Big Data approaches for prediction of clinically relevant outcomes

Nielsen, Rikke Linnemann

Publication date:
2020

Document Version
Publisher's PDF, also known as Version of record

Ph.D. Thesis
Doctor of Philosophy

Big Data approaches for prediction of clinically relevant outcomes

Rikke Linnemann Nielsen

January 2020

DTU Health Tech
Department of Health Technology

DTU Health Tech
Department of Health Technology
Section of Bioinformatics
Group for Disease Data Intelligence
Technical University of Denmark

Ørsteds Plads
Building 345C
2800 Kongens Lyngby, Denmark
Phone +45 4525 2525
healthtech-info@dtu.dk
www.healthtech.dtu.dk

Preface

The PhD thesis was prepared as part of the double PhD degree programme offered through the Sino-Danish Center of Research and Education (SDC) between the Technical University of Denmark (DTU) and the University of Chinese Academy of Sciences (UCAS) in partial fulfilment of the requirements of acquiring a PhD degree. The PhD project was funded by DTU and SDC.

The PhD project was supervised by Ramneek Gupta (DTU), Kjeld Schmiegelow (Copenhagen University Hospital and the University of Copenhagen), Anders Gorm Pedersen (DTU) and Xiu-Jie Wang (UCAS).

The content of the thesis presents work carried out between November 2016 and January 2020. The thesis deals with prediction models using machine learning methodologies to predict health and disease outcomes with potential applications in precision medicine. Three different areas of health and disease were explored including metabolic health, type 2 diabetes and childhood acute lymphoblastic leukemia. The PhD thesis contains an introduction with concepts to understand the scope of the thesis followed by three research papers as well as future prospects. For a full list of contributions in the PhD, see page vii.

Kongens Lyngby, January 31, 2020

Rikke Linnemann Nielsen

Summary

In the past decades, technological advances have provided the ability to collect, store and analyse Big data in disease biology. The amount and complexity of Big data requires us address the challenges posed by the large volume, veracity and heterogeneity. This includes, for instance, data integration across high-throughput omics data, electronic medical records and environmental characteristics. Interpretation of this heterogeneous data promises to derive value towards understanding health, disease and treatment outcomes at an individual patient level. The promise of precision medicine is to use individual variability in genomics, environment and lifestyle to guide personalized prevention and treatment strategies. Current methods in genomics employ association testing of single genetic variants in large cohorts. Even though challenged by the requirement of very large patient cohorts, these methods have produced several valuable insights over the past decade. Their limitation is in not providing a straightforward strategy for correlation of multiple features to exploit the complex interactions that exist in biology, and the difficulty in clinically translatable insights given the low effect sizes of individually associated variants. Machine learning approaches facilitate analysis of large and complex biological data by detection of non-linear interactions between heterogenous features and offer strategies to handle high-dimensional datasets. Furthermore, machine learning offers individual-level predictions which can be translated into personalized recommendations for prevention or treatment of disease outcome. These methods certainly have their own challenges, however are beginning to show promise in smaller patient cohorts. Eventually, the goal is to derive information from rich datasets, that may inform clinical practice, and step towards precision medicine.

This PhD thesis consists of three research projects where prediction models of different health and disease outcomes were developed using machine learning approaches with potential applications in precision medicine.

The first project explored predictive modelling of weight loss in dietary interventions of eight weeks with a whole grain-rich diet, a low-gluten diet or a refined grain in apparently healthy Danish individuals ($N = 102$) at cardiometabolic risk. The individuals were deeply phenotyped with several phenotypic and biochemical characteristics as well as genotype, gut microbiome and urine metabolome data. Given the challenge of a small cohort with several high-dimensional data for machine learning, several data transformations and reductions were made on the features to improve predictive power such as polygenic risk scores. In addition, feature engineering such as modelling variability of longitudinal post-prandial response, improved eventual predictive outcome. Finally,

an ensemble model capturing different aspects of biology was combined from the most predictive individual models. The developed model may function as an early screening tool when determining individual weight loss strategies.

The second project established prediction models of the time to insulin in type 2 diabetes patients. Artificial neural networks were used to integrate longitudinal biomarkers of drug prescriptions, biochemistry, anthropometry and blood pressure collected in electronic medical records for up to 20 years follow-up as well as information on lifestyle, social deprivation and genotype. Electronic medical records contain irregular sampled measurements with varying length of individual patient trajectories. Thus, these were formatted into a structured representation of data (both by a single time point and a longitudinal approach). By assessing individual risk of insulin requirement across approximately 6000 patients, the model may eventually assist in reduction of clinical inertia in very high risk patients or reduce health care interventions and costs by identifying very low risk patients.

The third project focus on prediction of asparaginase-associated pancreatitis (AAP), a serious treatment toxicity in childhood acute lymphoblastic leukemia (ALL) ($N = 1390$, 205 AAP cases). It is currently difficult to predict which patients are at risk of AAP. Even more difficult is the prediction of patients that will develop a second AAP following re-exposure to asparaginase after the first AAP event. Machine learning algorithms was used to establish SNP-based prediction models allowing to capture interactions between genetic variants for AAP and second AAP. One path of least disruptive implementation could be to only use the extremes in the model output where both low risk and high risk patients are identifiable at very high confidence. For instance, identification of high-risk patients of AAP can influence decision for increased monitoring, while patients at low risk of a second AAP can influence the decision to re-expose patients to asparaginase. The models may eventually assist in further stratification of patients and adjustment of treatment protocols of childhood ALL.

In summary, this thesis demonstrates the feasibility of machine learning approaches to integrate different types of health and disease data as well as their utility for subgroup and individual-level predictions of complex disease outcomes. It also explores how value can be derived from combinations of heterogeneous and longitudinal patient data. Eventually, the goal of these studies is to find a path for implementation and translation of findings from machine learning models to personalized strategies in prevention or treatment of disease.

Dansk resume

Teknologiske fremskridt har igennem de sidste årtier muliggjort indsamling, opbevaring og analyse af Big data indenfor sygdomsbiologi. Mængden og kompleksiteten af Big data kræver, at vi adresserer de udfordringer, der er forbundet med dets mængde, veracitet og heterogenitet. Dette inkluderer f.eks. data integration af high-throughput omics data, elektroniske patientjournaler og andre eksterne miljø faktorer. Øget forståelse af disse heterogene data er værdifuldt, da det kan medføre indsigt i den enkelte patients sundhed, sygdoms- og behandlings-udfald. Forståelse af individuel variation i genomet, livsstil og eksterne miljø faktorer er lovende for præcisionsmedicin, som vil benytte variationer i disse faktorer til at guide personaliserede forebyggelses- og behandlings-strategier. Nuværende metoder i genom forskning tester associationer af individuelle genetiske varianter i store kohorter. Disse metoder har givet flere værdifulde resultater igennem det sidste årti på trods af deres krav til meget store patient kohorter. Metoderne har også deres begrænsninger. De giver ikke en strategi for korrelering af flere features og tager ikke hensyn til de komplekse interaktioner der findes i biologi. Desuden har enkelte genetiske varianter ofte en begrænset effekt på det kliniske udfald. Metoder indenfor maskinlæring kan forbedre analyser af store of komplekse biologiske data, da de giver mulighed for ikke-lineære interaktioner imellem heterogene features og tilbyder strategier til at håndtere datasæts med store dimensioner. Desuden giver maskinlæring mulighed for individuelle forudsigelser, som kan give personaliserede anbefalinger til forebyggelse eller behandling af sygdomme. Maskinlæring metoder har også deres egne udfordringer, men er begyndt at vise lovende fremskridt i mindre patient kohorter. Formålet er at generere information fra datasæts som kan informere klinisk praksis og fremme præcisionsmedicin.

Denne Ph.d.-afhandling introducerer tre forskningsprojekter hvor forudsigelsesmodeller for sundhed- og sygdoms-udfald er udviklede ved brug af maskinlæring med mulig anvendelse indenfor præcisionsmedicin.

Det første projekt udviklede vægttabs-forudsigelsesmodeller for umiddelbart sunde danskere ($N = 102$) som havde øget risiko for kardiometaboliske sygdomme, der modtog otte ugers diæt interventioner med en diæt rig på fuldkorn, en diæt med lavt glutenindhold eller en raffineret korndiæt. Dybe fænotypiske data var målt i studieindividerne, som inkluderede flere fænotypiske og biokemiske egenskaber, genotype, tarm mikrobiom og urin metabolomics. For at udnytte den relativt lille kohort med mange højdimensionelle features i maskinlæring blev flere data både transformeret og reduceret (såsom score for polygen risiko) for at forbedre den forudsigelige power i studiet. Udsving i det postprandiale respons blev modelleret til nye features, som forbedrede forudsigelsen af vægttabsmodellen. De mest forudsigelige individuelle modeller (med forskellige bio-

logiske aspekter) blev kombineret i en ensemble-model, som kan fungere som et tidligt screeningsværktøj, når man bestemmer individuelle vægttabsstrategier.

Det andet projekt etablerede forudsigelsesmodeller for tiden til insulin i type 2 diabetes patienter. Neurale netværk blev brugt til at integrere biomarkører af medicinrecepter, biokemi, antropometri og blodtryk fra data samlet i op til 20 år i elektroniske patientjournaler såvel som information angående livsstil, social berøvelse og genotype. Elektroniske patientjournaler indeholder uregelmæssige prøvemålinger med varierende længde af individuelle patientforløb. Derfor blev data formateret til et mere struktureret dataformat (både af en 'enkelt tidspunkts' tilgang og en 'hel-forløbs' tilgang). Modellen vurderede individuel risiko for ca. 6000 patienters behov for insulin, som muligvis kan reducere klinisk inerti i høj-risiko patienter eller identificere lav-risiko patienter og derved muligvis reducere omkostninger i sundhedsvæsenet i forbindelse med regelmæssige patientscreeninger.

Det tredje projekt fokuserer på forudsigelse af asparaginase-associeret pankreatitis (AAP), en alvorlig toksicitet i forbindelse med behandling af akut lymfoblastisk leukæmi (ALL) hos børn ($N = 1390$, 205 med AAP). Det er på nuværende tidspunkt svært, at forudsige hvilke patienter har risiko for AAP og om endnu sværere, at forudsige hvilke patienter, der vil udvikle endnu en AAP ved asparaginase reeksponering efter det første AAP-tilfælde. Maskinlærings algoritmer blev brugt til at etablere SNP-baserede forudsigelses modeller som tillod interaktioner imellem genetiske varianter for AAP og endnu en AAP ved asparaginase behandling efter det første AAP-tilfælde. Et mål for mindst forstyrrende implementering af modellerne var ved, at identificere både lav- og høj-risikogrupper. F.eks. kan identificering af høj risiko AAP-patienter påvirke beslutninger ang. Øget overvågning, mens patienter med lav risiko af endnu en AAP kan påvirke beslutningen om hvorvidt det er sikkert at reeksponere patienten med asparaginase. Modellerne kan derfor, på lang sigt, påvirke stratificering af patienter og justering af behandlingsprotokoller i ALL hos børn.

Denne afhandling demonstrerer muligheden for maskinlæringsmetoder til at integrere forskellige typer af sundhedsdata samt deres anvendelighed for, at forudsige undergrupper og individer med komplekse sygdomsudfald. Afhandlingen undersøger også, hvordan heterogene og tidserie data kan skabe værdi. Formålet med disse studier er, at finde en strategi for implementering og oversættelse af maskinlæringsmodellerne til personaliserede strategier i forbindelse med forebyggelse eller behandling af sygdom.

List of contributions

The following scientific papers have been prepared during the PhD period.

Papers included in the thesis

- **Rikke L. Nielsen***, Marianne Helenius*, Sara Garcia, Henrik M. Roager, Derya Aytan-Aktug, Lea B.S. Hansen, Mads V. Lind, Josef K.Vogt, Marlene D. Dalgaard, Martin I. Bahl, Cecilia B. Jensen, Rasa Muktupavela, Christina Warinner, Vincent Appel, Rikke Gøbel, Mette Kristensen, Hanne Frøkiær, Morten H. Sparholt, Anders F. Christensen, Henrik Vestergaard, Torben Hansen, Karsten Kristiansen, Susanne Brix, Thomas N. Petersen, Lotte Lauritzen, Tine R. Licht, Oluf Pedersen, Ramneek Gupta. **Data integration for prediction of weight loss in randomized controlled dietary trials**. *Manuscript submitted to Nature Metabolism.*

- **Rikke L. Nielsen**, Louise Donnelly, Agnes M. Nielsen, Kaixin Zhou, Adem Dawed, Konstantinos Tsirigos, Bjarne Ersbøll, Line Clemmensen, Ewan R. Pearson, Ramneek Gupta. **Prediction of time to insulin requirement in patients with type 2 diabetes using artificial intelligence: A GoDARTS study**. *Manuscript submitted to Diabetes Care.*

- **Rikke L. Nielsen**, Benjamin O. Wolthers, Marianne Helenius, Birgitte K. Albertsen, Line Clemmensen, Kasper Nielsen, Jukka Kanerva, Riitta Niinimäki, Thomas L. Frandsen, Andishe Attarbaschi, Shlomit Barzilai, Antonella Colombini, Gabriele Escherich, Derya Aytan-Aktug, Hsi-Che Liu, Anja Möricke, Sujith Samarasinghe, Inge M van der Sluis, Martin Stanulla, Morten Tulstrup, Ester Zapotocka, Kjeld Schmiegelow and Ramneek Gupta. **Prediction of asparaginase-associated pancreatitis cases in childhood acute lymphoblastic leukemia.** *Manuscript submitted to Haematologica.*

(*) These authors contributed equally.

Papers not included in the thesis

- Kirsten Brunsvig Jarvis, Marissa LeBlanc, Morten Tulstrup, **Rikke L. Nielsen**, Birgitte Klug Albertsen, Ramneek Gupta, Pasi Huttunen, Olafur Gisli Jonsson, Cecilie Utke Rank, Susanna Ranta, Ellen Ruud, Kadri Saks, Sonata Saulyte Trakymiene, Ruta Tuckuviene, and Kjeld Schmiegelow. **Candidate single nucleotide polymorphisms and thromboembolism in acute lymphoblastic leukemia – A NOPHO ALL2008 study.** *Thrombosis Research 2019;184:92-98.*

- Sara L. Garcia*, Jakob Lauritsen*, Zeyu Zhang*, Mikkel Bandak, Marlene D. Dalgaard, **Rikke L. Nielsen**, Gedske Daugaard, Ramneek Gupta. **Prediction of nephrotoxicity associated with cisplatin-based chemotherapy in testicular cancer patients.** *Manuscript under revision for re-submission to JNCICS: Journal of the National Cancer Institute Cancer Spectrum.*

- Agnes M. Nielsen, **Rikke L. Nielsen**, Louise Donnelly, Kaixin Zhou, Anders B. Dahl, Ramneek Gupta, Bjarne K. Ersbøll, Ewan Pearson, Line Clemmensen. **A Comparison of Methods for Disease Progression Prediction Through a GoDARTS Study.** *Manuscript prepared for submission to BMC Medical Research Methodology.*

- Henrik Munch Roager*, Josef K. Vogt*, Mette Kristensen, Lea Benedicte S. Hansen, Sabine Ibrügger, Rasmus B. Mærkedahl, Martin Iain Bahl, Mads Vendelbo Lind, **Rikke L. Nielsen**, Hanne Frøkiær, Rikke Juul Gøbel, Rikard Landberg, Alastair B. Ross, Susanne Brix, Jesper Holck, Anne S. Meyer, Morten H. Sparholt, Anders F. Christensen, Vera Carvalho, Bolette Hartmann, Jens Juul Holst, Jüri Johannes Rumessen, Allan Linneberg, Thomas Sicheritz-Pontén, Marlene D. Dalgaard, Andreas Blennow, Henrik Lauritz Frandsen, Silas Villas-Bôas, Karsten Kristiansen, Henrik Vestergaard, Torben Hansen, Claus T. Ekstrøm, Christian Ritz, Henrik Bjørn Nielsen, Oluf Borbye Pedersen, Ramneek Gupta, Lotte Lauritzen, Tine Rask Licht. **Whole grain-rich diet reduces body weight and systemic low-grade inflammation without inducing major changes of the gut microbiome: a randomised cross-over trial.** *Gut 2019;68:83–93.*

- Graciela Gonzalez, Zhiyong Lu, Robert Leaman, Davy Weissenbacher, Mary Regina Boland, Yong Chen, Jingcheng Du, Juliane Fluck, Casey S. Greene, John Holmes, Aditya Kashyap, **Rikke L. Nielsen**, Zhengqing Ouyang, Sebastian Schaaf, Jaclyn N. Taroni, Cui Tao, Yuping Zhang, Hongfang Liu. **Text Mining and Machine Learning for Precision Medicine.** *Pacific Symposium on Biocomputing 2019;24:449-454.*

(*) These authors contributed equally.

Acknowledgements

Throughout the PhD project, I have received a lot of support, guidance and assistance from a number of people who have helped me turn this thesis into reality. Thank you to all of you.

A huge thank you should be given to my supervisor Ramneek Gupta. You have throughout the three years provided constant encouragements, opportunities and support as well as shared your knowledge and countless ideas. Thanks for helping me make this PhD project succeed.

Thank you to all my colleagues at the section of bioinformatics at the Technical University of Denmark for creating a friendly work environment. A special thanks should be given to all previous and current members of the Data Disease Intelligence (DDI) group. Marianne Helenius, thank you for a great collaboration, I look forward to continuing this in the coming years.

Thank you to Poul V. Andersen's foundation for providing a cross-sectional PhD collaboration grant between DTU Health Technology and DTU Compute in the beginning of the PhD project and thank you Agnes M. Nielsen for our close collaboration. Also, thank you Bjarne Ersbøll and Line Clemmensen for sharing ideas on the projects.

The projects presented in this thesis would not have succeeded without external collaborators. Thanks to everybody on the gut, grain and greens project and especially Marianne, Sara, Derya, Cecilia, Rasa, Henrik, Mads and Tine. Thank you to the people at the division of Population Health & Genomics, School of Medicine, University of Dundee; Louise, Ewan, Kaixin and Adem for introducing me to the world of diabetes research during my external stay in Scotland, as well as the collaboration throughout the PhD project. Finally, a huge thanks to Kjeld Schmiegelow and Benjamin O. Wolthers from the Department of Paediatrics and Adolescent Medicine at The Juliane Marie Centre, Rigshospitalet for a close collaboration on the pancreatitis project.

Thank you to Xiu-Jie Wang's research group for hosting me during my research stay in Beijing and helping me navigate the Chinese system and getting settled. Living in China, truly opened my eyes to a different world and a new range of challenges I did not imagine. This experience would not have been the same without Christina and Birte and our excessive dumpling consumption and Jing-A sessions. Also thank you to Christel and Bent for keeping an eye out on me!

Finally, a huge thanks to my family, friends and Thomas for always being there and believing in me. It truly means the world to me.

Contents

Abbreviations

3G	Gut, Grain and Greens
AAP	Asparaginase-associated pancreatitis
AI	Artificial intelligence
ALL	Acute lymphoblastic leukemia
ANN	Artificial neural network
CPU	Central Processor Unit
dbGaP	The database of Genotypes and Phenotypes
DNA	Deoxyribonucleic acids
EHR	Electronic health record
EMA	European Medicines Agency
EMR	Electronic medical record
FN	False negative
FP	False positive
FPR	False positive rate
GC-MS	Gas chromatography-mass spectrometry
GoDARTS	Genetics of Diabetes Audit and Research in Tayside Scotland
GPU	Graphical Processor Unit
GRS	Genetic risk score
GWAS	Genome-wide association studies
HR	High risk
IR	Intermediate risk
LASSO	Least absolute shrinkage and selection operator

LC-MS	Liquid chromatography-mass spectrometry
LIME	Local interpretable model-agnostic explanations
MCC	Matthews correlation coefficient
MRD	Minimal residual disease
NOPHO	Nordic Society of Pediatric Hematology and Oncology
NPV	Negative predictive value
PPV	Positive predictive value
PRS	Polygenic risk score
RCT	Randomized controlled trial
ROC-AUC	Area under the receiver operating characteristic curve
rRNA	Ribosomal RNA
SNP	Single nucleotide polymorphism
SR	Standard risk
T2D	Type 2 diabetes
TCGA	The Cancer Genome Atlas
TN	True negative
TP	True positive
TPR	True positive rate

CHAPTER 1

General introduction

1.1 Opportunities for Big data in health and disease

Large-scale studies of health and disease have focused on identifying associations between biomarkers and disease outcomes by studies of a single independent data type, where individual features from high-dimensional data have been tested separately [1–3]. As an example, genome-wide association studies (GWAS) test single nucleotide polymorphisms (SNPs) for association to a given phenotype [4]. However, analysis of a specific component in a biological system provides limited understanding of etiology and treatment response in complex diseases [5] such as type 2 diabetes (T2D) or cancer. Complex diseases are influenced by individual differences in the genome, lifestyle and environmental factors as well as their interactions [6–12].

During the past decades, technological advancements have allowed generation of various sources of Big data including SNP-chip profiling, molecular omics such as genomic, transcriptomic, proteomic, metabolomic and metagenomic data [13–16]. Furthermore, phenotypic characteristics can be collected from wearable sensors and electronic records of a patient's medical history [17–19]. As data continues to grow in complexity and size, it is required to develop and gain more experience with systematic and integrative approaches for data analyses. In addition, being able to utilize prior knowledge of biology contributes towards data reduction and provides a handle on data-driven insight of complex biological systems [13].

Today, the power of Big data can be utilized to leverage correlations across multiple types of data over time. This provides an opportunity to identify novel biomarkers from complex biological relationships that can be used to predict patient subgroups or individuals at greatest risk. This allows establishment of predictive models for health, disease or treatment outcomes that can assist in optimization of individual patient care [20, 21].

1.2 Motivation for predictive modelling with machine learning

While providing many opportunities, Big data also introduces several challenges including how to derive meaningful information for precision medicine [21]. In order to succeed, there is a need for more complex data analyses to handle and make sense of large volumes of biological data. Machine learning, a discipline within artificial intelligence (AI), has shown prospects in the analysis of large and complex data [3, 13].

Machine learning techniques are powerful as these can extract both linear or non-linear correlations across heterogeneous data such as multi-omics and other types of contextual data such as patient characteristics [13, 22]. Machine learning can deal with high-dimensional datasets either by using prior knowledge or by providing data-driven methods for feature selection [23]. It is also possible to do model-based integration of data by combining several models into more accurate ensemble models [13]. Machine learning models provide insight into feature importance which allow improved understanding of underlying biological systems. Machine learning models are furthermore useful as they provide individual-level risk predictions rather than population-based recommendations that are not always predictive of individual disease risk [6].

In order to realize the full potential of how machine learning algorithms can integrate Big data and predict outcomes such as disease predisposition, progression or treatment response, several challenges need to be addressed with respect to data handling and preprocessing, use of prior knowledge, modelling, translation and applications. Big data and machine learning approaches are believed to be useful in guiding clinical decision making in precision health and medicine [3, 20, 21, 24].

1.3 Aim of the PhD project

The aim of the PhD project was to develop prediction models of health and disease outcomes as well as disease progression in the field of precision medicine. The studies conducted as part of the PhD project have explored i) predictive models with machine learning algorithms, ii) data-driven elucidation of biomarkers predictive of outcomes, and iii) clinical utility of machine learning models by complementing decision making in precision medicine.

1.4 Thesis overview

The thesis is composed by eight chapters. *Chapter 1* to *Chapter 3* provide a motivation for the PhD project and a background on Big data, precision medicine and machine learning. More specifically, *Chapter 1* provides a brief overview, rationale and scope of the thesis. *Chapter 2* covers an introduction to precision medicine, Big data, study

designs and data analyses for precision medicine. *Chapter 3* introduces key concepts of machine learning methodologies used in this thesis.

The following three chapters introduce the research projects included in the thesis. The three projects explore different areas of health and disease including metabolic health, T2D and childhood acute lymphoblastic leukemia (ALL), and different events in these areas throughout an individual's disease journey (Figure 1.1).

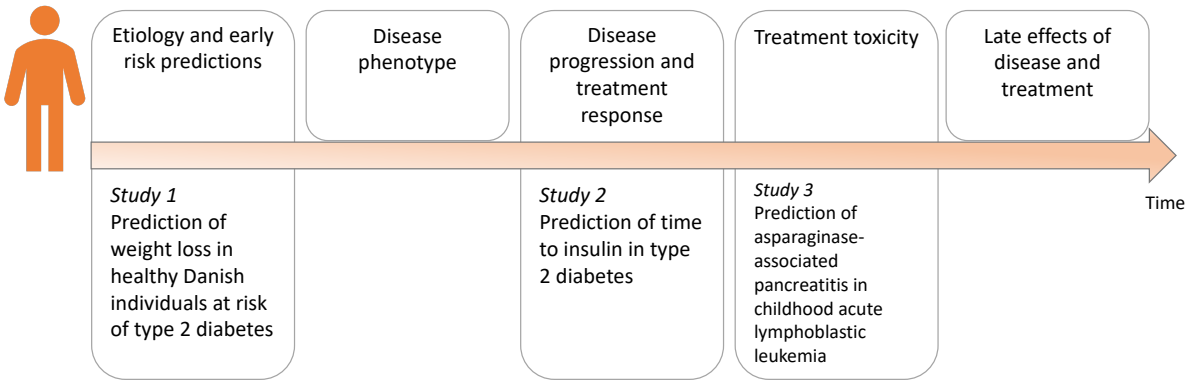


Figure 1.1: Precision health and medicine throughout a person's lifetime. This PhD thesis explores prediction of metabolic health, disease progression and treatment toxicity.

