates that are recorded, which is arbitrarily set at three or more by the regulators,[19,23] and the formula that is deployed to correct the QT-interval for heart rate.[24,25]

In **Chapter 3** a machine learning approach has been used to investigate ECG changes during aging. Other readout measures include the RR interval, PR interval and QRS duration. Typically, the pharmacological treatment effects are mediated by recognized channels on the cardiac surface.[26] However, there are cardiac effects that require a longer period of time to become visible on the surface ECG, such as aging induced cardiac fibrosis, and it is largely unknown if these subtle effects can be visualized on a surface ECG.[27,28] There has been a number of recent investigations regarding the prediction of physiological age – in contrast to actual chronological age – using medical records, vital signs, laboratory data, or epigenetic changes.[29,30] These investigations indicate the existence of a gap between predicted physiological age and actual chronological age. Exploration of this gap is clinically important as a serious gap difference has been shown to be associated with higher risks of all-cause mortality, cardiovascular disease, obesity, earlier menopause, and frailty.[30–35]

Chapters 4 and **5** contain two machine learning studies on data from abnormal driving behaviour. Research into abnormal driving behaviour is needed as car-drivers have a potential risk to become involved in a crash and compromise traffic safety of others and themselves. Traffic accidents associated with drug use have significantly increased over the past two decades.[36] Driving simulators provide a safe means of studying drug effects on the ability of proper car-driving[37]. Currently, many researchers use the standard deviation of the lateral position (SDLP) as a measure to quantify driving quality[37,38]. Although several studies have shown that the SDLP is sensitive to drug-induced changes in driving behaviour[38–41], it is highly unlikely that SDLP by itself is able to distinguish between numerous different aspects of driving. Altogether, it is questionable whether the SDLP alone is a good benchmark for safe driving.

Improving assessment of driving behaviour may be achieved by combining more parameters such as the mean lateral position (MLP), mean speed (MS), and the standard deviation of speed (SD-Speed) using machine learning, employing a specific algorithm[1,2,42,43]. Such algorithms may not only improve the recognition of impaired driving behaviour but may also explain how and to which extent the driving behaviour is affected. An algorithm combining multiple parameters may improve early recognition of the driving parameters affected by new drugs.

In **Chapter 4** two machine learning models are built to assess driving behaviour after intake of alcohol and alprazolam using all parameters. Subsequently, we compared the performance of these models with models using the golden standard (SDLP) alone.

The ultimate goal of the construction of these models are to test new drugs or interventions with (a selection of) models trained on distinguishing interventions already known to impair driving behavior. A model for detection of sleep-deprived driving may be a good first test in a battery of tests that can evaluate the effect of new drugs on driving behaviour, as sleepiness is also known to affect driving behavior.[44–47] Sleep deprivation can serve as a surrogate of sedation caused by sedative drug effects.[48]

Although drowsy drivers are as dangerous as drivers with unlawful blood alcohol levels they cannot be caught in a police checkpoint, but only in case of a perceived dangerous driving situation.[49] A sufficiently accurate model could be used to detect drug or food induced sleepiness, allowing either dose adjustment or adequate warning notes.

In **Chapter 5** a machine learning model is created to detect sleep-deprived driving. This model is used to investigate if it can predict sleep-deprived driving characteristics after intake of alcohol or alprazolam.

In **Chapter 6** machine learning in microbiome data of patients with skin disease is explored. The skin is the largest organ of the human body and is colonized by a wide range of microorganisms.[50] Many of the micro-organisms living on the skin (its microbiome) are harmless and, in some cases, provide vital functions.

At present, the skin microbiome is known to be involved in several skin diseases.[51] This breakthrough has led to additional knowledge on specific microorganisms that play a role in some of these skin disorders, for instance the role of *Staphylococcus aureus* in atopic dermatitis and *Cutibacterium acnes* in acne vulgaris. However, the role of microorganisms that are less abundant is still largely unknown. It is plausible that the presence of a combination of several different organisms forming a specific microbial profile might also contribute to the development and subtype of skin disease. Machine learning may offer a solution because the underlying computational analyses may facilitate the identification of specific patterns of microorganisms that are discriminative for a specific type of skin disease.[52]
Modelling of the human microbiome by machine learning offers the potential to identify specific microbial biomarkers and may aid in the diagnosis of many clinical diseases. As a result, machine learning may be highly informative for the development of therapeutic modalities to ameliorate the microbial imbalance and to counteract certain pathogens.

CONDENSED OBJECTIVES OF THIS THESIS

In **Chapter 1** we present an introduction to various machine learning techniques which can be applied in early phase drug development.

In **Chapter 2** a classical analysis on ECG recordings is presented, obtained in a placebo-controlled phase I single ascending dose trial with a compound that prolongs the QT interval. In **Chapter 3** a machine learning approach is used to investigate ECG changes during aging.

In **Chapter 4** we analysed the effects of alcohol and alprazolam on car driving behaviour. Using machine learning we aimed for improved assessment of aberrant driving by including multiple parameters derived from a driving simulator. We aimed to develop an algorithm to explain in what way and to which extent the driving behaviour was affected by alcohol and alprazolam.

In **Chapter 5** an attempt was made to develop another model allowing to characterize sleep-deprived driving behaviour. We aimed to demonstrate how driving behaviour after intake of alcohol or alprazolam is similar to sleep-deprived driving behaviour, in order to validate the use of the model for characterization of a new drug.

In **Chapter 6** we employed machine learning to predict disease from the microbiome dataset in patients with skin disorders. We tried to identify discriminative biomarkers in the microbiome of patients with seborrheic dermatitis versus healthy controls. We hypothesized that the microbiome-based biomarkers alone can be used to predict the correct diagnosis. Modelling of the human microbiome by machine learning offers the potential to identify specific microbial biomarkers, useful for new drug development.

Chapter 7 presents a general discussion on the main findings of this PhD thesis. In the section we discuss potential next steps in using machine learning during early phase drug development.

REFERENCES

1 Deo, R., *Machine learning in medicine.* Circulation, 2015. 132(20): p. 1920-1930.

2 Obermeyer, Z., Ezekiel, JE, *Predicting the future big data machine learning and clinical medicine* The New England journal of medicine, 2016. 375(13).

3 van EsDonk, M.J., *The quantification of growth hormone secretion: application of model-informed drug development in acromegaly*. 2019, Leiden University.

4 Cohen, A., et al., *The use of biomarkers in human pharmacology (Phase I) studies.* Annual review of pharmacology and toxicology, 2015. 55: p. 55-74.

5 Haq, K.T., S.J. Howell, and L.G. Tereshchenko, *Applying artificial intelligence to ECG analysis: promise of a better future*. 2020, Am Heart Assoc. p. e009111.

6 Cuocolo, R., et al., *Machine learning in oncology: a clinical appraisal.* Cancer letters, 2020. 481: p. 55-62.

7 Zhang, D., D. Shen, and A.s.D.N. Initiative, *Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease.* NeuroImage, 2012. 59(2): p. 895-907.

8 Deshpande, G., et al., *Identification of neural connectivity signatures of autism using machine learning.* Frontiers in human neuroscience, 2013. 7: p. 670.

9 Fekete, T., et al., *Multiple kernel learning captures a systems-level functional connectivity biomarker signature in amyotrophic lateral sclerosis.* PloS one, 2013. 8(12): p. e85190.

10 Sutherland, A., et al., *Development and validation of a novel molecular biomarker diagnostic test for the early detection of sepsis.* Critical care, 2011. 15(3): p. 1-11.

11 Kohli, A., E.A. Holzwanger, and A.N. Levy, *Emerging use of artificial intelligence in inflammatory bowel disease.* World Journal of Gastroenterology, 2020. 26(44): p. 6923.

12 Hathaway, Q.A., et al., *Machine-learning to stratify diabetic patients using novel cardiac biomarkers and integrative genomics.* Cardiovascular diabetology, 2019. 18(1): p. 1-16.

13 Fortino, V., et al., *Machine-learning–driven biomarker discovery for the discrimination between allergic and irritant contact dermatitis.* Proceedings of the National Academy of Sciences, 2020. 117(52): p. 33474-33485.

14 Johansson, H., et al., *A genomic biomarker signature can predict skin sensitizers using a cell-based in vitro alternative to animal tests.* BMC genomics, 2011. 12(1): p. 1-19.

15 Marcos-Zambrano, L.J., et al., *Applications of machine learning in human microbiome studies: a review on feature selection, biomarker identification, disease prediction and treatment.* Frontiers in Microbiology, 2021. 12: p. 313.

16 Lanzi, P.L. *Machine Learning and Data Mining: 14 Evaluation and Credibility*. Machine Learning and Data Mining: 2007 Apr. 16[cited 2022 01/02]; Available from: https://www.slideshare.net/pierluca.lanzi/machine-learning-and-data-mining-14-evaluation-and-credibility.

17 Siontis, K.C., et al., *Artificial intelligence-enhanced electrocardiography in cardiovascular disease management.* Nature Reviews Cardiology, 2021. 18(7): p. 465-478.

18 Straus, S.M., et al., *Non-cardiac QT$_c$-prolonging drugs and the risk of sudden cardiac death.* European heart journal, 2005. 26(19): p. 2007-2012.

19 CHMP, C.I., *E14: The Clinical Evaluation of QT/QT$_c$ Interval Prolongation and Proarrhythmic Potential for Nonantiarrhythmic Drugs. 2005*.

20 Darpo, B. and C. Garnett, *Early QT assessment–how can our confidence in the data be improved?* British Journal of Clinical Pharmacology, 2013. 76(5): p. 642-648.

21 Guideline, I.H.T., *The clinical evaluation of QT/QT$_c$ interval prolongation and proarrhythmic potential for non-antiarrhythmic drugs E14.* Recommended for Adoption at Step, 2006. 4.

22 Taubel, J., et al., *Shortening of the QT interval after food can be used to demonstrate assay sensitivity in thorough QT studies.* The Journal of Clinical Pharmacology, 2012. 52(10): p. 1558-1565.

23 Shah, R.R., J. Morganroth, and R.B. Kleiman, ICH E14 *Q&A (R2) document: commentary on the further updated recommendations on thorough QT studies.* British Journal of Clinical Pharmacology, 2015. 79(3): p. 456.

24 Guideline, I.E., *The Clinical Evaluation of QT/QT$_c$ Interval Prolongation and Proarrhythmic Potential for Non-Antiarrhythmic Drugs Questions & Answers (R3).* 2015.

25 Vandenberk, B., et al., *Which QT correction formulae to use for QT monitoring?* Journal of the American Heart Association, 2016. 5(6): p. e003264.

26 van Dam, P.M., et al., *The relation of 12 lead ECG to the cardiac anatomy: The normal CineECG.* Journal of Electrocardiology, 2021.

27 Biernacka, A. and N.G. Frangogiannis, *Aging and cardiac fibrosis.* Aging and disease, 2011. 2(2): p. 158.

28 Hayashi, H., et al., *Aging-related increase to inducible atrial fibrillation in the rat model.* Journal of cardiovascular electrophysiology, 2002. 13(8): p. 801-808.

29 Wang, F., T. Syeda-Mahmood, and D. Beymer. *Information extraction from multimodal ECG documents*. in *2009 10th International Conference on Document Analysis and Recognition*. 2009. IEEE.

30 Roetker, N.S., et al., *Prospective study of epigenetic age acceleration and incidence of cardiovascular disease outcomes in the ARIC study (Atherosclerosis Risk in Communities).* Circulation: Genomic and Precision Medicine, 2018. 11(3): p. e001937.

31 Breitling, L.P., et al., *Frailty is associated with the epigenetic clock but not with telomere length in a German cohort.* Clinical epigenetics, 2016. 8(1): p. 21.

32 Perna, L., et al., *Epigenetic age acceleration predicts cancer, cardiovascular, and all-cause mortality in a German case cohort.* Clinical epigenetics, 2016. 8(1): p. 64.

33 Wang, Z., et al., *Predicting age by mining electronic medical records with deep learning characterizes differences between chronological and physiological age.* Journal of biomedical informatics, 2017. 76: p. 59-68.

34 Horvath, S., et al., *Obesity accelerates epigenetic aging of human liver.* Proceedings of the National Academy of Sciences, 2014. 111(43): p. 15538-15543.

35 Levine, M.E., et al., *Menopause accelerates biological aging.* Proceedings of the National Academy of Sciences, 2016. 113(33): p. 9327-9332.

36 Ravera, S., et al., *A European approach to categorizing medicines for fitness to drive: outcomes of the DRUID project.* British journal of clinical pharmacology, 2012. 74(6): p. 920-931.

37 Liguori, A., *Simulator studies of drug-induced driving impairment.* Drugs driving and traffic safety, 2009: p. 75-82.

38 Verster, J., Roth, T, *Standard operation procedures for conducting the ontheroad driving test and measurement of the standard deviation of lateral position* SDLP International journal of general medicine, 2011. 359(4).

39 Mets, M., Kuipers, E, de Senerpont, DLM et al., *Effects of alcohol on highway driving in the STISIM driving simulator.* Human Psychopharmacology Clinical and Experimental, 2011. 26(6): p. 434-439.

40 Guo, F., Fang, Y, *Individual driver risk assessment using naturalistic driving data.* Accident; analysis and prevention, 2013. 61: p. 3-9.

41 Darby, P., Murray, M, Raeside, R, *Applying online fleet driver assessment to help identify target and reduce occupational road safety risks* Safety Science 2009. 47(3): p. 436-442.

42 Hegde, J., Rokseth, B., *Applications of machine learning methods for engineering risk assessment–A review.* Safety science, 2020. 122(104492).

43 Paltrinieri, N., Comfort, L., Reniers, G., *Learning about risk: Machine learning for risk assessment.* Safety science, 2019. 118: p. 475-486.

44 Soares, S., S. Ferreira, and A. Couto, *Driving simulator experiments to study drowsiness: a systematic review.* Traffic injury prevention, 2020. 21(1): p. 29-37.

45 Gaspar, J.G., et al., *Evaluating driver drowsiness countermeasures.* Traffic injury prevention, 2017. 18(sup1): p. S58-S63.

46 Schwarz, C., et al., *The detection of drowsiness using a driver monitoring system.* Traffic injury prevention, 2019. 20(sup1): p. S157-S161.

47 Koopmans, I., et al., *Sensitivity and validity of on-the-road and simulated driving test to measure impaired driving behaviour: effect of sleep deprivation.* Sleep Medicine (in preperation), 2020.

48 Van Steveninck, A., et al., *The sensitivity of pharmacodynamic tests for the central nervous system effects of drugs on the effects of sleep deprivation.* Journal of Psychopharmacology, 1999. 13(1): p. 10-17.

49 Haraldsson, P. and T. Akerstedt, *Drowsiness--greater traffic hazard than alcohol. Causes, risks and treatment.* Lakartidningen, 2001. 98(25): p. 3018-3023.

50 Grice, E.A. and J.A. Segre, *The skin microbiome.* Nature reviews microbiology, 2011. 9(4): p. 244-253.

51 Zeeuwen, P.L., et al., *Microbiome and skin diseases.* Current opinion in allergy and clinical immunology, 2013. 13(5): p. 514-520.

52 Leclercq, M., et al., *Large-scale automatic feature selection for biomarker discovery in high-dimensional OMICs data.* Frontiers in genetics, 2019. 10: p. 452.

CHAPTER 2

NUMBER OF ECG REPLICATES INFLUENCES THE ESTIMATED QT PROLONGING EFFECT OF A DRUG

H.E.C. van der Wall[1,3], P. Gal[1], M.J.B. Kemme[1,2], G.J.P. van Westen[3], J. Burggraaf[1,3,4]

Published in *Journal of Cardiovascular Pharmacology*, 2019
Volume 73, Issue 4, pp 257-264

1 Centre for Human Drug Research, Leiden, NL
2 VU Medical center, Department of Cardiology, Amsterdam, NL
3 Leiden Academic Centre for Drug Research, Leiden, NL
4 Leiden University Medical Center, Leiden, NL

ABSTRACT

INTRODUCTION The present analysis addressed the effect of the number of ECG replicates extracted from a continuous ECG on estimated QT interval prolongation for different QT correction formulas.

METHODS For one hundred healthy volunteers, who received a compound prolonging the QT interval, 18 ECG replicates within a 3 minute window were extracted from 12-lead Holter ECGs. Ten QT correction formulas were deployed and the QT_C interval was controlled for baseline and placebo and averaged per dose level.

RESULTS The mean prolongation difference was >4 ms for single and > 2 ms for triplicate ECG measurements compared to the 18 ECG replicate mean value. The difference was <0.5ms after 14 replicates. In contrast, concentration-effect analysis was independent of replicate count and also of QT correction formula.

CONCLUSION The number of ECG replicates impacted the estimated QT interval prolongation for all deployed QT correction formulas. However, concentration-effect analysis was independent of both the replicate number and correction formula.

INTRODUCTION

Drugs can be associated with cardiac arrhythmias and subsequent sudden cardiac death.[1] Careful cardiac assessment of the drug's effect on the ventricular repolarization has therefore become mandatory.[2] The effect on the ventricular repolarization manifests itself as morphological changes in the ST segment of the surface ECG and a prolongation of the QT-interval.[3] The ICH E14 guideline[4] covers the regulator's requirements on the assessment of the compound's QT interval prolonging effect as a proxy for (polymorphic) ventricular arrhythmia, which includes a thorough QT (TQT) study. A TQT study is a study specifically designed to evaluate the QT interval prolonging effect of a novel compound and consists of a placebo-controlled, cross-over study with a positive control.[4] Although many of these have been performed since the introduction of the guideline,[5] the TQT study is still under debate. The scientific value of the TQT remains subject of discussion, as the study exposes additional healthy volunteers or patients to the novel compound, and the costs are high.[5–7]

Several studies have evaluated novel approaches to assess a QT prolonging effect of novel compounds. Dense ECG recording that was implemented into phase I single ascending dose and multiple ascending dose studies showed that is possible in this context to reliably assess QT interval prolonging effects.[8,9] In addition, implementation of a concentration-effect analysis may improve the assessment of the QT prolonging effect even further.[8,10]

However, several elements in current practice to measure a compound's QT prolonging effect are not underpinned by peer-reviewed scientific data. This includes the number of ECG replicates that are recorded, which is arbitrarily set at three or more by the regulators,[4,11] and the QT correction formula that is deployed.[12,13] Therefore, we performed an analysis on ECG recordings obtained in a placebo-controlled phase I single ascending dose trial with a compound that prolonged the QT interval.

Aim of the study

The aim of the present analysis was to demonstrate the feasibility of a novel approach in which several epochs extracted from a continuous

ECG recording were used to assess the compound's effect on the QT interval. The optimal number of ECG epochs (replicates) required to assess this effect was investigated with the FDA recommended approach and the concentration-effect analysis.

METHODS

The present analysis was performed on a placebo-controlled, double-blind, single ascending dose study that was conducted at our center in 2016. The analysis was performed on this study because of the implementation of a Holter ECG in the study and the dose-dependent QT interval prolonging effect of the investigated compound. The study consisted of 10 consecutive cohorts of 10 volunteers of whom, at each dose level, eight received the active compound and two volunteers matching placebo. The dose of the investigated compound increased with each cohort, as is typical for a phase I single ascending dose trial. All subjects consented to their data being registered and the study was performed in accordance to Dutch law on medical-scientific research.

Data acquisition

All subjects were equipped with a 12-lead Holter ECG (Holter H12+ recorder, Mortara instruments BV, Milwaukee, WI, USA), which was mounted just before the dose administration until 24 hours after the dose administration. Standard electrode positioning was used. Subjects were in a supine position and in a calm, relaxed state for at least 5 minutes before any 5 minute window of continuous ECG recording. The ECG recordings from the Holter ECG were extracted during the latter 5 minutes. The protocol was approved by the Dutch health authorities and by the local ethics committee, Foundation Beoordeling Ethiek Biomedisch Onderzoek. Extractions were performed on a single time point which was associated with the largest QT interval prolongation observed using standard 12-lead ECGs made in triplicate. The Holter ECG strips were analyzed by Intermark ECG Research Technology BV (Someren, the Netherlands), who were blinded to treatment, using LabChart v8.1.3 (ADInstruments, Sydney, Australia) with a validated algorithm (ECG analysis module v2.4; ADInstruments, Sydney, Australia)., Per subject, 18 ECG epochs could be extracted and optimized for signal quality from the 5 minute window. The QT and RR interval were measured with the algorithm and manually adjusted when necessary as recommended by the E14 R3 guideline.[11]

QT_c formulas

The corrected QT (QT_c) interval was calculated based on the QT and RR interval, in addition to patient characteristics for selected QT formulas.

ECG extraction within window

ECGs in the present analysis were extracted without a time interval between the ECGs. In order to simulate a clinical situation, ECG recordings for each replicate count were selected in such a way to mimic a time interval in between the recording of these ECGs, as would be the case in a clinical situation. Table 1 displays the scheme that was used for our analysis.

ΔBaselineQT_c calculation

Per subject the QT_c interval for all evaluated QT correction formulas and number of ECG replicates was calculated. This generated 180 QT_c intervals, with 10 different formulas and a total of 18 ECG replicates per subject. The subject's baseline mean QT_c value was then subtracted from all calculated QT_c interval values, resulting in a QT_c change form baseline (ΔQT_c) for all 10 QT_c formulas and the 18 ECG replicates.

ΔplaceboΔBaselineQT_c calculation

The mean ΔQT_c from the subjects in the placebo group was subtracted from the ΔQT_c of the subjects who received the active compound, resulting in 180 placebo-corrected ΔQT_c (ΔplaceboΔBaselineQT_c , $\Delta\Delta QT_c$) per subject. The calculation for the $\Delta\Delta QT_c$ was performed in accordance with the E14 guideline.[4]

Table 1 *Table displaying the (randomized) selection pattern of ECG windows used for QT analysis. The main goal of the selection method was to mimic a time interval between recordings. Fields in grey are selected ECG replicates for a given experiment. For example, for experiments based on 3 ECG replicates, ECG replicates 1, 8, and 15 were used. And, ECG number 3 is used in the experiments based on 4, 6, 7, 10, 11, 12, 14, 15, 17, or 18 ECGs.*

Nr of replicates ECG	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1			▨				▨			▨	▨		▨	▨	▨	▨	▨	▨
2					▨			▨	▨			▨	▨	▨	▨	▨	▨	▨
3				▨		▨	▨			▨	▨	▨		▨	▨		▨	▨
4									▨		▨	▨	▨	▨	▨	▨	▨	▨
5		▨			▨			▨		▨			▨		▨	▨	▨	▨
6						▨				▨	▨	▨		▨		▨	▨	▨
7				▨			▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
8	▨		▨						▨				▨		▨	▨	▨	▨
9					▨			▨		▨	▨	▨	▨	▨	▨	▨	▨	▨
10						▨	▨		▨	▨		▨		▨				▨
11								▨			▨	▨	▨	▨	▨	▨	▨	▨
12									▨	▨	▨			▨	▨	▨	▨	▨
13				▨	▨		▨	▨		▨	▨		▨		▨	▨	▨	▨
14		▨				▨						▨		▨	▨	▨	▨	▨
15			▨				▨	▨	▨			▨	▨	▨	▨	▨	▨	▨
16						▨			▨	▨		▨	▨	▨	▨	▨	▨	▨
17					▨						▨		▨	▨		▨	▨	▨
18				▨		▨	▨	▨	▨		▨	▨	▨		▨	▨	▨	▨

Δ18 replicatesΔplaceboΔbaselineQT_c calculation

Since the true value of the $\Delta\Delta QT_c$ is unknown, the best estimate of the $\Delta\Delta QT_c$ for each formula was considered to be the mean $\Delta\Delta QT_c$ of 18 ECG replicates. The difference between the mean $\Delta\Delta QT_c$ of each replicate count (1 to 18) and the mean $\Delta\Delta QT_c$ of 18 ECG replicates was calculated, this results in a Δ18 replicatesΔplaceboΔbaselineQT_c ($\Delta\Delta\Delta QT_c$). The results of this analysis were displayed as a heat map (Figure 1).

Δ18 replicates 90% CI ΔbaselineQT_c calculation

The difference between the range of the 90% CI of the ΔQT_c of each replicate count and the range of the 90% CI of the ΔQT_c of 18 ECG replicates was calculated and averaged per cohort and then averaged over all 10 cohorts (Δ18 replicates90%CI ΔbaselineQT_c), as displayed in Figure 2.

Concentration-effect analysis

The concentration of the drug at the time of the ECG recording was derived from the concentration time profile of the compound using the Logarithmic Trapezoidal method[14].

A concentration-effect analysis was performed as previously described by Darpo et al.[8]. In short, subjects were divided into 10 groups based on the drug estimated investigated medicinal product concentration. These were plotted against the mean $\Delta\Delta QT_c$ for all QT_c formulas and number of ECG replicates.

Statistical analysis

Data are depicted as mean ±their standard deviation or percentages where appropriate. Python v3.5.2 (Wilmington, DE, USA) was used for statistical analysis. For concentration-effect analysis, a linear regression was used.

RESULTS

A total of 100 subjects were included initially. One subject, who received active treatment in cohort 2, was omitted because of insufficient data quality and the final analysis was performed on data of 99 subjects. Twenty subjects received placebo and were pooled into the placebo cohort. Ten other cohorts, where the dose was increased in successive cohorts, consisted of eight healthy volunteers each on active treatment. Baseline characteristics are displayed in Table 2.

The mean QT interval and RR interval per cohort at baseline and at the time of the Cmax are displayed in Table 3.