The three projects use machine learning to establish predictive models of individual outcomes. The prediction models include heterogeneous data integration of genetics and clinical characteristics including both single time point and longitudinal data. The applications of the models were evaluated in order to translate predictions into guidelines that potentially can influence intervention or treatment considerations to improve individual outcomes. The use of machine learning for data integration and prediction of disease outcomes presents specific challenges that are illustrated and addressed in the specific studies. Each chapter thus introduces a background, the study and bioinformatic challenges prior to the manuscript. The bioinformatic challenges are possibly better understood after reading the manuscript.

Chapter 4 presents a study on prediction of weight loss in apparently healthy Danish individuals with a cardio-metabolic risk profile undergoing dietary interventions with a whole grain-rich diet, a low-gluten diet or a refined grain diet in two randomized cross-over trials. The duration of the intervention periods were eight weeks separated by a minimum six weeks wash-out period. The study participants ($N = 50$ in the whole grain study, $N = 52$ in the gluten study) were deeply phenotyped with data on anthropometric traits, immunological and physiological biomarkers, host genotype, the gut microbiome, urinary metabolomics, a study diary on diet consumption, gastrointestinal transit time, a self-reported questionnaire on comfort, well-being and gastrointestinal

symptoms and a standardized meal test on postprandial response in the beginning and end of the two intervention periods. The project addresses the challenges of how to integrate multi-dimensional heterogeneous datasets and perform feature selection for prediction of weight loss on a cohort with a relatively small number of samples. The most predictive features for weight loss were identified in random forest models across different domains of data. The most predictive models were merged into an ensemble of models to provide more accurate predictions of individuals who would or would not benefit from a dietary intervention as a weight loss strategy.

Chapter 5 and *Chapter 6* present two studies based on patient data from T2D and childhood ALL, respectively. As an introduction to the disease areas, both chapters provide a background on these disease areas by introducing etiology, epidemiology and the treatment journey for the average patient.

Chapter 5 further introduces the second study on prediction of time to insulin in ∼6000 patients diagnosed with T2D. Patient data were available for up to 20-years follow-up in electronic medical records (EMRs) and genotype from the Genetics of Diabetes Audit and Research in Tayside Scotland (GoDARTS) cohort [25]. The clinical EMR data contained longitudinal irregular sampled measurements of anthropometry, biochemical data and drug prescription data which were extracted by two approaches before its integration with a social deprivation score and genotype. Artificial neural networks were used to predict the time horizon of insulin requirement 1 to 4 years ahead each year following confirmed T2D diagnosis for up to 10 years. Clinical application of the models of time to insulin can potentially reduce clinical inertia, which reflect poor management of hyperglycemia increasing a patient's overall risk of vascular diabetes complications. Identification of high risk patients for insulin requirement may motivate patient behavior i.e. a healthier lifestyle or compliance to therapy or motivate more frequent glycemic measurements of a patient. Alternatively, the model can also be used to identify patients with well-managed glycemia that could reduce health care costs from clinical interventions.

Chapter 6 introduces a study on prediction of asparaginase-associated pancreatitis (AAP) in childhood ALL. Asparaginase is a drug that has contributed to improved survival rates of childhood ALL but results in acute and persistent complications for some patients which are difficult to predict. Even more challenging is determining which patients will develop a second AAP following re-exposure to asparaginase after the first AAP event. The study thus explored machine learning methodologies to predict these outcomes. Data was collected by the Ponte di Legno toxicity working group and represents the current largest dataset available for AAP in childhood ALL. We integrated information on age, sex and ∼1.4M SNPs in 1390 childhood ALL patients (205 had AAP) at the age of 1–17.9 years into machine learning models to build robust classifiers of AAP risk. Different machine learning models, encoding of genetic features and feature selection strategies including prior knowledge from previous pancreatitis studies and on adult pancreatitis pathways assisted in prioritizing predictive SNPs of AAP. Combinations of predictive models of AAP were combined for development of a person-

alized artificial ensemble model that were evaluated for clinical utility in the context of childhood ALL treatment.

As a perspective on the three research projects, *Chapter 7* presents learnings on machine learning and AI's clinical utility. Finally, *Chapter 8* presents conclusions and future prospects of the three presented projects and on machine learning-based predictions for precision medicine. In some of the projects presented in this thesis, the predictive performance of the developed models is credible enough to lead us to believe that the path to implementation in clinical applications may not be in a too distant future.

CHAPTER 2

Big data and precision medicine

2.1 Precision medicine

Randomized controlled trials (RCTs) have long been the the gold standard in clinical research to determine the appropriate treatment for a given disease [26–28]. This have resulted in routine care that provide successful treatment for most patients. However, a group of patients may not benefit from the recommended treatment, while another subgroup of individuals will experience adverse effects from the given treatment in a heterogeneous population [29] (Figure 2.1).

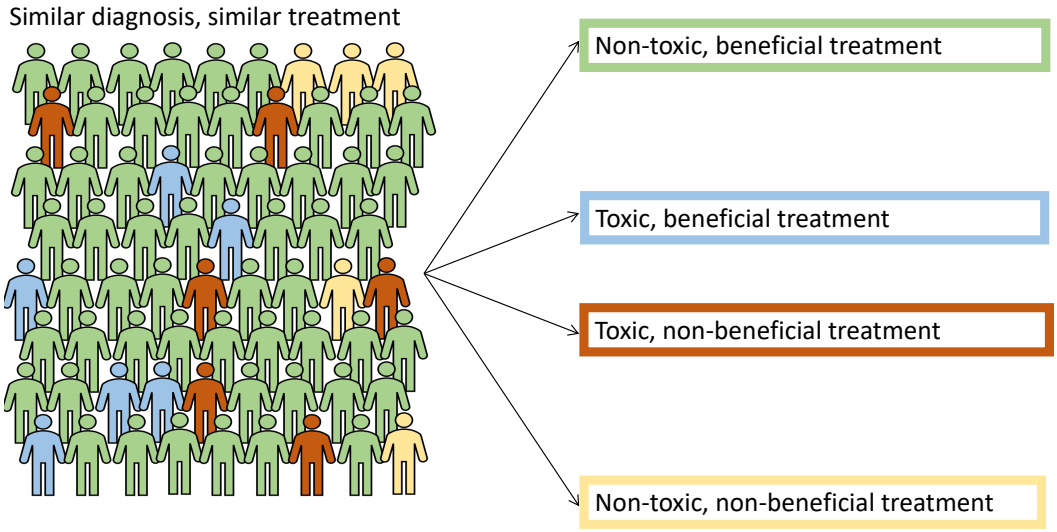


Figure 2.1: Results from randomized controlled trials have defined treatment guidelines at population level. As an example, in cancer therapy, the treatment efficacy may not be similar for all patients and some patients may experience treatment-associated toxicities.

Precision medicine (or personalized medicine) is an approach that utilizes the variability of genes, lifestyle and environment to determine tailored interventions or treatments specific for the individual patient to prevent, prolong or treat diseases [20, 30, 31]. In the case of Figure 2.1, precision medicine aims to understand and identify biomarkers

of the treatment response in order to provide individualized treatment that optimizes the therapeutic effect and reduces adverse treatment events. Precision medicine offers to optimize health care in several different scenarios by; preventing disease, improving understanding of disease etiology, defining disease phenotype and diagnosis, understanding disease progression, determining the most efficient treatment of disease or avoiding treatment-associated toxicities and late effects to maintain the patient's quality of life [9, 10, 20, 32]. It is now increasingly clear that we are likely to see data-driven advances in precision medicine with the generation of publicly available biological databases, technological advances in methods to characterize biological features (such as high-throughput omics platforms, cellular assays and wearables) as well as computational methodologies for analyzing Big data [33].

2.2 Big data

The term 'Big data' is widely used to describe data that is to large and complex to process by traditional software and hardware [21, 34]. Big data has been defined by several V's; Volume, velocity, variety, variability, veracity and value [35, 36]. The most accepted V's in the Big data definition is volume, velocity and variety [21]. Volume refers to the amount of the data. Velocity indicates how quickly the data is generated and available for data analysis. Variety covers the heterogeneity that is the different types of data. Other V's, such as variability refers to the overall quality of the data such as errors that may have occurred in the experimental analysis of the data or in the collection of data as well as overall 'messiness' of the data, as data can appear in structured or unstructured formats, with and without missing data. Veracity refers to the trustworthiness of the data. Finally, value refers to the worth of the data where a key promise of precision medicine is to turn Big data into actionable value for the benefit of patients [20, 37].

Several types of Big data can be collected for insight in designing precision medicine for the individual patient. This, amongst others, involves multi-omics approaches, wearable devices and clinical phenotyping (including medication, electronic medical or health records) [10, 12, 17, 20, 21, 38, 39]. The establishment of large networks and biobanks such as the UK Biobank [40], the database of Genotypes and Phenotypes (dbGaP) [41] and The Cancer Genome Atlas programme (TCGA) [42] provide opportunities to advance discoveries across human diversity and implement findings for precision medicine [15, 37]. However, there is a societal focus on data privacy, which in some countries is beginning to limit data sharing. Even though beyond the scope of this thesis, it should be mentioned that to keep the precision medicine momentum, we need large bodies of shared data, and thus need to develop and demonstrate secure and efficient systems for handling and processing data.

2.3 Omics

The suffix ”-omics” describes studies in biology ranging across several fields including genomics (DNA), transcriptomics (RNA), proteomics (protein), metabolomics (metabolites) and metagenomics (microbiome) etc. Omics data is generated using high-throughput screening techniques that results in large amounts of data given analysis of several biomarkers [43]. The following sections only describe omics data that have been analyzed in projects of this thesis including; genomics, the gut microbiome and the urine metabolome.

2.3.1 Genomics

The human genome refers to the complete set of genetic information coded by double-stranded deoxyribonucleic acids (DNA) composed of nucleotides across 23 pairs of chromosomes (22 pairs of autosomes and sex-chromosomes). The human genome project completed a reference of the human genome in 2003 [44]. Since then several GWAS and next-generation sequencing studies have generated data to explore human genetic variations. A primary aim of these studies has been to identify and understand how genetic variation relates to individual risk of disease or disease progression which potentially could assist in personalized medicine [4, 21, 45, 46]. The genetic variation amongst any two unrelated individuals is very similar (99.9%) across our $3.2\text{x}10^9$ base-pair long genome [47, 48]. The variation in the remaining 0.1% of the genome arise due to mutations and errors in recombination events. Genomic variations include SNPs, small indels (insertions and deletions), larger structural variants (such as copy number variations) and tandem repeat variations [49].

The most common type of genetic variations are SNPs. A SNP is a common single base-pair difference in the DNA that appear in at least 1% of a population [50, 51]. SNPs can be detected by genotyping platforms which are continuously evolving. Genotyping arrays can now detect up to ∼5 million SNPs [52]. The impact of SNPs in complex traits is explored in GWAS [11], which will be described in detail later in this chapter. Next-generation sequencing has provided massive amounts of data by whole exome sequencing or whole genome sequencing that allow analysis of all types of genetic variations within the human genome. The cost of assays for genotype chips (40-100$) is still much lower compared to whole genome sequencing (∼1000$) or whole exome sequencing (a few hundred dollars) [11, 45, 53] making genotype arrays a cost-effective option to study genetic variations. Even though these arrays have limitations, they still provide the ability to characterize larger numbers of samples within single studies. Genetic analyses in this thesis were all explored with genotype data from SNP arrays.

2.3.2 The gut microbiome

Our 'second' genome, the human microbiome is the genetic material from microorganisms (microbiota) that colonize across our body, including skin and mucosal ecosystems [14, 54]. Studies of the human gut microbiome have especially caught attention in human health as dysbiosis of the gut microbiome is associated with a wide range of diseases such as inflammatory bowel disease, cancer, obesity and T2D among others [14, 55, 56]. The genomic content of microorganisms colonized in the human gastrointestinal tract is estimated at 100 times the amount of the human genome [57]. The composition of the human gut microbiome is unique across individuals with large inter-individual variability [14, 58, 59]. The human gut microbiome composition in an adult remains relatively stable, but can be influenced by external factors such as diet, infections, antibiotic treatment, lifestyle or surgery [54, 59, 60]. The gut microbiome have several functions in maintenance of human health [56] and assist in protection against pathogens, degradation of food and production of short-chain fatty acids along with other metabolites that interact with the immune, endocrine and nervous systems [14, 57]. Studies of the human microbiome is facilitated by 16S ribosomal RNA (rRNA) gene amplicon or whole-metagenomic sequencing. 16S rRNA gene amplicon sequencing provides a convenient and accurate method to establish taxonomic classification at the genus level, while shotgun metagenomic sequencing offers detailed taxonomic and functional classification [61].

2.3.3 Metabolomics

Metabolomics is the study of metabolites which are small molecules reflecting the substrates, intermediates or products in molecular metabolic pathways [62]. Metabolites and their interactions are generally referred to as the metabolome. The study of metabolomics can infer biochemical profiles in different biological samples such as cells, tissues or fluids. The molecular phenotyping of metabolites in biological samples is analyzed by nuclear magnetic resonance spectrometry or mass spectrometry such as gas chromatography-mass spectrometry (GC-MS) or liquid chromatography-mass spectrometry (LC-MS). These methods can be targeted for specific metabolites e.g. cholesterol measurements or untargeted metabolomics to capture as many metabolites as possible [62, 63]. The study of metabolites is considered highly dynamic, as biochemical processes constantly occur. Metabolomics thus provide insight of biochemical processes at a given time point which can be useful for understanding disease evolution [63, 64].

2.4 Clinical study designs

Data on human health and disease can originate from a wealth of clinical study designs. Study designs and their analysis and interpretation have decades of experience and clinical tuition behind them [28, 65]. In this thesis, machine learning methodologies have been explored on data obtained from different clinical study designs, and this is thus by no means a comprehensive introduction to the area. This section will instead

briefly introduce study designs and how these relate to the projects presented in this thesis.

Clinical study designs can either be retrospective or prospective referring to the timing of the studies. Retrospective studies investigate previous collected data either from records or by interviewing study participants about previous exposures, while prospective studies follow a population (also referred to as study cohort) from the time of enrollment in the study over a time period to determine the occurrence of outcomes [65, 66]. Furthermore, the clinical study design can be divided into observational (epidemiological) or interventional studies. In observational studies, there is no active intervention, whereas interventional studies test the effect of an intervention in study participants [65, 67]. Observational study designs included in this thesis are case-control studies and cohort studies. In case-control studies, the difference between the cases and controls are compared to identify different risk factors between the groups. Case-control studies are retrospective, where cases and controls are collected so the study have sufficient statistical power to investigate statistical associations between groups with the outcome of interest [68]. In *Chapter 6*, a case-control study design is used on AAP in childhood ALL patients.

Cohort studies are observational longitudinal studies where a group of individuals ('the cohort') with shared characteristics are monitored to infer information over time. The information is collected in cross-sectional intervals. Cohort studies can be retrospective by using information collected previously or prospective by following the cohort over time and collect new data [68, 69]. The GoDARTS study was started in 1996 to identify patients with diabetes and study disease onset, progression and response to treatment [25]. This cohort is linked to EMRs from 1994 for some patients. The GoDARTS cohort has been used to study the time to insulin in T2D patients in *Chapter 5*.

Interventional studies are prospective experimental studies, where the intervention is changed in one or more groups of study participants to study outcomes relating to the intervention [65]. Interventions test the effect of diagnostic tests, diet, drug treatments etc. and compare outcomes between study participants with changed and unchanged variables [65]. Randomized controlled trials (RCTs) are the most common interventional study design [65].In RCTs, a homogeneous population is defined given specific inclusion criteria for the study. The study cohort is then randomized into two groups, where one group will be exposed to the intervention and the other group is used as a control group. The randomization step should ensure no selection bias, meaning that the only difference between the groups is the tested intervention, and thus allow researchers to infer causal implications of the intervention [65].

Randomized controlled cross-over trials test the impact of the intervention in all individuals. Study participants are usually randomized and will receive all tested treatments i.e. study participants will first receive one treatment before crossing over to the other treatment. The cross-over study design has a wash-out period between the interventions to allow the study participants to return to a 'baseline' state and avoid any carry-over effect from the first intervention. The cross-over design should be able

to compensate for unsuccessful randomization of groups as study participants serve as their own control [65, 70]. A study on weight loss predictions using data obtained from two randomized controlled cross-over trials is presented in *Chapter 4*.

Another RCT approach are trials for individualized care, the 'N-of-1' trials, which are multiple-period cross-over experiments where the treatment effect is compared within a single individual. This helps identifying the most efficient treatment in diseases with slow progression or chronic diseases for one single patient by assessing clinical outcomes repeatedly [71, 72].

RCTs have long been the gold standard of clinical trials, however the increase in Big data from electronic health records (EHRs) and patient registries may introduce more personalized approaches to clinical trial designs [28]. It is thus exciting to see where machine learning approaches and 'N-of-1' ideas will find their way into assisting clinical care and achieving individual-level patient attention through the utilization of previously hidden information such as the patient's genome sequence.

2.5 Electronic medical records and electronic health records

EMRs and EHRs function as observational databases that provide an opportunity to infer information from the entire medical history of a patient [25, 73]. EMRs data is collected within a single clinical practice, whereas EHRs are designed to be shared with multiple clinical providers [74]. EMRs and EHRs contain longitudinal information about medical diagnoses, drug prescriptions, imaging and results from laboratory tests as well as more unstructured information such as clinical notes [15, 21, 25, 45, 75]. Recordings in EMRs and EHRs are intended for use in clinical care, however they can be re-purposed for findings in precision medicine where the longitudinal aspect linked with clinical monitoring allows an approach to study several disorders and their progression as well as define disease subtypes [15, 76, 77]. As an example, data from EHRs has recently been used for prediction of gestational diabetes [78]. Electronic recordings are made when a patient interacts with a clinical facility (e.g. physicians or hospital). Since the electronic records not only contains information of a patient's health, but also contains information about a patient's interaction with the health care system, it is important to keep this in mind when analyzing the data. As an example, the date a patient was diagnosed with T2D is perhaps not the date when the disease developed. Furthermore, the recorded test is always ordered by clinical staff. 'Missing' information in electronic records can therefore provide information that the test was not needed, thus providing information about the clinical decision making process [73, 79]. Clinical decisions are dynamic and dependent on the current guidelines of care [73] which should be considered when analyzing the electronic record data and how this may influence the analysis.

2.6 Data-driven explorations for precision medicine

In the era of precision medicine, an increasing focus is to understand predisposing features of disease in order to prevent, predict or prolong disease initialization as well as influencing care pathways [30]. The link between a genomics biomarker or other high-throughput data and disease outcomes has for years been inferred at a population level by associations and statistical analyses of a single data type [1]. These analyses are performed assuming independence between variables where the analyses sometimes are adjusted for other variables or effect factors [65, 67]. These analyses have lead to important biological findings, and biomarkers for clinical translation. However, the number of identified biomarkers were less than originally expected [24]. A possible reason relates to the fact that biological systems are complex which makes the interplay between variables important to consider [11]. Developments in analytic methodologies have emerged with a shifting focus towards predictive modelling of subgroup or individual outcomes to complement association analyses at a population level. The following subsections introduce a background on methods and concepts that are used to analyze signals in biological data considering findings at population, subgroup or individual levels as well as data integration.

2.6.1 Moving from GWAS to disease prediction

The field of genomics is considered one of the foundations for developments in precision medicine [5, 30, 80]. Several methodologies are available to study genomic variants and their impact on personalized outcomes. In GWAS, SNPs are statistically tested for association to the phenotype in affected and non-affected individuals yielding a P-value for each SNP [4]. Thus, GWAS perform several independent statistical tests, which require multiple testing correction. The Bonferroni correction is often used in GWAS, where the genome-wide significance level, $P < 5x10^{-8}$, has been used for new genetic discoveries [11, 81, 82]. When an association is identified by GWAS, further analyses are needed to follow up on the signal to gain insight to biology. This relates to identification of the causal SNPs in disease mechanisms due to the design of the genotype array where a tag-SNP functions as a proxy of adjacent SNPs. Thus, GWAS does not necessarily detect the causal variant or variants, but rather a region on the genome in linkage disequilibrium (non-random co-inherited alleles) with the causal variant or variants. Fine mapping is therefore required to establish the causal variant [4]. Furthermore, it is important to gain understanding of a SNP's functional consequence. This is dependent on the genomic region where the SNP is located. In the coding region of the genome, the SNP can be synonymous or non-synonymous. The SNPs in coding regions are often synonymous [50] and have low functional impact as they do not change the translated amino acid. Non-synonymous SNPs, resulting in missense or non-sense variants, have a high functional impact on the amino acid sequence. The majority of SNPs are located in the non-

coding region of the genome and a larger proportion of these SNPs are suggested to be in regulatory elements [45]. Databases such as Ensembl have predictive tools such as the Variant Effect Predictor to evaluate the functional consequence of a given SNP [83].

GWAS have reported several genetic associations in complex diseases and traits. As of January, 2020 the NHGRI-EBI GWAS Catalog counts 172351 variant-trait associations from 4410 publications [84, 85]. These associations have provided insight of novel disease-causing mechanisms in various diseases and traits as well as clinical applications [11]. Despite these findings, it is argued that GWAS have only provided limited success given its original promises for precision medicine [45, 86].

The immediate clinical utility of many GWAS findings are challenged by the realization that the identified variants typically only explain modest fractions of complex heritable traits with average odds ratios $\sim 1.1-2$ [11, 24]. The 'missing' heritability of GWAS may be explained by gene-gene interactions and/or gene-environment interactions [11]. These more complex interactions are missed in single-marker association studies. Given the multi-factorial nature of many diseases and treatment response, it is relevant to consider the complex interplay in biological systems and individual variability across data types. Furthermore, it is possible that SNPs with weaker effect sizes do not pass the genome-wide significance threshold, but still contributes to the phenotypic trait through SNP interactions [45, 87]. Rare variants are also suggested to explain a part of the 'missing' heritability of common diseases [45]. Rare variants have limited statistical power for GWAS and are thus less well-studied. However, use of large cohorts, newer versions of GWAS genotyping arrays with a range of low-frequency variants and use of population reference genomes such as the 1000 Genome Project for imputation of genotype data have improved studies of rare variants [11].

To advance GWAS findings for clinical translation, polygenic risk scores (PRS) (also refered to as genetic risk scores (GRS)) are being applied [88]. PRS exploit that complex diseases often are polygenic and increase the predictive power by combining the effect from several genetic variants [24, 89]. PRS can either by calculated by an unweighted or weighted score. The unweighted PRS sum the risk-increasing alleles thereby giving all alleles an equal impact on phenotype. In weighted PRS, the effect sizes or odds ratios derived from GWAS are used to weight the risk-increasing alleles similar to linear regression models. PRS are easy to interpret and apply, however they are still limited by missing out on more complex genetic interactions as well as environmental interactions [24].

In order to leverage genomic information from complex biological systems to improve the predictability of subgroup or individual outcomes, more sophisticated methods are emerging. Machine learning methodologies are considered useful for genotype-phenotype explorations due to their ability to detect complex non-linear multivariate data interactions without underlying assumptions of the data distribution [24] as well as their ability to integrate genetic and environmental parameters. Furthermore, there are flexibility in how a machine learning model is setup, which allows for addressing some of the challenges in working with heterogeneous high-throughput data and integration of diverse

data types. The presented methodologies in this section provide their own strengths and limitations when exploring biology.

2.6.2 Data integration

The fact that biology is complex makes it attractive to integrative several data types to understand complex relationships in biology [90]. Data integration is however challenged by the noisy nature of biological data and its heterogeneity, sparsity, multicollinearity, and high-dimensionality [3]. Thus, there is a need to explore how to prepare different domains of biological data for integration. Machine learning methodologies present an approach to integrate Big data in biology as the algorithms can leverage existing heterogeneous data by feature selection strategies and thereby prioritize the most predictive features. Systems biology can also provide useful concepts that help incorporate prior knowledge into predictive models. This can be especially important when working with e.g. large genetic datasets where scaffolding data into biological pathways or protein-protein interactions, can help reduce the number of data points included in any model [91]. Machine learning models can during training take individual effects into account, where interaction of features may only be present in some subgroups [23]. Furthermore, machine learning models provide a hypothesis free testing strategy towards identifying predictive features that have the potential to provide information on new biological knowledge across diverse data types [15].

2.6.3 Longitudinal Big data

Some biomarkers remain mostly static over time such as germline genetic variation, whereas more dynamic biomarkers in omics, physiology and environmental factors (diet, behavior, sleep, smoking, lifestyle etc.) can change over time. Longitudinal studies employ continuous monitoring or measures to follow changes in disease progressions over time across a cohort but also within an individual. This can also provide information on how phenotypic characteristics are changing from a health state towards a disease state [15, 92]. Screening studies that follow changes over time within a single individual have already assisted in providing individual preventive strategies to improve health while also providing insight to early disease mechanisms [10, 20, 32]. This longitudinal approach has also been applied to study the stability of human health given extreme environmental exposure in the NASA twin study where multi-omics measures and cognition were targeted as a focus for monitoring how humans adapt to space that can be used for future space flights [93]. Using longitudinal variations of biomarkers within an individual in Big data screening approaches have been criticized to result in overdiagnosis of disease abnormalities without the ability to predict which abnormalities will cause health deterioration [80]. However, these studies are important for studying disease progression and in identification of early disease biomarkers. It remains to be seen, albeit promising, how far individual-level approaches and understanding will complement current clinical practice.

2.6.4 Artificial intelligence for disease predictions

There are exciting opportunities to exploit and create knowledge on disease onset, disease progression, treatment response as well as other health outcomes based on the Big data available in biology. AI presents an approach for integration of heterogeneous data and prediction of disease outcomes at the individual patient level. This has potentially transformative impact on the way clinical decision making is currently done towards moving from the 'one-size-fits-all' recommendations to more tailored guidance for individual patient care. This vision for AI in human medicine is supported by a report by the European Medicines Agency (EMA), where AI is envisioned as a future tool for decision making [94]. More recently, the U.S. Food and Drug Administration (FDA) published policies on software functions and medical applications as well as a draft policy on clinical decision support software in September, 2019 [95, 96]. The fact that regulatory agencies are paying close attention to AI and machine learning technologies as potential tools in clinical decision guidance underlines the large interest in these methodologies. However, the adoption of AI in health care has been expected to be difficult and slow according to a recent Harvard Business review as there is still a need to develop these regulatory frameworks further and deal with the 'black box' nature of machine learning models [97]. Furthermore, there are several data analytical challenges that need to be addressed when developing predictive machine learning models; some of which are addressed in the research projects of this thesis. Key concepts of machine learning methodologies are introduced in the following chapter.