Figure 1 *Average of the mean ΔΔQTc compared to the mean ΔΔQTc of 18 ECG replicates (mean ΔΔΔQTc) of all cohorts for every correction method in absolute values (milliseconds). The mean ΔΔQTc deviates with more than 0.5ms (10% of the safety limit) from the most accurate measurement when it is based on less than 14 ECG replicates and more than 1ms when it is based on less than 5 replicates.*

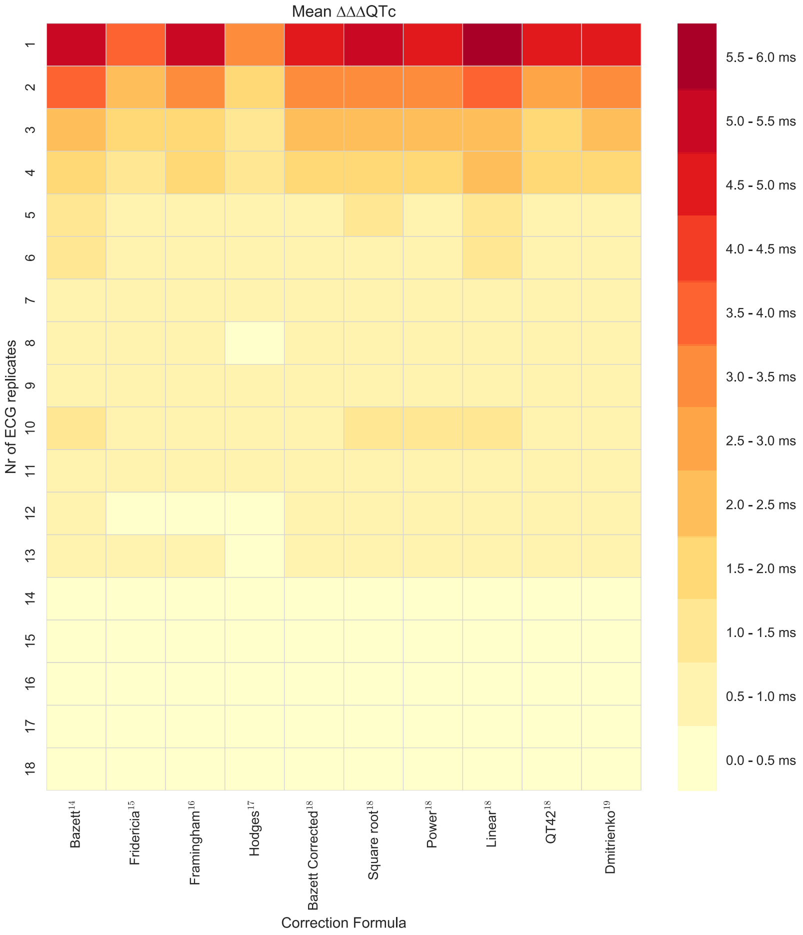


Figure 2 *Average upper limit of the 90% confidence interval of ΔΔQTc compared to the upper limit of the 90% confidence interval of ΔΔQTc of 18 ECG replicates (mean Δ18 replicates 90%CI ΔbaselineQTc) of all cohorts for every correction method in absolute values (milliseconds). For 7 out of 10 correction formulas, the 90% confidence interval of the ΔΔQTc within a cohort increases by more than 0.5 ms (10% of the safety limit) when it is based on less than 11 ECGs per subject compared to a ΔΔQTc based on 18 ECGs per subject.*

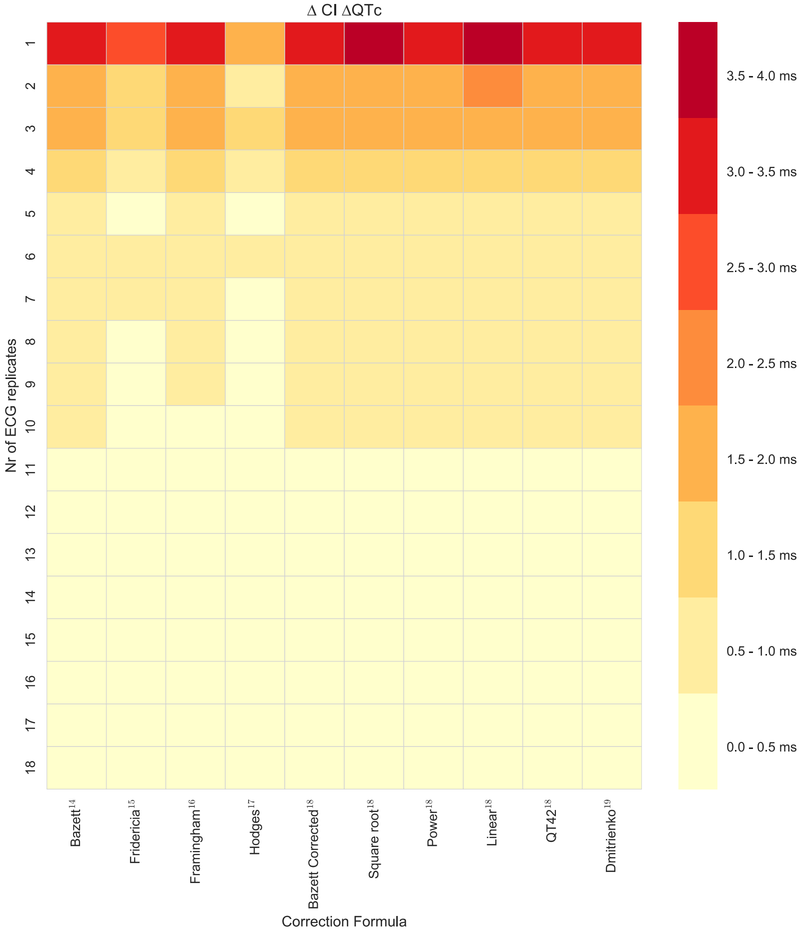


Table 2 *Baseline data. Average values with standard deviation or percentages where appropriate.*

Age (Years)	24.2 ± 4.8
Gender (Male)	100%
Systolic Blood Pressure (mmHg)	121.1 ± 9.2
Diastolic Blood Pressure (mmHg)	72.89 ± 8.05
Heart Rate (min^{-1})	59.9 ± 8.4
BMI (kg/m^2)	23.0 ± 2.9
Temperature (°C)	36.6 ± 0.36
Alcohol Usage (Units / Day)	1.1 ± 1.0
Smoking History (Cigarettes / Day)	0.0 ± 0.0
Cafeine Usage (Units / Day)	1.56 ± 1.16
HbA1c (%)	32.63 ± 2.6
ALAT (U / L)	25.84 ± 12.28
ASAT (U / L)	27.72 ± 7.16
Total Cholesterol (mmol / L)	4.2 ± 0.77
Creatinin (μmol / L)	81.03 ± 8.59
Glucose (mmol / L)	4.67 ± 0.45
PR Interval (ms)	149.13 ± 19.94
QRS Duration (ms)	101.0 ± 8.39
QT interval (ms)	405.89 ± 23.69

Mean and upper limit of 90%CI of $\Delta\Delta QT_c$

The variability of the mean $\Delta\Delta QT_c$ reduced substantially with each additional ECG replicate and remained within 0.5 ms (10 % of the safety limit of 5 ms) after 14 ECG replicates for all QT correction formulas. In Figure 1, the mean $\Delta\Delta\Delta QT_c$ for each number of ECG replicates for each QT correction formula is displayed. In addition, Figure 3 displays the results for a single cohort, with green squares that indicate a $\Delta\Delta QT_c$ prolongation <5 ms and red squares that indicate a $\Delta\Delta QT_c$ prolongation of >= 5 ms.

Table 3 *Estimated mean investigational medicinal compound concentration and the estimated QT prolongation using 3, 5 and 18 ECG replicates corrected with the Fridericia formula per decile with the standard deviation and with corresponding slope. The dose effect relation hardly changes with the increase in the number of ECG replicates measured.*

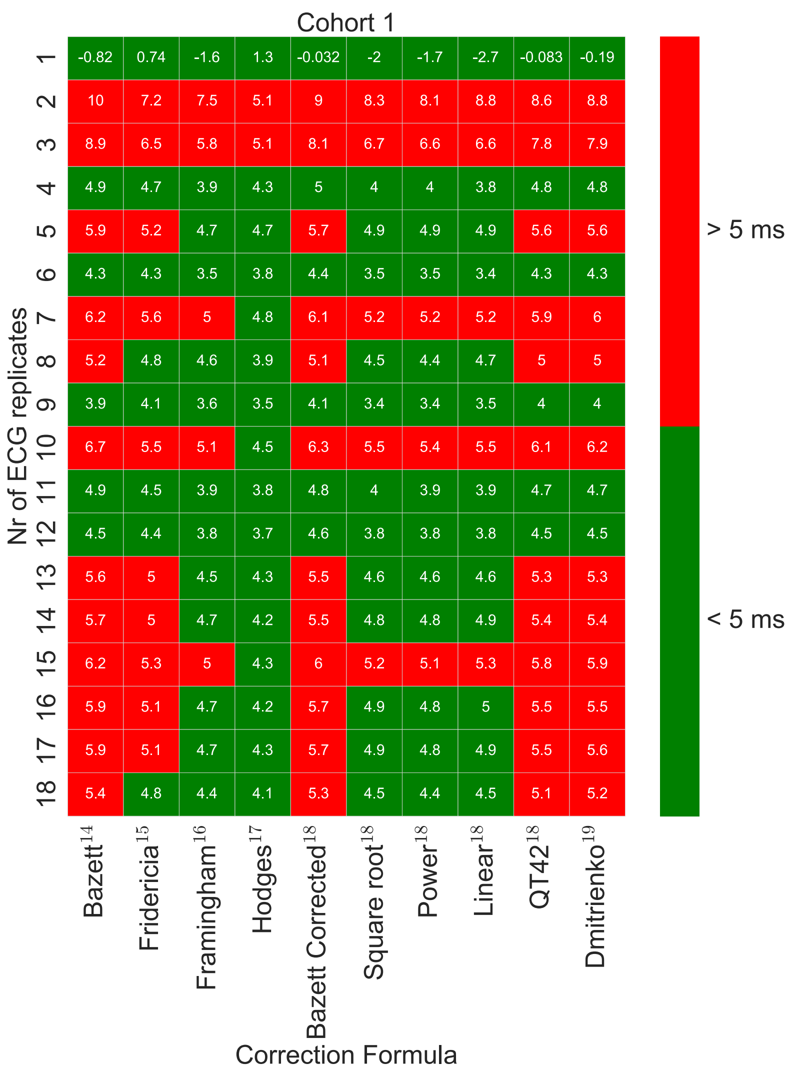
Decile	Estimated mean ± SD investigational medicinal compound concentration (ng/mL)	Mean ± SD QT prolongation (ms) using 3 ECG wreplicates	Mean ± SD QT prolongation (ms) using 5 ECG replicates	Mean ± SD QT prolongation (ms) using 18 ECG replicates
1	7.6 ± 2.5	6.51 ± 16.59	5.21 ± 12.47	4.84 ± 11.54
2	23.2 ± 3.1	6.08 ± 7.13	8.37 ± 5.63	7.31 ± 5.2
3	59.6 ± 10.7	-1.04 ± 10.79	0.45 ± 14.15	0.83 ± 13.11
4	119.6 ± 18.8	5.93 ± 11.59	8.78 ± 10.08	6.53 ± 9.6
5	181.3 ± 12.8	0.81 ± 9.06	2.82 ± 6.54	3.55 ± 7.93
6	238.5 ± 22.7	9.74 ± 13.30	9.01 ± 11.84	9.28 ± 12.15
7	335.3 ± 30.2	16.61 ± 13.63	15.65 ± 12.52	15.11 ± 11.96
8	397.9 ± 16.2	16.12 ± 18.56	14.56 ± 13.02	15.42 ± 12.72
9	485.3 ± 32.0	5.06 ± 13.22	7.46 ± 13.38	6.77 ± 13.71
10	616.1 ± 55.5	19.40 ± 13.37	20.17 ± 9.01	19.78 ± 10.98
Slope (ml*ng-1*ms)		0.022492	0.021380	0.022055
R^2		0.462857	0.539141	0.583485
p-value		0.030387	0.015601	0.010115

The variability of the range of the 90% CI of the $\Delta\Delta QT_c$ also reduced substantially with additional (>1) ECG replicates and remained within 0.5 ms after 11 ECG replicates for all QT correction formulas. Different QT correction formulas and the ECG replicates are displayed in Figure 2 for the range of the 90% CI of the $\Delta\Delta QT_c$.

Concentration-effect analysis of $\Delta\Delta QT_c$

The result of the assessment of the effect of the number of ECG replicates on the concentration-effect analysis is shown in Table 3.

Figure 3 *Mean $\Delta\Delta QT_c$ in milliseconds of an example cohort (Cohort 1) for each number of ECG replicates for every correction method. In this Figure the variation between the number of ECG replicates and between the correction formulas can be clearly seen.*



Nr of ECG replicates	Bazett[14]	Fridericia[15]	Framingham[16]	Hodges[17]	Bazett Corrected[18]	Square root[18]	Power[18]	Linear[18]	QT42[18]	Dmitrienko[19]
1	-0.82	0.74	-1.6	1.3	-0.032	-2	-1.7	-2.7	-0.083	-0.19
2	10	7.2	7.5	5.1	9	8.3	8.1	8.8	8.6	8.8
3	8.9	6.5	5.8	5.1	8.1	6.7	6.6	6.6	7.8	7.9
4	4.9	4.7	3.9	4.3	5	4	4	3.8	4.8	4.8
5	5.9	5.2	4.7	4.7	5.7	4.9	4.9	4.9	5.6	5.6
6	4.3	4.3	3.5	3.8	4.4	3.5	3.5	3.4	4.3	4.3
7	6.2	5.6	5	4.8	6.1	5.2	5.2	5.2	5.9	6
8	5.2	4.8	4.6	3.9	5.1	4.5	4.4	4.7	5	5
9	3.9	4.1	3.6	3.5	4.1	3.4	3.4	3.5	4	4
10	6.7	5.5	5.1	4.5	6.3	5.5	5.4	5.5	6.1	6.2
11	4.9	4.5	3.9	3.8	4.8	4	3.9	3.9	4.7	4.7
12	4.5	4.4	3.8	3.7	4.6	3.8	3.8	3.8	4.5	4.5
13	5.6	5	4.5	4.3	5.5	4.6	4.6	4.6	5.3	5.3
14	5.7	5	4.7	4.2	5.5	4.8	4.8	4.9	5.4	5.4
15	6.2	5.3	5	4.3	6	5.2	5.1	5.3	5.8	5.9
16	5.9	5.1	4.7	4.2	5.7	4.9	4.8	5	5.5	5.5
17	5.9	5.1	4.7	4.3	5.7	4.9	4.8	4.9	5.5	5.6
18	5.4	4.8	4.4	4.1	5.3	4.5	4.4	4.5	5.1	5.2

The mean IMP concentration per decile is displayed together with the estimated QT prolongation measured using 3, 5 and 18 ECG replicates corrected with the Fridericia formula and corresponding slope. For all QT correction formulas, a significant association was found in the concentration-effect analysis. This was also observed for all numbers of ECG replicates.

DISCUSSION

Based on our analysis we showed that the number of ECG replicates in QT studies has a substantial effect on the interpretation of a compound's QT interval prolonging potential for all deployed QT_c formulas. We observed an effect on the mean QT_c interval prolongation and on the range of the 90% confidence interval of the QT_c interval prolongation – parameters that are required by the regulators. To the best of our knowledge this is the first study to address the influence of the number of ECG replicates on the QT prolongation.

The ICH E14 document[4] dictates that, for accurate assessment of the QT interval, at least triplicate ECGs are implemented although evidence for this is limited. The specified cut-off for a positive TQT is 5 ms for mean $\Delta\Delta QT_c$ prolongation. The present analysis showed that all QT correction formulas have a mean difference of 1 ms when triplicate ECGs were extracted compared to 18 ECG replicate extraction. This implies that triplicate ECG extractions are likely to results in inaccurate QT-estimation and can only be used as exploratory method, but not to unambiguously quantify a QT prolonging effect.

The concentration-effect analysis has recently gained more attention in assessing the QT prolonging effect of a compound.[8] The present analysis corroborates these observations, as the concentration-effect analysis was substantially more robust in detecting a QT prolonging effect of the investigated compound as it was independent from the QT correction formula that was used and the number of ECG replicates. It is shown also here that the difference in QT prolongation between subjects becomes less when more QT replicates are measured. This can be deduced from the standard deviations, the R^2 and the p-values. However, despite the decrease in variance in QT prolongation with an increase in the number of ECG replicates, the dose-effect relationship (slope) hardly changes. Noteworthy, applying Hodges' QT correction formula underestimated the drug plasma concentration that would result in a 10 ms QT interval prolongation.

Several studies have compared the agreement of multiple QT correction formulas in large datasets that were collected in healthy volunteers.[12,13] In those studies it was reported that the agreement between the most frequently deployed QT correction formulas is limited (Bazett's and Fridericia's correction formulas). The two main issues with QT correction for RR interval are 1) the intrinsic variability of QT_C interval due to the beat-to-beat RR interval variation, and 2) the absence of a gold standard – which makes complete validation of QT correction formulas virtually impossible. Other studies have suggested that an individual QT/RR interval calculation may provide the best RR correction of the QT interval.[15,16] Unfortunately we could not confirm this in the current work due to limitations of the data set, requiring a wider range of RR intervals to be available for analysis.

The present analysis shows that the variability of mean $\Delta\Delta QT_C$ for all QT formulas exceeds 0.5ms until 14 ECGs have been recorded and included in the analysis. This finding indicates that on average, the mean $\Delta\Delta QT_C$ deviates by more than 10% of the safety limit from the best measured mean $\Delta\Delta QT_C$ (based on 18 replicates per subject), when based on fewer than 14 replicates per subject. This underlines the previously identified issues with correction of QT for the RR interval, but also indicates that the performance of these QT correction formulas is comparable. The present analysis, in line with previous studies, confirms the suitability of a phase I SAD study as replacement for a TQT.[8,9] in particular with implementation of a 24 hour 12-lead Holter ECG. This provides optimal flexibility to accurately assess the effect of a compound on the QT interval. Furthermore, the analysis on a large volume of ECG replicates can be performed after the compound's development has been moved into a later stage and can be cancelled in case the development of the compound is abandoned, thereby saving resources.

Limitations

The current analysis is a retrospective analysis with its inherent limitations. In addition, the concentration of the investigational compound was not assessed at the same time point as the ECGs were extracted. It was therefore necessary to estimate the compound concentration at the time point the ECGs were extracted. However, since any overestimation or underestimation of the compound concentration will be similar for all subject, the presented slopes will deviate very little from the actual slopes.

CONCLUSION

The number of ECG replicates impacted the estimated QT interval prolongation for all deployed QT correction formulas. In contrast, concentration-effect analysis provides robust data on QT interval prolongation independent of the formula and number of replicates.

REFERENCES

1 Straus, S.M.J.M. et al. Non-cardiac QT$_c$-prolonging drugs and the risk of sudden cardiac death. European Heart Journal 26, 2007-12 (2005).

2 Darpo, B., Nebout, T., Sager, P.T. Clinical Evaluation of QT/QT$_c$ Prolongation and Proarrhythmic Potential for Nonantiarrhythmic Drugs: The International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use E14 Guideline. The journal of Clinical Pharmacology 46, 498-507 (2006).

3 Clancy, C.E., Kurokawa, J., Tateyama, M., Wehrens, X.H.T. & Kass, R.S. K+ Channel Structure-Activity Relationships and Mechanisms of Drug-Induced QT Prolongation. Annual Review of Pharmacology and Toxicology 43, 441-61 (2003).

4 (CHMP), C.f.M.P.f.H.U. ICH E14: The clinical evaluation of QT/QT$_c$ interval prolongation and proarrhythmic potential for nonantiarrhythmic drugs. <https://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E14/E14_Guideline.pdf> (2005). Accessed 09-Feb-2018.

(5 Darpo, B. & Garnett, C. Early QT assessment--how can our confidence in the data be improved? British journal of clinical pharmacology 76, 642-8 (2013).

6 Taubel, J., Wong, A.H., Naseem, A., Ferber, G. & Camm, A.J. Shortening of the QT interval after food can be used to demonstrate assay sensitivity in thorough QT studies. Journal of clinical pharmacology 52, 1558-65 (2012).

7 Mehrotra, D.V., Fan, L., Liu, F. & Tsai, K. Enabling robust assessment of QT$_c$ prolongation in early phase clinical trials. Pharmaceutical statistics 16, 218-27 (2017).

8 Darpo, B. et al. Results from the IQ-CSRC prospective study support replacement of the thorough QT study by QT assessment in the early clinical phase. Clinical pharmacology and therapeutics 97, 326-35 (2015).

9 Ferber, G., Zhou, M. & Darpo, B. Detection of QT$_c$ effects in small studies-implications for replacing the thorough QT study. Annals of noninvasive electrocardiology: the official journal of the International Society for Holter and Noninvasive Electrocardiology, Inc 20, 368-77 (2015).

10 Shah, R.R., Morganroth, J. & Kleiman, R.B. ICH E14 Q&A(R2) document: commentary on the further updated recommendations on thorough QT studies. British journal of clinical pharmacology 79, 456-64 (2015).

11 (CHMP), C.f.M.P.f.H.U. ICH guideline E14: the clinical evaluation of QT/QT$_c$ interval prolongation and proarrhythmic potential for nonantiarrhythmic drugs (R3) – questions and answers. <http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500002878.pdf> (2016). Accessed 09-Feb-2018.

12 Vandenberk, B. et al. Which QT Correction Formulae to Use for QT Monitoring? Journal of the American Heart Association 5, (2016).

13 Luo, S., Michler, K., Johnston, P. & Macfarlane, P.W. A comparison of commonly used QT correction formulae: the effect of heart rate on the QT$_c$ of normal ECGs. Journal of electrocardiology 37 Suppl, 81-90 (2004).

14 Yeh, K.C. & Kwan, K.C. A comparison of numerical integrating algorithms by trapezoidal, Lagrange, and spline approximation. Journal of pharmacokinetics and biopharmaceutics 6, 79-98 (1978).

15 Bazett, H.C. The time relations of the blood-pressure changes after excision of the adrenal glands, with some observations on blood volume changes. The Journal of physiology 53, 320-39 (1920).

16 Fridericia, L.S. Die systolendauer im elektrocardiogramm bei normalen menschen und bei herzkranken. Acta Med Scand 53, 469-86 (1927).

17 Sagie, A., Larson, M.G., Goldberg, R.J., Bengtson, J.R. & Levy, D. An improved method for adjusting the QT interval for heart rate (the Framingham Heart Study). The American journal of cardiology 70, 797-801 (1992).

18 Hodges, M.L. Bazett's correction formula reviewed: evidence that a linear QT correction method is better. Journal of the American College of Cardiology 1, 694 (1983).

19 Rautaharju, P.M. & Zhang, Z.M. Linearly scaled, rate-invariant normal limits for QT interval: eight decades of incorrect application of power functions. Journal of cardiovascular electrophysiology 13, 1211-8 (2002).

20 Dmitrienko, A.A. et al. Electrocardiogram reference range derived from a standardized clinical trial population. Drug Information Journal 39, 395-405 (2005).

CHAPTER 3

CARDIAC AGE DETECTED BY MACHINE LEARNING APPLIED TO THE SURFACE ECG OF HEALTHY SUBJECTS: CREATION OF A BENCHMARK

Hein E.C. van der Wall MSc[a,b], Gert-Jan Hassing MSc[a], Robert-Jan Doll PhD[a], Gerard J.P. van Westen PhD[b], Adam F. Cohen MD PhD[a,b,c], Jasper L. Selder MD PhD[d], Michiel Kemme MD PhD[d], Jacobus Burggraaf MD PhD[a,b,c], Pim Gal MD PhD[a,c]

Published in *Journal of Electrocardiology*, 2022
Volume 72, pp 49-55

a Centre for Human Drug Research, Leiden, NL
b Leiden Academic Centre for Drug Research, Leiden, NL
c Leiden University Medical Center, Leiden, NL
d Amsterdam University Medical Cente, Leiden, NL r

ABSTRACT

OBJECTIVE The aim of the present study was to develop a neural network to characterize the effect of aging on the ECG in healthy volunteers. Moreover, the impact of the various ECG features on aging was evaluated.

METHODS & RESULTS A total of 6228 healthy subjects without structural heart disease were included in this study. A neural network regression model was created to predict age of the subjects based on their ecg; 577 parameters derived from a 12-lead ECG of each subject were used to develop and validate the neural network; A tenfold cross-validation was performed, using 118 subjects for validation each fold. Using SHapley Additive exPlanations values the impact of the individual features on the prediction of age was determined. Of 6228 subjects tested, 1808 (29%) were females and mean age was 34 years, range 18 – 75 years. Physiologic age was estimated as a continuous variable with an average error of 6.9±5.6 years (R^2= 0.72 ± 0.04) . The correlation was slightly stronger for men (R^2= 0.74) than for women (R^2= 0.66). The most important features on the prediction of physiologic age were T wave morphology indices in leads V4 and V5, and P wave amplitude in leads AVR and II.

CONCLUSION The application of machine learning to the ECG using a neural network regression model, allows accurate estimation of physiologic cardiac age. This technique could be used to pick up subtle age-related cardiac changes, but also estimate the reversing of these age-associated effects by administered treatments.

KEYWORDS Aging, ECG, Machine Learning, Healthy Volunteers, Artificial Intelligence

INTRODUCTION

Surface electrocardiograms (ECGs) are used frequently in routine clinical care, but also in investigational studies examining the effects of pharmacological and non-pharmacological treatments on the heart. Readout measures include the RR interval, PR interval, QRS duration and (corrected) QT interval. The ECG has long offered valuable insights into cardiac and non-cardiac health and disease, its interpretation requires considerable human expertise. Typically, the pharmacological treatment effects are mediated by recognized channels on the cardiac surface.[1] However, there are cardiac effects that require a longer period of time to become visible on the surface ECG, such as aging induced cardiac fibrosis, and it is largely unknown if these subtle effects can be visualized on a surface ECG.[2,3] Advanced AI methods, such as deep-learning convolutional neural networks, have enabled rapid, human-like interpretation of the ECG, while signals and patterns largely unrecognizable to human interpreters can be detected by multilayer AI networks with precision, making the ECG a powerful, non-invasive biomarker.[4]

There has been a number of recent investigations regarding the prediction of physiological age using medical records, vital signs and laboratory data, or epigenetic changes.[5–7] The likelihood of having a ‘normal’ ECG decreases with age. The most common findings are left ventricular hypertrophy pattern, leftward axis deviation and QRS widening.[8] Some of these abnormalities were significantly associated with all-cause death.[9,10] These investigations also indicated the existence of a gap between predicted physiological age and actual chronological age. Exploration of this gap is clinically important as a serious gap difference has been shown to be associated with higher risks of all-cause mortality, cardiovascular disease, obesity, earlier menopause, and frailty.[6,11–15] Various previous studies have already shown that the 12-lead ECG can be a reliable tool to estimate physiological aging.[6,11–20]

Previous studies have applied artificial intelligence to the raw ECG data, allowing estimation of physiologic ECG age, which was found to reflect aging and comorbidities.[21] However, these algorithms were based on large hospital datasets, thus including patients that may have disease-induced abnormalities in their ECGs, which makes the outcome difficult to interpret when applied to a healthy volunteer. Therefore, the aim of the present analysis was to develop a neural network in healthy volunteers to characterize the effect of aging on the ECG.

METHODS AND MATERIALS

Population

All data were collected at the Centre for Human Drug Research in Leiden, the Netherlands, a clinical research organization specialized in early phase drug development studies. Data collected during the mandatory medical screening to verify study eligibility for enrolment in the early phase drug development studies as a volunteer between 2010 and 2019 were included in the present analysis. The medical screening consisted of a single visit to the clinical unit where a detailed anamnesis, a physical examination, vital signs including blood pressure, temperature, weight and height measurement, body mass index calculation, and a twelve-lead ECG were recorded. Ethical approvals from the Medical Ethical Review Committee for the included studies were acquired and informed consent documents were signed by the volunteers prior to any data collection. The present study was performed in accordance to local regulations. All activities were performed in accordance with applicable standard operating procedures.

The medical screening consisted of a single visit to the clinical unit where a detailed history, a physical examination, vital signs including blood pressure, temperature, weight and height measurement, body mass index (BMI) calculation, and a 12-lead ECG were recorded. Additionally, haematology and chemistry blood panels, urine dipstick, and a urine drug test were analysed.

Data collection for the model

ECG parameters of 6228 subjects with an age between 18 and 75 years were included in the present study. All subjects that were used in this dataset were considered healthy, none of them had known cardiovascular risk factors, and all ECGs were considered normal, or abnormal but without clinical significance. The ECG reviews were performed manually, using standard MUSE cardiology terms. From each subject ECG, 574 features were extracted by the MUSE system. Additionally, gender was used as a feature. The age of the subjects was rounded in whole years. At least ten EGGs were available for each age.

In supplementary Tables 1 and 2, 54 features present in most leads and other ECG features used for the machine learning model are shown, respectively. In addition, gender of each subject was also included in the model.

Data pre-processing and selection

As validation set two subjects of each age were kept apart as final test set. The rest of the data was used as the training set.

To create a balanced training set the Synthetic Minority Oversampling Technique (SMOTE) algorithm was applied on the training set to create 'synthetic' subjects for the less populated age groups based on the values in the concerning age groups.[22]

Machine learning

A neural network was used as a machine learning model. The keras module v. 2.4.3 in python 3.8.5 was used to build a model. Before training, internal cross validation (three-fold) within the training set was used to optimize the model. The network was optimized for number of layers, number of nodes per layer, activation function per layer for each layer and learning rate. A batch size of 300 was used. The number of epochs (defined as the number of cycles through the full training dataset) for internal validation was determined based on validation performance in the internal validation set. The number of epochs for final validation was based on the median of the optimal number of epochs for the internal cross validations. This process of optimization, training, and validation was repeated 10 times with different training and test sets. The optimal models were evaluated on the test set with the R^2 score and mean absolute error. We also evaluated the model performance with respect to gender.

To gain insight into the impact of the individual features on the predicted age, each fold SHapley Additive exPlanations (SHAP) values were calculated[23] based on the training set. The importance of the features was validated by means of permutation importance (defined as the decrease in a model score when a single feature value is randomly shuffled).[24]

RESULTS

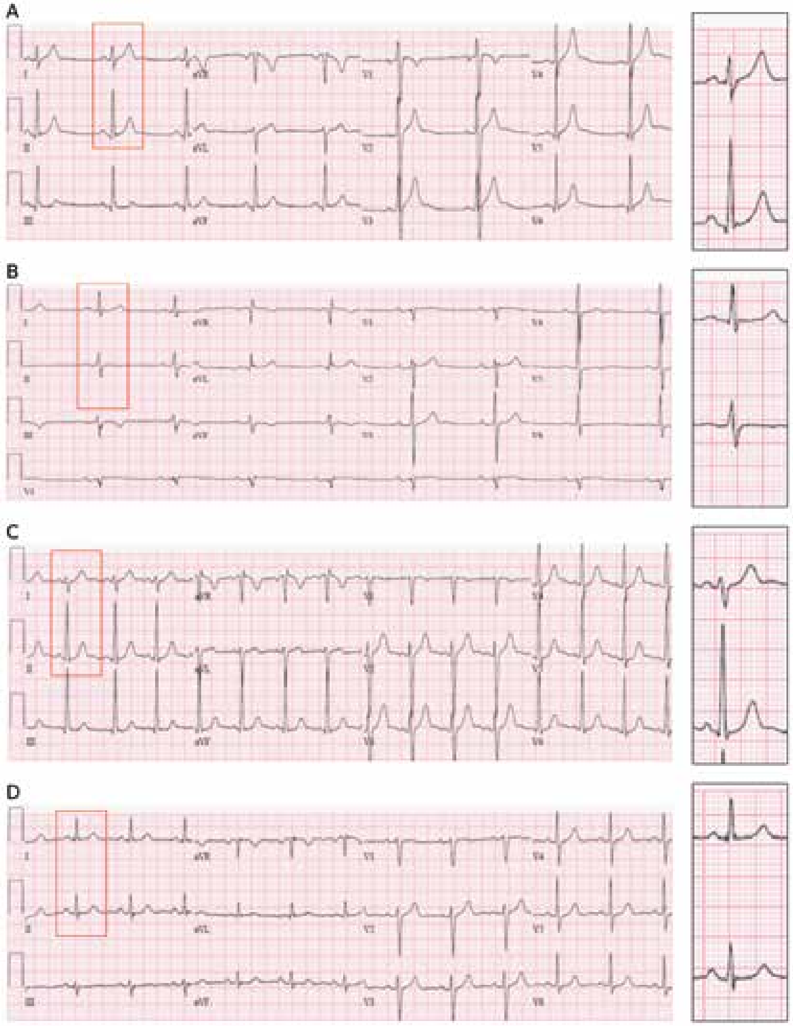
The clinical characteristics of the 6228 included subjects are displayed in Table 1. The study population was divided into ten chronological age groups of 6 years, starting from the age of 18 years. Each age group contained at least 194 subjects, and younger age groups comprised up to 2282 subjects. A total of 1808 (29 %) volunteers were female.

Table 1 Age and gender characteristics of the 6228 healthy subjects.

Subject age (years)	N	% female
18 – 23	2282	29
24 – 29	1563	26
30 – 35	449	20
36 – 40	247	17
41- 46	245	24
46 – 52	339	35
53 – 57	241	38
58- 63	194	42
63 – 69	393	40
69 – 75	275	35

The relation between the (predicted) physiologic age and the chronological age was assessed in 10 sets of 116 subjects. In Figure 2A, the relation between predicted physiologic age and chronological age of all 10 test sets is shown. The average relationship of the models showed an R^2 of 0.72 ± 0.04 (mean ± SD). The mean absolute error of all predictions was 6.9 ± 5.5 years. The average predicted physiologic age was 0.3 years younger than the average chronological age of the subjects. The median deviation of all predicted ages was 5.6 years from the actual age, indicating that half of the predictions was within the range of 5.6 years of chronical age.

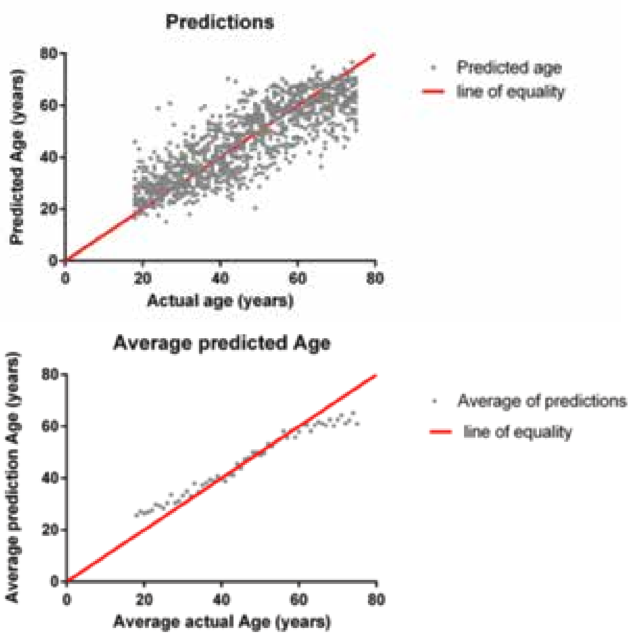
Figure 1 *ECG samples of young and elderly male and female healthy subjects.A: ECG of a young 18 year old male subject. B: ECG of an elderly 74 year old male subject.C: ECG of a young 19 year*



ECG examples of a young 18 year old male (1A) and an elderly 74 year old male (1B) are shown in Figure 1. Figure 1C shows an ECG of a young 19 year old female and Figure 1D shows an ECG of an elderly 74 year old female subject. Several differences between the young and older healthy subjects were discernable. In elderly persons the heart rate was lower, the T wave had a lower (absolute) amplitude in leads I,II,III,AVR, and AVL and the P-wave duration seemed shorter. However, these ECG differences showed considerable variations in the healthy population.

The average predicted age of all subject is presented in Figure 2B. The average predicted age of the 20 subjects per chronological age had a mean absolute error of 3.4 ± 3.0 years (R^2= 0.93). For subjects between 30 and 60 years old the mean absolute error of the average predicted age per chronological age was 1.6 ± 1.1 years.

Figure 2 *Relationship between (predicted) physiologic age and chronological age for 1180 healthy adults (10 test sets, Figure 2A, top). The average predictions for each age are shown in Figure 2B (bottom).*



In order to study gender differences, the predicted physiological ages of the male and female subjects in the test sets were separated and are presented in Figure 3. The predicted ages of the male subjects were more accurate (R^2= 0.74) than the predictions of the female subjects (R^2= 0.64). the mean absolute error in women of the predictions was 7.5 ± 5.9 years, significantly higher than that in men (6.8 ± 5.3 years, p = 0.03).

Figure 3 *Predictions of 819 male subjects (A, top) and 361 female subjects (B, bottom).*

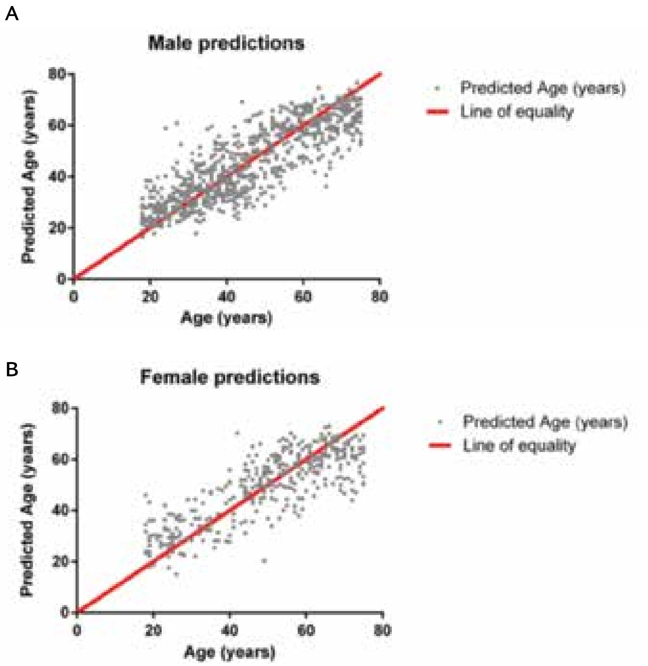


Figure 4 shows the SHAP values of the 40 most important ECG features used in the prediction model. So, the impact of each individual feature on the model output and physiologic aging can be seen. Some of the most important features on the prediction of physiologic age were T top abnormalities in leads V4 and V5, P top amplitude in leads AVR and II and atrial rate.

An increase of P peak amplitude in lead II for example, indicates a younger physiological age (a long red bar to the left). A longer PR interval both indicate an older physiologic age (longer red bar to the right). A higher atrial rate indicates a younger physiologic age (large red bar to the left). The impact of gender was only of minor importance with SHAP values ranging from -1.2 to 0.9. The order of the feature permutation importance is similar to the order of the SHAP values, confirming the impact of the features.

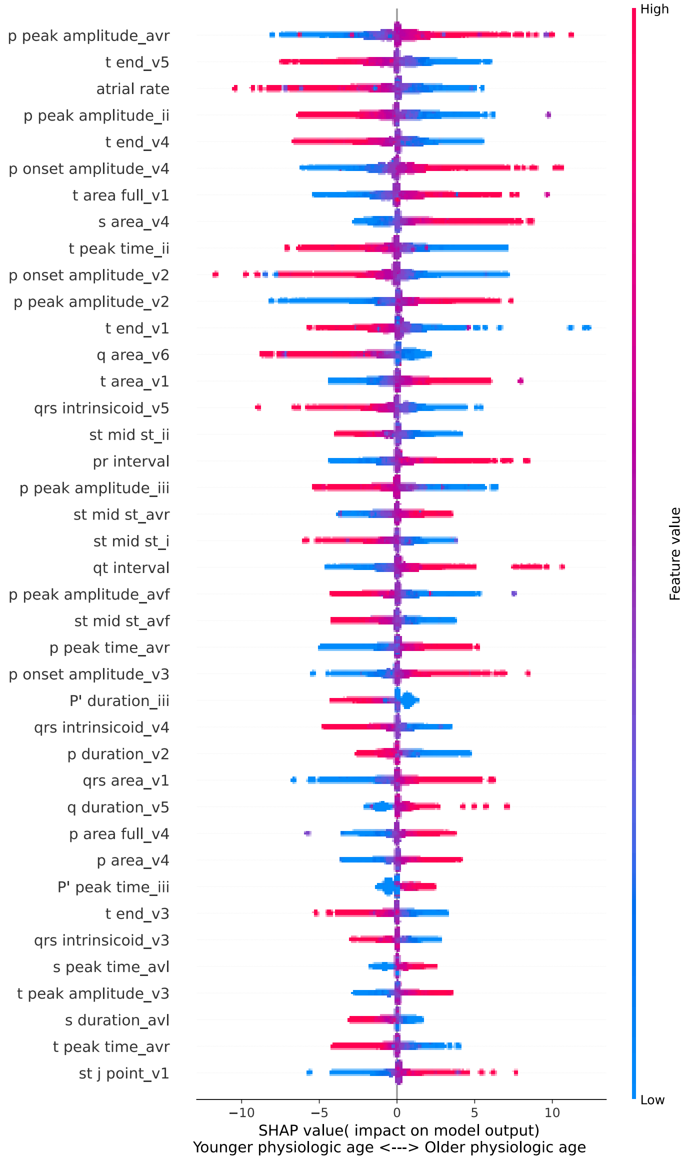
DISCUSSION

In this study we developed machine learning models that allow accurate prediction of physiologic cardiac age of healthy subjects based on 12-lead surface ECG parameters. Using a neural network we were able to estimate the age of a healthy subject with an error of 6.9 years and to analyze the impact of the ECG features. As the models were trained using only healthy subjects, we can assume that the predicted actual age is equal to the cardiac age. we also believed that a psychologically older heart is an unhealthier heart. The created models of the present study may serve as a benchmark for testing the effects of new pharmacological drugs on potential decline or improvement of physiologic health of the heart.

Application of Machine Learning

Attia et al. recently sought to determine whether the application of machine learning algorithms, including convolutional neural networks, to a large ECG patient dataset would be capable of predicting age and sex reported by patients, independent of additional clinical data[21]. They further investigated whether discrepancies between ECG age and chronological age might be a marker of physiological health. When the convolutional neural network-predicted age exceeded a patient's actual age by at least 7 years, there was a higher incidence of cardiovascular comorbidities, potentially suggesting that the convolutional neural network-predicted age from 12-lead ECGs may correlate with physiological health. Their findings suggested that physiological age is distinct from chronological age, and may have

Figure 4 *SHAP values of the 40 most important features for predicting physiologic age. High values of the features are represented in red. Low values are represented in blue. On the x-axis, the predicted physiologic age. Shorter bars mean less impact on physiologic aging.*



useful clinical applications. For example, if a patient's biologic age is 60 but their ECG age predicts that they are 70, it may indicate underlying cardiovascular disease and potential risk. A limitation of their study was, as also recognized by the authors, that all individuals included were patients, and thus an ECG was obtained for a certain clinical indication. It was questioned by the authors whether their results are similarly accurate among an ostensibly healthy population is unknown, and revalidation in such a cohort will therefore be critical.

The same holds true for the study by Hirota et al., who studied biological age, physiological age, and all-cause mortality by 12-lead ECG in patients without structural heart disease.[25] Their data showed that the gap between ECG-predicted physiological and biological age allowed estimation of increased risk of all-cause mortality. Although their study subjects were assumed to have no structural heart diseases, it was stated by the authors that it will be necessary to validate the results of their study in populations of healthy subjects. In our study, we only studied healthy individuals, giving the advantage of being a much needed benchmark study, which enables the validation of future studies in patients versus our data.

Performance of the model

The relation between chronological and predicted physiologic age was associated with an $R^2$of 0.72. Although with a smaller dataset than used by Attia et al., our predictions have a similar performance, probably because of the healthy population in our study, which we expect reduces the variability of the association. Given the large number of influencing factors that can affect ECG parameters the R^2 of 0.72 of our models seems sufficient to detect a pharmacodynamic effect in a cohort of subjects. Use of the entire dataset with a larger number of subjects may improve future performance of the model. In the present study,the impact of physiologic aging on the various ECG features was analyzed using SHAP values. Several changes are clearly visible in the ECG Figures. Some of these are already well known in clinical practice, such as prolongation of PR and QT interval and deceleration of heart rate.[16] Other changes, however, could only be recognized by using machine learning, while these may be evenly important Moreover, when multiple features change at the same time, it becomes difficult to judge whether the change in the ECG is good or bad without using machine learning. By means of machine learning techniques a combination of various ECG changes allows a more accurate insight into the physiologic health changes of the heart.

Gender differences

The accuracy of predicting physiologic age was found to be higher in males than in the female subjects. This may be due to the somewhat smaller female study population, but it may also reflect the atypical ECG repolarization patterns which are known to occur frequently in women.[26] The SHAP values show that impact of gender on physiologic age prediction was only of minor importance. Future studies, analyzing sex- and age- interaction could clarify this.

Pharmaceutical drug testing and potential implications

The prediction of the physiologic age for one single person is less relevant in this model. However for larger groups or cohorts of multiple subjects, the prediction could be more accurate. For example, for a group of 30 test subjects, the average deviation is only less than two years from average physiologic age. Therefore, our models could be particularly suitable as benchmark for testing new pharmaceutical drugs or other interventions which may have impact on cardiac health in the near future. Differences between physiologic ECG age and chronological age have been shown to predict all-cause and cardiovascular mortality and reflect physiologic age, cardiovascular health and long term outcomes.[27] It has also been found that a difference in predicted (cardiac) age and chronological age (higher cardiac age) was greater in patients with peripheral microvascular endothelial function.[28] Additionally, patients with an ECG-age more than 8 years greater than chronological age had a higher mortality rate.[29] Our models, trained with healthy subjects, would therefore be a good benchmark and could be used to predict the mean cardiac age of a cohort before (baseline) and after an intervention to determine its effect on the heart.

The proper use of a model – trained on the entire dataset – in early drug development can provide important information that can be used to make a go/no-go decision regarding further development of new drugs. Similarly, this can be used to guide the decision-making process regarding the dosage range to be used in phase II studies, determining a therapeutic window, and even identifying the target study population[30]. This way novel pharmacological drugs could be tested for effect on cardiac physiologic aging in the early phase of development.

Limitations

Our population consisted of only 29% female subjects. This may have influenced the accuracy of the model, but SHAP value analysis showed that gender only had a minimal impact on the predictions of physiologic age.

No data from children wer available for the present study. At the moment, legal age determination of children is a common probable and depends on imaging techniques that use radiation. Future studies, including age determination in children using ECG could be very useful.

The models have been trained on a limited range of ages. Therefore, the models are limmitted to predict inside this range, which means that wrong predictions among the youngest participants are always higher and among the oldest participants always lower. This is clearly visible in Figure 2. Future studies, including data with a bigger age range, might reduce these limitations.

ECG changes do not need to have a purely cardiac cause, but they may also be caused by effects of age on the position of the heart in the thorax, the presence of fat layers around the heart, and the shape of the thorax shape. Therefore, the found relationship does not necessarily mean older heart per se, but can also mean an older body.

CONCLUSION

The application of machine learning to the ECG using a neural network regression model, allows estimation of physiologic cardiac age. This technique could be used to pick up subtle age-related cardiac changes, but also estimate the reversing of these age-associated effects by administered treatments.

REFERENCES

1 van Dam, P.M., et al., *The relation of 12 lead ECG to the cardiac anatomy: The normal CineECG.* Journal of Electrocardiology, 2021.

2 Biernacka, A. and N.G. Frangogiannis, *Aging and cardiac fibrosis.* Aging and disease, 2011. 2(2): p. 158.

3 Hayashi, H., et al., *Aging-related increase to inducible atrial fibrillation in the rat model.* Journal of cardiovascular electrophysiology, 2002. 13(8): p. 801-808.

4 Siontis, K.C., et al., *Artificial intelligence-enhanced electrocardiography in cardiovascular disease management.* Nature Reviews Cardiology, 2021. 18(7): p. 465-478.

5 Wang, F., T. Syeda-Mahmood, and D. Beymer. *Information extraction from multimodal ECG documents*. in *2009 10th International Conference on Document Analysis and Recognition*. 2009. IEEE.

6 Roetker, N.S., et al., *Prospective study of epigenetic age acceleration and incidence of cardiovascular disease outcomes in the ARIC study (Atherosclerosis Risk in Communities).* Circulation: Genomic and Precision Medicine, 2018. 11(3): p. e001937.

7 Kistler, P.M., et al., *Electrophysiologic and electroanatomic changes in the human atrium associated with age.* Journal of the American College of Cardiology, 2004. 44(1): p. 109-116.

8 Vicent, L. and M. Martínez-Sellés, *Electrocardiogeriatrics: ECG in advanced age.* Journal of electrocardiology, 2017. 50(5): p. 698-700.

9 Lu, T.-P., et al., *Develop and apply electrocardiography-based risk score to identify community-based elderly individuals at high-risk of mortality.* Frontiers in cardiovascular medicine, 2021. 8.

10 Molander, U., et al., *ECG abnormalities in the elderly: prevalence, time and generation trends and association with mortality.* Aging clinical and experimental research, 2003. 15(6): p. 488-493.

11 Breitling, L.P., et al., *Frailty is associated with the epigenetic clock but not with telomere length in a German cohort.* Clinical epigenetics, 2016. 8(1): p. 21.

12 Perna, L., et al., *Epigenetic age acceleration predicts cancer, cardiovascular, and all-cause mortality in a German case cohort.* Clinical epigenetics, 2016. 8(1): p. 64.

13 Wang, Z., et al., *Predicting age by mining electronic medical records with deep learning characterizes differences between chronological and physiological age.* Journal of biomedical informatics, 2017. 76: p. 59-68.

14 Horvath, S., et al., *Obesity accelerates epigenetic aging of human liver.* Proceedings of the National Academy of Sciences, 2014. 111(43): p. 15538-15543.

15 Levine, M.E., et al., *Menopause accelerates biological aging.* Proceedings of the National Academy of Sciences, 2016. 113(33): p. 9327-9332.

16 Rijnbeek, P.R., et al., *Normal values of the electrocardiogram for ages 16–90 years.* Journal of electrocardiology, 2014. 47(6): p. 914-921.

17 Macfarlane, P., et al., *Effects of age, sex, and race on ECG interval measurements.* Journal of electrocardiology, 1994. 27: p. 14-19.