CHAPTER 3

Machine learning

Machine learning is an area within AI that combines methodologies originated from mathematics, statistics and computer science. One of the main focus points in machine learning is development of prediction models [13, 98]. Machine learning has been applied in several disease areas to predict outcomes such as T2D complications [99], breast cancer screenings [100], cardiovascular disease risk [101] and treatment response in rheumatoid arthritis [102] as a few examples.

The typical workflow of a machine learning analysis following data collection is composed by data exploration and preparation, modelling and evaluation of results. In data exploration, several operational data challenges are dealt with to prepare data for modelling. It is worth spending time on data preparation, as the quality of the data has a large impact on the final quality of the predictive model. In the modelling step, machine learning models are trained. Training involves cross-validation, fitting of parameters, tuning of hyperparameters as well as feature selection. Finally, the machine learning model is evaluated on its performance and validated before applying the model to predict outcomes on new data [13, 98] (Figure 3.1).

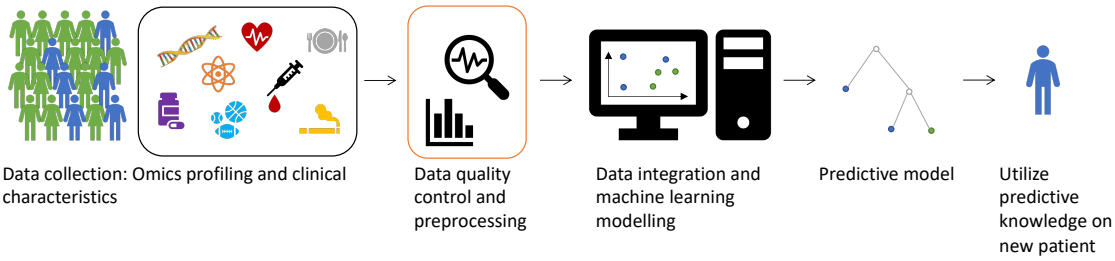


Figure 3.1: Workflow for developing a predictive machine learning model.

3.1 Operational data challenges

Any dataset is given by a set of samples and features. Every feature can consist of different types of values. The features collected in a dataset can be dichotomous/binary, discrete or continuous. Dichotomous/binary features can only take up two values usually either 1/0 (e.g. 'True/False', 'yes/no', 'male/female'). Discrete features can take a

finite value where there is a distance to the next value (e.g. the number of days since a treatment started or the number of drugs prescribed to a patient). Continuous values can however be of infinite number of values (e.g. body weight or height). Before data can be presented to machine learning models, thorough exploration of the data is needed to understand the dataset and its features, size, heterogeneity and irregularity. Operational data challenges that need to be considered in any data-driven approach are:

- **Data redundancy and irrelevant features.**

  Irrelevant features such as the personal identifier and redundant features should be removed prior to modelling.

- **Outliers.**

  Investigation of a feature's summary statistics and distribution assists in the detection of errors due to mis-reported measurements in data that can be corrected upon further data analysis. In patient data, measurements are often reported by a person, and errors can thus easily occur, which should be checked carefully.

- **Longitudinal data**.

  Longitudinal data presents challenges for non-deep machine learning methods which requires data presented in a structured format [103]. Longitudinal data can be presented by data extracted at specific time points or by features modelled to capture longitudinal information or trends in data [103]. Including longitudinal data in machine learning models have showed to improve performance in prediction of cardiovascular disease compared to single time point values [104].

- **Feature transformation**.

  Data may have to be formatted or normalized across data types to a common scale prior to modelling [3].

- **Missing data**.

  See subsection 3.1.1

- **Feature reduction and selection**.

  See subsection 3.1.2

3.1.1 Missing data

Missing data is common in any data analysis. No good solution exists for dealing with missing data; either missing data (features or samples) can be removed for complete case analysis, or missing values can be substituted with values by imputation. Assumptions regarding the nature of the missing data are made given three different mechanisms; i) data is missing completely at random (reason for missing data is completely unrelated to the dependent variable), ii) missing at random (reason for missing data does not depend

on the unobserved data) or iii) missing not at random (reason for missing data does depend on the unobserved data) [105]. If data is missing completely at random, data with missing values can be omitted [106]. Other assumptions about the missing data pattern allow for missing values to be imputed e.g. by regression estimating the missing value by multiple imputation. It should be noted that several imputation strategies are available. The validity of the assumption on the missing data pattern can rarely be determined from the data. The sensitivity of the imputation on study conclusions should thus be assessed, as inappropriate imputation may introduce bias in the downstream analysis [105]. The decision of whether or not to impute data should be evaluated given the potential impact on further modelling by loss of information, bias and power for a study.

3.1.2 Feature reduction and selection

Big data collected from omics and clinical phenotypes often result in high-dimensional datasets (i.e. a high number of features) given a limited number of samples. This is referred to as the 'curse of dimensionality' [3]. The risk of overfitting a model increases when the number of features (M) >> the number of samples (N) [13, 15]. Overfitting is a term describing when the model is fitted to the training data, but does not generalize well to new data. In order to deal with the curse of dimensionality in a data analysis, the number of features can be reduced or selected. In feature reduction, new features are created through data transformations e.g. a principal component analysis [98] or weighted correlation network analysis [107].

In feature selection, the goal is to identify which features are most predictive in a model of a given outcome [23]. Feature selection is useful in data integration as this allows detection of predictive signals across different data types. Two overall approaches are available for feature selection in machine learning modelling including prior knowledge or data-driven approaches [23, 91, 108, 109]. These are also known as filter feature selection (prior knowledge-driven) or wrapper and embedded feature selection (data-driven).

In the filter methods [23], features are selected based on information in the data such as variance, correlation or statistical associations given a pre-defined threshold for filtering [23]. Another prior knowledge method is using systems biology functional domains of ontologies, pathways, or protein-protein interactions to prioritize features [3, 110]. Prior knowledge of systems biology can improve the predictive performance in machine learning modelling [13, 91]. The strength of this approach is that the selected features can be compared to previous knowledge which provides a biological validation of the underlying mechanism. The limitation of the prior knowledge approach for feature selection is that this approach is possibly leaving out new discoveries of predictive features or interactions between features important for prediction [23].

Several data-driven feature selection approaches are available within the wrapper methods and embedded methods. In wrapper methods, a search algorithm iteratively

selects a subset of features to test and evaluates the predictive performance of the given feature combination [23]. Examples of wrapper methods are sequential forward selection or recursive feature elimination. In embedded feature selection methods, the machine learning model has a built-in feature selection strategy e.g. L1 regularization models or decision trees [108, 111, 112]. Wrapper and embedded feature selection methods have resulted in better predictive performance compared to filter methods [23]. Data-driven feature selections can assist in the discovery of predictive features as well as correlations between features. This ultimately allow for generation of new hypotheses of biomarkers important for disease and treatment response. However, a downside to the data-driven approach for feature selection, is that it may miss out on weaker signals in data and poses a higher risk of overfitting as well as increased computational run-time during training of the machine learning models.

3.2 Machine learning algorithms

The most common learning tasks in machine learning algorithms are supervised and unsupervised learning [3]. In unsupervised learning, the input data is unlabeled and the training of the models focuses on learning descriptions of the underlying structure within the input data by clustering, outlier detection, density estimation or association mining [98]. In contrast, in supervised learning, the input data is labeled with a prediction outcome which is known in the training of the model. Supervised learning typically involves prediction by regression (quantitative outcome e.g. HbA1c (%)) or classification (qualitative outcomes e.g. diseased or not diseased). In this thesis, classification models were developed which provide an individual prediction score corresponding to the probability of the outcome ranging 0 to 1. Supervised learning algorithms range from simple linear models to complex non-linear models. The number of samples available for modelling influence how complex a model can be fitted to the data. As a general rule, the more data available, the more complex algorithm can be trained without overfitting to the training dataset [113, 114]. Similar, a model can be underfitted if a too simple model is applied or parameters of a model is not properly tuned. In this thesis, popular algorithms in bioinformatics were used, including logistic regression, random forests and artificial neural networks (ANNs), for modelling tasks across different areas of health and disease. Regression models, random forests and ANNs are introduced in the following sections.

3.2.1 Logistic regression

Logistic regression is a generalized linear model used for classification of dichotomous prediction outcomes i.e. the dependent variable either belongs to a negative class or a positive class. In logistic regression, the outcome variable is represented by a probability between 0 and 1 given a Bernoulli distribution for the outcome distribution. Thus, the probability of the positive class can be expressed as p and the probability of the negative

class is $1-p$. In logistic regression, a linear combination of the independent variables is mapped through a link function i.e. the logit function (the logarithm of the odds ratio):

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + x_1 \cdot \beta_1 \tag{3.1}$$

Where $\beta_0 + x_1 \cdot \beta_1$ corresponds to a linear regression model with coefficients β and input feature x. The inverse logit is the logistic sigmoid function that can transform the log-odds to a probability ranging 0 to 1 of either class prediction $\sigma(z)$:

$$\sigma(z) = \frac{1}{1+\exp(-z)} \tag{3.2}$$

Where $z(x) = \beta_0 + x_1 \cdot \beta_1$ [98, 115], for a simple linear model. Logistic regression can also use multiple linear regression which increases the number of parameters in the model. The coefficient parameters of the linear model in the logistic regression are estimated by a cost function that evaluates how well the model is fitted by minimizing the overall error between the observed and predicted values. This is done iteratively through an optimization algorithm such as maximum likelihood [115].

3.2.2 Regularized regression

Logistic regression with several parameters increases the model complexity and thus the risk of overfitting. The risk of overfitting can be reduced by regularization. The two most common regularization methods in regression are ridge regularization (also called L2 regularization), or least absolute shrinkage and selection operator (LASSO) regression (also called L1 regularization) [24]. In both methods, the coefficients of the linear model can be regularized by a penalty λ. In ridge regression the constraint λ is set on the squared coefficient estimates in the cost function, which decreases the value of the coefficients towards 0. In LASSO regression, the penalty λ is added to the absolute value of the coefficients in the cost function which both shrinks the coefficients and can assist in feature selection as some coefficients can be set to 0 [116, 117]. Finally, the elastic net regularization combines the ridge and LASSO penalties [117].

3.2.3 Decision trees

Decision trees are hierarchical models that apply a branching strategy in prediction problems. The decision tree works by setting decision boundaries through multiple if/else statements in the splitting of groups of samples for the prediction outcome. This results in a 'tree-like' structure of the model, where the root node represents all available samples for training. When it is not possible to split a group further, meaning the node only contains one group, the group ends in a 'pure leaf'. In each split, features are tested for which feature leads to the optimal splitting by a split-criterion that compares

impurity before and after a split [98]. The split-criterion can be evaluated using different measures of impurity including entropy, Gini or classification errors [98]. The decision tree continues to grow until each class ends up in a pure leaf. Decision trees are thus easily overfitted to the training dataset. In order to overcome this, a decision tree's structure can be controlled by early stopping through controlling the depth of the tree, a minimum number of samples within a leaf to continue splitting, or based on a minimum purity gain or by pruning after the tree is built [98]. Benefits of decision trees include that it is not necessary to do any scaling or normalization of data prior to modelling, as each feature is evaluated separately, and the decision-making is easy to visualize. Drawbacks of decision trees are that they contain high variance and are easily overfit to the training data thus failing to generalize to new independent datasets. To overcome this challenge, multiple decision trees are combined in a random forest model [113].

3.2.4 Random forests

Random forests are ensemble models of multiple uncorrelated decision trees to improve the predictive performance by reducing the variance from individual decision trees. To ensure the decision trees are different, random forest uses two strategies. One strategy is bagging (bootstrap aggregation) where subsamples of the original datasets are created with replacement. The other strategy is setting the number of randomly sampled features available before splitting at each node [118] (Figure 3.2). The random forest can further be controlled during training by the number of trees to grow, the number of samples within the final leaves, the maximum number of terminal nodes and the number of samples for modelling in the random forest [119]. The random forest uses each tree to make a prediction. In classification, the final class is determined by majority voting from all the trees in the forest.

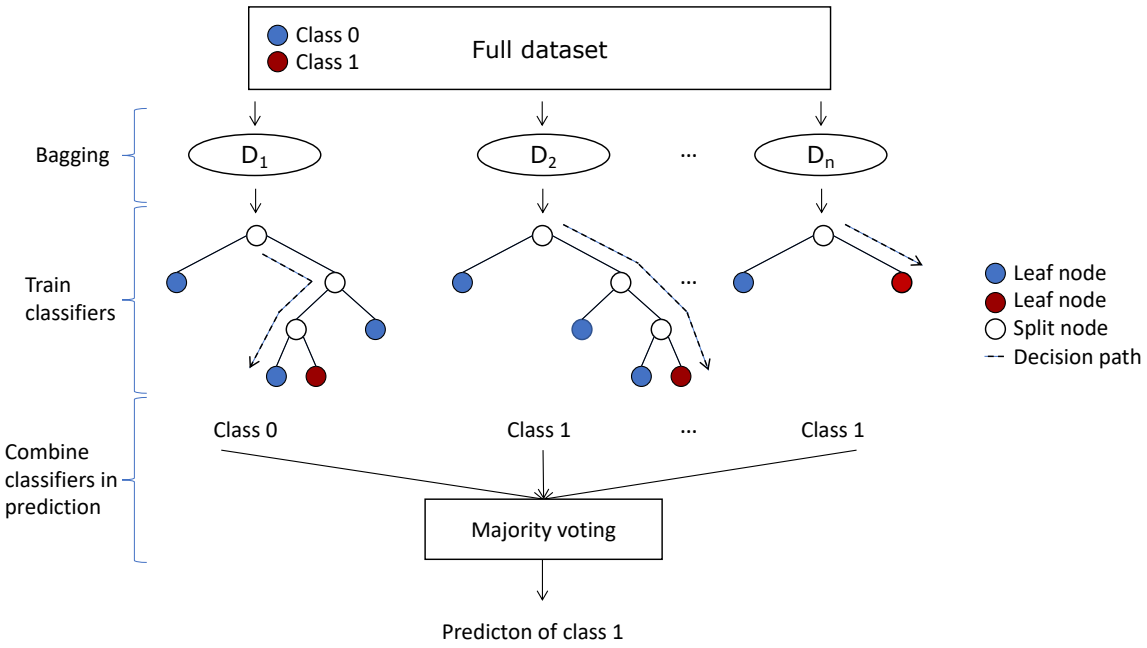


Figure 3.2: Random forest model with n trees are in the forest. D_1 to D_n are subsamples of the original dataset. The dashed lines illustrate the decision path when the random forest is making a prediction on a new sample.

3.2.5 Artificial neural networks

ANNs are machine learning models whose structure is strongly inspired by neurons in the human brain [120]. ANNs have gained their popularity as they are capable of learning highly complex and non-linear patterns in data due to their structure. ANNs are composed by three types of layers; an input layer, a hidden layer and an output layer. Each of these layers contains nodes (or neurons) that are fully connected by weights except for the bias node (Figure 3.3). The input layer's nodes are a vector of the features available for modelling. An ANN is trained to infer the relationship between the input features and the prediction outcome by optimizing the value of the weights by forward-backward propagation through several iterations to minimize the error between the observed and predicted value [121]. In the first iteration, the weights in the ANN are randomly initialized, usually around zero [116]. Given these weights, the feedforward propagation calculates the predicted value through parameters in the ANN. The information is processed sequentially, where the input features from the input layer are forwarded by a linear combination of weights to each neuron in the hidden layer, and from the hidden layer by a linear combination of weights to the output layer. The nodes in the hidden and output layer receive the linear signal from the previous layer and translate the signal through a non-linear activation function which typically includes the logistic sigmoidal function [98]. Each ANN typically contain multiple hidden nodes to

transform the linear signals. The prediction produced in the output layer is compared to the observed value using a cost function. This function evaluates the error between the observed and predicted values, often using the cross-entropy for classification tasks [121]. The error of the cost function is backpropagated from the output layer to the input layer. Using a weight optimization algorithm such as gradient descent, the weights are updated in the direction that minimizes the error of the cost function [121]. The forward-backward propagation is iterated until the model converges. To avoid overfitting of ANNs, early stopping or regularization, such as weight decay, can be implemented in the training of the neural network. Following training of the ANN, one round of forward propagation is run with the optimized weights to make predictions given the input data. ANNs are extremely flexible for modelling tasks, as they combine the output from neurons in the hidden layers that allow for detections of non-linear correlations in data. The flexibility offers great advantage in modelling of complex research questions and powerful predictions. However, given too many nodes, the ANN can be so complex it will overfit to the noise in the data [121]. Another important aspect of ANNs is the fact that the feature importance is difficult to interpret for clinical applications [122].

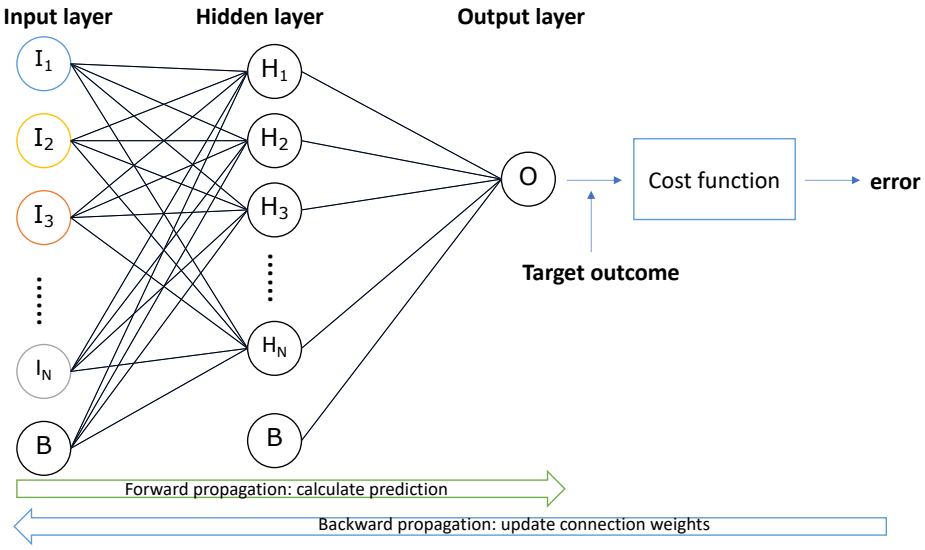


Figure 3.3: A one-layer artificial neural network and forward-backward propagation.

3.3 Deep learning

Deep learning is a subtype within machine learning which typically utilizes ANNs designed with multiple hidden layers and connection loops. This structure allows extremely complex modelling by transformation of input features during training of the model and detection of underlying data structure that is missed by non-deep machine learning models [13]. Challenges for deep learning models are that they require large amounts of samples for training and the feature importance from the trained model can be difficult

if not impossible to understand [3, 13, 104]. Deep learning models require more computational power, which has become feasible by the use of Graphical Processor Units (GPUs). GPUs can accelerate processing up to a hundred times compared to central processing units (CPUs) [15]. As an example in precision medicine, recurrent neural networks have been used to predict patients' future diagnoses and/or medication based on their EHRs with 8 years of longitudinal data [123]. The papers presented in this thesis have not focused on the use of deep learning, but nonetheless, deep learning has promising applications for genomics and precision medicine as the amount of Big biological data continues to increase [3, 13].

3.4 Performance measurements

The best performance measurement should always be evaluated given the type of prediction task (regression or classification) and the desired performance for deployment of the machine learning model. This thesis has focused on binary classification tasks, where several performance measurements exist. These are calculated by evaluating how many times the model made a correct or incorrect classification using thresholding of the predicted class probability. This prediction threshold can be adjusted dependent on the clinical utility of the model. Typically, a prediction threshold of 0.5 is used in classification, where scores > 0.5 are classified as belonging to the positive class, and scores ≤ 0.5 as belonging to the negative class. This results in four outcomes as given in the confusion matrix (Table 3.1).

	Predicted Positive class	Predicted Negative class
Actual Positive class	TP	FN
Actual Negative class	FP	TN

Table 3.1: Confusion matrix for a two-class classification problem with true positives (TP), false positive (FP), true negatives (TN) and false negative (FN).

The values in the confusion matrix can be used to calculate accuracy (Equation 3.3), sensitivity (Equation 3.4), specificity (Equation 3.5), the positive predictive value (PPV, Equation 3.6) and the negative predictive value (NPV, Equation 3.7).

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{3.3}$$

$$True\ positive\ rate = Sensitivity = recall = \frac{TP}{TP + FN} \tag{3.4}$$

$$True\ negative\ rate = Specificity = \frac{TN}{TN + FP} \tag{3.5}$$

$$Positive\ predictive\ value = precision = \frac{TP}{TP + FP} \tag{3.6}$$

$$Negative\ predictive\ value = \frac{TN}{TN + FN} \tag{3.7}$$

Accuracy is a measurement of how many classes were correctly predicted out of the total number of samples [98]. Sensitivity is a measurement of the proportion of the positive class being correctly predicted, while specificity is a measurement of the proportion of the negative class being correctly predicted. The PPV and NPV instead reports the proportion of positive predictions that are truly positive and the proportion of negative predictions that are truly negative [124].

A common challenge for predictive modelling with health and disease is data imbalance i.e. the number of the positive and negative classes are unevenly distributed. This occurs as a disease event may only impact few individuals compared to controls. In this case, machine learning algorithms are biased towards the majority class and accuracy is no longer an appropriate performance measurement. Two commonly used methods to deal with data imbalance for performance evaluation are resampling or using more appropriate performance measurements. In resampling, the dataset used for training of the model is modified. In down-sampling, samples of the over-represented class are removed at random, so the positive and negative classes are of similar sizes, whereas in up-sampling, samples of the under-represented class are sampled with replacement at random, so the positive and negative classes are of similar sizes. Hybrid methods of down- and up-sampling dealing with class imbalance also exist. The resampling approaches are easy to implement. However, as the sampling is random, this can produce model uncertainty which can impact performance, and some information for the training of the model is either under- or over-represented given the sampling strategy [98, 125].

Performance measurements such as the area under the receiver operating characteristic curve (ROC-AUC) or Matthews correlation coefficient (MCC) are performance measurements that are reported as being invariant to class imbalance [98, 126]. Yet, the ROC-AUC can still result in an overly optimistic performance in case of a large difference in the class distribution [127]. ROC-AUC is the area under the receiver operating characteristics curve, which is generated by the true positive rate (TPR, same as sensitivity) and the false positive rate ($FPR = 1 - specificity$) at varying classification thresholds. The TPR and FPR are normalized by the number of observations in the positive and negative class, respectively. A ROC-AUC of 0.5 indicates random model performance and that no class separation can be made, while a ROC-AUC of 1 indicates perfect classification by the model [98]. Considering the ROC curve at different thresholds, this can be used as a diagnostic plot. For example, if the classification threshold is decreased, more positive classes are predicted correctly by the model but at a cost

of decreasing specificity. ROC-AUC can be interpreted as a probabilistic performance measurement of identifying a positive class out of any random positive-negative class pair [128], where e.g. ROC-AUC: 0.65 translates to a 65% chance that the model can classify any random positive-negative class pair correctly.

MCC (Equation 3.8) ranges between -1 and 1, where -1 indicates that the classifier is predicting everything belonging to the opposite class, 0 means that the predictions are random, and 1 indicates a perfect agreement between the predicted and actual values [126].

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}} \tag{3.8}$$

3.5 Cross-validation

Cross-validation is a technique applied in machine learning to evaluate generalization performance of a trained model when used on new unseen data. In brief, cross-validation splits the entire dataset into training and test datasets, where the training dataset is used for training of the machine learning model and performance is evaluated on the test dataset. Multiple cross-validation strategies exist and the most common techniques are described.

The hold-out cross-validation strategy refers to the use of a single training and test dataset to evaluate performance. However, the split of data into a training and test dataset may have been poor. K-fold cross-validation reduces bias from the selection of training data compared to the hold-out cross-validation and provides more effective use of the available dataset. In K-fold cross-validation, the total dataset available for modelling is split into K sets, where K models will be trained across different dataset and tested on K test datasets ($K = 5$, Figure 3.4). The performance is reported across K test datasets, allowing evaluation of the model generalization across all samples in the dataset. Typically, the mean performance and standard deviation or standard error, or the 95% confidence interval performance is reported to evaluate model's performances across different splits. It is possible to apply stratified K-fold cross-validation, which splits the classes in equal proportions across the K datasets.

In leave-one-out cross-validation, K is set to be equal the number of samples in the dataset (N), $K = N$, where training is done on $K - 1$ samples and the model is tested on one single sample. Leave-one-out cross-validation maximizes the power for training of the machine learning model but increases the computational burden.