18 Mason, J.W., E.W. Hancock, and L.S. Gettes, *Recommendations for the standardization and interpretation of the electrocardiogram: part II: Electrocardiography diagnostic statement list: a scientific statement from the American Heart Association Electrocardiography and Arrhythmias Committee, Council on Clinical Cardiology; the American College of Cardiology Foundation; and the Heart Rhythm Society: endorsed by the International Society for Computerized Electrocardiology.*

Circulation, 2007. 115(10): p. 1325-1332.
19 Kligfield, P., et al., *Recommendations for the standardization and interpretation of the electrocardiogram: part I: the electrocardiogram and its technology a scientific statement from the American Heart Association Electrocardiography and Arrhythmias Committee, Council on Clinical Cardiology; the American College of Cardiology Foundation; and the Heart Rhythm Society endorsed by the International Society for Computerized Electrocardiology.* Journal of the American College of Cardiology, 2007. 49(10): p. 1109-1127.
20 Khane, R.S., A.D. Surdi, and R.S. Bhatkar, *Changes in ECG pattern with advancing age.* 2011.
21 Attia, Z.I., et al., *Age and sex estimation using artificial intelligence from standard 12-lead ECGS.* Circulation: Arrhythmia and Electrophysiology, 2019. 12(9): p. e007284.
22 Chawla, N.V., et al., *SMOTE: synthetic minority over-sampling technique.* Journal of artificial intelligence research, 2002. 16: p. 321-357.
23 Lundberg, S.M. and S.-I. Lee. *A unified approach to interpreting model predictions*. in *Advances in neural information processing systems*. 2017.
24 Altmann, A., et al., *Permutation importance: a corrected feature importance measure.* Bioinformatics, 2010. 26(10): p. 1340-1347.
25 Hirota, N., et al., *Prediction of biological age and all-cause mortality by 12-lead electrocardiogram in patients without structural heart disease.* BMC Geriatrics, 2020. 21(460).
26 Okin, P.M., *Electrocardiography in women: taking the initiative*. 2006, Am Heart Assoc.
27 Ladejobi, A., et al., *ECG-Derived Age And Survival: Validating The Concept Of Physiologic Age Detected By ECG Using Artificial Intelligence.* Journal of the American College of Cardiology, 2020. 75(11 Supplement 1): p. 3469.
28 Toya, T., et al., *Vascular Aging Detected by Peripheral Endothelial Dysfunction Is Associated With ECG-Derived Physiological Aging.* Journal of the American Heart Association, 2021. 10(3): p. e018656.
29 Lima, E.M., et al., *Deep neural network estimated electrocardiographic-age as a mortality predictor.* medRxiv, 2021.
30 Groeneveld, G.J., Hay, J. L., Van Gerven, J. M., *Measuring blood–brain barrier penetration using the NeuroCart, a CNS test battery.* Drug Discovery Today: Technologies, 2016. 20: p. 27-34.

SUPPLEMENTARY TABLES

Table S1 *ECG features present in most leads extracted by the Muse system included in the model. P′, R′, S′, and T′ indicate the second components of P, R, S, and T wave, respectively, which could be positive or negative polarity.*

Features per lead	
MAX R AMPLITUDE	R PEAK TIME
MAX S AMPLITUDE	R’ AREA
MAXIMUM ST LEVEL	R’ DURATION
MINIMUM ST LEVEL	R’ PEAK AMPLITUDE
P AREA	R’ PEAK TIME
P AREA FULL	S AREA
P DURATION	S DURATION
P OFFSET	S PEAK AMPLITUDE
P ONSET	S PEAK TIME
P ONSET AMPLITUDE	SPECIAL T
P PEAK AMPLITUDE	ST END ST
P PEAK TIME	ST J POINT
P’ AREA	ST MID ST
P’ DURATION	S’ AREA
P’ PEAK AMPLITUDE	S’ DURATION
P’ PEAK TIME	S’ PEAK AMPLITUDE
Q AREA	S’ PEAK TIME
Q DURATION	T AREA
Q PEAK AMPLITUDE	T AREA FULL
Q PEAK TIME	T DURATION
QRS AREA	T END
QRS BALANCE	T PEAK AMPLITUDE
QRS DEFLECTION	T PEAK TIME
QRS INTRINSICOID	T’ AREA
R AREA	T’ DURATION
R DURATION	T’ PEAK AMPLITUDE
R PEAK AMPLITUDE	T’ PEAK TIME

Table S2 *Other ECG features extracted by the Muse system included in the model.*

Features per ECG
ATRIAL RATE
P AXIS
Q OFFSET
QT INTERVAL
QTC BAZETT
Q ONSET
R AXIS
NUMBER OF QRS COMPLEXES
T AXIS
T OFFSET
T ONSET

CHAPTER 4

THE USE OF MACHINE LEARNING IMPROVES THE ASSESSMENT OF DRUG-INDUCED DRIVING BEHAVIOUR

van der Wall, H.E.C.[1,2], Doll, R.J.[1], van Westen, G.J.P.[2], Koopmans, I.[1,3], Zuiker, R.G[1], Burggraaf, J.[1,2,3], Cohen, A.F.[1,2,3]

Published in *Accident Analysis & Prevention*, 2020
Volume 148, p.105822

1 Centre for Human Drug Research, Leiden, NL
2 Leiden Academic Centre for Drug Research, Leiden, NL
3 Leiden University Medical Center, Leiden, NL

ABSTRACT

RATIONALE Car-driving performance is negatively affected by the intake of alcohol, sleep deprivation, tranquillizers, and sedatives. Although several studies have shown that the standard deviation of the lateral position (SDLP) is sensitive to drug-induced changes in a simulated driving performance test, this parameter might not fully assess and quantify deviant driving.

OBJECTIVE Using machine learning we aimed for a better assessment of driving performance by including multiple parameters derived from a simulator rather than the SDlp alone. We specifically analysed the effects of alcohol and alprazolam on car driving behaviour.

METHODS The data used in this study were collected during a previous study that was a single-centre, randomized, double-blind, double-dummy, placebo-controlled, four-way crossover-study with alcohol and alprazolam in 24 healthy subjects (12 M, 12 F, mean age 26 years, range 20-43 years). Using a gradient boosting-Classifier, quantification of the factors influencing driving performance after administration of alcohol or alprazolam was performed to assist in designing a predictive model.

RESULTS Adding additional features besides the SDlp increased the model performance from an accuracy of 65% to 83% for prediction of alprazolam intake and from 50% to 76% for prediction of alcohol ingestion. Analysis of other parameters such as the steering behaviour of the driver appears to be an important contributor to the improvement of the accuracy of the models.

CONCLUSION Machine learning using multiple driving features in addition to the SDLP improves the assessment of drug-induced driving behaviour. These algorithms may serve as a benchmark in the development and application of psychopharmacological medicines.

INTRODUCTION

Drugged driving crashes have significantly increased over the past two decades. Car-driving behaviour is negatively affected by the intake of alcohol, sleep deprivation, tranquillizers, and sedatives (Arnedt 2000; Mets 2011). Driving simulators provide a safe means of studying drug effects on car-driving (Liguori 2009). While the assessment of deviant driving behaviour is difficult, many researchers use the standard deviation of the lateral position (SDLP) as a measure to quantify driving quality (Liguori 2009; Verster 2011). Although several studies have shown that the SDLP is sensitive to drug-induced changes in driving behaviour (Mets 2011; Guo 2013; Darby 2009; Verster 2011), it is highly unlikely that it is able to distinguish between numerous different aspects of driving. Various kinds of medical drugs may impair the ability of car-driving in a different way. Altogether, it is questionable whether the SDLP alone is a good benchmark for safe driving.

Improving assessment of driving behaviour may be achieved by combining more parameters such as the mean lateral position (MLP), mean speed (MS), and the standard deviation of speed (SD-Speed) using machine learning, employing a specific algorithm (Obermeyer 2016; Deo 2015; Hegde 2020; Paltrinieri 2019). Such algorithms may not only improve the recognition of impaired driving behaviour, but may also explain in what way and to which extent the driving behaviour is affected. In earlier studies machine learning has been applied on driving behaviour (Yang 2015; Chen 2017; Dong 2016; Dogan 2011), but none of these studies concerned the recognition of the intake of drugs based solely on driving parameters. Such an algorithm could improve early recognition of the way new drugs affect driving behaviour.

Using machine learning we aimed for a better assessment of aberrant driving by including multiple parameters derived from a driving simulator rather than the SDLP alone. We specifically analysed the effects of alcohol and alprazolam on car driving behaviour as these effects have shown to have the highest frequencies among fatally injured drivers (Bunn 2019). We aimed to develop an algorithm to explain in what way and to which extent the driving behaviour was affected.

MATERIAL & METHODS

Data Collection

The data used in this study were collected during a previous study from our group (Huizinga 2019). For a detailed description of the study design see the above-mentioned study from our institution. In short, this was a single-centre, randomized, double-blind, double-dummy, placebo-controlled, four-way crossover-study with alcohol and alprazolam in 24 healthy subjects (12 males, 12 females, age range 20-43 years), while performing neurocognitive and psychomotor tests on the NeuroCart® and a driving simulator (Green Dino BV, Wageningen, The Netherlands). The interventions consisted of intravenously administered alcohol using a validated clamping protocol to obtain concentrations of 0.5 g $L^{-1}$and 1.0 g L^{-1}, and alprazolam which was given orally in a dose of 1 mg. Driving tests and laboratory tests were done at regular time intervals during a study day. In the current analysis the driving parameters from the study days with 1.0 g L^{-1}alcohol, alprazolam and placebo were considered. Because the pharmacodynamical effects for alcohol and alprazolam varied during one single occasion, measurements at 2- and 4-hours post dose were used for the classification of alprazolam. Measurements at 5- and 6-hours post dose, were used for the alcohol classification. The lane position was calculated using the strip-index parameter, which is the lateral position on the entire highway. The speed (km/h) was derived from the mean-speed parameter at various time intervals, according to the following formula: *speed at t = mean speed at t x t – mean speed at t-1 x (t-1) (t= dimensionless time point)*

All used parameters are listed in Table 1.

Analysis

Pipeline Pilot 2018 (BIOVIA 2018) was used for all analyses and calculations performed in the present study. A gradient boosting model from the scikit-learn module v0.21.2 (Pedregosa 2011) in python 3.6.7 was used for classification.

Table 1 *List of used driving parameters with their desciptions.*

Parameters	Description
Strip-index	lateral position on the entire highway
Lane Position	lateral position in the lane
Speed	speed
Steer	steer-position
Steer-speed	speed of steering to the right
Front-distance-meters	distance to the car in front in meters

Baseline-corrected features

To create an algorithm, features for every observation were required. The first 5 and the last 10 minutes of each measurement were removed from the dataset which left 15 minutes (from 5 to 20 minutes) of driving data per measurement. Contrary to the previous analyses as reported by Huizinga et al., lane switches were included in the dataset. For each parameter time series the following features were calculated:

- Mean: the mean of the whole time series
- Std: the standard deviation of the whole time series
- Diff: the average absolute difference between successive time points in a time series
- Intensity: the highest intensity of the power spectrum of the Fourier transform (sampling frequency of 10 Hz) of the time series corrected with the mean value of the time series
- Frequency: the frequency with the highest intensity of the power spectrum of a Fourier transform (sampling frequency of 10 Hz) of the time series corrected with the mean value of the time series
- For the speed, steer-speed and front-distance-meters, also the following was calculated:
- Min: the minimum of the time series
- Max: the maximum of the time series

In addition, the following features obtained from the original study of Huizinga et al. (Huizinga 2019) – after cleaning the data (including removal of lane switches) – were used: the standard deviation of the lateral position (SDLP, also referred to as GD_SLDP2), the mean lateral position (GD_lane_mean), the mean speed (GD_SPD_mean), the standard deviation of speed (GD_SDspeed). A list of all features is shown in Table 2.

To obtain baseline corrected values, the mean of all baseline values (of all treatment arms) of the subject was subtracted from the values after drug ingestion. Finally, when two features had a high correlation (> 0.9 or < -0.9), only the most important one – based upon the feature importance of fitting the algorithm on the training set – was used for our analysis.

Table 2 *Overview of all features.*

Feature Name	mean	std	diff	intensity	freq (f)	min	max
Strip-index	x	x	x	x	x		
Lane Position	x	x	x	x	x		
Speed	x	x	x	x	x	x	x
Steer	x	x	x	x	x		
Steer-speed	x	x	x	x	x	x	x
Front-distance-meters	x	x	x	x	x	x	x
Features from original study							
GD_SDLP2							
GD_lane_mean							
GD_SPD_mean							
GD_SDspeed							

Machine Learning

First, it was studied whether the administration of alprazolam or alcohol could be distinguished from placebo treatment using only the SDLP obtained from the original analysis. Next, it was studied if this could be performed using all features.

The data sets were randomly split into a training set, consisting of 80% of the subjects, and a test set containing the other 20%. The features in the training and test set were normalized.

The training and testing of the algorithm were repeated five times for both the data set with only SDLP and the full data set with all features. The model performance was evaluated by assessing accuracy, specificity, sensitivity, positive predictive values (PPV), and negative predictive values (NPV). All data were presented as mean ± SD. Also, the probability/continuous scores of the predictions – ranging from 0 (placebo) to 1 (intervention) – were extracted to show how the models could be used for distinguishing abnormal from normal driving behaviour.

RESULTS

Data from all 24 subjects were used for our analysis, but not every subject performed all study days. The data set comprised a total of 80 test drives from 20 study days with placebo treatment; 40 of these placebo tests were used to create and validate the model for alprazolam and the other 40 test drives for optimization and validation of the alcohol model. The effect of alprazolam was assessed using 44 test drives from 22 study days and for the evaluation of the effect of alcohol 35 test drives from 18 study days were available. We ensured that the external test set only contained subjects who joined both the drug administration days and the placebo study days.

Alprazolam

Figure 1 shows the accuracy, specificity, and sensitivity for alprazolam usage versus placebo of SDLP alone (black bars), and of the models using all driving features on predicting alprazolam ingestion (grey bars). These driving features have been listed in Table 2. Figure 1 clearly shows that the addition of other driving features considerably improved the prediction model compared to the performance of the model using SDLP only. The accuracy improved from 65 ± 0% to 83 ± 4%, the specificity from 50 ± 0% to 82 ± 7%, and the sensitivity from 80 ± 0% to 83 ± 6%. For the models using all features, the PPV and NPV were 83 ± 6% and 84 ± 4%, respectively, versus 62 ± 0% and 71% ± 0%, respectively, for the model using SDLP only.

Figure 1 *Performances with standard deviation of the models using only the standard deviation of the lateral position (SDLP, black bars) and the models using all features on predicting ingestion of alprazolam (grey bars). PPV, positive predictive values. NPV, negative predictive values.*

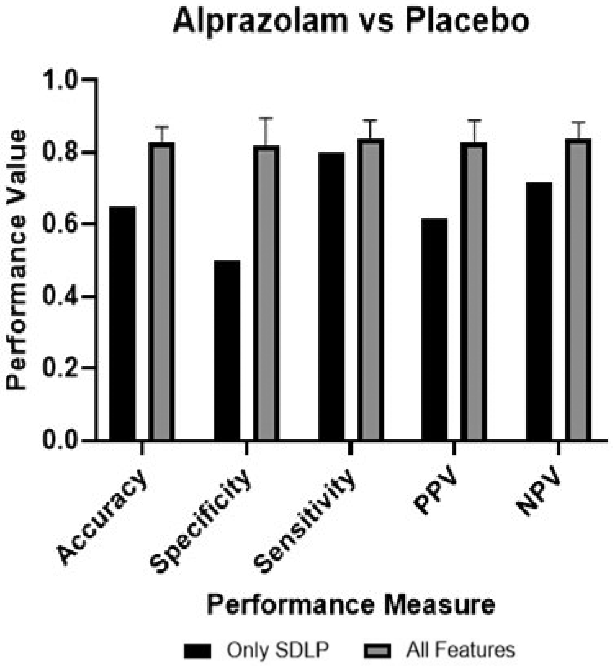
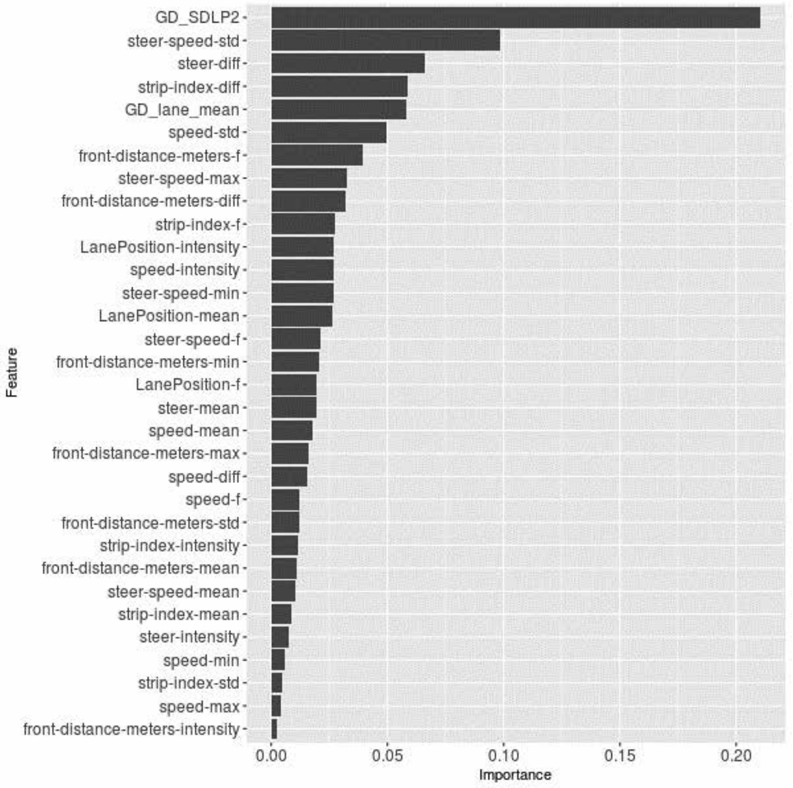


Figure 2 shows the average feature importance of the models based on all features included in our analyses. The most important feature for predicting whether a subject had used alprazolam was the GD_SDLP2, which represents the standard deviation in lateral position after removal of lane switches. By contrast, the maximal speed was only of minor importance in predicting the usage of alprazolam.

In Figure 3, boxplots are shown containing the continuous (probability) predictions of one of the repetitions for both alprazolam models. It is clearly shown that the difference in prediction score between alprazolam and placebo is significantly larger when using multiple features.

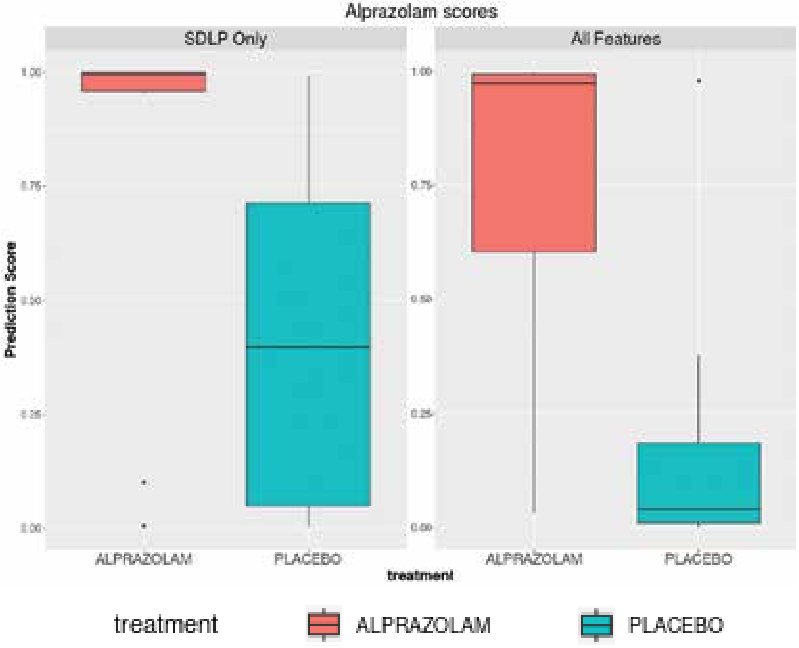
Figure 2 *Average feature importance of the models using all parameters predicting ingestion of alprazolam using all features.*



Alcohol

Figure 4 shows the accuracy, specificity, and sensitivity for alcohol usage versus placebo of both SDLP alone (black bars) and the models using all driving features on predicting alcohol intake (grey bars). In terms of performance, the accuracy improved from 50 ± 0% to 76 ± 4%, the specificity from 60 ± 0% to 82 ± 7%, and the sensitivity improved from 40 ± 0% to 70 ± 0%. For the models using all features, the PPV and NPV were 80 ± 7% and 73 ± 2%, respectively, versus 50 ± 0% and 50 ± 0%, respectively, for the model using SDLP only.

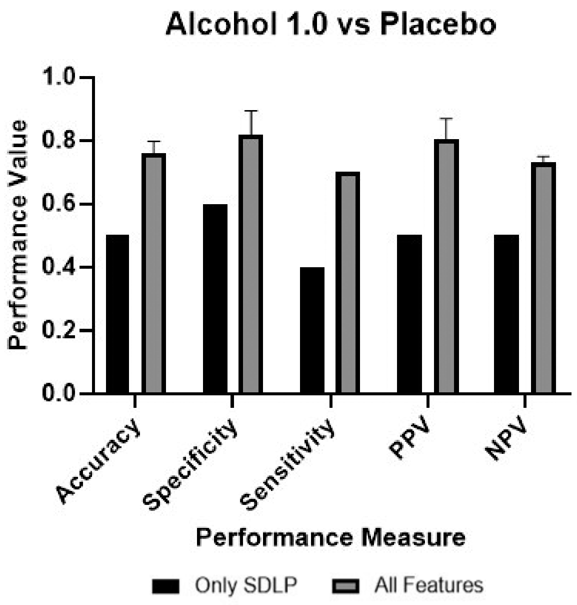
Figure 3 *Boxplots of the probability predictions of one of the repetitions of alprazolam intake. Left: the model using standard deviation of the lateral position (SDLP) only. Right: the model using all features.*



Similar to the results with alprazolam, it is clear that the addition of driving features substantially improved the performance of the model predicting alcohol ingestion. In Figure 5 the relevance of the various features that were used in the analyses on alcohol intake is shown. The most important feature for predicting the presence of alcohol was – similar to the results for alprazolam prediction – the SDLP (GD_SDLP2). Conversely,the meanspeed (after removal of lane switches,GD_SpeedMean) wasonly of minor importance.

In Figure 6, boxplots are shown containing the continuous (probability) predictions of one of the repetitions for both alcohol models. It is clearly shown that the difference in prediction score between alcohol and placebo is significantly larger when using multiple features.

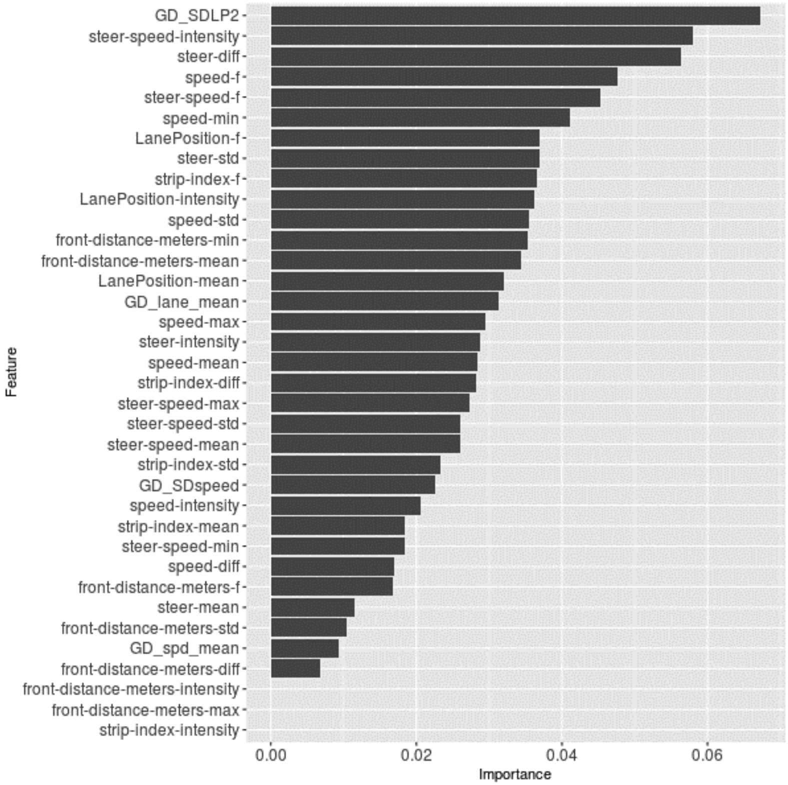
Figure 4 *Performances with standard deviation of the models using only the standard deviation of the lateral position (SDLP, black bars) and the models using all parameters on predicting alcohol intake (grey bars). PPV, positive predictive values. NPV, negative predictive values.*



DISCUSSION

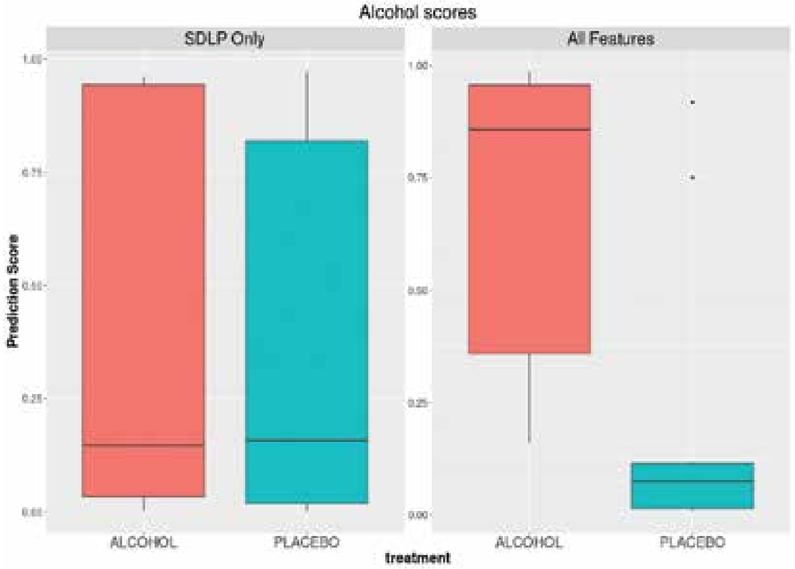
Sedative drugs and alcohol are well known to significantly influence driving behaviour, which can be evaluated by driving parameters such as SDLP (Verster 2011). Accurate knowledge of these side-effects is of crucial importance in the development and application of new psychoactive medicines. This is the first study to create an algorithm using machine learning to detect driving impairment due to the use of drugs with inclusion of multiple driving features rather than the SDLP alone. These algorithms provided improved insight into the way driving behaviour was affected by alcohol and alprazolam. These models may serve as a benchmark for analysis of newly developed drugs.

Figure 5 *Average feature importance of the models using all parameters predicting alcohol intake*



The previous study from our group had already shown that alprazolam and alcohol significantly affected the main parameters of driving in the simulator and affected scores of safe driving (Huizinga 2019). To extend these findings, the current study showed that, if only the SLDP was taken into account, machine learning models, trained on the data of 80% of the subjects, could predict the intake of alprazolam or alcohol in the remaining 20% of the subjects with an accuracy of only 65% and 50%, respectively. These relatively low percentages are probably due to the high inter-subject variability. Since the effect of a medicine may vary substantially for each subject and the number of subjects in the datasets is relatively small, the change in SDLP may differ for subjects in a training set compared to subjects in a test set. In our dataset used for 'the alprazolam model', the prediction of alprazolam use was difficult when based on the training set using only SDLP. Adding more features to train the model, the performance increased the predictive accuracy of alprazolam intake to over 80%. The higher accuracy is associated with a clearer distinction between placebo and drug intake in the continuous / probability predictions.

Figure 6 *The probability predictions of one of the repetitions of alcohol intake. Left: the model using standard deviation of the lateral position (SDLP) only. Right: the model using all features.*



The predictive accuracy for alcohol ingestion increased to 76%. These percentages are slightly higher than observed previously (Chen 2017), where the authors only evaluated the effects of alcohol and additionally used physiological measurements in their model. In their study the authors could successfully distinguish drunk driving from normal driving with an accuracy of 70%. Our results are likely to be more accurate, because we used a correction for baseline measurements of each subject. Part of the inter-subject variability can significantly be reduced by correcting the driving results for the baseline measurements of the same subject. This will make it substantially easier to evaluate the effect of a drug on driving behaviour.

When using multiple driving features rather than only SDLP as read-out, the performance of the models improved with 18 and 26 percentage points for alprazolam and alcohol, respectively. For both the alprazolam and alcohol models the SDLP was a major determinant. However, adding analyses of other parameters such as the steering behaviour substantially increased the capacity to distinguish between drug usage and placebo. This observation emphasizes the importance of analysing multiple features rather than SDLP alone. Previously, such features where difficult or impossible to obtain from simulators or on-road tests, but the current generation of cars allows such data to be relatively easily collected. In this manner it may be possible to develop systems that learn normal driving behaviour of an individual and detect abnormalities for that particular driver. This may be a substantial advantage particularly when assessing the effect of drugs or alcohol.

Limitations

In daily practice, a car accident caused by drug or substance is the true endpoint, but this is difficult to assess. It would seem a reasonable assumption that abnormal driving behaviour is a proxy for this endpoint. Preferably this proxy should be as predictive as possible.

Although the use of an ensemble machine model, such as the gradient boosting model used in this study, is more accurate and robust (Mesquita 2017), this model is accompanied by lack of interpretability (Wang 2015). The importance of the features can be extracted after training the model, but it is not directly clear how the features are being used by the model. The PPV and NPV are quite high – 83% and 84% for the alprazolam model, respectively, and 80% and 73% for the alcohol model, respectively, showing the reliability of the models. However, the model was tested on the measurements of 5 subjects in this study. Also, it has not yet been analysed how the model performs in other interventions.

In the future, more prediction models for impaired driving have to be created, that detect aberrant driving characteristics. It would be interesting to test for instance the effects of a cognitive disorder or sleep deprivation. These new prediction models, can be used as some kind of 'test battery' to create a unique 'fingerprint' (profile) with respect to both desired and undesired effects on driving. However, any inability to detect deviant driving behaviour may be related to limitations of this test battery to detect novel driving behaviour abnormalities, and further studies may still be warranted. With these considerations, the proper use of created algorithms in early drug development can provide important information that can be used to make a go/no-go decision regarding further development. Similarly, they can be used to guide the decision-making process regarding the dosage range to be used in phase II studies, determining a therapeutic window, and even identifying the target study population (Groeneveld 2016). This way novel psychopharmacological drugs could be tested on driving behaviour in the early phase of development. Using these algorithms adequate probability scores can be given to test-drives, which provide an indication about the way and the extent to which these drugs are modifying driving behaviour.

CONCLUSION

In our study we showed how machine learning may improve the assessment of drug-induced driving behaviour. In particular, the inclusion of multiple driving features rather than SDLP alone improved the performance of an algorithm predicting the way driving behaviour was affected by alcohol or alprazolam. These algorithms may serve as a benchmark in the development and application of psychopharmacological medicines.

REFERENCES

- Arnedt, JT, Wilde, GJS, Munt, PW et al. 2000. 'Simulated driving performance following prolonged wakefulness and alcohol consumption: separate and combined contributions to impairment', *Journal of Sleep Research*, 9: 233-41.
- BIOVIA, Dassault Systèmes. 2018. 'Pipeline pilot (version 2018).', *Biovia*.
- Bunn, TSM, Chen, IC. 2019. 'Use of multiple data sources to identify specific drugs and other factors associated with drug and alcohol screening of fatally injured motor vehicle drivers.', *Accident Analysis & Prevention*, 122: 287-94.
- Chen, H, Chen,L. 2017. 'Support vector machine classification of drunk driving behaviour', *International journal of environmental research and public health*, 1: 108.
- Darby, PWM, Murray, M, Raeside, R. 2009. 'Applying online fleet driver assessment to help identify target and reduce occupational road safety risks ', *Safety Science*, 47: 436-42.
- Deo, RC. 2015. 'Machine learning in medicine', *Circulation*, 132: 1920-30.
- Dogan, Ü, Edelbrunner, J, Iossifidis, I. 2011. 'Autonomous driving A comparison of machine learning techniques by means of the prediction of lane change behavior.', IEEE *Conference on Robotics and Biomimetics* IEEE.
- Dong, W, Li, J, Yao, R et al. 2016. 'Characterizing driving styles with deep learning', *arXiv preprint arXiv:1607.03611*.
- Groeneveld, G. J., Hay, J. L., Van Gerven, J. M. 2016. 'Measuring blood–brain barrier penetration using the NeuroCart, a *CNS* test battery', *Drug Discovery Today: Technologies*, 20: 27-34.
- Guo, F, Fang, Y. 2013. 'Individual driver risk assessment using naturalistic driving data', *Accident; analysis and prevention*, 61: 3-9.
- Hegde, J., Rokseth, B. 2020. ' Applications of machine learning methods for engineering risk assessment–A review.', *Safety Science*, 122.
- Huizinga, CRH, Zuiker, RG, de Kam, ML et al. 2019. 'Evaluation of simulated driving in comparison to laboratory-based tests to assess the pharmacodynamics of alprazolam and alcohol', *Journal of Psychopharmacology*, 33: 791-800.
- Liguori, A. 2009. 'Simulator studies of drug-induced driving impairment', *Drugs driving and traffic safety*: 75-82.
- Mesquita, D.P., Gomes, J.P. and Junior, A.H.S. 2017. 'Ensemble of efficient minimal learning machines for classification and regression', *Neural Processing Letters*, 43: 751-66.
- Mets, MA, Kuipers, E, de Senerpont, DLM et al. 2011. 'Effects of alcohol on highway driving in the STISIM driving simulator', *Human Psychopharmacology Clinical and Experimental*, 26: 434-39.
- Obermeyer, Z, Ezekiel, JE. 2016. 'Predicting the future big data machine learning and clinical medicine ', *The New England journal of medicine*, 375.
- Paltrinieri, N., Comfort, L., Reniers, G. 2019. 'Learning about risk: Machine learning for risk assessment.', *Safety Science*, 118: 475-86.
- Pedregosa, F, Varoquaux, G, Gramfort, A et al. 2011. 'Scikit-learn: Machine learning in Python.', *Journal of machine learning research*, 12: 2825-30.
- Verster, JC, Roth, T. 2011. 'Standard operation procedures for conducting the ontheroad driving test and measurement of the standard deviation of lateral position SDLP', *International journal of general medicine*, 359.
- Wang, J, Ryohei, F, and Yosuke ,M. 2015. 'Trading interpretability for accuracy: Oblique treed sparse additive models.', *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*: 1245-54.
- Yang, Y, Sun, H, Liu, T et al. 2015. 'Driver workload detection in on-road driving environment using machine learning', *Proceedings of ELM-2014*, 2: 389-98

CHAPTER 5

USING MACHINE LEARNING TECHNIQUES TO CHARACTERIZE SLEEP-DEPRIVED DRIVING BEHAVIOR

van der Wall, H.E.C.[1,2], Doll, R.J.[1], van Westen, G.J.P.[2], Koopmans, I.[1,3], Zuiker, R.G[1], Burggraaf, J.[1,2,3], Cohen, A.F.[1,2,3]

Published in *Traffic injury prevention,* 2021
Volume 22, Issue 5, pp.366-371

1 Centre for Human Drug Research, Leiden, NL
2 Leiden Academic Centre for Drug Research, Leiden, NL
3 Leiden University Medical Center, Amsterdam, NL

ABSTRACT

OBJECTIVE Sleep deprivation is known to affect driving behaviour and may lead to serious car accidents similar to the effects from e.g., alcohol. In a previous study, we have demonstrated that the use of machine learning techniques allows adequate characterization of abnormal driving behaviour after alprazolam and/or alcohol intake. In the present study, we extend this approach to sleep deprivation and test the model for characterization of new interventions. We aimed to classify abnormal driving behaviour after sleep deprivation, and, by using a machine learning model, we tested if this model could also pick up abnormal driving behaviour resulting from other interventions.

METHODS Data were collected during a previous study, in which 24 subjects were tested after being sleep-deprived and after a well-rested night. Features were calculated from several driving parameters, such as the lateral position, speed of the car, and steering speed. In the present study, we used a gradient boosting model to classify sleep deprivation. The model was validated using a 5-fold cross validation technique. Next, probability scores were used to identify the overlap of driving behaviour after sleep deprivation and driving behaviour affected by other interventions. In the current study alprazolam, alcohol, and placebo are used to test/validate the approach.

RESULTS The sleep deprivation model detected abnormal driving behaviour in the simulator with an accuracy of 77 ± 9%. Abnormal driving behaviour after alprazolam, and to a lesser extent also after alcohol intake, showed remarkably similar characteristics to sleep deprivation. The average probability score for alprazolam and alcohol measurements was 0.79, for alcohol 0.63, and for placebo only 0.27 and 0.30, matching the expected relative drowsiness.

CONCLUSION We developed a model detecting abnormal driving induced by sleep deprivation. The model shows the similarities in driving characteristics between sleep deprivation and other interventions, i.e. alcohol and alprazolam. Consequently, our model for sleep deprivation may serve as a next reference point for a driving test battery of newly developed drugs.

INTRODUCTION

Research into abnormal driving behaviour is needed as car-drivers have a potential risk to become involved in a crash and compromise traffic safety of others and themselves. The risk on abnormal driving behaviour and ensuing car accidents depends on numerous factors, such as predisposing driving style and individual characteristics (e.g., age, gender) of the car-driver and intake of alcohol. (Irwin et al. 2015; Sagberg et al. 2015) CNS -active medicines and recreational substances may also negatively affect car-driving behavior. (Arnedt et al. 2000; Houwing et al. 2012; Mets et al. 2011; Robertson et al. 2017)

To quantify abnormal driving behaviour many researchers have used the standard deviation of the lateral position (SDLP) of the car on the road as a valid measure.(Darby et al. 2009; Mets 2011; Verster and Roth 2011) In a recent study we have shown that by using machine learning, a more sensitive driving measure based on multiple driving features can be created.(van der Wall et al. 2020) In that study, two models were developed that were able to classify driving behaviour affected by either alcohol or by alprazolam. Moreover, our results suggested that a series of these machine learning models could evolve to a test battery, allowing a more precise and accurate evaluation of abnormal driving behaviour in the process of new drug development. However, the generalizability of such a model is still unknown. At the moment, it has only been shown that such a model can recognize solely the drug that has been used for the development of the model. However, the ultimate goal would be to test new drugs or interventions with (a selection of) these models. A model for detection of sleep-deprived driving would be a good first in a battery of tests that can evaluate the effect of new drugs on driving behavior, as sleep deprivation can serve as a surrogate of sedation caused by sedative drug effects.(Van Steveninck et al. 1999)

In the current study, we attempted to create a model to evaluate the effect of sleep deprivation on driving behaviour as sleepiness is also known to affect driving behavior.(Gaspar et al. 2017; Koopmans et al. 2020; Schwarz et al. 2019; Soares et al. 2020) Although drowsy drivers are as dangerous as drivers with unlawful blood alcohol levels they cannot be caught in a police checkpoint, but only in case of a perceived dangerous driving situation.(Haraldsson and Akerstedt

2001) Such a model, when sufficiently accurate could be used to detect drug or food induced sleepiness, allowing either dose adjustment or adaquate warning notes.

The aim of the current study was twofold: 1) to develop a new model allowing to characterize sleep-deprived driving behavior, and 2) to demonstrate how driving behaviour after intake of alcohol or alprazolam is similar to sleep-deprived driving behavior, in order to validate the use of the model for characterization of a new drug.

METHODS AND MATERIALS

Data collection for the model

All Data used in the present study were collected during two previous studies. (Huizinga et al. 2019; Koopmans et al. 2020) In both studies subjects were healthy adults who were in possession of a valid drivers license. They were active and skilled drivers with a minimum mileage of 3000 km per year. Subjects were instructed to drive in a driving simulator (Green Dino BV, Wageningen, The Netherlands) with a steady lateral position in the right-hand lane of a 30 min dual-carriageway highway scenario similar to the one being used during on-road tests; overtaking other vehicles was allowed. The simulators have a non-moving base and consist of a mock-up car with three pedals (clutch, brake and gas), manual shift, steering wheel, safety belt, indicators and hand brake. The controls are linked to a dedicated graphics computer that simulates road environment and dynamic traffic. The driving simulators have a wide view display, made with three LCD (24") flat panel monitors positioned side by side. The total LCD monitor surface is 0.48 m^2.

Data used to create the model that allows the characterization of sleep-deprived driving behaviour were collected during a previous study. (Koopmans et al. 2020) In short, this was an exploratory single-center cross-over study in 24 healthy male subjects, 23 to 35 years of age, to investigate the effects of sleep deprivation on driving.
Subsequently, this model was used to demonstrate sleep-deprived driving characteristics in subjects after intake of alprazolam or alcohol. The effect of alprazolam and alcohol on driving was previously studied in our institution by Huizinga et al. (2019). (Huizinga et al. 2019) In short, this was a single-center, randomized, double-blind, double-dummy, placebo-controlled, four-way crossover-study on alcohol and alprazolam in 24 healthy subjects (12 males, 12 females, age range 20-43 years), while performing neurocognitive and psychomotor tests on the NeuroCart® – a comprehensive battery that can test all functional domains of the central nervous system (CNS) – and a driving simulator (Green Dino BV, Wageningen, The Netherlands). The interventions consisted of intravenously administered alcohol to obtain steady state concentrations of 0.5 g L-1 and 1.0 g L-1, oral administration of 1 g alprazolam, or placebo. (Zoethout et al. 2012) Driving and laboratory tests were performed at regular time intervals during a study day. As the pharmacodynamic effects for alcohol and alprazolam varied during one single study period, measurements at 2- and 4-hours post-dose were used for the characterization of alprazolam. Measurements at 5 and 6 hours post dose were used for the characterization of alcohol. All used parameters are listed in Table A1.

Feature pre-processing

All measurements were corrected for baseline, by subtracting the mean of all baseline values of all treatment arms in the alcohol and alprazolam datasets, of the subject from the values after drug ingestion.

Due to the nature of the study the sleep-deprivation dataset had no baseline measurement shortly before the intervention. Therefore, the morning measurement in the well-rested occasion was used as baseline. The feature values of this measurement were subtracted from the sleep-deprived measurement and the afternoon well-rested measurement (which was used as control).

The features required to develop a model were created in a similar way as described in the article by van der Wall et al. (van der Wall et al. 2020) In short, the mean, the standard deviation, and the mean absolute difference between consecutive time points were calculated for all parameters. In addition, the minima and maxima for the speed, steer-speed and distance to the car in front were calculated. Additionally, the maximum intensity of the power spectrum of a Fourier transform at low frequency (< 0.05 Hz) and high frequency (> 0.05 Hz) were calculated for all parameters.

Finally, some of the features were obtained from the original study, which were calculated after cleaning the data (including removal of lane switches): the standard deviation of the lateral position (SDLP), the mean lateral position (MLP), the mean speed (MS) and the standard deviation of speed (SDS). A list of all features is show in Table A2.

Feature selection

When two features had a high correlation (> 0.9 or < -0.9), only the most important one – based upon the feature importance of fitting the model on the training set was used for final validation.(van der Wall et al. 2020)

Machine learning

In our previous study two linear and two non-linear models were tested on driving simulator data after intake of alcohol or alprazolam. The linear models showed accuracies of 67% and 54% for the alprazolam training set (logistic regression and Support Vector Machine, respectively), and 60% and 52% for the alcohol training set. The non-linear models gave the best performances. Random forest and gradient boosting models both showed an accuracy 81% for the alprazolam training set, and 65% and 68% for the alcohol training set, respectively. (van der Wall et al. 2020) Since it was shown to work best overall on this type of driving simulator data, a gradient boosting classifier with a subsample rate of 0.5 was used as a model, which was obtained from the scikit learn module version 0.23.1 in python 3.7.3.

A cross validation was performed to evaluate the performance of the algorithm. The features of 80% of randomly selected subjects was used five times to train the model. Subsequently, the model was tested on the other 20%. The features were standardized based on the data in the training set. The model performance was evaluated by assessing accuracy, specificity, sensitivity, positive predictive values (PPV), and negative predictive values (NPV). (Wong and Lim 2011) Data were presented as mean ± SD. The performance when using all driving features was compared with the performance of the model when using SDLP only.

To demonstrate sleep-deprived driving characteristics in subjects after intake of alprazolam or alcohol, the model was trained on the entire sleep deprivation data set (all measurements and all features) and tested on the alprazolam and alcohol data sets.

Since a continuous 'sleep-deprived' score was preferred to get an indication of how similar the effects of the drug was to the sleep deprivation effects, a simple (binary) prediction from a classifier did not seem useful. Therefore, the measurements in the alcohol and alprazolam datasets were given probability predictions.

Probability scores indicate the likelihood that the outcome of the model is positive for a particular intervention, in this case sleep-deprivation. Therefore, these values may provide information about the similarity to the sleep-deprived effects. A higher score means a higher probability of being sleep-deprived and therefore the driving pattern is more similar to a sleep-deprived driving pattern. In the current study, probability scores can be used to demonstrate the similarity between the effects of sleep deprivation and alcohol/alprazolam intake on driving behavior. A high probability score means that a test subject using these compounds acting on the central nervous system shows driving characteristics very similar to those of sleep deprived subjects.

Finally, the model was also fitted on the alprazolam and alcohol sets to collect the feature importance for detection of these interventions, whereupon it was compared with the feature importance of the sleep-deprivation model. The importance of the features derived from the same parameter were added together because of the high correlations between these features.

Statistical analysis

Treatment effects on the probability score were analysed with a mixed model analysis of variance (ANOVA) with treatment, measurement, occasion in the cross-over experiment and treatment by measurement as fixed factors and subject, subject by treatment and subject by measurement as random factors. This was similar to the statistical analysis of the SDLP in the study of Huizinga et al. For this analysis, R version 3.6.1 was used. The difference was considered significant if the p-value was below 0.05.

RESULTS

From the sleep-deprived driving dataset in 24 subjects, one subject did not complete the 'baseline' measurement and therefore the data of this subject were not taken into consideration. Additionally, one subject did not finish the control measurement. Therefore, 45 baseline corrected measurements, consisting of 23 sleep-deprived and 22 controls, were used for our analyses.

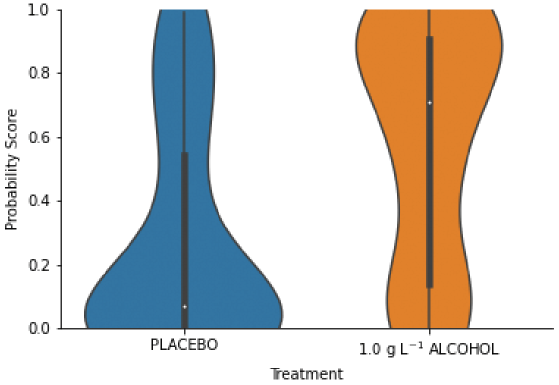
Model performance

The model had an accuracy of 77± 9 %, a specificity of 77 ± 23 %, and a sensitivity of 76 ± 9 %. The PPV and NPV of the model were 83 ± 17 % and 74 ± 5 %. Supplementary Figure 1 shows the average feature importance of the repeated model fitting. The most important feature for predicting whether a subject was driving sleep-deprived was the SDLP. The maximum steering speed was only of minor importance for predicting sleep-deprived driving behavior.
When using only SDLP the accuracy of the model was 70 ± 10 %, the sensitivity 80 ± 14 % and the specificity 58 ± 13 %. In this case the PPV and NPV were 69 ± 7 % and 74 ± 18 %.

Characterization of alprazolam or alcohol affected driving behavior

For the alcohol and alprazolam data, the same measurements were used as in our previous study in which the models for detection of abnormal driving was introduced.(van der Wall et al. 2020) This concerned 80 placebo measurements (40 to compare with alprazolam and 40 to compare with alcohol), 44 measurements after alprazolam intake and 36 measurements after 1 g/l alcohol intake.

Figure 1 *Sleep deprived probability score for placebo and 10 g/L alcohol. Violin plot of the probability scores of the measurements in the 10 g/L alcohol data set, indicating the distribution of the probability scores. The left violin plot shows the scores for the placebo measurements. The right violin plot shows the scores for the 10 g/L alcohol measurements. The width of the violin reflects the relative number of measurements with that score.*



Additionally, in the current study also 42 measurements after 0.5 g/l alcohol were used to compare with placebo on sleep-deprived driving characteristics.

In Figure 1 a violin plot is shown containing the probability scores of the model on the 1.0 g/l alcohol dataset. It demonstrates the similarity between the effects of sleep deprivation and alcohol intake on driving behavior. The mean probability score for the measurements after 1.0 g/l alcohol intake was 0.63 ± 0.37, substantially higher than the score of 0.30 ± 0.36 for the measurements after placebo intake. This finding indicates a substantial similarity in the driving characteristics after sleep deprivation and those after 1.0 g/L alcohol intake. Statistical analysis of the scores revealed that there was a significant effect of treatment (p = 0.0014).

The violin plot in Figure 2 shows the probability scores of the model on the 0.5 g/l alcohol dataset. The mean probability score for the measurements after 0.5 g/l alcohol intake was 0.41 ± 0.40, compared to 0.30 ± 0.36 of the placebo measurements. The differences in these scores were insignificant (p = 0.2434).

Figure 2 *Sleep deprived probability score for placebo and 0.5 g/L Alcohol Violin plot of the probability scores of the measurements in the 0.5 g/L alcohol data set, indicating the distribution of the probability scores. The left violin plot shows the scores for the placebo measurements. The right violin plot shows the scores for the 0.5 g/L alcohol measurements. The width of the violin reflects the relative number of measurements with that score.*

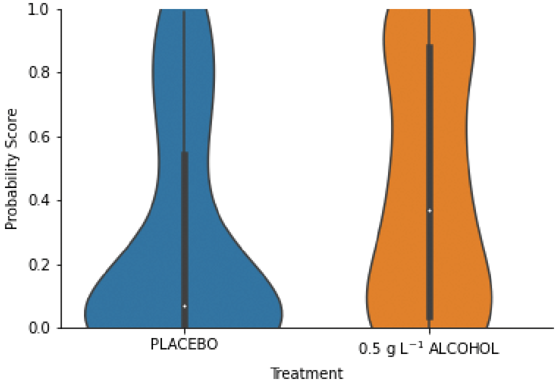
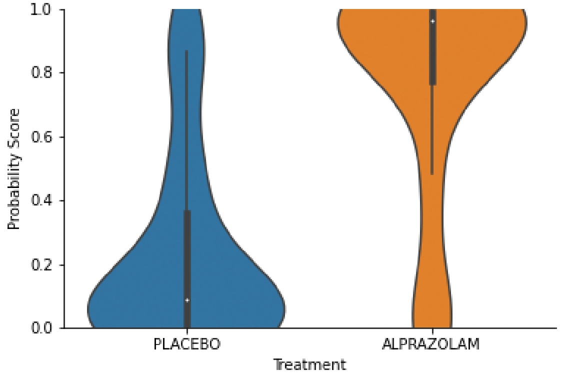


Figure 3 *Sleep deprived probability score for placebo and alprazolam. Violin plot of the probability scores of the measurements in the alprazolam data set, indicating the distribution of the probability scores. The left violin plot shows the scores for the placebo measurements. The right violin plot shows the scores for the alprazolam measurements. The width of the violin reflects the relative number of measurements with that score.*

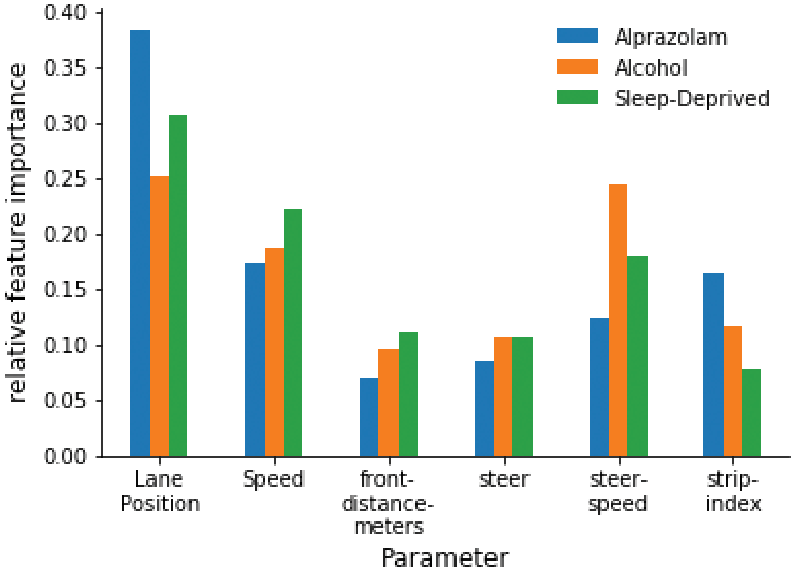


The probability scores on the alprazolam data set are shown in the violin plot of Figure 3. The mean probability score for the measurements after alprazolam intake was as high as 0.79 ± 0.32, versus 0.27 ± 0.32 for the placebo measurements. So, after alprazolam intake subjects showed car-driving characteristics which were quite similar to sleep-deprived driving characteristics. As expected, the ANOVA yielded a significant effect of treatment ($p < 0.0001$).