Different machine learning models have distinct model hyperparameters that can be tuned, and when evaluating the performance on the test dataset, there is a risk of overfitting these parameters during training. In order to optimize independent training of these, nested cross-validation can be used. Nested cross-validation have two layers of cross-validation where the dataset is split into three parts of a training dataset, a

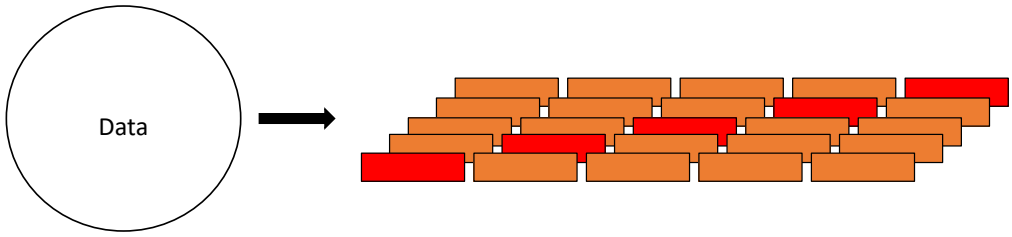


Figure 3.4: Five-fold cross-validation ($K = 5$), red = test dataset, orange = training dataset.

validation dataset and a test dataset. An example of nested cross-validation ($K_{outer} = 5$, $K_{inner} = 5$) is given in Figure 3.5. In the inner layer, the learning process is repeated to estimate the optimal set of hyperparameters as well as feature selection given the predictive performance evaluated on the validation set [13].

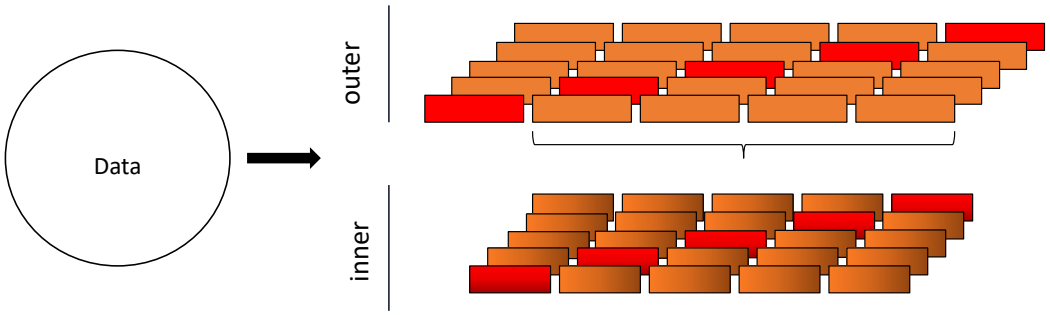


Figure 3.5: Nested five-fold cross-validation, outer red = test dataset, outer orange = training dataset, inner red = validation dataset, inner orange = training dataset.

3.6 Model robustness

The performance of a machine learning model with random initialization of parameters (e.g. bias and weights in an ANN) is dependent on the random model initialization seed on which the model was trained. The choice of random seed should be set to ensure reproducibility of the model. Since the randomization of model parameters influences training and thereby potentially model performance, it is useful to repeat training of the machine learning models across several seeds to ensure the model performance is robust. This performance can be further cross-checked against a permuted label setup in permutation analyses. Permutation analyses allow statistical comparison of the performance distribution of classifiers over multiple model initializations trained on permuted and non-permuted prediction labels to evaluate the robustness of the models [129].

3.7 Feature importance

An important aspect of machine learning modelling is understanding each feature's contribution in prediction of the outcome. The feature importance across a logistic regression, a random forest and an ANN is evaluated by different methods due to the parameters of the models. The simplest model to understand is logistic regression, whereas the most difficult feature importance to interpret is given by the ANN.

The feature importance in a logistic regression model is given by its coefficients, where a larger value indicates a more significant importance. The positive or negative sign of the coefficients allows insight to how this impact class prediction. For random forest models, the feature importance has been evaluated using the Gini index, which refers to the split criteria of the random forest [116]. The feature importance thus reflects how much a given variable contributed to a decrease in node impurity. The importance is dependent on the number of features in the model, as the feature importance sum to 1. If all the features are equally important, this is assigned the value of $\frac{1}{M}$, where M is the number of features. In both decision trees and random forest, the impurity-based feature importance is always represented by a positive number, and is thus not indicative for what class the feature is most predictive of [113]. In ANNs, the weights provide insight to the relative feature importance. Olden's algorithm evaluates feature importance in ANNs by a sum of the product between the weights in the input-hidden layer and hidden-output layer per input feature [130]. However, as multiple nodes often exist in the hidden layer, the relationship between the importance and the outcome is not a simple relationship. To understand how the feature importance influences predictions in a single sample, local interpretable model-agnostic explanations (LIME) can be explored [122].

It can be challenging to compare feature importance across different types machine learning models. In order to compare the feature importance across different types of models, a 'leave-one-feature-out' approach was applied in this thesis. In this case, the model was re-trained setting each feature's value to zero. The impact of the feature in the overall model was then compared by the change in predictive performance such as ROC-AUC.

3.8 Ensemble classifiers

Several predictive model classifiers can be combined into an ensemble model. In ensemble models, the predictions from individual models are combined sample-wise by different scoring methods e.g. the mean to make a final ensemble prediction for each sample. An ensemble of multiple independent machine learning models generally improves the predictive performance and results in more robust predictions rather than a single machine learning model [13]. The use of ensemble models can overcome challenges from missing data and assist in dealing with high-dimensional heterogeneous data by providing a simple framework that combines models without increasing the complex-

ity of individual models. Ensemble models thus provide an approach towards improved predictive performance and stability of the final prediction as well as data integration.

3.9 Introduction to machine learning applications

The three research projects presented in the following chapters used machine learning methodologies to develop predictive models with potential applications in precision medicine. The research projects applied concepts presented in this chapter to deal with different data challenges. These involved handling large volumes of diverse types of Big data in both small and large study cohorts.

CHAPTER 4

Health application: Prediction of weight loss in dietary clinical trials

4.1 Metabolic health, diet and weight loss

Worldwide, more than 1.9 billion adults were overweight in 2016 - a number which is on the rise [131, 132]. Overweight and obesity are serious public health challenges and are associated to metabolic disorders such as T2D, cardiovascular diseases and certain types of cancer [133, 134]. Weight loss can reduce the risk of metabolic dysfunctions resulting in lifestyle-related comorbidities in overweight and obese individuals [135]. Thus, there is a considerable interest identifying factors that are predictive of weight loss [136]. Various dietary interventions have been suggested to improve chances of weight loss [60, 133, 137, 138]. At the same time, it has also been argued that no single diet can be useful for weight loss across all individuals [139]. Individual predisposition to weight loss in response to dietary interventions are still relatively unknown, and the great heterogeneity in factors determining metabolic response through complex interactions between the human genome, gut microbiome and dietary interventions make it challenging to address. Integrating this information in machine learning models can assist in picking out correlations in data types that unravel signatures predictive of weight loss at the individual level. As an example, the postprandial glucose response has a high variability between individuals consuming similar meals [140]. Predictive models based on clinical factors and gut microbiota composition have been established to predict and guide personalized dietary interventions for lower postprandial blood glucose response [140–142].

4.2 Study introduction

In this project, weight loss was predicted in non-diabetic, middle-aged Danes with a cardiometabolic risk profile that were enrolled in two dietary randomized cross-over trials. The two trials were carried out by research groups participating in the Gut, Grain and Greens (3G) project (`http://www.3g-center.dk/`). The trials tested the impact of dietary interventions with a whole grain-rich diet or a low-gluten diet for eight weeks compared to a refined grain diet (low whole grain content and high gluten content) separated by a minimum six weeks wash-out period. Diet was consumed *ad libitum* during the intervention periods (Figure 4.1).

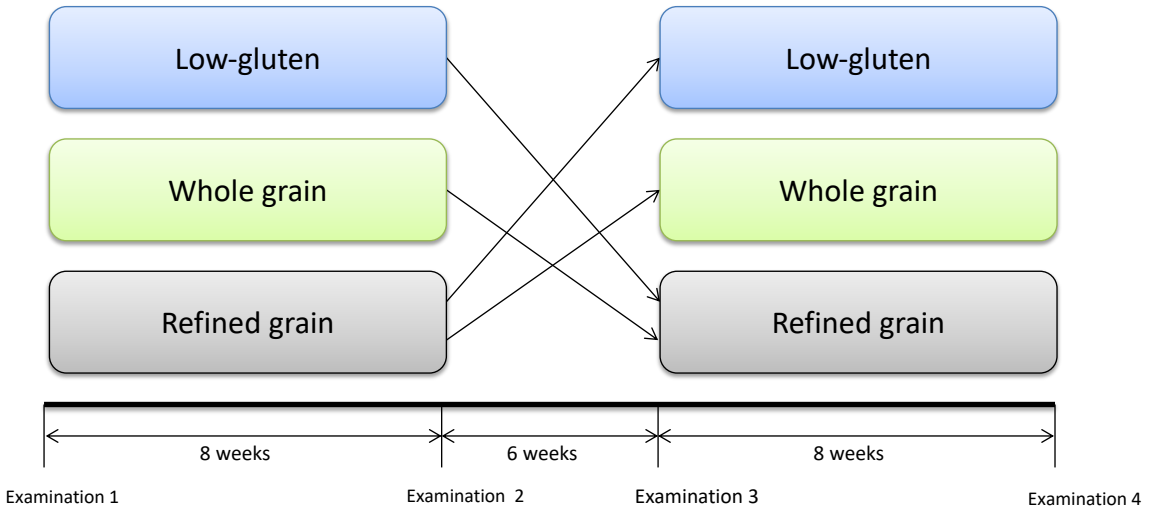


Figure 4.1: The whole grain and gluten study designs. The whole grain study tested a whole grain-rich diet versus a refined grain diet. The gluten study tested a low-gluten diet versus a refined grain diet. The refined grain diet was similar in both trials. `https://clinicaltrials.gov` Whole grain study ID-no: NCT01731366, gluten study ID-no: NCT01719913)

Both studies recruited 60 individuals. In total, 50 and 52 study participants completed the whole grain and gluten trial, respectively. The study participants were deeply phenotyped in the beginning and end of the intervention periods (four examination days in total).

Data was collected on genotype (Infinium CoreExome-24 BeadChip array platform), the gut microbiome (16S rRNA amplicon sequencing and whole metagenomic shotgun sequencing), the urine metabolome (GC-MS and LC-MS), blood pressure and anthropometrics, biochemical measurements from blood samples, gut permeability, gastrointestinal transit time, a self-reported questionnaire of overall well-being and gastrointestinal symptoms, a study diary on consumption of study products, breath hydrogen excretion and a postprandial meal test on glucose, insulin, free fatty acids and GLP-2.

The dietary interventions' impact on the gut microbiota composition as well as biomarkers of metabolic health was previously investigated as the primary and secondary outcomes of both clinical trials. Both studies reported a significant weight loss in response to a whole grain-rich diet or a low-gluten diet compared to a refined grain diet at the population-level using a linear mixed model [60, 138]. Considering the weight loss at the individual level ($N = 102$) following any dietary intervention period with a whole grain-rich diet, a low-gluten diet or a refined grain diet, it was clear that the weight loss was not a universal response across all individuals. The weight changes were also observed at varying magnitude at the individual level.

Thus, in this study, we wished to leverage the rich data collection available in the two clinical trials to predict weight loss using baseline information prior to any dietary intervention to identify predisposition factors for weight loss in combination with a given intervention diet. The refined grain diet was considered an intervention diet rather than a control diet, as it was in the original clinical trials, since the content of whole grain was very low and the content of gluten was very high compared to a Danish habitual diet [143]. We established classification models of weight loss (responders) or weight gain (non-responders) with random forest models. To focus on different aspects of biology for the weight loss prediction, each available datatype was modelled separately to identify the most predictive data types of weight loss. Individual models were guided by feature representations, prior knowledge and feature selection. Following identification of the most predictive data types, different data types were combined. We identified that a feature combination of the type of diet, gut microbiota composition and urine metabolites identified by LC-MS were most predictive of weight loss compared to other data combinations. This learning would be interesting to apply in larger studies to confirm these findings of weight loss predictability. To improve the predictions without increasing the complexity of individual models and allow for missing data across data types, models trained on different types of data were integrated in an ensemble model. The ensemble model was scored only on confident predictions and was finally applied to identify non-responders with a high degree of certainty. Identification of people that are unlikely to benefit from a dietary intervention can be informed early and prioritize other weight loss strategies.

4.3 Bioinformatics challenges: Machine learning with multi-dimensional heterogeneous datasets given limited cohort sizes

The study explores data integration with machine learning for multi-dimensional heterogeneous datasets on limited cohort sizes. Machine learning algorithms' ability to learn

patterns from data and generalize to new datasets generally improve given a higher number of samples being available for training of the model. From a methodological point of view, a key challenge in this study was; How could very heterogeneous data types with a large number of features successfully be integrated into robust prediction models of weight loss given a limited number of samples? We selected a random forest algorithm for modelling as this would not require any scaling or normalization across data types and could assist in feature selection during model development. All models were tested for robustness by performance evaluation on repeated cross-validation and shuffled cross-validation sets. To further assist in model development, we considered solutions for exploiting the study design to increase the number of individual samples, cross-validation, feature engineering, reduction and selection as well as data integration strategies.

4.3.1 Study power

The randomized cross-over study design provided an opportunity to double the number of samples available since prediction of weight loss was done using features available at baseline ($N = 203$, one individual had missing measurement of body weight and was only included in one intervention period). The models thus always included the type of dietary intervention that study participants would receive i.e. a whole grain-rich diet, a low-gluten diet or a refined grain diet to control for the double-inclusion of study participants. In randomized cross-over studies, there is always a risk of carry-over effects from the first intervention, which ideally should be controlled for by a sufficient long wash-out period between the interventions. As the machine learning models always included the type of diet to control for the two baseline measurements used in an individual, we included the order in which the dietary intervention was given as well as the actual length of the wash-out period to correct for potential carry-over effect. However, these features either confused or did not influence the predictive performance and was thus not included in the final models.

4.3.2 Cross-validation

To provide a robust estimate of the performance of each model, the same model initialization across 50 class-stratified shuffle-split five-fold cross-validations was used. This ensured different samples were used for training of the model, which would allow to maximize training of the model from different splits of data. Alternatively, a leave-one-out cross-validation scheme could be used for handling small datasets.

4.3.3 Feature engineering, reduction and selection

The post-prandial meal test response was measured over five timepoints (eight timepoints for breath hydrogen). In the analysis performed in the original trial studies, the time-

series data was represented by the area under the postprandial curve. However, this does not capture information about the dynamic volatility that may be very different despite a similar area under the curve. Thus, we developed an image analysis-inspired algorithm that worked by gridding the postprandial curve to capture information about dynamics in data by different fluctuation calculations (Figure 4.2).

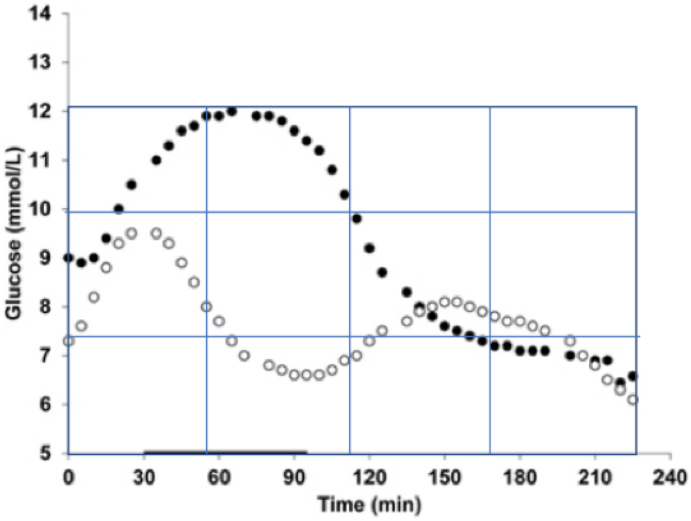


Figure 4.2: Example of the image analysis-inspired algorithm where gridding of images was used to extract information in the postprandial response variables. To capture information about volatility, we summarized the information captured in the vector-grids by three different approaches described further in the manuscript. Data is simulated for this plot.

PRS of SNPs associated with obesity, body weight and sagittal abdominal diameter were developed traits to maximize power of genetic variants. Prior knowledge of genes in the host genome involved in metabolic pathways, inflammation and gut microbiota composition was used to prioritize functional SNP dataset by SNP-to-gene annotations. Similarly, a functional catalogue of butyrate-producing species was identified in literature and used as a feature selection strategy of microbial species. Finally, data-driven feature selection approaches were applied including forward selection strategies testing pair and triplet combinations as seed and ReliefF [144–146].

4.3.4 Data integration

Predictive models of weight loss based on different biological features were integrated in an ensemble model to i) improve the confidence of the predictions, ii) allow for missing baseline data and iii) allow for several biological aspects to count in the final weight loss prediction without increasing the complexity of any individual trained model given a relatively small number of samples. Models were selected for the ensemble across different data types that were more predictive of weight loss compared to a baseline model

only with features on the type of dietary intervention (ROC-AUC > 0.62). The final prediction from the ensemble model was calculated by only averaging very confident individual prediction scores of no weight loss (prediction score ≤ 0.25) or weight loss (prediction score ≥ 0.75) to limit the possibility of false negative or false positive predictions. The final ensemble model was used to evaluate responders and non-responders of weight loss. However, we considered the main application of the prediction model to be identification of non-responders to dietary interventions as this would provide the lowest risk in guidance of personalized weight loss strategies.

4.4 Manuscript

The following manuscript is submitted to *Nature metabolism*. The supplementary material is seen in Appendix A.

Data integration for prediction of weight loss in randomized controlled dietary trials

Rikke Linnemann Nielsen[1,2*], Marianne Helenius[1*], Sara Garcia[1], Henrik M. Roager[3,4], Derya Aytan-Aktug[1,4,] Lea Benedicte Skov Hansen[1], Mads Vendelbo Lind[3], Josef K.Vogt[6], Marlene Danner Dalgaard[1], Martin I. Bahl[4], Cecilia Bang Jensen[1] , Rasa Muktupavela[1], Christina Warinner[5], Vincent Appel[6], Rikke Gøbel[6], Mette Kristensen[3], Hanne Frøkiær[7], Morten H. Sparholt[8], Anders F. Christensen[8], Henrik Vestergaard[6], Torben Hansen[6], Karsten Kristiansen[9], Susanne Brix[10], Thomas Nordahl Petersen[4], Lotte Lauritzen[3**], Tine Rask Licht[4**], Oluf Pedersen[6**], Ramneek Gupta[1**].

* These authors contributed equally

** Corresponding authors

Affiliation

1. Department of Health Technology, Technical University of Denmark, Denmark.
2. Sino-Danish Center for Education and Research, University of Chinese Academy of Sciences, Beijing, China.
3. Department of Nutrition, Exercise and Sports, University of Copenhagen, Denmark.
4. National Food Institute, Technical University of Denmark, Denmark.
5. Department of Anthropology, Harvard University, Cambridge, USA 02138.
6. The Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, DK-2200, Copenhagen, Denmark.
7. Institute for Veterinary and Animal Sciences, University of Copenhagen, Denmark.
8. Department of Radiology, Bispebjerg Hospital, Copenhagen, Denmark.

9. Laboratory of Genomics and Molecular Biomedicine, Department of Biology, University of Copenhagen, DK-2100, Copenhagen, Denmark.
10. Department of Biotechnology and Biomedicine, Technical University of Denmark, Denmark.

Counts

Title length: 86/90.

Abstract: 150/150.

Introduction: 462/500.

Figure and table count: 7/8.

Supplementary material: 6/8.

Word count: 3351/5000 (Result + discussion).

References in main paper: 42/60.

Abstract

Diet is an important component in weight management strategies, but heterogeneous responses to the same diet makes it difficult to foresee individual outcomes. Omics-based technologies now allow for analysis of multiple factors for weight loss prediction at the individual level. Here, we classify weight loss responders (N=106) and non-responders (N=97) of overweight non-diabetic middle-aged Danes over eight weeks in two earlier reported dietary trials by integrating gut microbiome, host genetics and urine metabolome together with measures of physiology and anthropometrics into random forest models. The most predictive models for weight loss included features of diet, gut bacterial species and urine metabolites (ROC-AUC:0.84-0.88) compared to a diet-only model (ROC-AUC:0.62). 64% of the non-responders were identifiable with 80% confidence. A model ensemble omitting microbiome and metabolome profiles but using physiology, host genetics and gastrointestinal transit-time lead to ROC-AUC:0.72. Such models will assist in selecting appropriate weight management strategies, as individual predispositions vary.

There is considerable interest in identifying markers that can predict responsiveness to various weight loss interventions[1]. Weight loss modelling has previously focused on energy intake and expenditure[2,3], macronutrient balance[4], anthropometrics[5], glycemic and insulinemic statuses[6,7] and gut microbiome profiles by the *Prevotella*-to-*Bacteroides* ratio[8].

Multi-omics data has shown promise in improving the understanding of complex phenotypes such as metabolic heath[9,10], which reflect an interplay between physiology, genome and exposome (diet, microbiome, metabolome) of a given individual. At the cohort level, associations to obesity have been found in the human gut microbiome[11], the plasma metabolome[12] and the host genome[13]. Integration of multiple omics has recently been applied for unravelling weight changes in insulin sensitive and insulin resistant individuals[14,15]. Results from these studies show progress towards signatures of weight loss, although inter-individual heterogeneity still leaves a challenge in individual level predictions. In general, computational integration of multi-omics data is challenging due to data heterogeneity, a large number of variables, small sample sizes and missing data[16]. Machine learning methods have shown some progress in this area, especially when coupled with adequate data handling and relevant feature reduction strategies[10,17] and have been applied in prediction of the personal glycemic response to diet[18–20]. Further, ensemble methodologies have demonstrated improved stability on machine learning predictors[16,21].

We previously investigated the impact of eight weeks dietary interventions on human metabolic health outcomes in two Danish randomized cross-over trials with a whole grain-rich diet or low-gluten diet, associated with a beneficial and non-beneficial impact on metabolic health, respectively[22–24] , and identified weight loss as a response to each of the interventions relative to a refined grain diet[25,26]. It has however been argued that no single dietary strategy would be appropriate for all individuals and that certain biomarkers can be important in relation to predisposition for weight loss[7]. This study investigates the use of machine learning to predict which individuals who will experience weight loss during the eight weeks of dietary

interventions with whole grain, low gluten or refined grain. We present random forest-based data integration of anthropometry, blood serum markers, gut microbiome markers, urine metabolomics and host genomics to investigate, if the weight loss response can be predicted based on randomisation onto dietary intervention and biomarkers at baseline prior to any dietary intervention. Models were guided by prior knowledge as well as feature selection and representation strategies to improve predictability with limited cohort sizes (N=102). Performance and robustness were estimated through cross-validation and shuffling cross-validation sets, respectively. By identifying the propensity of study participants likely to experience weight loss, a more effective individual targeting of dietary interventions can be facilitated, eventually in concert with comprehensive population weight loss strategies. Furthermore, understanding predictive features of weight loss response will drive improved understanding of the interplay between gut microbiota, diet and individual predisposition.

Results

Personal weight loss response to whole grain-rich, low-gluten and refined grain diets

In our previous whole grain diet study, participants with a cardiometabolic risk profile were randomized to follow either a whole grain-rich diet or a refined grain diet for eight weeks before cross-over to the other study diet after a six weeks wash-out period[25]. In a similarly designed study, a low-gluten diet was compared to the same refined grain diet[26], which was designed to have high gluten and low whole grain content. An overview of the study design and collection of data of anthropometry markers of physiology, urine metabolites, gastrointestinal transit time, faecal stool samples for gut microbiome analyses and host genomics is shown in Figure 1. In total, 102 individuals completed the whole grain and gluten clinical trials which, across the two intervention baselines allowed 204 samples for modelling. One individual in the gluten trial had missing values of body weight in an intervention which was excluded. Thus, 203 samples were used to model individual body weight response. We investigated weight loss responders (N=106) or non-responders (N=97) by machine learning-modelling of biomarkers measured at the start of the intervention periods. An individual with any degree of weight loss compared to baseline body weight was considered a responder (range: -0.06% to -10.43 %), whereas no change or weight gainers were classified as non-responders independent of the dietary study arm. In the whole grain-rich diet intervention, low-gluten diet intervention or refined grain diet intervention, study participants experienced relative body weight changes ranging between ±5% after eight weeks of dietary changes relative to their baseline body weight. Weight loss responders experienced a relative body weight decrease of on average 1.67%±1.42%, while weight gain non-responders had an average relative body weight increase of 1.39%±1.2% (Figure 2a for the whole grain trial and Figure 2b for the gluten trial).

Predictability of weight loss using diet information alone

Random forest models were modelled to predict weight loss responders and non-responders, where we ensured that the same model initialization across 50 shuffle-split five-fold cross-validations was used (Supplementary Material 1). This machine learning setup was used for all trained models reported in this paper. A baseline performance of the area under the receiver operating characteristic curve (ROC-AUC): 0.62 was established (N=203) using information only about the type of diet. Inclusion of the accurate continuous whole grain intake (2.0−210.28g/day) and gluten intake (3193.85−22961.47mg/day) at baseline as a potential predictor of weight loss together with the type of diet resulted in ROC-AUC: 0.63 (N=201). Thus, the type of dietary intervention and the habitual whole grain and gluten intake is not sufficient to predict weight loss for all individuals during the dietary intervention.

Prior knowledge feature development for machine learning-based data integration

We integrated information about heterogeneous biomarkers of metabolic health into machine learning models in order to understand the individual level omics and physiological predisposition profiles to weight loss. We had in total 288,938 features available for modelling (Table 1); 28 anthropometric and physiological features (Clinical), gastrointestinal transit time (TransitTime), 10,093 16S-based OTUs (16S), 4,808 shotgun sequenced species (mapped to taxa by MGmapper (MGm) or as metagenomic species (MGS)), 1,370 urinary metabolites (analysed by gas-chromatography mass spectrometry (GC-MS) and liquid chromatography mass spectrometry (LC-MS)) and 272,588 single nucleotide polymorphisms (SNPs) from the host genome (LithPath, LithPathLD and GRS).