In Figure 4 the sum of all feature importances for all parameters is shown for the alprazolam, 1.0 g/l alcohol and sleep-deprived dataset.

For all datasets, the sum of the importances of the features derived from the lane position was the highest. For the 1.0 g/l alcohol set, the steer-speed features were also of high importance.

Figure 4 *Relative feature importance. Sum of the relative feature importances grouped by type for all parameters (sum of all feature importances is 1). An overview of all features is shown in Table A2.*



DISCUSSION

Sleep deprivation is known to impair driving performance (Gaspar et al. 2017; Peters et al. 1999; Philip et al. 2005; Schwarz et al. 2019; Soares et al. 2020). The current study has shown that, by using machine learning, the effect of sleep deprivation on driving behaviour can be classified. The created model performed with an average accuracy of 77%. The current study did also show that driving behaviour after sleep deprivation had a great similarity with driving behaviour after alprazolam intake and to a lesser extent after alcohol intake of 1 g/l. For the first time, we developed a model to characterize abnormal driving behaviour for a single intervention, which can also be used for characterization of other/new interventions. Using this model, the effects of a newly developed drug on driving behaviour can be compared with the effects of sleep deprivation. In this way, a series of these machine learning models could evolve to a test battery, which allows a more precise and accurate evaluation of abnormal driving behaviour by creating a predictive effect profile for a medicine.

The performance of the created model is similar to what we have shown for the classification of alcohol and alprazolam in our previous study.(van der Wall et al. 2020) These earlier models, which performed with an accuracy of around 80%, could characterize abnormal driving behaviour solely for the drug that had been used for the development of that particular model. In the present study we have shown that such a model may serve to characterize driving behaviour after a diversity of interventions. Given the large number of influencing factors that can affect driving behaviour the accuracy of 77% of our model seems satisfying. Larger number of subjects may improve future performance of the model.

When using all driving features the performance of the model is, although not significantly, higher then when using only SDLP. This is in line with the results of previous study.(van der Wall et al. 2020)

Also, when using SDLP only, the specificity is much lower and thereby susceptible to false positive results.

The SDLP remains the most important feature for distinguishing abnormal driving behaviour and has rightfully been used as a standard measure. However, by combining all driving features in a model the way in which driving behaviour deviates can be determined, so that a more informative assessment can be made.

The probability scores for the placebo measurements in the alprazolam dataset (3 and 4 hours after intake) and alcohol datasets (5 and 6 hours after intake) were very similar. This means that the average control measurement would get a probability score somewhat below 0.3. An average sleep-deprived score of 0.3 can be considered high for a subject driving under 'normal' circumstances but considering the relatively small training set this score is reasonable. When more data will become available, the recognition of control measurements will improve and the probability scores for these measurements will decrease.

The current study has also shown that driving behaviour after sleep deprivation shows characteristics, which are quite like those after alprazolam intake and, but to a lesser extent, after alcohol intake of 1 g/l. An explanation for this can be found in the analysis of feature importance. The features derived from the lane position, which also includes the SDLP, are most important. However, for assessing the effects of alcohol, the features calculated from the steer-speed parameter are also of high importance, while these are of minor importance in the detection of alprazolam and sleep deprivation.

As alprazolam is a sedative, great similarities with drowsiness were expected, which was confirmed in our study by the highest sleep deprived probability scores after intake of this drug. After 1.0 g/L alcohol intake also a substantial similarity of the abnormal driving characteristics was found, but to a lesser extent. After 0.5 g/l alcohol intake, which is just under the legal limit, there was almost no sleep-deprived effect in driving behavior. The probability scores seem to match the relative drowsiness expected from the interventions, indicating that the model shows the degree of sleep deprivation in other interventions.

Limitations

Although PPV and NPV of the model were quite high (around 75%), these predictive values were still suboptimal, possibly due to the relatively small number of subjects. This information can be derived from the violin plots, where still a great variation of scores can be observed. Therefore, more subjects are needed to be tested in order to obtain a reliable characterization of driving behaviour of a drug, and to neutralize the errors of the model. Part of the variation can also be explained by inter-subject differences. The effect of an intervention may vary substantially for each subject and the number of subjects in the datasets is relatively small. Moreover, the change in driving performance may also differ for subjects in a training set compared to subjects in a test set. Currently, the training set only contains male subjects. As we don't know the difference of the effect of sleep-deprivation on driving behaviour between males and females, we cannot say anything about the effect of sleep-deprivation in female subjects. However, female subjects have been involved in several previous car driving studies.(Åkerstedt et al. 2010; Banks et al. 2004)

It must be kept in mind that baseline measurements are still required for an accurate evaluation. In this way, the inter-subject variation may be reduced by correction for the baseline variation in driving style. The accuracy might improve when individual normal driving behaviour could be learned based on multiple control measurements in one subject.

It is difficult to estimate how dangerous driving behaviour is with sleep-deprived driving characteristics. Because sleep deprivation may lead to abnormal driving and may cause accidents, it seems reasonable to assume that with a higher score on sleep-deprived driving, driving performance is more impaired. On the other hand, a small overlap with driving after sleep deprivation does not mean that driving is safe. Driving behaviour might be negatively influenced in a different way. In the current study the alcohol intake well over the legal dose (1.0 g L^{-1}) appears to have less impact on the driving performance than alprazolam, while previous studies have shown that alcohol greatly impairs driving behavior. (Arnedt et al. 2000; Bunn et al. 2019; Huizinga et al. 2019; Irwin et al. 2017; Mets et al. 2011). In our previous study we have shown that driving behaviour after alcohol intake could be assessed with an accuracy of more than 80%.(van der Wall et al. 2020) Therefore, a test battery with multiple models is needed to give a good indication of the way driving is affected by drugs. This test battery consists of a set of models, characterizing the effect of CNS- active medicines on driving behavior.

The order of sleep deprivation for the interventions tested in the current study was as could have been anticipated (alprazolam highest, then 1.0 g/l alcohol, then 0.5 g/l alcohol, then placebo). Therefore, the model developed in the current study can be used to identify how interventions such as the intake of drugs may influence the driving performance in a sleep-deprived way. Therefore, this can serve as a benchmark in a test battery to characterize how drugs affect driving performance. In this way psychopharmacological drugs could be tested for effects on driving behaviour in an early stage of development.

REFERENCES

- Åkerstedt T, Ingre M, Kecklund G, Anund A, Sandberg D, Wahde M, Philip P, Kronberg P. 2010. Reaction of sleepiness indicators to partial sleep deprivation, time of day and time on task in a driving simulator–the drowsi project. Journal of sleep research. 19(2):298-309.
- Arnedt JT, Wilde G, Munt PW, MacLean AW. 2000. Simulated driving performance following prolonged wakefulness and alcohol consumption: Separate and combined contributions to impairment. Journal of Sleep Research. 9(3):233-241.
- Banks S, Catcheside P, Lack L, Grunstein RR, McEvoy RD. 2004. Low levels of alcohol impair driving simulator performance and reduce perception of crash risk in partially sleep deprived subjects. Sleep. 27(6):1063-1067.
- Bunn T, Singleton M, Chen I-C. 2019. Use of multiple data sources to identify specific drugs and other factors associated with drug and alcohol screening of fatally injured motor vehicle drivers. Accident Analysis & Prevention. 122:287-294.
- Darby P, Murray W, Raeside R. 2009. Applying online fleet driver assessment to help identify, target and reduce occupational road safety risks. Safety Science. 47(3):436-442.
- Gaspar JG, Brown TL, Schwarz CW, Lee JD, Kang J, Higgins JS. 2017. Evaluating driver drowsiness countermeasures. Traffic injury prevention. 18(sup1):S58-S63.
- Haraldsson P, Akerstedt T. 2001. Drowsiness--greater traffic hazard than alcohol. Causes, risks and treatment. Lakartidningen. 98(25):3018-3023.
- Houwing S, Mathijssen R, Brookhuis K. 2012. In search of a standard for assessing the crash risk of driving under the influence of drugs other than alcohol; results of a questionnaire survey among researchers. Traffic injury prevention. 13(6):554-565.
- Huizinga CR, Zuiker RG, de Kam ML, Ziagkos D, Kuipers J, Mejia Y, van Gerven JM, Cohen AF. 2019. Evaluation of simulated driving in comparison to laboratory-based tests to assess the pharmacodynamics of alprazolam and alcohol. Journal of Psychopharmacology. 33(7):791-800.
- Irwin C, Iudakhina E, Desbrow B, McCartney D. 2017. Effects of acute alcohol consumption on measures of simulated driving: A systematic review and meta-analysis. Accident Analysis & Prevention. 102:248-266.
- Irwin C, Monement S, Desbrow B. 2015. The influence of drinking, texting, and eating on simulated driving performance. Traffic injury prevention. 16(2):116-123.
- Koopmans I, Heima H, Doll RJ, van der Wall HEC, de Kam ML, Groeneveld GJ, Cohen AF, Zuiker RG. 2020. Sensitivity and validity of on-the-road and simulated driving test to measure impaired driving behaviour: Effect of sleep deprivation. Sleep Medicine (in preperation).
- Mets M, Kuipers, E, de Senerpont, DLM et al. 2011. Effects of alcohol on highway driving in the stisim driving simulator. Human Psychopharmacology Clinical and Experimental. 26(6):434-439.
- Mets MA, Kuipers E, de Senerpont Domis LM, Leenders M, Olivier B, Verster JC. 2011. Effects of alcohol on highway driving in the stisim driving simulator. Human Psychopharmacology: Clinical and Experimental. 26(6):434-439.
- Peters RD, Wagner E, Alicandri E, Fox JE, Thomas ML, Thorne DR, Sing HC, Balwinski SM. 1999. Effects of partial and total sleep deprivation on driving performance. Public Roads. 62(4).
- Philip P, Sagaspe P, Moore N, Taillard J, Charles A, Guilleminault C, Bioulac B. 2005. Fatigue, sleep restriction and driving performance. Accident Analysis & Prevention. 37(3):473-478.
- Robertson RD, Hing MM, Pashley CR, Brown SW, Vanlaar WG. 2017. Prevalence and trends of drugged driving in canada. Accident Analysis & Prevention. 99:236-241.
- Sagberg F, Selpi, Bianchi Piccinini GF, Engström J. 2015. A review of research on driving styles and road safety. Human factors. 57(7):1248-1275.
- Schwarz C, Gaspar J, Miller T, Yousefian R. 2019. The detection of drowsiness using a driver monitoring system. Traffic injury prevention. 20(sup1):S157-S161.
- Soares S, Ferreira S, Couto A. 2020. Driving simulator experiments to study drowsiness: A systematic review. Traffic injury prevention. 21(1):29-37.
- van der Wall H, Doll R, van Westen G, Koopmans I, Zuiker R, Burggraaf J, Cohen A. 2020. The use of machine learning improves the assessment of drug-induced driving behaviour. Accident Analysis & Prevention. 148:105822.
- Van Steveninck A, Van Berckel B, Schoemaker R, Breimer D, Van Gerven J, Cohen A. 1999. The sensitivity of pharmacodynamic tests for the central nervous system effects of drugs on the effects of sleep deprivation. Journal of Psychopharmacology. 13(1):10-17.
- Verster JC, Roth T. 2011. Standard operation procedures for conducting the on-the-road driving test, and measurement of the standard deviation of lateral position (sdlp). International journal of general medicine. 4:359.
- Wong HB, Lim GH. 2011. Measures of diagnostic accuracy: Sensitivity, specificity, ppv and npv. Proceedings of Singapore healthcare. 20(4):316-318.
- Zoethout RW, de Kam ML, Dahan A, Cohen AF, van Gerven JM. 2012. A comparison of the central nervous system effects of alcohol at pseudo-steady state in caucasian and expatriate japanese healthy male volunteers. Alcohol. 46(7):657-664.

SUPPLEMENTARY TABLES

Table A1 *Overview of all parameters.*

Parameters	Description
Strip-index	lateral position on the entire highway
Lane Position	lateral position in the lane
Speed	speed in km/h
Steer	steer-position
Steer-speed	speed of steering to the right
Front-distance-meters	distance to the car in front in meters

Table A2 *Overview of all features.*

Feature Name	mean	STD	diff	Intensity at low frequency	Intensity at high frequency	min	max
Strip-index	x	x	x	x	x		
Lane Position	x	x	x	x	x		
Speed (km/h)	x	x	x	x	x	x	x
Steer	x	x	x	x	x		
Steer-speed	x	x	x	x	x	x	x
Front-distance-meters	x	x	x	x	x	x	x
Features from original study							
SDLP (standard deviation of the lateral position)							
MLP (mean lateral position)							
MS (mean speed)							
SDS (standard deviation of speed)							

CHAPTER 6

DISCRIMINATIVE MACHINE LEARNING ANALYSIS FOR SKIN MICROBIOME: DISCOVERING BIOMARKERS IN PATIENTS WITH SEBORRHEIC DERMATITIS.

Published in *Artificial Intelligence for Medical Sciences*, 2022
Volume 148, p. 1-7

van der Wall H.E.C.[1,2,*], Doll, R.J.[1], van Westen G.J.P.[2], Niemeyer-van der Kolk T.[1], Feiss, G.[6], Pinckaers, H.[3], van Doorn, M.B.A.[4], Nijsten T[4], Sanders M.G.H[4], Cohen, A.F.[1,2], Burggraaf, J.[1,2,5], Rissmann, R.[1,2], Pardo, L.M[4]

1 Centre for Human Drug Research, NL
2 Drug Discovery and Safety, Leiden Academic Centre for Drug Research, NL
3 Department of Pathology, Radboud University Medical Center, NL
4 Department of Dermatology, Erasmus Medical Center, NL
5 Department of Clinical Pharmacology, Leiden University Medical Center, NL
6 GL Feiss Consulting, LLC, Telford, Pennsylvania, US

ABSTRACT

In recent years the skin microbiome has taken center stage as drug target and as disease biomarker. Computational analyses of microbiome sequencing data from patients with skin diseases, for example seborrheic dermatitis, can be performed to identify discriminative biomarkers in the microbiome profile. The aim of the present study was twofold, namely to employ machine learning to predict disease from the microbiome dataset, and to identify discriminative biomarkers in the microbiome of patients with seborrheic dermatitis versus healthy controls using machine learning techniques.

The population consisted of 97 patients with seborrheic dermatitis and 763 healthy controls. Skin swabs were taken from naso-labial fold (lesional skin: n = 22; non-lesional skin: n = 75, controls: n = 763). Using an extra trees machine learning model, differences between the skin microbiome of patients with seborrheic dermatitis versus healthy controls were characterized. Subsequently, the most important microorganisms for discrimination were determined by feature analysis and SHapley Additive exPlanations (SHAP) values.

The accuracy of the prediction models to discriminate between skin affected by seborrheic dermatitis and facial skin from healthy subjects was 77 % and the ROC-AUC was 83 %. Next to C*utibacterium* and S*taphylococcus*, the most important organisms for discrimination had a relatively low occurrence.

Our study showed that machine learning can be utilized to identify discriminating biomarkers in the microbiome skin. This indicates that machine learning can be of major importance in basic skin research, and in the discovery and development of new individualized therapies, involving the microbiome.

INTRODUCTION

The skin is the largest organ of the human body and is colonized by a wide range of microorganisms.[1] Many of the micro-organisms living on the skin (its microbiome) are harmless and, in some cases, provide vital functions.

Despite the great interest of the skin as an ecosystem during the past decade, the study of the skin microbiome was until recently restricted by the low host-commensal cell ratio and the high taxonomical divergence among skin sites.[2] This changed by the introduction of methodology to remove microbial DNA from low biomass skin samples such as described by Garcia-Garcera et al. in 2013. The authors utilized a combination of molecular techniques that involved standard, quantitative PCR and amplicon sequencing of 16S rRNA, which significantly improved the field of skin microbiome research. At present, the skin microbiome is known to be involved in several skin diseases.[3] This breakthrough has led to additional knowledge on specific microorganisms that play a role in some of these skin disorders, for instance the role of *Staphylococcus aureus* in atopic dermatitis and *Cutibacterium acnes* in acne vulgaris. However, the role of microorganisms that are less abundant is still largely unknown. It is plausible that the presence of a combination of several different organisms forming a specific microbial profile might also contribute to the development and subtype of skin disease. A challenge to explore this hypothesis is however hampered by the magnitude of the data which analysis is frequently beyond conventional data analyses. Machine learning may offer a solution because the underlying computational analyses may facilitate the identification of specific patterns of microorganisms that are discriminative for a specific type of skin diseases.[4]

Machine learning has been rapidly adopted in microbiome studies for diagnosing clinical diseases. Modelling of the human microbiome by machine learning offers the potential to identify specific microbial biomarkers and may aid in the diagnosis of many clinical diseases. For instance, machine learning has already shown its ability to identify key features (markers) and modelling predictive biomarker signature in a variety of fields, including oncology[5–7], neurology[5–7], immunology[8], gastroenterology[9], diabetes[10], and skin diseases.[11,12] The advantages of machine learning techniques over classical statistical models are to infer relationships between variables for automatic pattern discovery and handling with multi-dimensional data.[13] As a result, machine learning may be highly informative for the development of therapeutic modalities to ameliorate the microbial imbalance and to counteract certain pathogens. By training a highly accurate model it is easy to find out which features are most informative for classification. For a dataset with many different features, in this case more than 600, machine learning is therefore saving time and effort, compared to existing statistical methods. In addition, the benefits of machine learning comprise flexibility and scalability compared with conventional statistical approaches, which makes it deployable for several tasks, such as diagnosis and classification, and survival predictions.[14] As a result, machine learning may be highly informative for the development of therapeutic modalities to ameliorate the microbial imbalance and to counteract certain pathogens.

The aim of the present study was twofold, namely to employ machine learning to predict disease from the microbiome dataset, and to identify discriminative biomarkers in the microbiome of patients with seborrheic dermatitis versus healthy controls using machine learning techniques. We hypothesized that the microbiome-based biomarkers alone can be used to predict the diagnosis. These models can then be employed toidentify discriminative biomarkers in the microbiome.

METHODS & MATERIALS

All data used in the present study were obtained from a previous study performed in participants from the Rotterdam Study.[15] This was a cross sectional study embedded in a population-based study. Skin swabs were taken from naso-labial fold from 97 participants with seborrheic dermatitis (lesional skin: n = 22; non-lesional skin n=75) and controls without skin conditions on the face or scalp (n = 763). Participants with seborrheic dermatitis and involvement of the nasolabial fold were considered lesional cases, and those without involvement of the nasolabial fold non-lesional cases. The median age was 53 years in the control group, 56 years for non-lesional cases and 68 years for lesional cases (for further details see Sanders, Nijsten[15]). The Rotterdam Study has been approved by the Medical Ethics Committee of the Erasmus MC (registration number MEC 02.1015) and by the Dutch Ministry of Health, Welfare and Sport (Population Screening Act WBO, license number 1071272-159521-PG). All participants provided written informed consent to participate in the study and to have their information obtained from treating physicians.

Data collection for the model

In all participants included in the study, the skin microbiome was analyzed by amplifying the V1 to V3 variable regions of the 16S rRNA gene using the 27F-519R primer pair and dual indexing. The genes were annotated using the Silva database. In the current study, microbiome data from the three categories of skin from the face were analyzed; facial skin from controls, (facial) non-lesional and lesional skin from patients with seborrheic dermatitis. To show the clearly visible differences between the skin categories, average microbiome profiles were created for each category of skin by taking the average occurrence of each bacterium present in all pertaining subjects. Subsequently, by using machine learning the lesional skin from patients with seborrheic dermatitis was characterized in order to discriminate it from the skin of healthy subjects. The occurrence of 686 organisms at genus level present in either one of both datasets were used as features. In addition, three well known alpha-diversity indices, the Simpson's diversity index, the Shannon diversity index, and the Chao1 index, were used as features.[16]

Data pre-processing and selection

As there were more observations available from healthy skin than from the skin affected by seborrheic dermatitis, there was an imbalance in the data set. A four-fold cross validation was used. Therefore, six microbiome profiles (close to 25 percent of the total number of affected profiles) were used for validation of the model each fold. The same number of healthy profiles were used to produce a balanced validation set. The remaining profiles were used to train the model (training set).

To create a balanced training set, the SMOTE (Synthetic Minority Oversampling Technique) algorithm was applied to produce 'synthetic' profiles of the skin affected by seborrheic dermatitis based on the values already present in these underrepresented microbiome profiles.[17]. Next, the features in both the training and validation sets were standardized based on the mean and standard deviation of the features in the training set.

Feature selection

As the last preprocessing step, feature selection is performed on all 689 features, including the diversity indices. When two features had a high correlation (> 0.9 or < -0.9), only the most important one – based upon the feature importance of fitting the model on the training set – was used.[18] The features were selected by the training set in unsupervised fashion.

Machine learning

Several different machine learning algorithms which were obtained from the scikit-learn module version 1.0.1 in python 3.7.9 were employed on the data. A DecisionTree Classifier, a RandomForest Classifier, a GradientBoostingClassifier, a Supprort Vector Classifier, Logistic Regression, and a Extra Trees Classifier were used to make an attempt to distinguish healthy from affected skin. Prior to training, a nested cross-validation (within the training set) was used to optimize the model hyperparameters. This process of oversampling, feature selection, optimization, training, and validation was repeated in each fold with different training and validation data. In each fold, the validation was performed with profiles that had not previously been in a validation set, so that all 22 different profiles from skin affected with seborrheic dermatitis were tested at least once. Two profiles were twice in the test set. The optimal models were evaluated on the validation fold with the accuracy, sensitivity, specificity,and Area Under the Receiver OperatingCharacteristic(ROC) Curves (AUC). An overview of the machine learning workflow is shown in supplementary Figure 2. The best performing model was used for further analyses. The performance of the optimized model using the selected features was compared with the performance of conventional logistic regression using all features. The performance of the optimized model using the selected features was compared with the performance of conventional logistic regression using all features.

To gain insight into the impact of the individual features on the predictions, SHAP(SHapley Additive exPlanations) values were calculated. {Lundberg, 2017 #16}[19] To validate the importance of the features, the feature values of the correctly and incorrectly predicted occasions were compared. The impact of the features was validated by means of the feature importance of the model.

RESULTS

The three skin categories showed many similarities in the microbiome. As expected, *Staphylococcus* and *Cutibacterium* showed the highest relative abundance (range 20–50%); *Cutibacterium* was highest on average in healthy profiles and in profiles of the non-lesional skin of patients with seborrheic dermatitis. *Staphylococcus* was highest in the skin affected by seborrheic dermatitis. Figure 1 shows the average skin microbiome profiles of lesional and non-lesional facial skin of patients with seborrheic dermatitis and of healthy skin. The non-lesional skin shows an average microbiome profile which is in between healthy and affected skin.

From all the models tested, the extra trees classifier performed best. The results of the employment of this model type are shown below.