To guide the machine learning models, we developed features and prioritised biomarkers for modelling by prior knowledge strategies (detailed information is given in Methods). We therefore ended up with eight clinical variables of glucose metabolism, inflammation, gut permeability and anthropometric traits (ClinicalA), 759 SNPs annotated to genes involved in selected metabolic pathways, inflammation and gut microbiome composition identified in

pertinent literature (LithPath and LithPathLD, Supplementary Material 2a), 250 most varying 16S-based OTUs during the dietary interventions (16S_B) and 30 shotgun sequenced faecal microbiome species features (mapped by MGmapper to butyrate-producing species (MGm and MGm_ABC, Supplementary Material 2b). We also considered information from the changes in MGS' in the previous clinical trial studies[25,26], where the changes in the relative abundance of MGS' when on refined grain diet was compared to a whole grain-rich diet or low-gluten diet. For the gluten study, 14 MGS' were found significantly changing in abundance[26], when comparing the changes in abundance for the two dietary interventions. No MGS' changed significantly in the whole grain study[25]. From both studies, the top 14 most significant MGS' were included for modelling of weight loss (topMGS, Supplementary Material 2c). In addition, we developed five genetic risk scores (GRS); three for BMI, one for body weight change and one for sagittal abdominal change after the whole grain intervention compared to the refined grain diet intervention (Supplementary Material 2d). Finally, we modelled features of the longitudinal measurements of the postprandial response including breath hydrogen and plasma free fatty acids, GLP-2, glucose and insulin by an image analysis approach of the postprandial dynamics to capture volatility in addition to the AUC (PostPran).

Weight loss prediction is improved by inclusion of gut microbiome and urinary metabolome features

To identify metabolic profiles predictors of weight loss following a whole grain-rich, low-gluten or refined grain dietary intervention, we tested the predictive performance of each data type separately and ensured that the same study participants were available across multiple data types to allow for comparison of models performance (N=130; 63 non-responders and 67 responders, referred to as comparison models). All comparison models included information of which type of diet the study participants were randomised to receive, since data from both baselines (start of intervention periods) in the cross-over studies were used for modelling of

weight loss responders and non-responders. On this subset, information only about diet type gave a ROC-AUC: 0.61 (Diet in Table 2). By adding features to this model, several models both applying all available data features for a given data type as well as the prior knowledge-developed datasets with and without data-driven feature selection methods were trained (See all combinations for N=130 samples in Supplementary Material 3, selected reported combinations in for N=130 in Table 2). The most predictive models were identified by data-driven selection of microbiome signatures of MGmapper species and top 250 most varying 16S-based OTUs when considering each type of biological information separately (Table 2; Diet.MGm_B with ROC-AUC: 0.82 and Diet.16S_B with ROC-AUC: 0.82, respectively). To explore the predictive signals identified in the intestinal gut microbiome features further, we integrated gut microbiome signatures from 16S-based OTUs from a pool of the top 250 species most varying during the dietary interventions (16S_B) or butyrate-producing species from the MGmapper Bacteria draft database (MGm_B1) along with urine metabolites identified by LC-MS (LC-MS). For the model Diet.16S_B.LC-MS, the type of diet was always included, while we forward selected features from the 16S-based OTUs and the LC-MS together. For the model Diet.MGm_B1.LC-MS, we only forward selected on the urine metabolites identified by LC-MS, while the type of diet and butyrate-producing species from the MGmapper Bacteria draft database always were included. This resulted in the best performing models for weight loss predictions with performance of ROC-AUC: 0.86 and ROC-AUC: 0.90, respectively (Diet.16S_B.LC-MS and Diet.MGm_B1.LC-MS in Table 2).

We ensured that models were not overfitted by performing a permutation analysis on the prediction outcome of weight loss responders or non-responders for a total of 50 times. Training the models on the randomly permuted target confirmed that the two best models performed better than random, since there was only a slight overlap between the test and random ROC-AUC performance distributions, which were significantly different ($p=8.86*10^{-10}$ and $p=2.58*10^{-9}$, Supplementary Material 4).

196

197 **Microbiome and metabolome association to weight loss**

198 After establishing that random forest models including features of the faecal microbiome, urine
199 metabolome and the type of diet (whole grain-rich, low-gluten or refined grain) were most
200 predictive of weight loss, we expanded the random forest models to include all samples that
201 were available for each given data type (N=147–203; 74-97 non-responders and 73–106
202 responders depending on data type, Table 3; all trained models in Supplementary Material 5).
203 The best performing models for weight loss predictions were again Diet.16S_B.LC-MS (ROC-
204 AUC: 0.84, N=169) and Diet.MGm_B1.LC-MS (ROC-AUC: 0.88, N=173).
205 The feature importance, represented by the Gini coefficient in the random forest models, was
206 reported for the selected intestinal microbiome features (16S-based OTUs or butyrate-
207 producing species from MGmapper species) and the urinary metabolomic features in the four
208 best random forest models (two models trained on N=130 for comparison of models, and two
209 models on N=169/173 for models including all available data (Figure 3)).
210 For forward selected features, only features selected in at least 15% of all trained random forest
211 models across the 50 shuffle-split five-fold cross-validations are reported. In all four models,
212 the type of diet was considered important from evaluation of the Gini coefficient (above the
213 red line in Figure 3). The main impact for classification related to whether study participants
214 received a refined grain diet, whole grain-rich or low-gluten diet, seen as the refined grain diet
215 was considered most important by all models (Figure 3). This was expected, as a statistically
216 significant relative weight loss was previously found after consuming the whole grain-rich or
217 the low-gluten diet compared to the refined grain diet in the two clinical trials[25,26]. All types of
218 diet were considered equally in the training of the random forest models. Summary statistics
219 between responders and non-responders as well as putative annotations for most important
220 urinary metabolite features identified by LC-MS and microbiome features are provided in
221 Supplementary Material 6. Metabolites were selected through a data-driven forward selection

approach from a total of 1,285 urinary metabolites identified by LC-MS. In the two models of Diet.16S_B.LC-MS, the taxonomies for the forward selected microbial species were from a pool of the 250 most varying 16S-based OTUs in the intervention periods together with urine metabolites identified by LC-MS. Only the family *Ruminococcaceae* and genus *Streptococcus* were selected in enough models (15%) to be considered important given the number of all selected features (ROC-AUC: 0.86 for N=130 and ROC-AUC: 0.86 for N=169, Table 2 and 3). *Ruminococcaceae* was most abundant in weight loss responders. *Streptococcus* was, by contrast, more abundant in non-responders. The species in the two Diet.MGm_B1.LC-MS models were pre-selected based on literature extracted butyrate-producing species (Supplementary Material 2a) and have been identified as being important for metabolic health, where the species *Faecalibacterium prausnitzii, Eubacterium ramulus* and *Roseburia faecis* are of special importance to the model with a selected range of urine metabolites as seen in Figure 3 (right column) (ROC-AUC: 0.90 for N=130 and ROC-AUC: 0.88 for N=173, Table 2 and 3).

An ensemble of multi-omics models is more robust to varying input data

We explored if the combination of multiple trained weight loss prediction models could improve prediction performance. Further, as given individuals may have missing data, the ensemble approach allows use of all available omics data. The ensemble was made from a selection of the different combinations of potential prediction models performing ROC-AUC > 0.62 (Diet baseline performance for all available models (N=203)). This approach resulted in an ensemble consisting of 334 models across seven different input data combinations. The seven data combinations include diet, forward selected clinical features, SNPs annotated to genes in metabolic pathways, inflammation and gut microbiome composition identified from a literature search, post-prandial response, gastrointestinal transit time, butyrate-producing

247 species, 16S-based OTUs and urinary metabolites identified by LC-MS. All included models
248 are marked in bold in Table 3.
249 The ensemble model was evaluated using different scoring methods, where the highest
250 performing ensemble achieved ROC-AUC: 0.86 by averaging prediction scores from the seven
251 original models, while performances of other ensemble models that include predictions of a
252 sufficiently high confidence ranges in ROC-AUC from 0.69 to 0.84 (Figure 5a). The best
253 ensemble model thus performs very similar to the Diet.16S_B.LC-MS (ROC-AUC: 0.84) and
254 Diet.MGm_B1.LC-MS (ROC-AUC: 0.88) models. Scoring the ensemble model only with
255 highly confident predictions excluding models using the gut microbiome and the urine
256 metabolome, resulted in ROC-AUC: 0.72 (Subset ensemble, Figure 4a). Therefore, clinical
257 information, host genotype, the post-prandial response features and gastrointestinal transit time
258 should be considered important in weight loss predictions as well, if the information of gut
259 microbiome and urinary metabolites is not available.
260 In order to only identify highly confident responders or non-responders, the score threshold
261 that divides the classes was varied for the models included in the ensemble (Figure 4b, 4c).
262 Using t=0.30 as the classification threshold, the ensemble model correctly classified 64% of all
263 individuals who gained or maintained body weight on a given diet (non-responders) where only
264 17% of these were false negative classifications (responders to diet). Conversely, by setting a
265 threshold of t=0.70 for detection of people who will experience a weight loss, the ensemble
266 model predicted 61% of the individuals at a cost of 26% false positive predictions.

Discussion

We investigated if it was possible to predict weight loss responders and non-responders following specific dietary interventions over a period of eight weeks in healthy Danish subjects with a cardio-metabolic risk profile. Both dietary trials previously reported significant weight loss after the intervention with whole grain-rich and low-gluten diets, respectively, compared to a refined grain diet with the use of linear mixed models[25,26]. Despite the overall statistical significance of weight loss, the individuals differed widely, and weight loss was reported on all three dietary interventions (whole grain, low-gluten and refined grain), ranging up to 5% of initial body weight for 98% of the study participants. This is in line with the observation that individuals participating in randomized controlled trials tend to lose weight, independent of the intervention arm, if the study participants have measurements of body weight during the intervention[27]. Baseline body weight may also play a role in this regard.

Random forest models were trained across 50 shuffle-split cross-validated models for robust performance estimation. We found that only including information about the type of diet (whole grain-rich, low-gluten or refined grain diet) lead to a predictive performance of ROC-AUC: 0.62 (N=203). Information on the type of diet is therefore only predictive of weight loss in some individuals. Despite a previously reported correlation between change in body weight and change in energy intake in the whole grain trial[25], the total energy intake did not improve the predictive performance of weight loss together with age, sex and the type of diet (ROC-AUC: 0.57). As other biomarkers may predispose individuals towards a weight loss, we integrated information about host genetics, urine metabolome, physiological measures, postprandial response, whole shotgun gut microbiome sequencing and 16S rRNA amplicon sequencing data measured before the start of the dietary interventions. These were integrated together with diet in random forest models, where rigorous feature selection assisted with heterogenous data integration.

Features of the intestinal microbiome and urine metabolome were the most predictive of weight loss in combination with the type of diet, and boosted performance from a diet-only model from ROC-AUC 0.62 to 0.86–0.90. Several models of weight loss responders and non-responders were trained with different feature combinations, where we evaluated the importance of the features selected in the highest-ranking models (selected in minimum 15% of all models). For the 16S-based OTUs, the family *Ruminococcaceae* and genus *Streptococcus* were the most important features. *Ruminococcaceae* is one of the most abundant firmicutes families in the human gut and metabolizes plant material into short chain fatty acids (SCFA)[28]. *Ruminococcaceae* has been observed in higher abundance in obese Mexican women[29], but also been associated with lower risk of obesity, cardiometabolic diseases[30] and lower BMI[31], and are found at lower abundance in Indian type 2 diabetes patients[32]. Decreased abundance of *Streptococcus* has been found in the development from glucose intolerance to type 2 diabetes[33], and higher abundance in normal weight vs obese Mexican woman[29]. Some species of *Streptococcus* have been associated with weight loss and reduced fat accumulation in mice[34]. The most important MGmapped gut microbiome species towards prediction included *F. prausnitzii*, *E. ramulus* and *R. faecis.* We pre-selected butyrate-producing species as these are associated to metabolic and intestinal health[35], of these, the most important included *F. prausnitzii*, *E. ramulus* and *R. faecis. F. prausnitzii* has been shown to be associated with metabolic health in various studies due to anti-inflammatory properties from butyrate production. It has also been linked to obesity where the abundance is lower than in the healthy metabolic states[36]. *E. ramulus* has been associated with insulin resistance or dyslipidaemia in obese postmenopausal women[37]. In combination with LC-MS characterised urine metabolites, these microbiome features were most predictive and would be recommended for follow-up as potential weight loss predictive signals in future studies. For most of these species, there were no significant difference in the prevalence between responders and non-responders indicating

non-linear combinations of features have been predictive of the weight loss response reflecting the nature of random forest models.

The implementation and use of an ensemble approach with heterogenous models, each requiring a different combination of input features, showed to be more resilient to missing data in the prediction of weight loss responders or non-responders, and had the highest performance of ROC-AUC: 0.86. In this regard, it was notable that omission of information of gut microbiome and urine metabolome features, resulted in a predictive performance of ROC-AUC: 0.72 using host genotype, gastrointestinal transit time and selected physiological features. Further, it allowed combining confident predictions per individual, thus using models that are most suited for that individual. We believe that such artificial intelligence (AI) frameworks can be useful as they integrate complex correlations across heterogenous data and facilitate discovery of signatures that potentially predispose to weight loss following a dietary intervention. AI frameworks may be developed to function as screening tools to assist in comprehensive strategies for weight management. For example, dietary interventions that are unlikely to benefit an individual may be deprioritised in favour of other weight loss strategies. Our AI models are able to identify 64% of the non-responders with 8 out of 10 correctly classified (NPV = 0.83), which is fairly promising.

Limitations of study

While there is some agreement that a 5% to 10% weight loss goal is considered successful long term weight loss, consensus is limited on what constitutes significant short-term weight loss at an individual level[5,38] and more importantly, normal body weight fluctuation is relatively unknown[39,40]. Thus, we recommend future studies of short-term weight loss to collect several longitudinal body weight measurements per individual in order to determine what can be deemed as significant personalised weight loss considering the daily fluctuations of an individual.

The study did not include information on exercise habits or the specific polysaccharide composition for starches and fibers, which are both known to have an impact on weight loss[41,42]. However, study participants were informed not to make life-style changes in the beginning of the clinical trials to avoid changes in exercise habits. Given day-to-day weight fluctuation, a single point cut-off definition of weight loss used in this study may insufficiently capture clinically significant weight loss for every individual. However, it was not possible to restrict the machine learning analysis to the responder extremes (e.g. by upper and lower quantiles) as there was too little data to run the models. The data used for modelling was obtained from clinical cross-over studies in 102 individuals with two baseline time points per individual. Although deeply phenotyped, this is considered a limited number of individuals for effective data integration and machine learning. The urine metabolomics clearly improved performance when coupled with microbial species, but most of the useful features lacked annotation from the metabolomics pipelines.

Finally, our robustness tests go some way in fairly assessing performance but due to limited sample size, the prediction label could not be permuted completely at random, which explains why we do not see a truly random ROC-AUC performance on permuted models. However, the performance of the non-permutated data was significantly higher and had a narrower ROC-AUC distribution for the best models compared to the permuted data. Throughout the various evaluation setups, the best features prevailed, and it was good confirmation to see prior knowledge in microbial species – the butyrate producers – contributing substantially to the highest performing models. Eventually, validation on independent cohorts would be meaningful to gain more confidence in the driving factors of individual weight loss.

Methods

Clinical studies design

The study protocol, randomisation, inclusion and exclusion criteria, randomisation, and study products in the clinical studies that intervened with a whole grain-rich diet (https://clinicaltrials.gov, ID-no: NCT01731366) or a low-gluten diet (https://clinicaltrials.gov, ID-no: NCT01719913) are described previously[22]. Both studies consisted of two 8 week-intervention periods separated by a 6 weeks washout period. The whole grain-rich intervention aimed at an intake of ≥75g whole grain per day and the target for the low-gluten diet was <2g/day of gluten. Both studies used the same refined grain diet as control, which was designed to contain <10g/day of whole grain and >20g gluten/day. The studies recruited a total of 120 healthy Danish men and women (60 subjects for each study), who were generally healthy, but should be overweight (defined by BMI or waist circumference) and have two other risk markers for the metabolic syndrome (high blood pressure, plasma glucose, or triglyceride or low HDL-cholesterol).

For detailed experimental produces and analyses of the collected data, we refer to the methods sections in previously published papers[25,26]. In brief, the study participants attended an examination before and after each intervention periods. The examinations were scheduled in the morning, where study participants were instructed to be fasted ≥10h overnight, to avoid tooth brushing and smoking, and to abstain from alcohol and exercise ≥24h. The participants had a physical examination and fasting blood and urine samples, as well as faecal samples were collected. The physical examination consisted of blood pressure measurements and anthropometrics including measurements of body weight, sagittal abdominal diameter, waist circumference, and body composition by bioelectrical impedance analysis.

The blood samples were analysed for various biomarkers of glucose and lipid metabolism, markers of inflammation and liver health, such as glucose, insulin, cholesterol, triglyceride, IL-6, CRP and alanine aminotransferase and aspartate aminotransferase. The physical examination

consisted of blood pressure measurements and anthropometrics including measurements of body weight, sagittal abdominal diameter, waist circumference, and body composition by bioelectrical impedance analysis. Urine samples, and faecal samples were collected. Gut permeability was assessed by lactulose and mannitol secretion in the urine, while transit time was measured by X-ray after ingestion of 24 radiopaque markers[25,26]. The subjects were also asked to fill out a self-reported questionnaire of overall well-being and gastrointestinal symptoms using a visual analogue scale (VAS) and to keep a study dairy to monitor dietary compliance. In addition, the study participants consumed a standardized breakfast[22] to assess their post-prandial response. The fasting blood sample was used as time 0 and samples were obtained again 30, 60, 120 and 180 minutes after the meal. The samples from all five time points in the time series were analysed with focus on glucose regulation and appetite hormones. Breath hydrogen (H_2) was measured twice at fasting and then seven times with 30-minute intervals after the meal, giving a total of eight time points.

Weight loss responders and non-responders outcome

We focused on identifying weight loss responders and non-responders. We considered body weight (kg) where individuals were grouped for classification by assessing the relative individual change between visit 1 and 2, and between visit 3 and 4 independent of the dietary study arm. The relative change is calculated as $\Delta_w = \frac{w_{after} - w_{before}}{w_{before}}$, where w is the body weight. The responders and non-responders to diet were defined by individuals losing weight or not during the dietary intervention periods, i.e. $\Delta_{weight} \geq 0$ are the non-responders, and $\Delta_{weight} < 0$ are the responders.

In this study, we investigated which data type using baseline biomarkers were predictive of weight loss and what features were most important by the machine learning models. In order to compare this across models, we thus only used data types available in all individuals

(N=130). Secondly, we aimed to train models with all available samples per data type and selected models for a model ensemble to capture individuals at a high certainty for not losing or gaining weight using all biological information from highly confident prediction models.

Gut microbiome

Faecal samples were sequenced by 16S rRNA amplicon sequencing and by shotgun sequencing. Taxonomies were annotated from the 16S data using QIIME2 tool[43] with default quality filtering parameters. The annotation process was completed in three steps: pre-processing, selection of representative sequences and assigning taxonomies. OTU clusters are generated using the Deblur sub-OTU method. As default, all the OTU clusters with abundances less than 2 or 0.005% were removed, then assigned to taxonomies using the SILVA 128 reference database[44]. 10,093 unique OTU clusters passed the quality control. The unique OTU clusters were assigned to different levels of taxonomy from kingdom to species. Relative taxonomy abundances were calculated as the ratio between the obtained taxonomy abundances in a sample and the total taxonomy abundances of the sample.

Shotgun metagenomic sequencing (Illumina paired 2x150nt) was used to generate metagenomic species (MGS) by mapping reads to human gut microbiome reference genes from the integrated gene catalogue (IGC) as previously described[25,26]. In addition, shotgun sequenced Illumina paired-end reads were mapped using MGmapper version 2.7 with five reference databases: Human Microbiome, Meta Hit Assembly, Bacteria, Bacteria Draft, Human and Fungi. Only taxonomical species annotated by the Human microbiome reads, bacteria reads and Bacteria Draft reads were used in the data analyses. The Meta Hit Assembly was skipped, after we found no butyrate-producing species in the catalog. The Human and Fungi catalogues were not used due to too few mapped reads. MGmapper uses similarity-based mapping with the BWA-mem algorithm to find taxonomies in a specified database with pre- and post-processing of the raw reads to lower the number of false positives in the taxonomy

annotation[45]. The mapping was compiled as species relative abundances, which are calculated as $S\_abundance = 100 \cdot \frac{ReadCount}{Size \cdot 2}$ for the paired-end reads, where the *Size* is the length of the reference sequence in base pairs.

Urine metabolomics

Urine samples were analysed by gas chromatography-mass spectrometry (GC-MS) and liquid chromatography–mass spectrometry (LC-MS) (in both positive and negative ionization mode) as previously reported[25,46]. Metabolites measured by LC-MS were putatively annotated using the metabolites' mass, retention time and mode by searching features of interest against the Human Metabolome Database[47] and Metlin Database[48] and annotated at level 3-4 as described by the Metabolomics Standard Initiative[49].

Genotype

DNA was extracted from human blood leucocyte nuclei and were genotyped by Infinium CoreExome-24 BeadChip (Illumina, San Diego, CA). Genotypes were called from Genome studio using the human genome assembly GRCh37 as calling reference. 117 study participants were genotyped for 547,644 single nucleotide polymorphisms (SNPs) after updating genome to build 37. Quality control and genome-wide association study (GWAS) was performed using PLINK1.9[50] for sample and SNP call-rates (98%), sex check, excess heterozygosity and homozygosity, inbreeding, pedigree (relatedness), removal of non-European ancestry, HWE (0.005) and minor allele frequency (MAF, 1%) resulting in 105 samples and 272,588 SNPs.

Genome-wide association study and genetic risk scores

To reduce the feature input space for the 272,588 genetic variants, we utilized three different approaches to prioritise SNPs for modelling of weight loss. First, we performed a literature

study on genes involved in metabolic pathways, inflammation and gut microbiome composition (Supplementary Material 2a). SNPs were annotated to genes using Ensembl variant effect predictor[51] (human build GRch37) resulting in 703 unique SNPs. The selected SNPs were linkage-disequilibrium pruned and binary encoded according to the presence of major and minor alleles. This resulted in 56 SNPs for modelling.

We also performed a GWAS using data from the whole grain trial. A linear GWAS was only done on the weight changes and sagittal abdominal changes $\Delta_{whole\ grain-rich\ diet} - \Delta_{refined\ grain\ diet}$. This phenotype assumes the changes are only caused by the dietary interventions to capture genetic predisposition to changes. If the phenotype for GWAS did not follow a normal distribution, it was converted into a z-score prior to association analysis by linear regression. Age, sex and randomisation order of the intervention treatments were included as co-variates.

In addition, to test the hypothesis that genetic risk variants for obesity, overweight, body weight and sagittal abdominal diameter are important in predisposing individuals to weight loss and maximize the power for genetic information, we developed 5 weighted genetic risk scores (GRS) with 4-10 SNPs in each based on a total of 32 SNPs (Supplementary Material 2d). The SNPs used to create the GRS's were found through a literature study on SNPs involved in metabolic health using the NHGRI-EBI GWAS catalogue[52] with the search keyword "obesity". The GRS were calculated as the sum of the number of minor alleles multiplied by the odds ratio (OR) from GWAS in literature or GWAS conducted on body weight and sagittal abdominal diameter only on the whole grain study cohort.

Postprandial response to standardized meal test

The postprandial response to a standardized test meal was measured using four biochemical markers from blood samples being; free fatty acids, GLP-2, glucose and insulin and in breath hydrogen. After a 10-hour fasting period, the first blood was sampled and samples were

obtained again at 30, 60, 120 and 180 minutes after a standardized breakfast, for a total of five time points in the time series. Breath hydrogen (H_2) was measured twice at fasting and averaged and then seven times every 30 minutes following the meal, giving a total of eight time points. Typically, the postprandial response is represented by area under the postprandial curve, which does however not capture information about temporal variation in data. We thus modelled the dynamics of the postprandial response by three new feature representations of volatility. First, the fluctuation was measured as movement of the values between time points by differencing the time points consecutively and summing up their absolute differences using following formula:

$$fluc1 = \frac{\sum_{i=2}^{len(x)} abs(x_i - x_{i-1})}{len(x)}$$

For this calculation, we used the normalized time series and we called the outcome measure fluc*1*. Secondly, we plotted the non-normalized time series of the measures for each patient and analysed the fluctuation by its graph. In the analysis of the graph, interpolation of the series was used to add more time points. If the series had one missing value the interpolation would be used to replace the missing value; if there were more than one missing value, the representations would be set as 0. The interpolation function interp1d from the SciPy (version 1.2.1) Python package was used with 100 interpolated points and a spline interpolation (kind="cubic"). Each graph was then divided into a grid of size 10x10 and 50x50 all with the same axis scale based on the maximum and minimum value in the data of the postprandial marker. From the grid division the 100 interpolated time points were interpreted into an image vector consisting of 1s and 0s for squares with and without points, respectively. This was done vertically with each column being interpreted from the lower boundary to the upper and concatenated into one vector. The image vector was summed, and we called the outcome

measure *fluc2*. The third measure we got from filtering the sum, thus it would only allow for addition between squares with two or more consecutive 1s. This measure we called *fluc3*.

Data integration and machine learning strategies

Data collected from before the two interventions for all participants was used to explore predictive biomarkers of weight loss in machine learning models. The type of diet (whole grain-rich, low-gluten or refined grain diet) was included in all models to differentiate between the two baselines for the same individual. We only included study completers for modelling. Random forest classifiers modelled the samples using 50 shuffle-split five-fold cross-validation stratified by target classes. The models were made in Python (version 3.7.1) using the package Scikit-learn (version 0.20.1) for machine learning methods, especially the class RandomForestClassifier(), which was used for modelling. The number of decision trees in the forest was fixed at n_estimators=50 and model initialization was fixed at random_state=42 for reproducibility. The number of features considered at each split was set as max_features=None, meaning that the random forest could use all features for a split. To limit the forest growing into overfit the minimum decrease in impurity required to make a split was set to min_impurity_decrease=0.01, meaning that there had to be at least 1% decrease in impurity[53].