Figure 2 shows the true labels (clinical diagnosis) versus the predicted labels for seborrheic dermatitis versus controls based on the skin microbiome. The models predicting seborrheic dermatitis had an overall accuracy of 77% (range 73 – 81%, compared to 48% ± 14% with conventional binary logistic regression). Out of the 24 profiles

Figure 1 *Average microbiome profiles of the facial skin of healthy subjects and of the non-lesional and lesional (facial) skin of patients with seborrheic dermatitis. The y-axis shows the relative occurrences in percentages. The profiles show organisms at genus level that had on average an occurrence of more than 1%. Other organisms are combined as 'other'.*

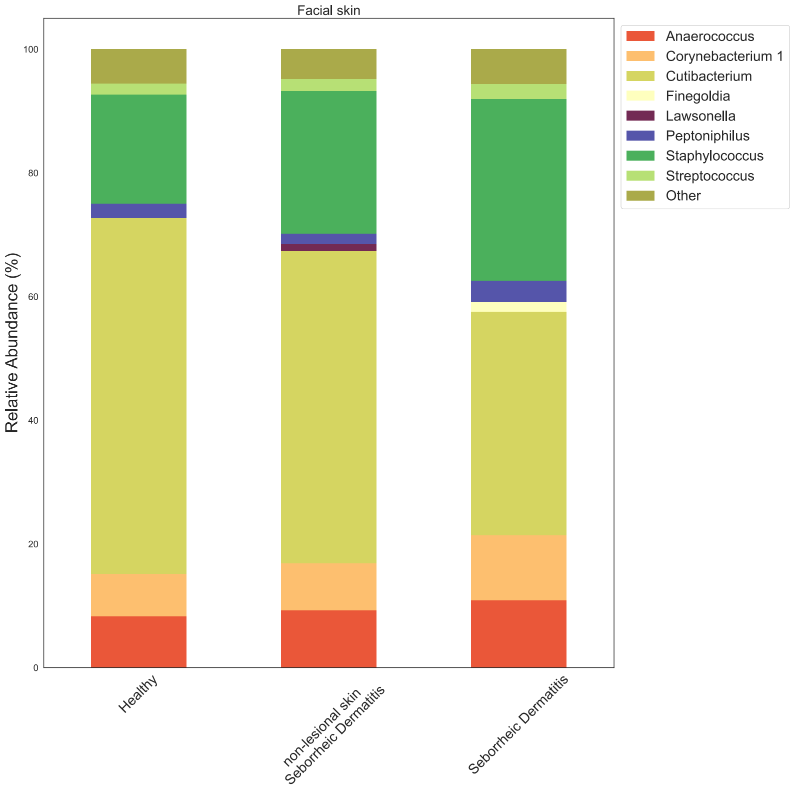


Figure 2 *Confusion matrix (predictive analysis tool) of the predicted diagnoses based on the skin microbiome profile of patients with seborrheic dermatitis by means of machine learning. The x-axis shows the predicted labels. The y-axis shows the true labels.*

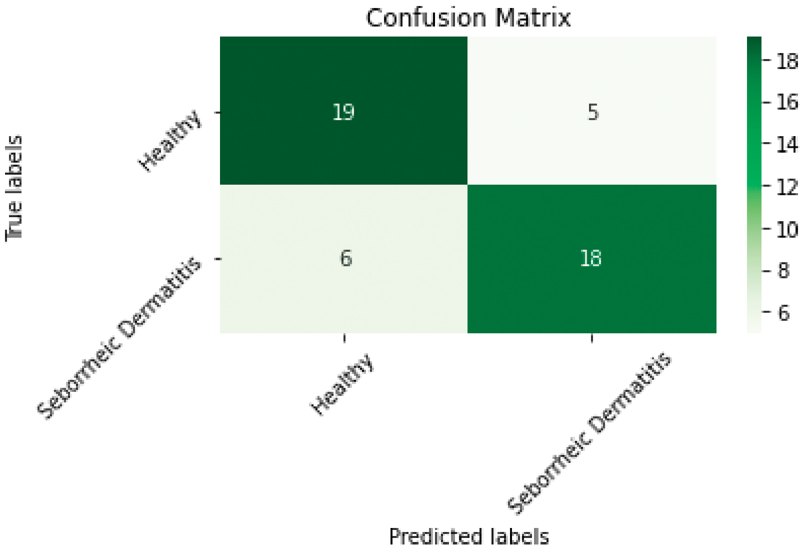


Figure 3 *Receiver Operating Characteristic (ROC) curve of the models predicting seborrheic dermatitis where the black line is the mean curve and the gray area is the standard deviation.*

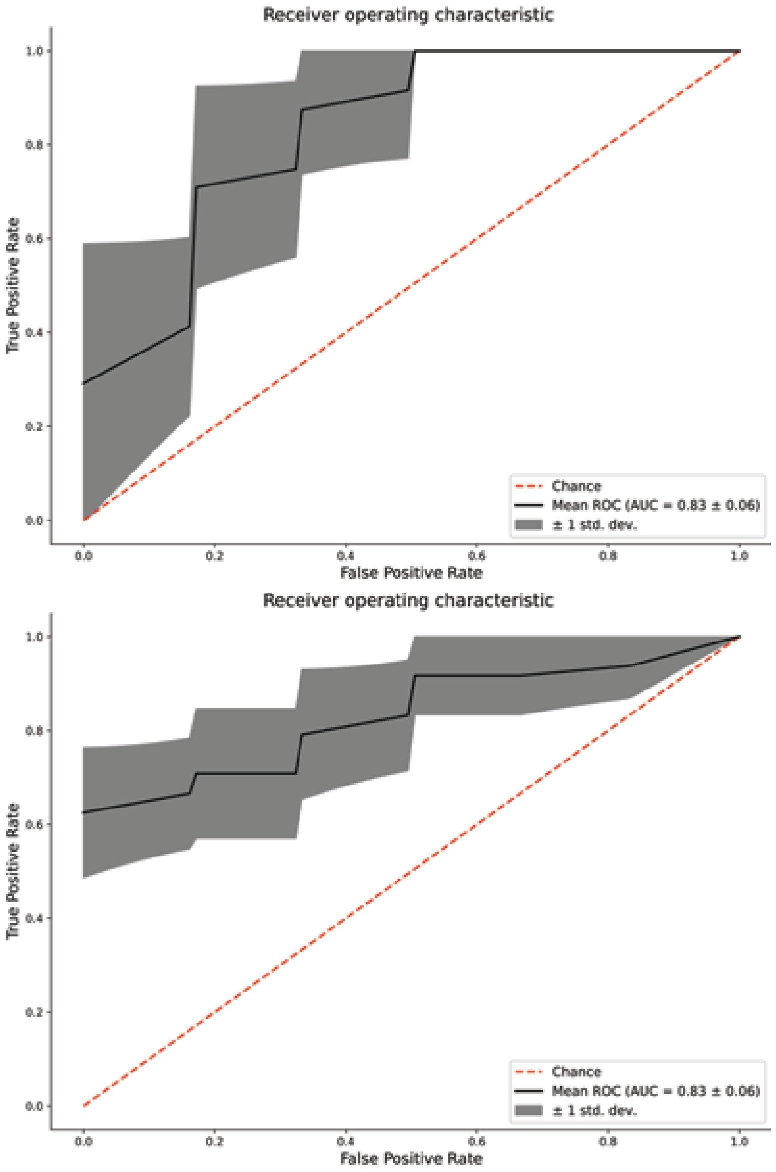
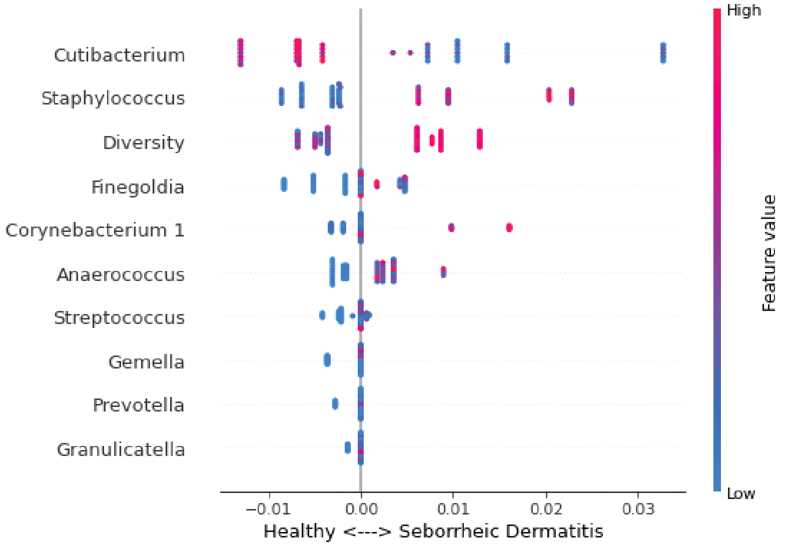


Figure 4 *SHAP values of the 10 features with the highest impact on the prediction of disease diagnosis for seborrheic dermatitis. A relatively high occurrence of a microorganism is shown in red, whereas a relatively low occurrence of a microorganism is shown in blue. The predicted diagnosis of control is on the left side of the x-axis, and of seborrheic dermatitis on the right side of the x-axis. Shorter bars mean less impact on diagnosis.*



with seborrheic dermatitis, 18 were correctly predicted, indicating a sensitivity of 75% (range 65-85%). Of the 24 healthy profiles, 19 were correctly predicted, indicating a specificity of 79% (range 71-87%). The average AUC of the models was 83% (range 77-89%, Figure 3).

Figure 3 shows the impact of the occurrence of each organism on the predictions of seborrheic dermatitis versus healthy control. A low occurrence of *Cutibacterium* and a high occurrence of *Staphylococcus* was shown to be most predictive for the diagnosis of seborrheic dermatitis. It can be observed that the other micro-organisms, which had any impact in the discrimination of seborrheic dermatitis, showed a relatively low occurrence.

The boxplots of the standardized values of the most important organisms for the correct and wrongly predicted profiles are shown in the supplementary material. These Figures confirm the findings of the SHAP values.

DISCUSSION

In the present study we demonstrated that machine learning- based models may facilitate the identification of discriminative biomarkers in the microbiome of patients. These findings are particularly important for skin diseases, in which the microbiome has not been fully elucidated. Modulations of skin microbiome composition to restore host–microbiome homeostasis could become important future strategies to treat or prevent skin disease.[3] This highlights the potential important role of machine learning in the discovery of targets for new medical therapies.

Machine learning has been recently applied in microbiome studies for diagnosing clinical diseases in various fields, such as oncology, neurology, immunology and dermatology.[4] In the present study we aimed to develop a machine learning model to predict the diagnosis using skin microbiome profiles from patients with seborrheic dermatitis. The created models provided a unique insight into the types and complex patterns of micro-organisms involved this skin condition. Although many data are available on the skin microbiome in seborrheic dermatitis, our results show that bacteria with a low abundance are also valuable for disease discrimination. The factor of low *Cutibacterium* contributes to our models, being consistent with previous reports.[20]

Many factors are known to significantly affect the skin microbiome, including weather conditions and washing behaviour, but also skin diseases and the use of therapeutic agents.[21–24] The current study showed that, based on the microbiome alone, machine learning models could predict a diagnosis of seborrheic dermatitis with an accuracy of 77%. The area under the curve from our models was 83%. There was a strong predictive correlation between the microbiome profile and the specific dermatologic disease. Given the large number of influencing factors that can affect the microbiome (although excluded as much as possible in the clinical trial) perfect predictive power using the microbiome cannot be expected. But, because of their high performances, these models could potentially be used toidentify discriminative disease biomarkers in the microbiome.

Value of machine learning

Machine learning methods are being actively and widely used to elucidate the composition of microbiome and to investigate how they affect host phenotypes.[23,25] Various studies have already explored the power of machine learning to use microbiome patterns to predict host characteristics.[23,26–28] In addition, machine learning has earlier been applied on the skin microbiome to predict the postmortem interval.[29] In the current study we have shown that disease biomarkers can also be found using machine learning in the skin microbiome. Some of the parameters, such as a high abundance of Staphylococcus, and the diversity are already known to play a role in seborrheic dermatitis.[30–32]

Furthermore, machine learning identified distinctive organisms that are not initially considered important to investigate based on high occurrence rates. As shown in Figure 4, the occurrence of *Corynebacterium 1*, *Anaerococcus*, and *Finegoldia* also play a role in the distinction between facial skin from healthy individuals and subjects with seborrheic dermatitis. While the above-noted three organisms might have been identified through occurance alone (Figure 1), *Gemella*, *Prevotella* and *Granullicatella* would not have been identified as they occurred at less than one percent and are included in 'other'. Use of machine learning identified organisms which would otherwise have been overlooked.

Using the same dataset, conventional binary logistic regression produced results with lower discrimination ability.

The results of this proof of concept study indicate that machine learning can be a valuable tool to find organisms that distinguish diseased skin from healthy skin . While the microbiology of seborrheic dermatitis has been thought to be fairly well elucidated, the importance of low occurrence organisms was shown. This suggest that in skin diseases with a less known and/or more complex microbiome profile machine learning could be a valuable investigative tool.

Limitations

Some limitations of the study should be noticed. Apart from a relatively low number of patients, DNA from microbial eukaryotes, such as yeast or fungi, could unfortunately not be classified meaning by this 16S gene screening method and limited primer set, only prokaryotic DNA, known in the database, could be recognized. This precludes recognition of fungi, such as yeast and bacteria not known in the database, while the fungal genus Malassezia is also known to be a potential biomarker for seborrheic dermatitis.[33,34] For future studies, sequencing of the internal transcribed spacer (ITS) region[35], would enable the possibility of identifying fungi as potential biomarkers.

There is an age difference between the control group and lesional cases. This could be a factor of disease. Future studies, linking the organisms to an age group would be very interesting.

Genus level was the deepest screening level in this study, indicating a second limitation of screening by means of 16S DNA sequencing with a limited set of primers. Therefore, it has to be taken into account that the exact distribution of species within the genus is unknown. For example a large part of the 16S DNA recognized as Staphylococcus might not originate from the species S. aureus but from S. epidermidis, a species which is very common on healthy skin.[36] This should be investigated in future studies with novel techniques that have the ability to give deeper insight on species or even strain level.

CONCLUSION

In recent years, various elements in the skin microbiome have become of high interest for pharmaceutical companies as new drug targets. Despite the challenges and hurdles yet to overcome, it seems very likely that microbiome modulation will play a future role in the treatment of skin disease. From our study it has become clear that machine learning can be instrumental for the identification of biomarkers in the microbiome of skin. Consequently, machine learning can be of major importance in basic skin research, and in the discovery and development of new individualized therapies, involving the microbiome.

References

1 Grice, E.A. and J.A. Segre, The skin microbiome. Nature reviews microbiology, 2011. 9(4): p. 244-253.

2 Garcia-Garcerà, M., et al., A new method for extracting skin microbes allows metagenomic analysis of whole-deep skin. PloS one, 2013. 8(9): p. e74914.

3 Zeeuwen, P.L., et al., Microbiome and skin diseases. Current opinion in allergy and clinical immunology, 2013. 13(5): p. 514-520.

4 Leclercq, M., et al., Large-scale automatic feature selection for biomarker discovery in high-dimensional OMICs data. Frontiers in genetics, 2019. 10: p. 452.

5 Zhang, D., D. Shen, and A.s.D.N. Initiative, Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. NeuroImage, 2012. 59(2): p. 895-907.

6 Deshpande, G., et al., Identification of neural connectivity signatures of autism using machine learning. Frontiers in human neuroscience, 2013. 7: p. 670.

7 Fekete, T., et al., Multiple kernel learning captures a systems-level functional connectivity biomarker signature in amyotrophic lateral sclerosis. PloS one, 2013. 8(12): p. e85190.

8 Sutherland, A., et al., Development and validation of a novel molecular biomarker diagnostic test for the early detection of sepsis. Critical care, 2011. 15(3): p. 1-11.

9 Kohli, A., E.A. Holzwanger, and A.N. Levy, Emerging use of artificial intelligence in inflammatory bowel disease. World Journal of Gastroenterology, 2020. 26(44): p. 6923.

10 Hathaway, Q.A., et al., Machine-learning to stratify diabetic patients using novel cardiac biomarkers and integrative genomics. Cardiovascular diabetology, 2019. 18(1): p. 1-16.

11 Fortino, V., et al., Machine-learning–driven biomarker discovery for the discrimination between allergic and irritant contact dermatitis. Proceedings of the National Academy of Sciences, 2020. 117(52): p. 33474-33485.

12 Johansson, H., et al., A genomic biomarker signature can predict skin sensitizers using a cell-based in vitro alternative to animal tests. BMC genomics, 2011. 12(1): p. 1-19.

13 Marcos-Zambrano, L.J., et al., Applications of machine learning in human microbiome studies: a review on feature selection, biomarker identification, disease prediction and treatment. Frontiers in Microbiology, 2021. 12: p. 313.

14 Rajula, H.S.R., et al., Comparison of conventional statistical methods with machine learning in medicine: diagnosis, drug development, and treatment. Medicina, 2020. 56(9): p. 455.

15 Sanders, M.G., et al., Composition of cutaneous bacterial microbiome in seborrheic dermatitis patients: A cross-sectional study. PloS one, 2021. 16(5): p. e0251136.

16 Simpson, E.H., Measurement of diversity. Nature, 1949.

17 Chawla, N.V., et al., SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 2002. 16: p. 321-357.

18 van der Wall, H., et al., The use of machine learning improves the assessment of drug-induced driving behaviour. Accident Analysis & Prevention, 2020. 148: p. 105822.

19 Lundberg, S. and S.-I. Lee, A unified approach to interpreting model predictions. arXiv preprint arXiv:1705.07874, 2017.

20 Xu, Z., et al., Dandruff is associated with the conjoined interactions between host and microorganisms. Scientific reports, 2016. 6(1): p. 1-9.

21 Nørreslet, L.B., T. Agner, and M.-L. Clausen, The Skin Microbiome in Inflammatory Skin Diseases. Current Dermatology Reports, 2020. 9(2): p. 141-151.

22 Picardo, M. and M. Ottaviani, Skin microbiome and skin disease: the example of rosacea. Journal of clinical gastroenterology, 2014. 48: p. S85-S86.

23 Carrieri, A.P., et al., Explainable AI reveals changes in skin microbiome composition linked to phenotypic differences. Scientific Reports, 2021. 11(1): p. 1-18.

24 Shukla, S.K., N.S. Murali, and M.H. Brilliant, Personalized medicine going precise: from genomics to microbiomics. Trends in molecular medicine, 2015. 21(8): p. 461.

25 Namkung, J., Machine learning methods for microbiome studies. Journal of Microbiology, 2020. 58(3): p. 206-216.

26 Knights, D., E.K. Costello, and R. Knight, Supervised classification of human microbiota. FEMS microbiology reviews, 2011. 35(2): p. 343-359.

27 Zhou, Y.-H. and P. Gallins, A review and tutorial of machine learning methods for microbiome host trait prediction. Frontiers in genetics, 2019. 10: p. 579.

28 Moitinho-Silva, L., et al., Predicting the HMA-LMA status in marine sponges by machine learning. Frontiers in microbiology, 2017. 8: p. 752.

29 Johnson, H.R., et al., A machine learning approach for using the postmortem skin microbiome to estimate the postmortem interval. PloS one, 2016. 11(12): p. e0167370.

30 Tamer, F., et al., Staphylococcus aureus is the most common bacterial agent of the skin flora of patients with seborrheic dermatitis. Dermatology practical & conceptual, 2018. 8(2): p. 80.

31 Vicent, L. and M. Martínez-Sellés, Electrocardiogeriatrics: ECG in advanced age. Journal of electrocardiology, 2017. 50(5): p. 698-700.

32 Tanaka, A., et al., Comprehensive pyrosequencing analysis of the bacterial microbiota of the skin of patients with seborrheic dermatitis. Microbiology and immunology, 2016. 60(8): p. 521-526.

33 Hay, R., Malassezia, dandruff and seborrhoeic dermatitis: an overview. British Journal of Dermatology, 2011. 165: p. 2-8.

34 Gupta, A.K., et al., Skin diseases associated with Malassezia species. Journal of the American Academy of Dermatology, 2004. 51(5): p. 785-798.

35 Metcalf, J.L., et al., Microbial community assembly and metabolic function during mammalian corpse decomposition. Science, 2016. 351(6269): p. 158-162.

36 Seite, S. and T. Bieber, Barrier function and microbiotic dysbiosis in atopic dermatitis. Clinical, cosmetic and investigational dermatology, 2015. 8: p. 479.

SUPPLEMENTARY FIGURES

Figure 1 *Boxplots of the standardized values of the 10 most important features for distinguishing seborrheic dermatitis from healthy skin based on the microbiome in order of importance of the models. On the left side are the healthy profiles and right are profilesaffected by seborrheic dermatitis. Visualized in blue are the profiles predicted as healthy and the profiles predicted as affected are shown in orange. The y-axis shows the values of the standardized occurrences. The p values of an independent t-test between the right an wrong predicted profiles are shown.*

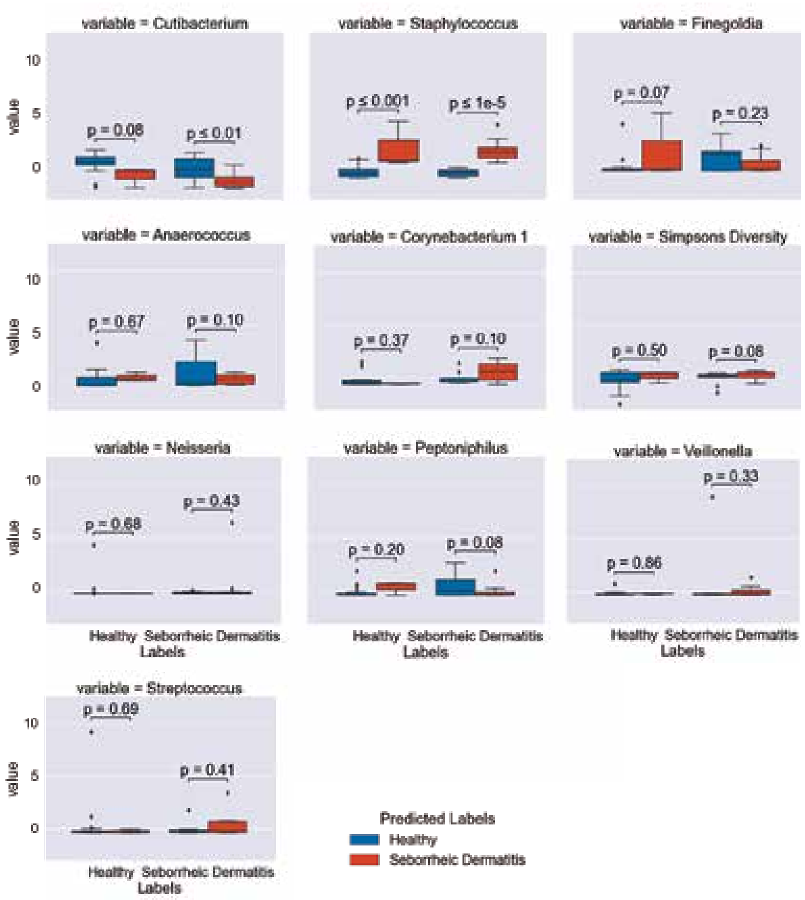
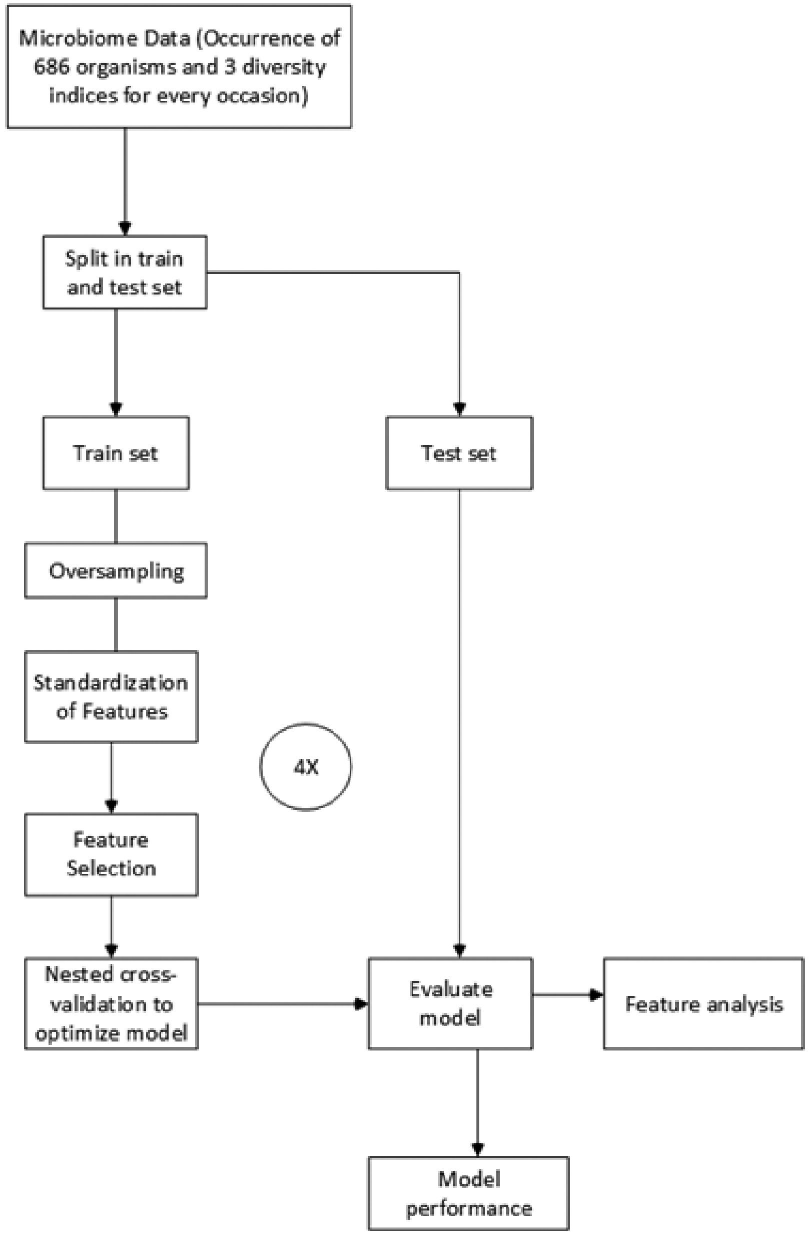


Figure 2 *Machine learning workflow (4 fold cross-validation).*


Microbiome Data (Occurrence of 686 organisms and 3 diversity indices for every occasion)
Split in train and test set
Train set
Test set
Oversampling
Standardization of Features
4X
Feature Selection
Nested cross-validation to optimize model
Evaluate model
Feature analysis
Model performance

CHAPTER 7

GENERAL DISCUSSION

In this PhD thesis machine learning techniques have been applied for data analysis on large data sets obtained in early stage clinical research projects. In drug development and clinical trials, biomarkersmay be used to help identify populations for a study, monitor therapeutic response, and identify side effects. Although an increasing number of analyses is performed, it is not always clear 1) how to handle the collected data, and 2) whether all these analyses are useful to study. We hypothesized that a good way to use these data is to employ a machine learning algorithm. The models built based on the data could then serve as new biomarkers to recognize the intended and unintended effects of (new) drugs. Also, after using a model, its predictions could be explained. It may lead to a better understanding of how the various features used in the algorithm influence the cause and effect of a specific drug. With the help of machine learning, the data collected at an early stage of clinical drug discovery is optimally exploited.