Model and feature importance evaluation

The predictive performance of the machine learning models was assessed as the area under the Receiver Operating Characteristic (ROC) curve (ROC-AUC). The ROC is a graph showing the true positive rate (TPR) against the false positive rate (FPR), when the threshold is varied for labelling a data point as either positive or negative in a binary classifier. In addition, we reported sensitivity, specificity, and Matthews Correlation Coefficient (MCC). The importance of features in the random forest models was evaluated using the Gini index[54]. The Gini index feature importance is part of the random forest algorithm, which evaluated how many times a

given feature was involved in a node split. This will be shown as the Mean Decrease in Impurity (MDI), i.e. how much a variable on average contributes to the decrease in node impurity. The averaged importance of a feature, if all are assigned similar importance, should be $imp = \frac{1}{M}$, where M is the number of features in a model, since the feature importance sum to 1.

Feature selection

For data types and combinations with higher dimensionality, additional means of feature selection were applied, which used both prior knowledge and data-driven approaches to lower the feature space when optimizing the models to avoid overfitting.

Prior knowledge feature selection

We prioritized features in the microbiome data from 16S taxonomies, MGmapped species, MGS' and in the genotype data. The prioritization and representation of genotype data is described previously in the Methods section "*Genome-wide association study and genetic risk scores*". For the microbiome data from 16S taxonomies, we assessed the prevalence and variance between visits and used this to select the top 10 and top 250 16S-based OTUs for the machine learning models. 16S-based OTUs present in at least 5 people were considered. The prevalence and variance were used as a proxy for selecting taxa with high information.

For the MGmapped gut microbiome species, we prioritized 17 butyrate-producing species identified in previous studies and which were available in the MGmapper datasets mapped against the catalogues Bacteria, Bacteria draft and Human Microbiome (Supplementary Material 2b). Of the 17 unique butyrate-producing species, the Bacteria catalogue had nine microbial species, the Bacteria draft catalogue had 11 microbial species and the Human Microbiome catalogue had 10 microbial species. *Anaerostipes caccae* was removed from analysis, since it was found to have an abundance of 0 in all study participants at the two baselines in the Bacteria draft catalog.

For the MGS', we prioritized the top altered species from the whole grain and gluten studies[25,26]. For the gluten study, 14 MGS' were significantly altered when comparing the changes to abundance on refined grain diet and low-gluten diet. In the whole grain study, no MGS' were significantly altered when comparing the changes to abundance on refined grain diet and whole grain-rich diet. The top 14 most altered MGS' were therefore selected from the whole grain study as well. This resulted in 28 pre-selected MGS'.

Data-driven feature selection

In addition, we did an exhaustive feature selection on the metabolome data. An exhaustive search was done with all possible subset pairs or triplets of metabolites in order to assess if any subset could improve predictive power. For that, the random forests were run with each subset and ROC-AUC performances compared.

Forward selection was also applied to many combinations of different data sets. This selection was performed by adding one feature at a time and then check which increased ROC-AUC. When multiple equally good features were found, all are first added to see if this performs better. If the new model is not better, the function will randomly select one of the equally good features. However, the other features are still in the pool and can be selected at a later iteration. The worst performing features were gradually removed at each iteration in order to save computation time. This continued until performance no longer increased, and the optimal model was saved. The feature selection has been made with a set of parameters which include a list of features to select from (selects), the maximum number of features to select (max_features), the initial fraction of features to remove at each iteration (frac) and the step size of removing features (step), which is updated after a feature is added. The list of features to select from depended on which data sets where included, and the features not shown in this list where added before the first iteration. The maximum number of features to select was set to max_features=8 or unlimited (giving ~5-15 selected features). The initial fraction for removing the lowest performing features was frac=0.4, meaning that the 40% worst performing

features are removed in the first iteration. The step size was set to step=0.1, thereby the number of removed features was 10% less at each iteration. Once the fraction became less than 0.1, the step size is changed to 0.01 automatically, and when this fraction results in 0, a single feature will hence forth be removed at each iteration.

Statistical analyses

Statistical testing of distributions (responder and non-responders and permutation analysis of the ROC-AUC distributions) were assessed by a two sampled unpaired t-test if data followed a Gaussian distribution or Mann Whitney test if non-Gaussian distribution. A p-value < 0.05 was considered significant.

Personalised artificial intelligence ensembles

The data types were combined into different sets in order to determine how they might capture different aspects of the data, which were reported by an ensemble model. The ensemble model is built based on the prediction scores from multiple models (50 shuffle-split five-fold cross-validation models). We created different ensemble models by different confidence predictive thresholds {t= [$\leq$0.30, $\geq$0.70], [$\leq$0.25, $\geq$0.75], [$\leq$0.20, $\geq$0.80]} and by four scoring methods for which each sample is evaluated across all states based on prediction scores. This yields a final set of scores or predictions per sample, for which the ensemble performance can be evaluated. The scoring methods are:

1) *Mean of scores:* The mean of prediction scores.
2) *Majority voting:* The prediction score for each model is rounded to either 0 or 1, for the classes non-responder and responder, respectively. The predicted class chosen by most models will be the ensemble prediction.
3) *Confident mean of scores:* The mean of the prediction scores that is considered "confident", based on a set threshold. If this was set to e.g. 0.7, then all samples with

prediction score equal to/below 1-0.7 = 0.3 or equal to/above 0.7 would be considered confident scores to be included in the mean score. If a sample has no confident scores, it is excluded from the performance calculation for the ensemble.

4) *Majority voting on confident scores:* A mixture of 2) and 3). The predictions that are considered "confident" based on a threshold are rounded to either 0 or 1 for the non-responder and responder classes. The predicted class is the one chosen by most models in the ensemble.

Prediction of individuals at high confidence of weight changes

The predictions made with these models are the probabilities of a sample belonging to either class 0 (non-responders) or class 1 (responders). The class probabilities for each tree are estimated as the fraction of samples belonging to the same class in each leaf of the tree. As the random forest consists of multiple decision trees, the class probabilities are predicted as a mean of the predicted class probabilities for each tree in the forest. By thresholding the prediction probabilities, we can at a given probability define the number of participants that we are sure will or will not experience weight loss. To evaluate this, we reported the sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV).

640

641 **Data availability**

642 The raw Illumina read data for all whole grain study samples have been deposited to the
643 Short Read Archive database with the accession number PRJNA395744.
644 The raw Illumina read data for all gluten study samples have been deposited to the Short
645 Read Archive the accession number PRJNA491335.

646

647

648 **References**

649 1. John, G. K. *et al.* Dietary alteration of the gut microbiome and its impact on weight
650 and fat mass: A systematic review and meta-analysis. *Genes (Basel)* **9**, E167 (2018).
651 2. Thomas, D. M., Gonzalez, M. C., Pereira, A. Z., Redman, L. M. & Heymsfield, S. B.
652 Time to correctly predict the amount of weight loss with dieting. *J. Acad. Nutr. Diet.*
653 **114**, 857–861 (2014).
654 3. Thomas, D. M. *et al.* A Simple Model Predicting Individual Weight Change in
655 Humans. *J Biol Dyn.* **5**, 579–599 (2011).
656 4. Chow, C. C. & Hall, K. D. The dynamics of human body weight change. *PLoS*
657 *Comput. Biol.* **4**, e1000045 (2008).
658 5. Finkler, E., Heymsfield, S. B. & St-Onge, M.-P. Rate of weight loss can be predicted
659 by patient characteristics and intervention strategies. *J Acad Nutr Diet.* **112**, 75–80
660 (2012).
661 6. Ritz, C., Astrup, A., Larsen, T. M. & Hjorth, M. F. Weight loss at your fingertips:
662 personalized nutrition with fasting glucose and insulin using a novel statistical
663 approach. *Eur. J. Clin. Nutr.* **73**, 1529–1535 (2019).
664 7. Hjorth, M. F., Zohar, Y., Hill, J. O. & Astrup, A. Personalized Dietary Management of
665 Overweight and Obesity Based on Measures of Insulin and Glucose. *Annu. Rev. Nutr.*

38, 245–272 (2018).

8. Hjorth, M. F. *et al.* Prevotella-to-Bacteroides ratio predicts body weight and fat loss success on 24-week diets varying in macronutrient composition and dietary fiber: results from a post-hoc analysis. *Int. J. Obes.* **43**, 149–157 (2019).

9. Zhou, W. *et al.* Longitudinal multi-omics of host–microbe dynamics in prediabetes. *Nature* **569**, 663–671 (2019).

10. Schüssler-Fiorenza Rose, S. M. *et al.* A longitudinal big data approach for precision health. *Nat. Med.* **25**, 792–804 (2019).

11. Le Chatelier, E. *et al.* Richness of human gut microbiome correlates with metabolic markers. *Nature* **500**, 541–546 (2013).

12. Cirulli, E. T. *et al.* Profound Perturbation of the Metabolome in Obesity Is Associated with Health Risk. *Cell Metab.* **29**, 488-500.e2 (2019).

13. Goodarzi, M. O. Genetics of obesity: what genetic association studies have taught us about the biology of obesity and its complications. *Lancet Diabetes Endocrinol.* **6**, 223–236 (2018).

14. Dao, M. C. *et al.* A data integration multi-omics approach to study calorie restriction-induced changes in insulin sensitivity. *Front. Physiol.* **9**, (2019).

15. Piening, B. D. *et al.* Integrative Personal Omics Profiles during Periods of Weight Gain and Loss. *Cell Syst.* **6**, 157–170 (2018).

16. Graim, K. *et al.* PLATYPUS: A Multiple–View Learning Predictive Framework for Cancer Drug Sensitivity Prediction. *Pac Symp Biocomput.* **24**, 136–147 (2019).

17. Wilmanski, T. *et al.* Blood metabolome predicts gut microbiome α-diversity in humans. *Nat. Biotechnol.* **37**, 1217–1228 (2019).

18. Popp, C. J. *et al.* The rationale and design of the personal diet study, a randomized clinical trial evaluating a personalized approach to weight loss in individuals with pre-diabetes and early-stage type 2 diabetes. *Contemp. Clin. Trials* **79**, 80–88 (2019).

19. Mendes-Soares, H. *et al.* Assessment of a Personalized Approach to Predicting Postprandial Glycemic Responses to Food Among Individuals Without Diabetes. *JAMA Netw. open* **2**, e188102 (2019).

20. Zeevi, D. *et al.* Personalized Nutrition by Prediction of Glycemic Responses. *Cell* **163**, 1079–1094 (2015).

21. Alghamdi, M. *et al.* Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford ExercIse Testing (FIT) project. *PLoS One* **12**, 1–15 (2017).

22. Ibrügger, S. *et al.* Two randomized cross-over trials assessing the impact of dietary gluten or wholegrain on the gut microbiome and host metabolic health Clinical Trials. *J. Clin. Trials* **4**, (2014).

23. Ye, E. Q., Chacko, S. A., Chou, E. L., Kugizaki, M. & Liu, S. Greater Whole-Grain Intake Is Associated with Lower Risk of Type 2 Diabetes, Cardiovascular Disease, and weight gain. *J Nutr* **142**, 1304–13 (2012).

24. Sapone, A. *et al.* Spectrum of gluten-related disorders: consensus on new nomenclature and classification. *BMC Med.* **10**, (2012).

25. Roager, H. M. *et al.* Whole grain-rich diet reduces body weight and systemic low-grade inflammation without inducing major changes of the gut microbiome: a randomised cross-over trial. *Gut* **68**, 83–93 (2019).

26. Skov, L. B. *et al.* A low-gluten diet induces changes in the intestinal microbiome of healthy Danish adults. *Nat. Commun.* **9**, 4630 (2019).

27. Johns, D. J., Hartmann-Boyce, J., Jebb, S. A. & Aveyard, P. Weight change among people randomized to minimal intervention control groups in weight loss trials. *Obesity (Silver Spring).* **24**, 772–780 (2016).

28. Biddle, A., Stewart, L., Blanchard, J. & Leschine, S. Untangling the genetic basis of fibrolytic specialization by lachnospiraceae and ruminococcaceae in diverse gut

communities. *Diversity* **5**, 627–640 (2013).

29. Chávez-Carbajal, A. *et al.* Gut microbiota and predicted metabolic pathways in a sample of Mexican women affected by obesity and obesity plus metabolic syndrome. *Int. J. Mol. Sci.* **20**, 1–18 (2019).

30. de la Cuesta-Zuluaga, J. *et al.* Gut microbiota is associated with obesity and cardiometabolic disease in a population in the midst of Westernization. *Sci. Rep.* **8**, 1–14 (2018).

31. Schwiertz, A. *et al.* Microbiota and SCFA in lean and overweight healthy subjects. *Obesity* **18**, 190–195 (2010).

32. Bhute, S. S. *et al.* Gut microbial diversity assessment of Indian type-2-diabetics reveals alterations in eubacteria, archaea, and eukaryotes. *Front. Microbiol.* **8**, 1–15 (2017).

33. Zhang, X. *et al.* Human Gut Microbiota Changes Reveal the Progression of Glucose Intolerance. *PLoS One* **8**, (2013).

34. Yoda, K. *et al.* A combination of probiotics and whey proteins enhances anti-obesity effects of calcium and dairy products during nutritional energy restriction in aP2-agouti transgenic mice. *Br. J. Nutr.* **113**, 1689–1696 (2015).

35. Nielson T. Baxter, Alexander W. Schmidt, Arvind Venkataraman, Kwi S. Kim, Clive Waldron, T. M. S. Dynamics of human gut microbiota and short-chain fatty acids in response to dietary interventions with three fermentable fibers. *Host-Microbe Biol.* **10**, 1–13 (2018).

36. Martín, R. *et al.* Functional characterization of novel Faecalibacterium prausnitzii strains isolated from healthy volunteers: A step forward in the use of F. prausnitzii as a next-generation probiotic. *Front. Microbiol.* **8**, 1–13 (2017).

37. Brahe, L. K. *et al.* Specific gut microbiota features and metabolic markers in postmenopausal women with obesity. *Nutr. Diabetes* **5**, e159-7 (2015).

38. Knell, G., Li, Q., Pettee Gabriel, K. & Shuval, K. Long-Term Weight Loss and

744 Metabolic Health in Adults Concerned With Maintaining or Losing Weight: Findings
745 From NHANES. *Mayo Clin. Proc.* **93**, 1611–1616 (2018).
746 39. Foreyt, J. P. *et al.* Psychological correlates of weight fluctuation. *Int. J. Eat. Disord.*
747 **17**, 263–275 (1995).
748 40. Adami, G. F., Campostano, A., Bessarione, D., Gandolfo, P. & Scopinaro, N. Weight
749 fluctuation due to reducing diet, resting energy expenditure and body composition in
750 obese patients. *Diabetes, Nutr. Metab. - Clin. Exp.* **9**, 18–21 (1996).
751 41. Dreher, M. L. Role of Fiber and Healthy Dietary Patterns in Body Weight Regulation
752 and Weight Loss. *Adv. Obesity, Weight Manag. Control* **3**, 244–255 (2015).
753 42. Thorogood, A. *et al.* Isolated aerobic exercise and weight loss: A systematic review
754 and meta-analysis of randomized controlled trials. *Am. J. Med.* **124**, 747–755 (2011).
755 43. Bolyen, E. *et al.* QIIME 2: Reproducible, interactive, scalable, and extensible
756 microbiome data science. *PeerJ Prepr.* (2018). doi:10.7287/peerj.preprints.27295
757 44. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: Improved data
758 processing and web-based tools. *Nucleic Acids Res.* **41**, 590–596 (2012).
759 45. Petersen, T. N. *et al.* MGmapper: Reference based mapping and taxonomy annotation
760 of metagenomics sequence reads. *PLoS One* **12**, e0176469 (2017).
761 46. Roager, H. M. *et al.* Colonic transit time is related to bacterial metabolism and
762 mucosal turnover in the gut. *Nat. Microbiol.* **1**, 16093 (2016).
763 47. Wishart, D. S. *et al.* HMDB 4.0: The human metabolome database for 2018. *Nucleic*
764 *Acids Res.* **46**, D608–D617 (2018).
765 48. Guijas, C. *et al.* METLIN: A Technology Platform for Identifying Knowns and
766 Unknowns. *Anal. Chem.* **90**, 3156–3164 (2018).
767 49. Sumner, L. W. *et al.* Proposed minimum reporting standards for chemical analysis
768 Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative
769 (MSI). *Metabolomics* **3**, 211–221 (2007).

50. Chang, C. C. *et al.* Second-generation PLINK: Rising to the challenge of larger and richer datasets. *Gigascience* **4**, 1–16 (2015).

51. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 1–14 (2016).

52. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).

53. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

54. Hastie, T., Tibshirani, R., & Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.).* (Springer, 2009).

Acknowledgments

The authors would like to thank all study participants involved in the study.

Author contributions

Conception or design of the work: RLN, JKV, MK, HF, HV, TH, TRL, LL, OP, RG

Acquisition of data: HMR, MVL, MDD, RG, MHS, AFC

Analysis of the data: RLN, MH, SG, HMR, DAA, LBSH, RM

Interpretation of the data: RLN, MH, SB, KK, RG, MIB, KK, SB

Creation of new software used in the work: SG, CBJ, RG, VA, CW, TNP

Drafted the manuscript: RLN, MH, RG

Revision of manuscript: All authors.

Competing interests

The authors declare no competing interests.

Figures

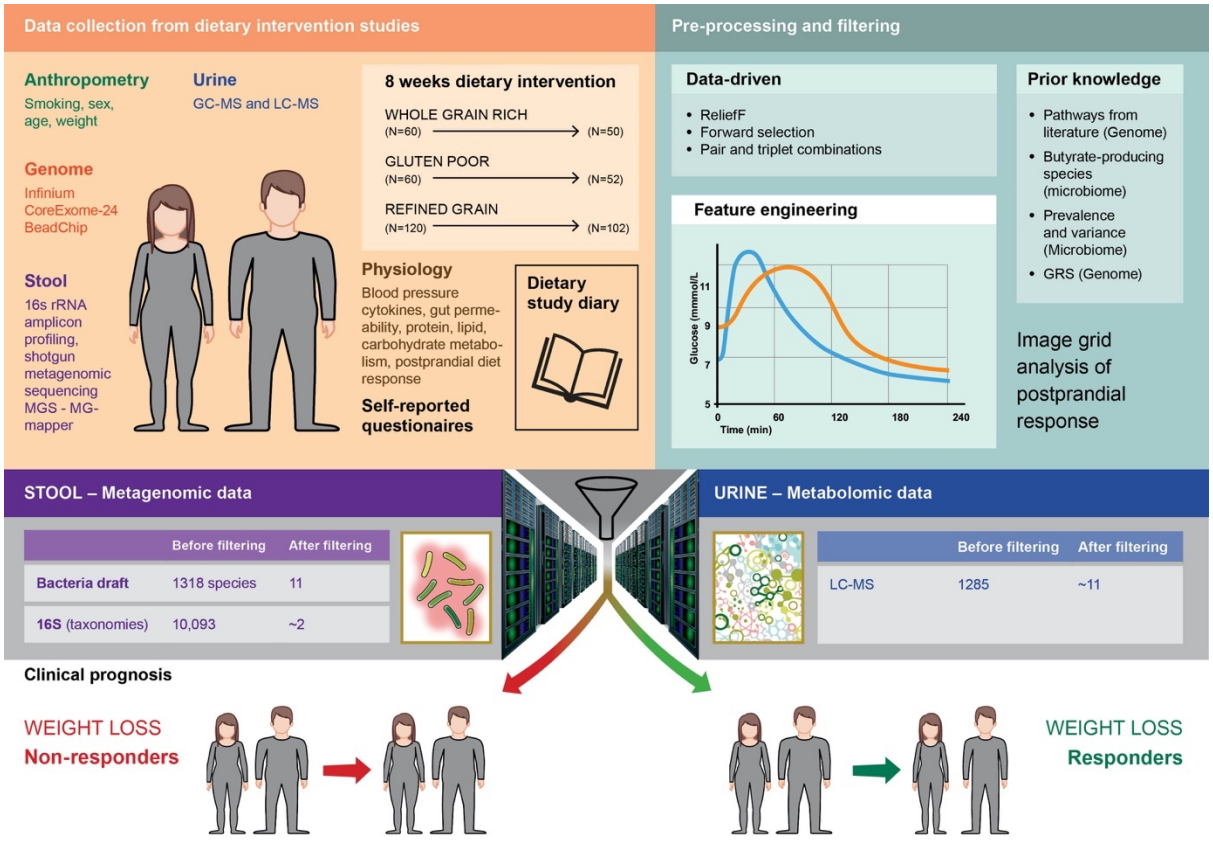


Figure 1: **Study design including data availability, feature development and selection, best features selected for model and clinical prognosis of weight loss**. Participants achieving any weight loss was considered as a responder. Different combinations of features were selected for modelling e.g. included the faecal stool samples by butyrate-producing species from MGmapper catalog Bacteria draft and by forward selected 16S taxonomies selected from a pool of the top 250 most varying. These were combined with forward selected urine metabolites identified by LC-MS.

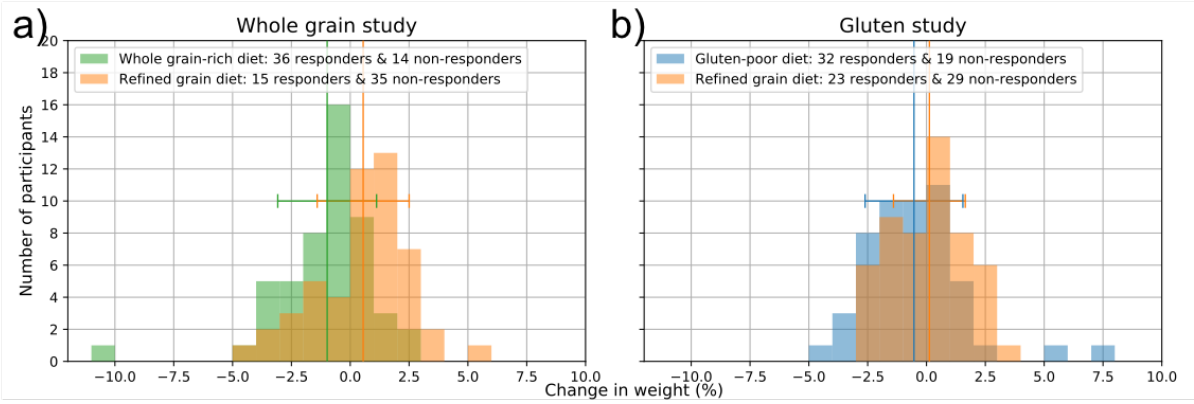


Figure 2: ***Weight changes in the two dietary intervention arms during eight weeks*** a) Distribution of percentage changes in body weight for the whole grain study. **b)** Distribution of percentage changes in body weight for the gluten study. The coloured lines denote mean and standard deviations for the diet groups (green=whole grain-rich diet, orange=refined grain diet, blue=low-gluten diet).

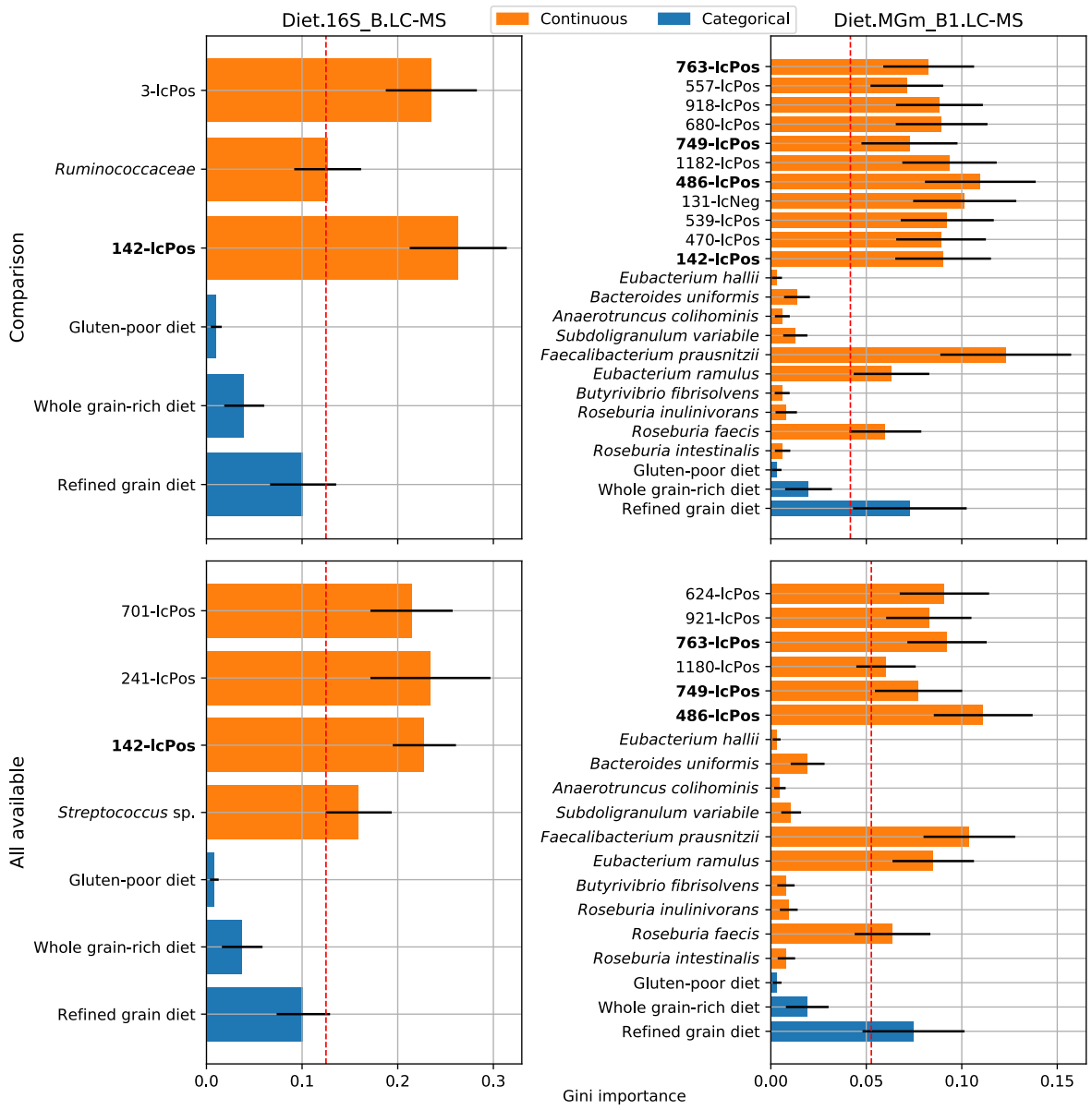
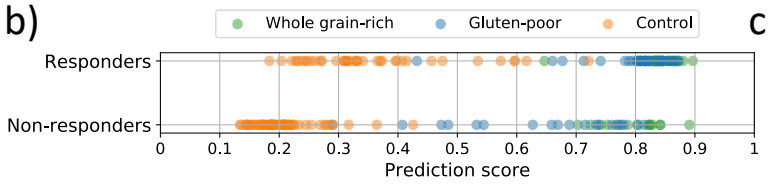


Figure 3: **Gini feature importance for the models**. Models have data combinations of the type of diet, forward selected 16S-based OTUs from a pool of the top 250 most varying (left, Diet.16S_B.LC-MS) or butyrate-producing species (right, Diet.MGm_B1.LC-MS) and forward selected urine metabolites identified by LC-MS for features selected minimum 15% across all trained models. The columns represent the two data combinations, and the rows represent the runs on 130 common samples (Comparison) as well as runs on all samples available for the data combination (All available). The red line marks features of highest importance given the relative Gini coefficient.

a)

Scoring method	Confidence	ROC-AUC	Sensitivity	Specificity	MCC
Mean	None	0.86	0.73	0.70	0.43
Majority voting	None	0.73	0.76	0.69	0.46
Mean of confident scores	t ≤ 0.30 & 0.70 ≤ t	0.84	0.69	0.70	0.39
Majority voting on confident scores	t ≤ 0.30 & 0.70 ≤ t	0.70	0.70	0.70	0.40
Mean of confident scores	t ≤ 0.25 & 0.75 ≤ t	0.83	0.69	0.70	0.39
Majority voting on confident scores	t ≤ 0.25 & 0.75 ≤ t	0.69	0.69	0.70	0.39
Mean of confident scores	t ≤ 0.20 & 0.80 ≤ t	0.82	0.69	0.71	0.40
Majority voting on confident scores	t ≤ 0.20 & 0.80 ≤ t	0.70	0.69	0.71	0.40
Subset ensemble: Mean of confident scores	t ≤ 0.25 & 0.75 ≤ t	0.72	0.69	0.66	0.35

b)



c)

Score	Sensitivity	PPV	Specificity	NPV
0.1	1.0	0.52	0.0	nan
0.2	0.99	0.63	0.37	0.97
0.3	0.88	0.73	0.64	0.83
0.4	0.74	0.7	0.66	0.7
0.5	0.69	0.72	0.7	0.67
0.6	0.65	0.72	0.72	0.65
0.7	0.61	0.74	0.76	0.64
0.8	0.56	0.86	0.9	0.65
0.9	0.0	nan	1.0	0.48

Figure 4: **Ensemble of weight loss models a)** Performances based on four scoring schemes and different classification thresholds for predictive models included in different personalised ensemble models. **b)** The prediction score of each sample are plotted against their true class with colours representing the type of dietary intervention. The scores shown are from ensemble scoring method mean of confident scores (t ≤ 0.25 & 0.75 ≤ t). **c)** The sensitivity, positive predictive value (PPV), specificity and negative predictive value (NPV) are calculated at different score thresholds to separate the classes for the ensemble model shown in b). Abbreviation: MCC: Matthews correlation coefficient.

Tables

Table 1: **Overview of feature selection for random forest models.**

Data type	**Data label**	**Number of features before filtering**	**Number of features after prior know-ledge filtering**	**Was data-driven feature selection applied (Y/N)**
Diet: Binary features that represent if the consumed diet is whole grain-rich, low-gluten or refined grain.	Diet	3	-	N
Anthropometrics and physiological	ClinicalA	28	8 (age, sex, BMI and blood CRP, IL-6, HbA1c, HOMA-IR and zonulin)	N
	ClinicalB		-	Y
Whole grain and gluten intake	ContinuousIntake	2	-	N
Gastrointestinal transit time	TransitTime	1	-	N
Self-reported	VAS	16	-	N
Postprandial response	PostPran	5	-	N
Genome data		272,588 SNPs (after QC)		
Literature pathways	LitPath		703 SNPs	Y
LD pruned literature pathways	LitPathLD		56 SNPs	Y

Genetic risk scores	GRS		5 GRS's of 32 SNPs	N
Metagenomic data: 16S (taxonomies)		10093		
Top 10 most variating	16S_A		10	N
Top 250 most variating	16S_B		250	Y
Prevalence	16S_C		3,321	Y
Metagenomic data: MGmapper (species)				
Bacteria catalogue	MGm_A	464	9	N
Bacteria draft catalogue	MGm_B	1318	-	Y
	MGm_B1		11	N
HumanMicrobiome catalogue	MGm_C	444	10	N
Butyrate-producing species from MGmapper catalogues	MGm	-	30	N
Metagenomic data: Metagenomic species	MGS	1264	-	Y
Top 14 from whole grain and gluten studies	topMGS		28	N
Metabolic data				
GC-MS	GC-MS	85	-	Y
LC-MS	LC-MS	1285	-	Y

Table 2: **Model performances for models run on a set of 130 samples present in all below data combinations**. This is reported as mean of five cross-validations repeated 50 times with random shuffles of the cross-validation splits. The blue-red colorbar is for area under the receiver operating characteristic curve (ROC-AUC), sensitivity and specificity, while the blue.yellow-red colorbar is for Matthews correlation coefficient (MCC). Diet represents the dataset consisting of the three features indicating which diet was consumed. EnergyIntake is the energy intake at baseline, while ContinuousIntake is the total intake of whole grain (g/day) and gluten (mg/day) at baseline. ClinicalA and B are both feature subsets selected by prior knowledge and forward selection, respectively, from the set of 28 anthropometric and physiological features. LithPathLD and GRS are subsets of genetic variants selected by prior knowledge, where LithPathLD also was subject to forward selection. 16S_B is the set of forward selected 16S-based OTUs selected from a pool of the top 250 most varying features. MGm_B and MGm_B1 are subsets of species mapped by MGmapper to the Bacteria draft catalogue, which are selected by forward selection and prior knowledge as butyrate-producing species, respectively. LC-MS[45-lcPos_142-lcPos] holds a pair of urine metabolites identified by LC-MS. PostPranFluc3_50 is the post prandial response features free fatty acids, GLP-2, glucose and insulin, which are represented by the third image analysis method with a grid-size of 50x50 (See Methods). Abbreviations for model combinations are explained in Table 1 and in the main text. Performances of all models run on the 130 samples are in Supplementary Material S.3.

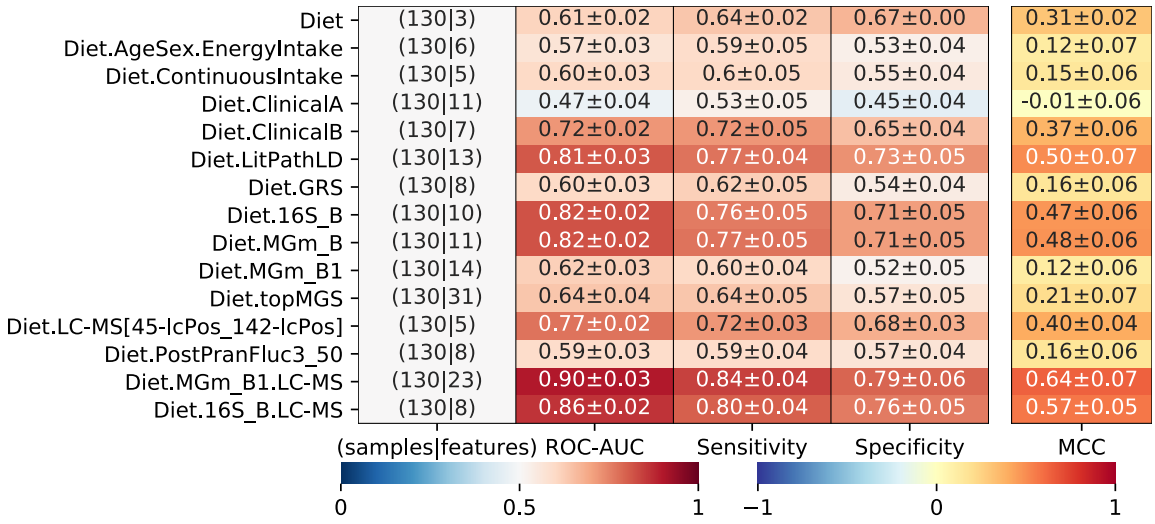
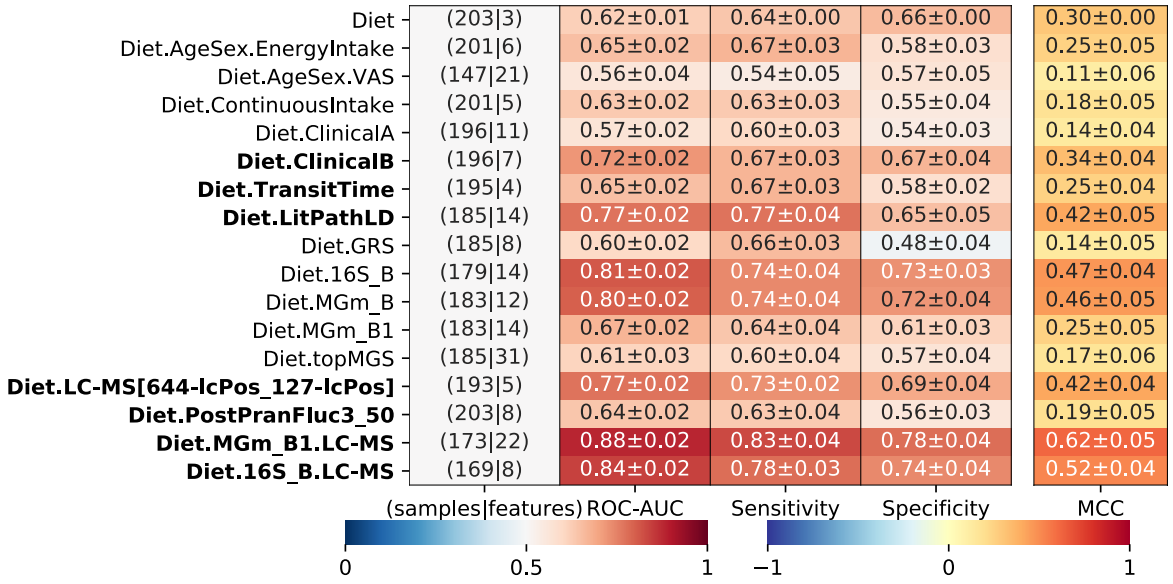


Table 3: **Model performances for models run on all samples available for a given data combination**. This is reported as mean of five cross-validations repeated 50 times with random shuffles of the cross-validation splits. Models in **bold** were included in an ensemble (ROC-AUC > 0.62). The blue-red colorbar is for area under the receiver operating characteristic curve (ROC-AUC), sensitivity and specificity, while the blue.yellow-red colorbar is for Matthews correlation coefficient (MCC). Diet represents the dataset consisting of the three features indicating which diet was consumed. EnergyIntake is the energy intake at baseline, while ContinuousIntake is the total intake of whole grain (g/day) and gluten (mg/day) at baseline. VAS represents the self-reported features measured by Visual Analogue Scale. ClinicalA and ClinicalB are both feature subsets

selected by prior knowledge and forward selection, respectively, from the set of 28 anthropometric and physiological features. TransitTime is the baseline transit time. LithPathLD and GRS are subsets of genetic variants selected by prior knowledge, where LithPathLD also was subject to forward selection. 16S_B is the set of forward selected 16S-based OTUs selected from a pool of the top 250 most varying features. MGm_B and MGm_B1 are subsets of species mapped by MGmapper to the Bacteria draft catalogue, which are selected by forward selection and prior knowledge as butyrate-producing species, respectively. topMGS is the top 28 selected MGSs from the whole grain and gluten studies. LC-MS[45-lcPos_142-lcPos] holds a pair of urine metabolites identified by LC-MS. PostPranFluc3_50 is the post prandial response features free fatty acids, GLP-2, glucose and insulin, which are represented by the third image analysis method with a grid-size of 50x50 (See Methods). Abbreviations for model combinations is explained in Table 1 and in main text. Performances of all models run are in Supplementary Material S.5.



CHAPTER 5

Clinical application I: Prediction of time to insulin in type 2 diabetes

5.1 Global diabetes prevalence

Diabetes is one of the most serious metabolic health challenges faced in the world today [147]. In 2019, 463 million people worldwide were suffering from diabetes, a number which is estimated to rise to 700 million in 2045 [148]. Additionally, 374 million people are estimated to have impaired glucose tolerance and are at risk of developing diabetes according to a report by the International Diabetes Federation [148]. Diabetes is a disorder characterized by an impaired glucose metabolism due to limited insulin secretion, insulin action or a combination of both resulting in high blood glucose levels (hyperglycemia) [149]. Diabetes is a highly heterogeneous disease and can be divided into several subtypes including type 1 diabetes (T1D), latent autoimmune diabetes in adults (LADA, also known as type 1.5 diabetes), maternally inherited diabetes and deafness (MIDD), maturity-onset diabetes of the young (MODY), neonatal diabetes, gestational diabetes, and type 2 diabetes (T2D) [149]. T1D and T2D are the most commonly occurring within the diabetes spectrum, whereof 90% of diabetes patients are estimated to be diagnosed by T2D [150]. T2D is the focus of this chapter.

5.2 Type 2 diabetes

T2D is a progressive metabolic condition driven from the loss of the pancreatic β-cell insulin secretory capacity and insulin resistance resulting in hyperglycemia [151, 152]. Insulin has a crucial impact on regulation of metabolic homeostasis. In an individual with normal glucose tolerance, insulin is produced and secreted from the β-cells of the pancreatic islets in response to increased blood glucose levels [153]. The blood glucose level is regulated at a whole-body level where insulin has metabolic effects in multiple tissues that facilitates blood glucose levels to return to normal concentrations. For example, insulin stimulates glucose uptake in muscle cells and adipose tissue, promotes glycogen formation of blood glucose in the liver and regulates mechanisms of appetite signaling

and energy expenditure in the brain amongst other functions [153, 154]. Failure to maintain these metabolic mechanisms in diabetes can result in uncontrolled hyperglycemia that over time increases diabetes patients' risk of several micro- or macro-vascular comorbidities [149, 155]. Well-established risk factors of T2D include unfavorable lifestyle choices (unhealthy diet, physical inactivity, smoking), obesity, older age, sleep deprivation, socioeconomic status, urbanization and genetic predisposition [156–159]. In addition, the gut microbiota may modulate metabolic health and possibly influence T2D development [160]. Despite T2D is defined as its own subtype, T2D is a highly heterogeneous disorder. T2D heterogeneity has been elucidated from clustering analyses, where genetic variants have been used for identifying different disease pathway clusters of T2D subtypes relating to insulin deficiency or insulin resistance pathways in different tissues [161]. Five other subgroups of T2D patients have been identified by clinical biomarkers [162]. Such studies illustrate that there are most likely several T2D subtypes.

5.3 The treatment journey in type 2 diabetes

Diabetes can be diagnosed by measurements of HbA1c, the fasting plasma glucose levels, the 2-hour plasma glucose levels or by an oral glucose tolerance test [163]. HbA1c is a measurement of the glycated hemoglobin and reflects the average glycemic level over the past two or three months (i.e. turnover rate of erythrocytes), thereby offering less perturbations in its measurement compared to fasting glucose test or the oral glucose tolerance test. A HbA1c measurement of 5.7%−6.4% indicates a pre-diabetic state, whereas HbA1c$\geq$ 6.5% is the diagnosis criteria of diabetes [163]. Following diagnosis, treatment of T2D focus on prevention or delay of diabetes-associated complications, while maintaining patients' quality of life [164]. Guidelines of T2D therapy are here described according to a consensus report on management of hyperglycemia by the American Diabetes Association and the European Association for the Study of Diabetes [164]. Treatment is aimed towards a glycemic target where a HbA1c $<$ 7% is suggested as a reasonable target for most adults. The glycemic target is evaluated per patient given several factors such as the duration of T2D, life expectancy, comorbidities and cardiovascular risk factors as well as the patient's preference for treatment [164, 165]. To meet the glycemic target, guidelines for T2D management typically involve lifestyle interventions including weight loss and exercise, and/or followed by pharmacological interventions where metformin is the typical mono-therapy [164, 166]. Following diabetes progression, reflecting the continuous deterioration of the insulin secretion function and insulin resistance, the treatment is intensified to maintain glycemic control. Sulfonylureas or dipeptidyl peptidase-4 (DPP-4) inhibitors in combination with other antidiabetic drugs are commonly used as dual therapy in glycemic control of T2D [164, 166, 167]. Finally, insulin will be prescribed to manage blood glucose levels (Figure 5.1) [164, 166].

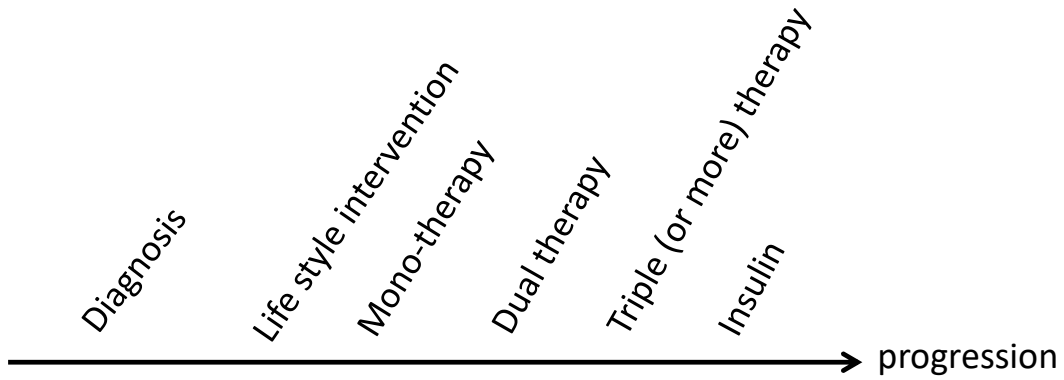


Figure 5.1: Type 2 diabetes treatment intensification following disease progression.

5.4 Clinical inertia

In T2D management, one aim of treatment is to delay the prescription of insulin as much as possible in order to minimize patient's treatment burden [167]. Despite close monitoring of blood glucose levels throughout disease progression, the treatment intensification with both oral antidiabetic drugs and insulin has been identified as being delayed and insufficient to maintain glycemic control for a substantial proportion of T2D patients [168]. This is known as clinical (or therapeutic) inertia which potentially increases risk of long-term diabetes complications [168, 169]. Clinical inertia in T2D is most profound at the intensification onto insulin [169]. Suggested causes of clinical inertia in T2D include both decisions regarding treatment made by the patients e.g. fear of hypoglycemia, route of treatment administration (oral or injection devices) as well as by the physicians e.g. patient risk given other comorbidities, side effects, overestimation of patient-adherence to treatment guidelines and evaluation of quality of life [164, 168, 169].

5.5 Variation in progression rates of time to insulin in type 2 diabetes

The time to insulin is not similar across all T2D patients and is difficult to determine for individual patients. In the GoDARTS cohort [25] used in this project, some T2D patients require insulin only one year after diagnosis whereas other patients can maintain sufficient glycemic target values on oral antidiabetic drugs for up to 20 years [151] (Figure 5.2).

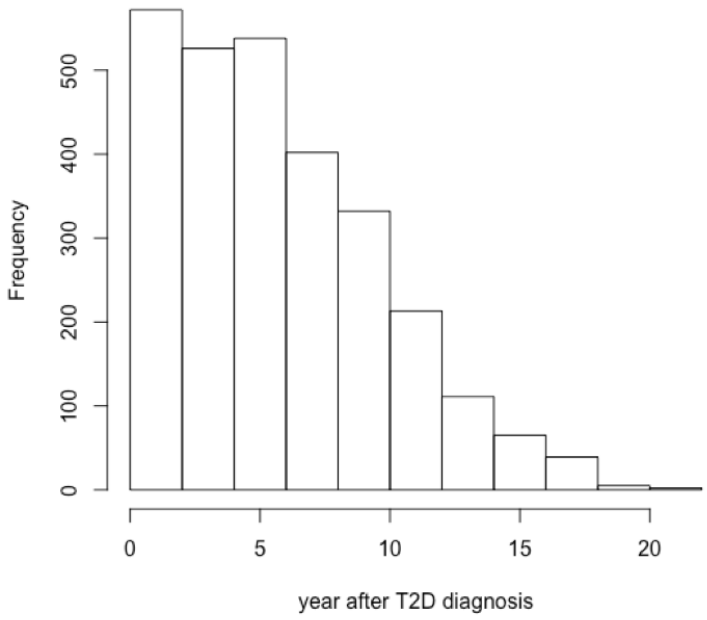


Figure 5.2: Distribution of the time to insulin requirement for type 2 diabetes (T2D) patients in the GoDARTS cohort [25]. The average time to insulin was 5.7±4.1 years.

5.6 Study introduction

In this project, prediction models of T2D patients' time to insulin requirement within the next 1 to 4 years ahead of any yearly time point after T2D diagnosis are presented. Approximately 6000 T2D patients from an observational cohort in Tayside, UK were available for modelling. Patient data on various clinical biomarkers including anthropometry, blood pressure, lifestyle, a social deprivation score derived from the Scottish Index of Multiple Deprivation, biochemistry data and drug prescriptions was obtained from EMRs. Patients were also genotyped on SNP platforms (Affymetrix Genome-Wide Human SNP Array 6.0 and Illumina HumanOmni Express platform [25]).

Due to the longitudinal irregular sampled clinical measurements, clinical features were extracted at 'baseline' time points up to 10 years follow-up after diagnosis by two approaches; i) a fixed time point approach and ii) an approach that developed longitudinal features by a linear, general and autoregressive trend.

First, a clinical baseline model without genotype data was trained. The time to insulin requirement was predictable in a substantial proportion of patients at any time point independent of duration of T2D using only the clinical data. The insulin requirement was especially predictable close to event (+1 year) compared to +2, +3 and +4 years ahead. The feature importance across all models for up to 10 years after diagnosis was guided by similar features. HbA1c was the most predictive feature and appeared

even more important when measured closer to the insulin requirement event. However, HbA1c was not as predictive of time to insulin when applying a linear model (logistic regression or LASSO regression) compared to an ANN which outperformed the linear model's performance when predicting 1 or 2 years ahead. This indicated a non-linear association between features for at least some patients, which would be worth exploring further given underlying T2D heterogeneity.

Genetic features were prioritized for modelling in 279 SNPs available out of 403 SNPs previously associated with T2D risk [159] as well as four GRS on β-cell function, insulin resistance, BMI and lipodystrophy. Integration of the genetic features did not improve performance of the clinical model.

The utility of the clinical models was assessed. We argue that an AI-based model providing data-driven risk assessment of the time to insulin based on a patient's treatment trajectory in EMRs can possibly reduce health care costs associated with glycemic surveillance for very low-risk patients. In high-risk patients, the prediction provide guidance that may influence patient behavior regarding lifestyle choices or compliance to therapy or assist in reducing clinical inertia through increased glycemic monitoring and identification of the need for treatment intensification.

5.7 Bioinformatic challenges: Disease progression modelling and genotype integration

In studies of disease progression, it is important to be able to predict how a patient's disease develops and when a given disease event will occur. In order to make a clinical useful model, the EMR and genotype data was extracted and transformed into features that could be modelled by machine learning models.

5.7.1 Longitudinal feature extraction

Each T2D patient had varying follow-up across irregularly sampled data from the EMRs. The clinical characteristics (anthropometry, blood pressure, and biochemistry data) thus appeared messy and unfit for traditional machine learning methodologies. Thus, two data extraction approaches were used; one approach was based on a single time point extraction and a second approach focused on modelling of new features given all longitudinal data available during disease progression. Features were extracted yearly between the time of confirmed T2D diagnosis and up to 10 years later (also referred to a baseline time points). Thus, in total, 10 datasets with clinical features were extracted. In the single time point approach, a measurement was extracted per clinical feature closest to the baseline time point within a ± 6 months window. The longitudinal features were

developed using a first order autoregressive model that for each clinical variable created three variables for the linear, general and autoregressive trend in the data as described in Nielsen *et al* [170] (*manuscript not included in the thesis*). A limitation of the clinical feature extraction approach was this introduced missing values which were imputed or discarded (if $> 80\%$ missing). The impact of imputation was tested by different imputation strategies and evaluated on feature importance and predictive performance. The imputation strategy did not influence these.

Drug prescriptions were focused only on diabetes drugs. Features of drug prescriptions were encoded by categories of 'no therapy', 'mono-therapy', 'dual therapy' or 'triple or more therapy' when considering the number of prescriptions extracted by the single time point approach within a ± 2 months window.