The use of machine learning on three different kinds of data derived from early phase clinical trials is explored; data based on electrocardiographic (ECG) measurements (classical data), data collected in driving performance test (innovative data), and data collected in microbiome studies (emerging data).

Chapter 1 comprises a general introduction and overview of the presented studies.

In **Chapter 2**, we showed that the number of ECG replicates in so-called QT studies has a substantial effect on the interpretation of a compound's QT interval prolonging potential for all used QT_c formulas. This analysis was performed after the prolongation was corrected for the heart rate with one of the several QT correction formulas. We observed effects on the mean QT_c interval prolongation and on the range of the 90% confidence interval of the QT_c interval prolongation—parameters that are required by the regulators. This analysis showed that for all QT correction formulas there is a mean difference of 1 ms when triplicate ECGs were extracted compared with 18 ECG replicate extraction. These findings imply that triplicate ECG extractions are likely to result in inaccurate QT estimation and can only be used as an exploratory method, but not to unambiguously quantify a QT prolonging effect. The analysis performed in this thesis on a large volume of ECG replicates can be performed after the compound's development has been moved into a later stage. It can be cancelled in case the development of the compound is abandoned, thereby saving resources.

By using all 12-lead surface ECG parameters, we developed machine learning models that allow prediction of physiologic cardiac age of healthy subjects in **Chapter 3.** We were able to estimate the age of a healthy subject with a mean absolute error of 6.9 years and to analyze the impact of the ECG features, using a neural network. Additionally, we demonstrated the impact of the ECG features on the predicted age.

The impact of physiologic aging on the various ECG features was analyzed using SHapley Additive exPlanations (SHAP) values. By means of machine learning techniques a combination of various ECG changes allowed a more accurate insight into the aging of the heart.

As the models were trained using only healthy subjects, we can assume that the predicted actual age is equal to the cardiac age. We also assumed that a decline in heart health occurs with age. Therefore, our models may serve as a benchmark for (new) pharmacological drugs on potential decline or improvement of physiologic health of the heart.

In **Chapter 4,** we presented a study to create an algorithm using machine learning to detect driving impairment due to the use of drugs. We included multiple driving features rather than the Standard Deviation of the Lateral Position (SDLP) alone. These algorithms provided improved insight into the way driving behaviour was affected by alcohol and the benzodiazepine alprazolam. For both the alprazolam and alcohol models the SDLP was a major determinant. However, when using multiple driving features rather than only SDLP as read-out, the performance of the models improved with 18 and 26 percentage points for alprazolam and alcohol, respectively. This observation emphasizes the importance of analysing multiple features rather than SDLP alone, when assessing drug effects on driving behaviour.

Over the years, intelligent driving assistance systems have been created to mitigate incidents because most of the accidents relate to driver performance[1] Therefore, it is important to assess the effect of drugs on car driving. It may be possible to develop systems that learn normal driving behaviour of an individual and to detect abnormalities for that driver. This may be a substantial advantage particularly in assessing the effect of drugs or alcohol. Ideally, one would use such models as created in Chapter 4 to evaluate new drugs/interventions.

We used machine learning in **Chapter 5** to classify the effect of sleep deprivation on driving behaviour. The created model performed with an average accuracy of 77%. It was shown that the created models in Chapters 4 and 5 can be used as some kind of 'test battery' to create a unique 'fingerprint' (profile) with respect to both desired and undesired effects on driving. Using this model, the effects of a newly developed drug on driving behaviour can be compared with the effects of sleep deprivation. In this way, a series of these machine learning models could evolve to a test battery, allowing a more precise and accurate evaluation of abnormal driving behaviour by creating a predictive effect profile for a specific drug. Using these algorithms, adequate probability scores can be given to test-drives, which provide an indication about the way and the extent to which these drugs are modifying driving behaviour.

In **Chapter 6**, we demonstrated that machine learning-based models may facilitate the identification of discriminative biomarkers in the microbiome of patients. These findings are particularly important for skin diseases, in which the microbiome has not been fully elucidated. The created models provided a unique insight into the types and complex patterns of micro-organisms involved in this skin condition.

Based on the microbiome alone, machine learning models could predict a diagnosis of seborrheic dermatitis with an accuracy of 77%. The area under the curve from our models was 83%. Machine learning identified distinctive organisms that were not initially considered important to investigate based on high occurrence rates. Use of machine learning identified organisms which would otherwise have been overlooked.

The results of this proof-of-concept study indicate that machine learning can be a valuable tool to find organisms that distinguish diseased skin from healthy skin. This suggests that in skin diseases with a less known and/or more complex microbiome profile machine learning could be a valuable investigative tool.

FUTURE PERSPECTIVES

Although the ultimate assessment of the effect of a drug or intervention will always need the judgement by individuals with a medical background, applying machine learning to the large amount of data derived from drug development can offer substantial support and may lead to new insights. This may be expected, as machine learning is already being used in decision making across a broad range of fields.[2] However, in drug development, every subject or patient is unique, and therefore the result of historical data cannot be completely reliable. There are still some problems to overcome, for example the pragmatic issue that our health system is reluctant to completely entrust a machine with a task that a human can do at higher accuracy, even in case of substantial cost savings.[2] Machines are not endowed with common sense, and therefore need many more examples than a human physician would. However, when physicians are unsure or when a decision has to be made with major consequences, machine learning models trained on historical data can help.[3] Moreover, it is impossible for a physician to learn from as many historical examples as a machine could. A machine learning model can learn from more cases than a physician would experience during his or her entire career. As more and more data of the early phases of drug development become available soon, more accurate estimates can be made about the effect of a new drug. This means that the magnitude of the effect can be determined more accurately. Considering the various topics of this thesis, several specific issues for future perspectives should be described.

When ECG data of large and diverse populations – including female, young and elderly – will be gathered, the physiologic cardiac age can be established with more accuracy. A higher accuracy of the algorithm models in larger datasets in more diverse populations also applies for car driving behaviour and microbiome results.

Specifically for car driving, the models could also trace more precisely in which the way the drug has an effect. A series of these machine learning models could evolve to a test battery, which allows a more precise and accurate evaluation of abnormal driving behavior. Driving data of patients with a cognitive disorder will extend the test battery and may allow distinction between driving-behaviour affected by sleep deprivation and a cognitive disorder.

Considering the microbiome data, addition of microbiome data from patients with other skin diseases, will make it possible to differentiate skin diseases based on the microbiome in case these skin diseases are visually difficult to distinguish or would require invasive diagnostics.

Some models also still need to be validated on external data (data not collected at the same facility) to make these generally suitable for testing interventions. For example, the neural network created in Chapter 3 on ECG features could also be tested on patients with structural heart disease.

The proper use of thoroughly validated machine learning models in early drug development can provide important information, which can be used to make a go/no-go decision regarding further development of new drugs.[4] For example, a machine learning model could determine, based on their ECGs, that a new drug has a negative effect on the age (i.e. health) of the hearts of subjects when a new drug is tested. A model could rate driving behaviour after intake of a newly developed drug as significantly more abnormal. When relevant for the risk profile, it may be concluded that this development should be halted at an early stage, which might save a lot of money.[5]

In this thesis the use of machine learning is only explored on ECG data, data derived from driving simulators, and microbiome data. It would be interesting to explore machine learning on other types of data derived from drug development, such as neurological data (e.g. EEG data), and psychological data. While nowadays more and more data is collected using wearables, providing easy collection without clinical visits, machine learning might be able to classify behaviour characteristics based on this data.[6,7]

Finally, as newer and more powerful computing hardware becomes available, advances are made with newer machine learning algorithms that produce more powerful models with fewer errors.[8] For example, more use could be made of ‘deep learning’, that has made huge progress in recent years.[9] It has produced major breakthroughs and is now used on billions of digital devices for complex tasks such as speech recognition, image interpretation, and language translation.[10] It would be interesting to study these the application of these techniques in the early phases of drug development.

CONCLUSION

In this PhD thesis, we showed that machine learning can be applied on data derived from early-stage clinical trials in order to create or to detect the effect of medicines or other interventions. This may inform the later stage of drug development about the effect of a drug. The proper use of thoroughly validated machine learning models in early drug development can provide important information, which can be used to make a go/no-go decision regarding further development of new drugs.[4] Similarly, these models can be used to guide the decision-making process regarding the dosage range to be used in phase II studies, determining a therapeutic window, and even identifying the target study population. In this way novel pharmacological drugs could be tested for effect on a subject or a patient in the early phase of development.

REFERENCES

1 Terán, J., et al., Intelligent driving assistant based on road accident risk map analysis and vehicle telemetry. Sensors, 2020. 20(6): p. 1763.

2 Deo, R.C., Machine learning in medicine. Circulation, 2015. 132(20): p. 1920-1930.

3 Rajkomar, A., J. Dean, and I. Kohane, Machine learning in medicine. New England Journal of Medicine, 2019. 380(14): p. 1347-1358.

4 Groeneveld, G.J., Hay, J. L., Van Gerven, J. M., Measuring blood–brain barrier penetration using the NeuroCart, a *CNS* test battery. Drug Discovery Today: Technologies, 2016. 20: p. 27-34.

5 Cohen, A., et al., The use of biomarkers in human pharmacology (Phase I) studies. Annual review of pharmacology and toxicology, 2015. 55: p. 55-74.

6 Saboor, A., et al., Latest research trends in gait analysis using wearable sensors and machine learning: A systematic review. IEEE Access, 2020. 8: p. 167830-167864.

7 Ghandeharioun, A., et al. Objective assessment of depressive symptoms with machine learning and wearable sensors data. in 2017 seventh international conference on affective computing and intelligent interaction (ACII). 2017. IEEE.

8 Vamathevan, J., et al., Applications of machine learning in drug discovery and development. Nature reviews Drug discovery, 2019. 18(6): p. 463-477.

9 Chen, H., et al., The rise of deep learning in drug discovery. Drug discovery today, 2018. 23(6): p. 1241-1250.

10 Hinton, G., Deep learning—a technology with the potential to transform health care. Jama, 2018. 320(11): p. 1101-1102.

APPENDICES

SUMMARY

As a result of major advances in technology in healthcare, an increasing amount of data is collected during clinical trials. It is essential to realize, however, that data by themselves are useless. To be useful, data must be analyzed, interpreted, and acted on. Machine learning strategies may provide helpful solutions. In **Chapter 1,** the concept of machine learning is explored. Machine learning focuses on the development of algorithms that can change when exposed to new data. It can also be used to obtain insights, predictions, and decisions from vast amounts of data by combining different parameters.

This thesis contains machine learning approaches on a variety of clinical data sets. The *classical* data consist of electrical signals from the ECG of healthy subjects, the *innovative* data originate from measurements in a driving simulator, and *emerging* data are derived from DNA analysis of the microorganisms living on the skin of patients with skin disease.

In **Chapter 2** we addressed the effect of the number of ECG replicates extracted from a continuous ECG on estimated QT interval prolongation for 10 different QT correction formulas. We showed that the number of ECG replicates impacted the estimated QT interval prolongation for all deployed QT correction formulas. So, for an accurate estimation of QT prolongation more than three ECG's are needed.

In **Chapter 3** we present a study in which we developed a neural network to characterize the effect of aging on the ECG in healthy volunteers. We used this model to predict the physiologic age of individual hearts based on their ECG; physiologic age was estimated as a continuous variable with an average error of 6.9±5.6 years (R^2= 0.72 ± 0.04). The correlation was slightly stronger for men (R^2= 0.74) than for women (R^2= 0.66). Using SHapley Additive exPlanations (SHAP) values the impact of the individual features on the prediction of age was determined. We concluded that this technique could be used to pick up subtle age-related cardiac changes; it also enabled us to estimate the reversing of these age-associated effects by administered treatments.

In **Chapter 4 & 5** machine learning was used for a better assessment of driving performance in drivers using drugs by including multiple parameters derived from a simulator rather than the Standard Deviation of the Lateral Position (SDLP) alone. We specifically analysed the effects of alcohol and alprazolam on car driving behaviour. Adding additional features besides the SDLP increased the model performance from an accuracy of 65% to 83% for prediction of alprazolam intake and from 50% to 76% for prediction of alcohol ingestion. Analysis of other parameters such as the steering behaviour of the driver appeared to be an important contributor to the improvement of the accuracy of the models. We extended this approach to sleep deprivation and tested the model for characterization of new interventions. A model detecting sleep deprivation based on driving behaviour provided an accuracy of 77 ± 9%. We identified the overlap of driving behaviour after sleep deprivation and driving behaviour affected by other interventions. Abnormal driving behaviour after alprazolam, and to a lesser extent after alcohol intake, showed remarkably similar characteristics as observed with sleep deprivation, matching the expected relative drowsiness. Consequently, our model for sleep deprivation may serve as a next reference point for a driving test battery of newly developed drugs.

In **Chapter 6** machine learning techniques were used to identify discriminative biomarkers in the microbiome of patients with seborrheic dermatitis versus healthy controls. The accuracy of the prediction models was 77% and the ROC-AUC was 83%. The most important microorganisms for discrimination – next to C*utibacterium* and S*taphylococcus* – had a relatively low occurrence, which can be easily overlooked in standard analyses. This indicates that machine learning can be of major importance in basic skin research, as well as in the discovery and development of new individualized therapies involving the microbiome.

In this PhD thesis, we showed that machine learning can be applied on data derived from early stage clinical trials in order to detect and evaluate the effect of drugs and other interventions.

NEDERLANDSE SAMENVATTING

Als gevolg van de grote technologische vooruitgang in de gezondheidszorg worden in toenemende mate gegevens verzameld tijdens de uitvoering van klinische onderzoeken. Het is evenwel essentieel om te beseffen dat gegevens op zich van weinig of geen waarde zijn. Ten behoeve van hun optimale bruikbaarheid dienen gegevens geanalyseerd, geïnterpreteerd en verwerkt te worden. Machine learning-strategieën kunnen hiertoe nuttige en adequate oplossingen bieden.

In **hoofdstuk 1** verkenden we het concept van machine learning. Machine learning richt zich op de ontwikkeling van algoritmen die zich kunnen aanpassen wanneer deze worden blootgesteld aan nieuwe gegevens. Tevens kan machine learning worden gebruikt om inzichten, voorspellingen en beslissingen te verkrijgen uit grote hoeveelheden gegevens door het combineren van diverse parameters.

Dit proefschrift bevat machine learning-benaderingen toegepast op verschillende klinische datasets. De *klassieke* gegevens bestaan uit elektrische signalen van het electrocardiogram (ECG) verkregen bij gezonde proefpersonen, de *innovatieve* gegevens zijn afkomstig van metingen in een rijsimulator, en de *opkomende* gegevens zijn afgeleid van DNA-analyse van de micro-organismen die op de huid voorkomen van patiënten met huidziekten.

In **hoofdstuk 2** hebben we het effect van het aantal ECG's – geëxtraheerd uit een continue ECG registratie op de geschatte verlenging van het QT-interval – voor 10 verschillende QT-correctieformules onderzocht. We toonden aan dat het aantal ECG's van invloed was op de nauwkeurigheid van geschatte verlenging van het QT-interval voor alle ingezette QT-correctieformules. Voor een nauwkeurige schatting van QT-verlenging zijn meer dan drie ECG's noodzakelijk.

In **hoofdstuk 3** presenteren we een onderzoek waarin we een neuraal netwerk ontwikkelden om het effect van veroudering op het ECG bij gezonde vrijwilligers te karakteriseren. We gebruikten dit model om de fysiologische leeftijd van individuele harten te voorspellen op basis van hun ECG; de fysiologische leeftijd werd geschat als een continue variabele met een gemiddelde foutmarge van 6,9 ± 5,6 jaar (R^2 = 0,72 ± 0,04). De correlatie was enigszins sterker voor mannen (R^2= 0,74) dan voor vrouwen (R^2= 0,66). Met behulp van SHapley Additive exPlanations (SHAP)-waarden werd de impact van de individuele kenmerken op de voorspelling van fysiologische leeftijd bepaald. We stelden vast dat deze techniek kan worden gebruikt om subtiele leeftijd-gerelateerde hartveranderingen op te sporen; het stelt ons ook in staat om het omkeren van deze leeftijdsgebonden effecten door toegediende behandelingen in te schatten.

In **hoofdstukken 4 en 5** maakten we gebruik van machine learning voor een betere beoordeling van de rijprestaties van bestuurders die medicijnen gebruikten. In plaats van uitsluitend de standaard deviatie van de laterale positie (SDLP) te analyseren werden meerdere parameters afgeleid van een simulator meegenomen. We analyseerden specifiek de effecten van alcohol en alprazolam op het rijgedrag van de bestuurder. Het toevoegen van extra parameters naast de SDLP verhoogde de prestaties van het model. De nauwkeurigheid van het model nam toe van 65% tot 83% na inname van alprazolam, en van 50% tot 76% na inname van alcohol. Analyse van andere parameters – zoals het stuurgedrag van de bestuurder – bleek een belangrijke bijdrage te leveren aan de verbetering van de nauwkeurigheid van de modellen. We breidden deze benadering uit naar slaaptekort en testten het model voor de karakterisering van nieuwe interventies. Een model dat slaaptekort opspoort op basis van rijgedrag leverde een nauwkeurigheid van 77 ± 9% op. Daarmee identificeerden we de overlap tussen rijgedrag na slaaptekort en rijgedrag beïnvloed door andere interventies. Abnormaal rijgedrag na gebruik van alprazolam – en in mindere mate na alcoholgebruik – vertoonde opmerkelijk vergelijkbare kenmerken als waargenomen bij slaaptekort, passend bij de verwachte relatieve slaperigheid. Daarom kan ons model voor slaapgebrek dienen als een volgend referentiepunt voor een rijtestbatterij van nieuw ontwikkelde medicijnen.

In **hoofdstuk 6** pasten we machine learning technieken toe om discriminerende biomarkers te identificeren in het microbioom van patiënten met seborroïsche dermatitis versus gezonde controles. De nauwkeurigheid van de voorspellingsmodellen was 77%, en de ROC-AUC was 83%. De belangrijkste micro-organismen voor discriminatie – naast *Cutibacterium* en *Staphylococcus* – kwamen relatief weinig voor, waardoor men deze micro-organismen in standaardanalyses eenvoudig over het hoofd kan zien. Dit geeft aan dat machine learning van groot belang kan zijn bij fundamenteel huidonderzoek. Dit geldt eveneens voor de ontdekking en ontwikkeling van nieuwe geïndividualiseerde therapieën waarbij het microbioom betrokken is.

In dit proefschrift hebben we aangetoond dat machine learning kan worden toegepast op gegevens die zijn afgeleid van klinische onderzoeken om in een vroeg stadium het effect van medicijnen en andere interventies op te sporen en te evalueren.

LIST OF PUBLICATIONS

Niemeyer-van der Kolk, T., **Van Der Wall, H.E.C.**, Balmforth, C., Van Doorn, M.B.A. and Rissmann, R., 2018. A systematic literature review of the human skin microbiome as biomarker for dermatological drug development. *British journal of clinical pharmacology*, *84*(10), pp.2178-2193.

van der Wall, H.E.C., Gal, P., Kemme, M.J., van Westen, G.J.P. and Burggraaf, J., 2019. Number of ECG replicates and QT correction formula influences the estimated QT prolonging effect of a drug. *Journal of Cardiovascular Pharmacology*, *73*(4), pp.257-264.

Hassing, G.J., **Van der Wall, H.E.C.**, Van Westen, G.J.P., Kemme, M.J.B., Adiyaman, A., Elvan, A., Burggraaf, J. and Gal, P., 2019. Body mass index related electrocardiographic findings in healthy young individuals with a normal body mass index. *Netherlands Heart Journal*, *27*(10), pp.506-512.

van der Kolk, T.N., Buters, T.P., Krouwels, L., Boltjes, J., de Kam, M.L., **van der Wall, H.**, van Alewijk, D.C., van den Munckhof, E.H., Becker, M.J., Feiss, G. and Florencia, E.F., 2020. Topical antimicrobial peptide omiganan recovers cutaneous dysbiosis but does not improve clinical symptoms in patients with mild-to-moderate atopic dermatitis in a phase 2 randomized controlled trial. *Journal of the American Academy of Dermatology*.

Hassing, G.J., **van der Wall, H.E.**, van Westen, G.J., Kemme, M.J., Adiyaman, A., Elvan, A., Burggraaf, J. and Gal, P., 2020. Blood pressure-related electrocardiographic findings in healthy young individuals. *Blood Pressure*, *29*(2), pp.113-122..

Hassing, G.J., Fienieg, B., **Van Der Wall, H.E.C.**, Van Westen, G.J.P., Kemme, M.J.B., Adiyaman, A., Elvan, A., Burggraaf, J. and Gal, P., 2020. The association between body temperature and electrocardiographic parameters in normothermic healthy volunteers. *European Heart Journal*, *41*(Supplement_2), pp.ehaa946-0377.

Niemeyer-van der Kolk, T., **van der Wall, H.**, Hogendoorn, G.K., Rijneveld, R., Luijten, S., van Alewijk, D.C., van den Munckhof, E.H., de Kam, M.L., Feiss, G.L., Prens, E.P. and Burggraaf, J., 2020.

Pharmacodynamic effects of topical omiganan in patients with mild to moderate atopic dermatitis in a randomized, placebo-controlled, phase II trial. *Clinical and translational science*, *13*(5), pp.994-1003.

van den Munckhof, E.H., Niemeyer-van der Kolk, T., **van der Wall, H.**, van Alewijk, D.C., van Doorn, M.B., Burggraaf, J., Buters, T.P., Becker, M.J., Feiss, G.L., Quint, W.G. and van Doorn, L.J., 2020. Inter-and intra-patient variability over time of lesional skin microbiota in adult patients with atopic dermatitis. *Acta Dermato-Venereologica*, *100*(1), pp.1-2.

van der Wall, H.E.C., Doll, R.J., van Westen, G.J.P., Koopmans, I., Zuiker, R.G., Burggraaf, J. and Cohen, A.F., 2020. The use of machine learning improves the assessment of drug-induced driving behaviour. *Accident Analysis & Prevention*, *148*, p.105822.

Van der Wall, H.E.C., Doll, R.J., van Westen, G.J.P., Koopmans, I., Zuiker, R.G., Burggraaf, J. and Cohen, A.F., 2021. Using machine learning techniques to characterize sleep-deprived driving behavior. *Traffic injury prevention*, *22*(5), pp.366-371.

van der Wall, H.E., Hassing, G.J., Doll, R.J., van Westen, G.J., Cohen, A.F., Selder, J.L., Kemme, M., Burggraaf, J. and Gal, P., 2022. Cardiac age detected by machine learning applied to the surface ECG of healthy subjects: Creation of a benchmark. *Journal of Electrocardiology*, *72*, pp.49-55.

van der Wall, H.E.C., Doll, R.J., van Westen, G.J.P., Niemeyer-van der Kolk, T., Feiss, G., Pinckaers, H., van Doorn, M.B.A., Nijsten, T., Sanders, M.G.H., Cohen, A.F., Burggraaf, J. Rissmann, R. and Pardo, L.M., 2022. Discriminative Machine Learning Analysis for Skin Microbiome: Observing Biomarkers in Patients with Seborrheic Dermatitis. Journal of Artificial Intelligence for Medical Sciences, pp.1-7.

CURRICULUM VITAE

Hein van der Wall was born on december 20th 1989 in Amsterdam. After finishing his high school, he started his bachelor studies in Life Science and Technology (Leiden University/Delft University of Technology) in 2009 which he finished with an internship at the Industrial Microbiology department of the TU Delft. He continued his studies with a the Master Life Science and Technology at the TU Delft, finished with a 6 months internship at the Cell Systems Engineering department and a 3 months internship at the Centre for Human Drug Research in Leiden.

In 2017, Hein started his PhD at the Centre for Human Drug Research supervised by prof. Koos Burggraaf and soon after he joined the Computational Drug Discovery group supervised by prof. Gerard van Westen. The research of his thesis focused on the exploration of machine learning in the context of early-stage clinical research.

ACKNOWLEDGEMENTS

This thesis would not have been completed without the help of a group of people.

Foremost, prof. Koos Burggraaf, thank you for giving me the opportunity to do my PhD at the Centre for Human Drug Research and for your supervision. Also thanks to prof. Gerard van Westen who made the collaboration with the LACDR and his support, suggestions and proposed changes to the thesis.

I would like to express my gratitude to dr. Robert-Jan Doll, for his daily supervision and his continuous efforts to improve my thesis.

I'm grateful to prof. Adam Cohen, for having the confidence to include me in his company in 2017. Thank you for your everlasting support during my stay at CHDR.

To my colleagues at the CHDR, thank you for all the collaboration and support. You helped me to gain insights in the data collection and the use of machine learning in the context of clinical research.

To my colleagues at the LACDR, you made my time as a PhD student maybe a little less efficient but also much more exciting. You insprired me to complete my thesis with the coffee breaks, drinks and other activities outside of work.

I would like to express my thanks to all the others with whom I collaborated, the people I discussed the application of my work in their field, the students I supervised.

Dear Roos, you always supported me while writing and kept pushing me to continue. I hope we will continue to enjoy each other for a long time.

Dear mom and dad, Sake and Lucas, you always gave me your unconditional love and always supported me and motivated me to continue my research and finish my thesis. You have been indispensable to me for my recovery and empowerment after my accident