5.7.2 Data imbalance

Drawbacks of converting the time to insulin into a binary outcome included introduction of an imbalanced prediction outcomes i.e. a limited number of patients developed the requirement for insulin within the one-year prediction window compared to those who did not when predicting +1, +2, +3 and +4 years ahead. Thus, it was necessary to balance the classes using down-sampling on the training data; both in the inner and outer level of the cross-validation. To increase the number of patients requiring insulin, the same prediction outcome with a two- or three-year prediction window was explored *(results not shown)*. This resulted in similar performance as presented with the one-year prediction window presented in the paper. However, this did not allow for an as timely guidance on the requirement of insulin and was thus not considered further.

5.7.3 Genotype

Genotype data was obtained from two different genotype arrays and was thus imputed to approximately 27.4M SNPs. To reduce the feature input space, we hypothesized that SNPs associated with the risk of T2D and biomarkers involved in T2D progression would assist in prediction of patients' time to insulin. 403 SNPs previously associated with T2D [159] in independent GWAS were selected. Furthermore, GRS of β-cell function, insulin resistance, BMI and lipodystrophy were used. To prioritize genetic features, these were added to the clinical model by a data-driven forward selection strategy.

5.8 Manuscript

The following manuscript is submitted to *Diabetes Care*. The supplementary material is given in Appendix B.

Prediction of time to insulin requirement in patients with type 2 diabetes using artificial intelligence: A GoDARTS study.

Rikke L. Nielsen MSc (1,2), Louise Donnelly MD (3), Agnes M. Nielsen PhD (4), Kaixin Zhou PhD (3), Adem Dawed (3), Konstantinos Tsirigos PhD (1), Bjarne Ersbøll PhD (4), Line Clemmensen PhD (4), Ewan R. Pearson FRCP PhD (3), Ramneek Gupta PhD (1).

Affiliations

(1) Department of Health Technology, Technical University of Denmark, Denmark.

(2) Sino-Danish Center for Education and Research, Eastern Yanqihu campus, University of Chinese Academy of Sciences, China.

(3) Division of Population Health & Genomics, School of Medicine, University of Dundee, Dundee, Scotland.

(4) Department of Applied Mathematics and Computer Science, Technical University of Denmark, Denmark.

Corresponding authors: Ewan Pearson, Ramneek Gupta

Contact information of the corresponding authors:

Ewan Pearson, Division of Population Health & Genomics, School of Medicine, University of Dundee, Dundee, Scotland. +44 (0)1382 383387, e.z.pearson@dundee.ac.uk

Ramneek Gupta, Department of Bio and Health Informatics, Technical University of Denmark, Denmark. +45 45 25 24 22, +45 45 93 15 85, ramg@dtu.dk

Short running title*:* Predicting time to insulin in type 2 diabetes

Word count: 3656/4000

References: 33/40

Tables and figures: 4/4

Abstract

Objective

The rate of progression in type 2 diabetes is heterogenous and hard to determine for individual patients. This study aims to understand how predictable time to insulin is for the individual patient using information from electronic medical records (EMR) with or without patient genotype, and to identify the most predictive factors for progression to insulin requirement.

Research Design and Methods

6333 type 2 diabetes patients from the GoDARTS cohort were included in the study. We used artificial neural networks with stringent cross-validation to predict patients' requirement for insulin 1 to 4 years ahead of any time during treatment from 1 to 10 years after diagnosis. Longitudinal EMR data and 279 type 2 diabetes genetic risk variants and four genetic risk scores were included in the models.

Results

The time to insulin requirement 1 to 4 years ahead identified patients whose progression was driven primarily by HbA1c, diabetes treatment and the age at diagnosis (ROC-AUC at: +1Y:0.83, +2Y:0.73, +3Y:0.69, +4Y:0.66). Genotype features did not improve performance. Artificial neural networks were used to pick up non-linear correlations in the data, as regression models were not able to achieve similar performance. 50% of the low-risk patients were identified with 99% confidence, while 50% of the high-risk patients of insulin requirement were identified with 23% confidence.

Conclusions

Non-linear machine learning methods were used for integration of longitudinal EMR data predicting time to insulin at the individual patient level which can provide clinically useful prediction for patients at low or high risk of progressing to insulin requirement.

Introduction

One aim of type 2 diabetes treatment is to control hyperglycemia in order to delay adverse effects associated with type 2 diabetes and potentially slow disease progression. However, some patients fail to maintain glycemic control due to inertia of insulin prescription or delayed treatment intensification[1,2]. The decision to intensify treatment medication is physician determined, and relies largely on HbA1c levels[3,4], cardiovascular history and other information such as life-style, age and gender[5] and patient choice. There is considerable variation in the rate at which treatment is intensified, reflecting heterogeneity in progression of the underlying disease, variation in patient lifestyle and variation in treatment inertia. Some patients require insulin within the first years following diagnosis, whereas other patients remain well controlled for up to 20 years without the need for insulin[6]. Studies of diabetes progression defined by insulin initiation have identified various biomarkers associated with progression: longer duration of type 2 diabetes[7,8], higher HbA1c[7,9], triglycerides[7,10]; lower HDL-C[7,10]; younger age[8–11]; more diabetes complications[11] and white ethnicity[11] have been associated with faster progression to insulin, whereas lower LDL-C has been associated with slower progression[7]. Obesity has been found to be associated with both fast[10] and slow progression[7].

The type 2 diabetes guidelines are based on trial evidence of the mean response of one treatment arm. The increased availability of different treatment options as well as the increasing data that can be obtained from electronic health records, life style and genetic profiling provides the potential to improve treatment stratification of patients, but also complicates clinical decision making[4]. Despite type 2 diabetes risk having a strong genetic component[12], the use of genetic markers for diabetes treatment is currently only implemented in clinic for monogenic types of diabetes[5] due to insufficient identified effect sizes associated with type 2 diabetes risk. The promise of data-driven decision making is realizable in part through advanced data integration methodologies, rigorous data cleaning and intelligent feature selection that drive predictors. Machine learning methods are flexible in being

able to utilize multivariate inputs from various heterogenous data types and identifying non-linear relationships by modelling prediction of outcomes[13]. Artificial intelligence methods are currently implemented on various clinical decisions for precision medicine[14] and the number of publications on machine learning in diabetes research have rapidly increased over the years[15]. Examples of prediction tasks in diabetes research involve the prediction of risk for individuals to develop type 2 diabetes after 1,3 or 8 years[16], treatment success or failure of metformin one year after treatment initialization[17], longitudinal changes in HbA1c values[18], events of hypoglycemia[19], diabetes complications 3,5 and 7 years after patients first visit to the hospital[20] and diabetic retinopathy[21].

In this study, we used information collected in electronic medical records (EMRs) and genotype from the GoDARTS cohort[22]. We utilized artificial neural networks (ANNs) to integrate heterogenous and longitudinal biomarkers to predict type 2 diabetes progression by classifying patients' time to insulin (TTI). Due to the benefits of controlling glycemic levels, it is of clinical interest to know ahead of time if and when patients will require insulin treatment initiation. We demonstrate that machine learning models can predict individual risk of clinical insulin requirement 1 to 4 years ahead of current prescription patterns, with the potential of reducing clinical inertia in a subset of type 2 diabetes patients.

Research Design and Methods

Electronic medical record data

Patient data was obtained from a cohort-based population in Tayside, Scotland recruited between 1999 and 2012 as part of the Genetics of Diabetes Audit and Research (GoDARTS) study as described previously[10,22]. All clinical information was obtained by EMRs. This included the year of diabetes diagnosis, information about life-style including a social deprivation score and smoking (ever/never), anthropometry including body weight, BMI, systolic and diastolic blood pressure, age and gender, biochemical blood measurements of HbA1c, creatinine, alanine transaminase (ALT), aspartate transaminase (AST), cholesterol, LDL, HDL, triglyceride, GAD antibody and drug prescriptions. The EMRs consist of patient information (N=10149) whereof 7238 were diagnosed with type 2 diabetes after 5/1994 with available data[22]. We only included patients where a social deprivation score was given, and patients diagnosed with type 2 diabetes (where HbA1c≥6.5% (48 mmol/mol) or first drug prescription) from 1994 to 2010 (Supplementary Material 1). Following filtering, 6333 patients were available from the GoDARTS cohort, which are summarized by the variable measured closest to diagnosis confirmed by HbA1c≥6.5% (48 mmol/mol) or first drug prescription ±6 months in Supplementary Material 2.

Longitudinal extraction of clinical EMR and drug prescription features

Observational measurements for individual patient journeys in their EMRs are irregularly sampled according to the time of physician consultation (Figure 1). Thus, we modelled the longitudinal information within 1–10 years of diagnosis by two approaches. The first focused on measurements available at a fixed-time point each year ranging from 1 year following diagnosis, up to 10 years. As the availability of measurements were dependent on physician/clinic visit, a window of ±6 months was allowed for measurement extraction. The completeness of biochemical biomarkers related to

how often a patient had engaged with the health care system (missingness estimates in Supplementary Material 3). The second approach utilized auto-regressive modelling to capture all longitudinal information available from diagnosis to the starting time point of interest (again, each year up to 10 years from diagnosis). We considered all measurements for each variable taken before a given year if this was reported minimum three times in the EMR as described by Nielsen et al[23]. The time-dependent behavior in the autoregressive models is described by three extracted features describing a general level in data, a linear trend in data and an autoregressive aspect in data.

Longitudinal extraction of drug prescriptions

Drug prescription data was focused only on diabetes medications (Supplementary material 4). Features of no diabetes treatment, monotherapy, dual therapy or triple or more therapy were summarized by counting the number of different diabetes drug prescriptions within a four-month window ($\pm$2 months around baseline) each year from 1 year following diagnosis (up to 10 years) and was merged into the fixed-time point data and the autoregressive data as features.

Diabetes progression definition by time to insulin

Diabetes progression was investigated using time to insulin (TTI) as the prediction outcome. TTI was defined as the first day of insulin treatment (confirmed by sustained use with at least 6 months' duration) or as the clinical requirement for insulin (HbA1c>8.5% (69 mmol/mol) treated with two or more non-insulin diabetes therapies) since time of confirmed type 2 diabetes diagnosis (first HbA1c$\geq$6.5% (48 mmol/mol) or prescription of first drug).

In this study, observation start points ('baseline') were defined 1–10 years from diagnosis. For each baseline, a separate model was constructed to predict TTI +1, +2, +3 and +4 years into the future. Only patients who did not have requirement of insulin prior to baseline were considered in each

model. In each model, a patient was defined as a case when they progressed onto insulin within a prediction window of 1 year after the +1/+2/+3/+4 year points respectively. Patients who did not have the requirement for insulin within the prediction window were included as controls. As an example, a patient with baseline year 1 after diagnosis will be defined as a case or control in a +1Y model depending on if the patient went onto insulin within year 1 and 2 from diagnosis. Similarly, for a +2 year prediction: between year 2 and 3, +3 year prediction: between year 3 and 4 and finally +4 year prediction: between year 4 and 5 (Figure 1). We tested how predictive the data extracted for year 1–10 by the fixed time point approach and autoregressive-modelled features was due to the irregular timed measurements of clinical features in the EMRs. The best data representation was assessed by ROC-AUC and was used for further development of models. The number of patients used for training of TTI models are seen in Supplementary Material 5. The models did not include competing risk of death within this prediction window and the percentage of patients dying within the prediction window was up to 2.8% for all models (Supplementary Material 6).

Machine learning modelling and performance

We used artificial neural networks (ANNs) for modelling of TTI which were trained and tested using a case/control-stratified two-level five-fold cross-validation. The outer level of the cross-validation setup thus functions as five independent validation sets. ANNs were modelled using the R packages nnet[24] and caret[25]. To determine if simpler models were just as effective, a logistic regression and a least absolute shrinkage and selection operator (LASSO) regression was subsequently modelled using the glmnet package[26] . The performance of the machine learning models was evaluated by area under the receiver operating characteristic (ROC) curve (ROC-AUC), sensitivity, specificity, Matthews correlation coefficient (MCC) and the Youden's J statistics (J). Feature importance in the ANNs was evaluated by Olden's relative importance[27]. Details of the ANN, LASSO and logistic regression

166 modelling and performance measurements are given in Supplementary Materials 7. Models were
167 tested for stability in performance by comparing ROC-AUC across 50 different model initialization
168 seeds to a permuted outcome where the cross-validation splits were maintained.
169
170 *Feature selection and integration of genetic variants*
171 5922 patients were genotyped of the 6333 patients included in the study. Genotype data from GWAS
172 arrays (Affymetrix Genome-Wide Human SNP Array 6.0 and the Illumina HumanOmni Express)
173 was imputed up to ca 80M SNPs as described previously[22]. 27.4M SNPs were available across both
174 platforms. In order to assess if genetic variants associated with type 2 diabetes and related traits would
175 be important for diabetes progression, we derived weighted genetic risk scores (GRS) for beta-cell
176 function[10,28], insulin resistance[29] and BMI[30] and an unweighted genetic risk score for lipodystrophy[31]
177 from previously published GWAS, as these are potential metabolic predictors of progression
178 (Supplementary Materials 8). The weighted GRS was constructed by summing the number of risk-
179 increasing alleles carried weighted by the logarithm of the allelic odds ratio of the SNP. We also
180 prioritized 403 SNPs associated with type 2 diabetes from literature[12] for modelling from the imputed
181 GWAS data. Of the available 402 SNPs, 349 passed filtering with an info score at 0.7. SNPs with
182 missing information due to equal probabilities of genotype in any sample were removed resulting in
183 279 SNPs. Genetic variants were integrated by a forward selection strategy where SNPs chosen if
184 ROC-AUC increased by minimum 0.01 when adding new features in an ANN with 10 hidden nodes
185 and a weight decay of 1. If two or more SNPs resulted in similar increase in ROC-AUC, all were
186 included in the model.

Results

Longitudinal information assessment

For modelling time to insulin (TTI), we incorporated longitudinal information extracted by the fixed time-point approach as well as autoregressive features across EMRs. Features extracted by the fixed time-point representation resulted in higher ROC-AUC compared to autoregressive modelling of features utilizing data from the entire period that the patient had been diagnosed for in the +1Y prediction models. No substantial differences in ROC-AUC performance were seen for the +2Y, +3Y and +4Y prediction models between fixed time-point and auto-regressive data (Supplementary Material 9). The fixed time point data was thus used for further analysis.

Modelling of time to insulin

The ANN models could predict time to insulin between 1 and 10 years after diagnosis using the fixed time point data with ROC-AUC 0.66–0.83. For all baseline starting points, highest performance was seen for the +1Y model with ROC-AUC: 0.83±0.04 and 2 to 4 years forward with ROC-AUC of 0.66–0.73 (Table 1). Integration of the genotype information from four GRS's of beta cell function, insulin resistance, BMI and lipodystrophy and genetic risk variants associated with type 2 diabetes did not improve prediction of the clinical +1Y, +2Y, +3Y and +4Y models (Table 1, Supplementary Material 10 shows performance on patients where genotype data were available (N=5922)).

Clinical features associated with time to insulin

The relative importance of features for the clinical models predicting TTI +1Y, +2Y, +3Y, and +4Y ahead in type 2 diabetes patients showed HbA1c was the most important feature. The highest impact of HbA1c was in the +1Y model and decreasing towards the +4Y model where other features became

more important for the prediction of TTI (Figure 2). Younger age drives progression and the type of diabetes medication at the time of interest, where no-drug or monotherapy is associated with lower risk of progression onto insulin, whereas dual therapy or triple or more therapy is associated with greater risk of progression onto insulin. In all four models (+1Y, +2Y, +3Y and +4Y), the narrow standard deviation of the relative feature importance suggests models learned similar features regardless of the baseline starting point. The models that included GRS and genetic risk variants for type 2 diabetes showed no significant importance for any of the host genotype features (Supplementary Material 11). Due to the high feature importance of HbA1c in the clinical models, we tested if less complex models could predict with a similar performance. LASSO regression models and logistic regression models with HbA1c, age, diabetes drug prescriptions, calendar year of diagnosis and BMI performed less well for the fast progression models (LASSO ROC-AUC: +1Y: 0.66±0.05 (95%CI [0.64–0.67]), +2Y: 0.64±0.05 (95%CI [0.62–0.65]), logistic regression ROC-AUC +2Y 0.66±0.04 (95%CI [0.64–0.67]), and ANN ROC-AUC +1Y: 0.83±0.04 (95%CI [0.82–0.84]), +2Y: 0.73±0.04 (95%CI [0.72–0.75])), but had similar ROC-AUC for the +3Y and +4Y models (Supplementary Material 12). This suggests that ANNs crucially capture HbA1c interactions with other clinical features for a substantial number of patients, in a way that is not captured in more traditional approaches. Finally, we found no major impact of clinical variable imputation on the feature importance in the ANN, LASSO regression or logistic regression models (Supplementary Material 13). Robustness of the TTI models with fixed time point EMR data was checked by testing the performance across 50 model initializations and cross-checked against a permuted label setup of TTI case/control outcome confirming high robustness of the +1Y, +2Y, +3Y and +4Y ANN models (Supplementary Material 14).

Patient utility of time to insulin model

The sensitivity, positive predictive value (PPV), specificity and negative predictive value (NPV) are summarized at different prediction thresholds in Table 2 for guiding clinical utility with respect to patient risk. For clinical implementation of such a machine learning framework, we considered the impact of model thresholding and defined <0.3 as a suitable prediction threshold for patients with low probability of requiring insulin and patients with a prediction score threshold >0.7 at a high risk of requirement insulin within the next year (+1Y) across all the 10 years. At a prediction threshold of 0.3, 50% of patients that will not require insulin will be correctly predicted at an NPV of 99%. Thus, only 1% of these patients would be FN corresponding to ca. 20 patients that should have received insulin but were not identified at being at risk in the model. On the other hand, a high prediction threshold of 0.7 would identify a subgroup of patients (ca. 400–700, ~9% of total number of patients) where one in four (avg. PPV= 23%) will develop the need for insulin. We investigated the HbA1c and diabetes drug in each of the prediction scoring tail distributions, which showed the model separated these patients based on HbA1c (low vs high) and the type of diabetes drug (Supplementary Material 15).

The potential to address misclassified type 2 diabetes patients

A second model was trained on all patients where the first clinical models' predictions were false (positive or negative) in order to understand predictability within this group of patients (Figure 1). The second clinical model achieved ROC-AUC of: 0.73 (+1Y), 0.70 (+2Y), 0.65 (+3Y) and 0.69 (+4Y). Host genotype features utilized in this study did not improve ROC-AUC (Supplementary Material 16). These patients were characterized by a lower HbA1c (+1Y: 8.14±1.17, +2Y: 7.56± 51.04, +3Y: 7.35±1.01, +4Y: 7.14±0.91) than the general group (+1Y: 9.32±1.64, +2Y: 8.54±1.56, +3Y: 7.90±1.35, +4Y: 7.68±1.28) and high HbA1c did not appear to be a major indicator of

progression. This group contains a more heterogenous biomarker profile compared to the first clinical model and markedly larger standard deviations (Feature importance in Supplementary Material 17 for +1Y, +2Y, +3Y, +4Y models and summary statistic differences in patient groups in the original and second clinical model in Supplementary Material 18). The second model appeared to be robust in terms of a clearly higher performance than with randomly permuted labels (Supplementary Material 14) demonstrating that we were able to re-classify several patients by use of a second model. This was not explored further within the scope of this study.

Conclusions

In this study, we applied artificial neural networks (ANNs) with rigorous cross-validation to predict progression onto insulin in the next 1, 2, 3 or 4 years in *ca.* 6000 type 2 diabetes patients from an observational cohort-based population with up to 10-year follow-up after diagnosis. We demonstrated that time to insulin (TTI) was predictable with ROC-AUC: 0.66–0.83 within a 4 year period from any time point after diagnosis. Highest prediction performance was for near-term predictions (+1Y ROC-AUC: 0.83). Even though HbA1c was a clear major driver for the predictions, LASSO regression models or logistic regression models performed substantially lower than ANN models for the +1Y and +2Y predictions, suggesting that a non-linear combination of information from other features is needed to predict progression. Despite the challenge that only ~3–6% of patients in any individual model progressed onto insulin, our results indicate it is possible at any given time in the patients' journeys to predict progression onto insulin within a year irrespective of their HbA1c levels.

By applying prior knowledge on genetic variants associated with the risk of developing type 2 diabetes[12], it was not possible to improve performance within these models. While individual genetic variants, owing in part to small effect sizes, are not predictive of type 2 diabetes or progression alone, there is focus on applying a number of genetic variants to construct genetic risk scores (GRS). In this study, the use of GRS associated with metabolic pathways involved in type 2 diabetes development did not improve the prediction models performance. This was not entirely unexpected as the use of GRS have previously shown limited capabilities of predicting individual risk to type 2 diabetes[12] or its progression[10]. In an experiment (*data not shown*) examining patients with low HbA1c, and controls defined as receiving lifestyle or monotherapy, inclusion of selected T2D risk variants did add to performance suggesting that the value of genetic information could be explored further in context of specific groups of patients.

We further explored if a second model could re-classify mis-classified patients from the original TTI models. This group of patients were characterized by a lower mean HbA1c and the re-trained models predicted progression onto insulin with ROC-AUC 0.65–0.73 (+1Y ROC-AUC: 0.73) suggesting that it was possible to correctly classify some of the 'difficult' patients that are likely also clinically less intuitive. This could be grounds for future investigation perhaps with deeper learning models and with wider incorporation of genetics.

The trained ANN models were found to be robust across multiple model initializations and against an output label permutation test. Missing data were imputed by extreme single values as a marker that the test had not been ordered in this period. This produce a biased estimate of the results[32], however it assumes that patients are comparable with each other and learns given an extreme value outside of any biological distribution that these patients are their own group. The motivation for extreme values was considered given biases in data analysis from EMRs[33]. Missing data for biochemical features were similar across cases and controls. LDL had the highest missingness of ~57–64% while the most predictive feature HbA1c had ~27–38% missing values. In addition, our models did not include competing risk of death. However, this was limited to a maximum 2.7% of cases in the machine learning models.

Despite stringent nested cross-validation and permuted labels experiments, any data-driven analysis has limitations, and thus an external validation perhaps on a different demographic would be warranted. The robustness in performance on the presented models provide reasonable grounds to apply these models on separate cohorts.

In conclusion, we here demonstrate that machine learning methods can integrate heterogenous data including clinical and genetic information which can help identify patients requiring insulin treatment. The time to insulin risk assessment models provide an individual probability of requiring insulin within the next +1, +2, +3 or +4 years ahead at any time point during type 2 diabetes treatment.

We believe machine learning models such as these presented in this paper will in the future assist in targeted patient intervention by identification of patients that require earlier treatment by insulin thereby addressing insulin inertia, increased monitoring of HbA1c or timely and more precise messaging to influence patient behavior that might delay insulin requirement. To determine the impact on clinical practice, the rate of false positives and false negatives needs to be considered carefully. We suggested clinically relevant prediction thresholding of the prediction scores that would offer useful thresholds (<0.3) at low risk of progression (with NPV of 0.99) and >0.7, with a 23% PPV for insulin requirement. Other thresholds can be set, depending on the clinical scenario. For example, setting a prediction threshold of 0.2 on the +1Y model identified 35% of patients that can be "sent home for a year", at no risk of missing out patients that will require insulin (NPV = 100%). In our data set, this corresponds to ca. 1400–2000 patients at any point in the 10-year period after diagnosis for whom this risk threshold could be incorporated into pathways of clinical care resulting in reduced health care intervention and health care costs. Identification of patients at high risk of requiring insulin could instead have increased monitoring, compliance or urgent lifestyle messaging warranted which can assist in prolonging the time to insulin. Trials of implementation of such an AI driven prediction tool would be required to establish such potential benefits.

Acknowledgments

The authors are grateful to all the participants who took part in this study; the general practitioners; the Scottish School of Primary Care for its help in recruiting the participants; and the whole team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists, and nurses.

Funding

RLN was supported by the Sino-Danish Center for Education and Research. RLN and AMN were supported by the Poul V Andersen Foundation. ERP holds a Wellcome Trust New Investigator award (102820/Z/13/Z).

Duality of Interest

Authors declare no conflict of interest.

Author Contributions

Study concept and design: ERP, RG

preparation of data from electronic medical records: RLN, LD, AMN

preparation of genetic data: RLN, LD, KZ, AD

performed data analyses: RLN, LD, AMN

interpreted the results: RLN, LD, ERP, RG

wrote the manuscript: RLN, RG

reviewed/edited the manuscript: RLN, LD, RG, ERP

Study supervision: LD, RG, ERP

Administrative, technical, or material support: KT

References

1. Khunti K, Wolden ML, Thorsted BL, Andersen M, Davies MJ. Clinical inertia in people with type 2 diabetes: A retrospective cohort study of more than 80,000 people. *Diabetes Care*. 2013;36(11):3411-3417. doi:10.2337/dc13-0331
2. Khunti K, Millar-Jones D. Clinical inertia to insulin initiation and intensification in the UK: A focused literature review. *Prim Care Diabetes*. 2017;11(1):3-12. doi:10.1016/j.pcd.2016.09.003
3. Hanefeld M. Use of insulin in type 2 diabetes: What we learned from recent clinical trials on the benefits of early insulin initiation. *Diabetes Metab*. 2014;40(6):391-399. doi:10.1016/j.diabet.2014.08.006
4. Davies MJ, D'Alessio DA, Fradkin J, et al. Management of hyperglycaemia in type 2 diabetes, 2018. A consensus report by the American Diabetes Association (ADA) and the European Association for the Study of Diabetes (EASD). *Diabetologia*. 2018;61(12):2461-2498. doi:10.1007/s00125-018-4729-5
5. Fitipaldi H, McCarthy MI, Florez JC, Franks PW. A global overview of precision medicine in type 2 diabetes. *Diabetes*. 2018;67(10):1911-1922. doi:10.2337/dbi17-0045
6. Jennison C, Donnelly LA, Doney ASF, Zhou K, Pearson ER, Franks PW. Rates of glycaemic deterioration in a real-world population with type 2 diabetes. *Diabetologia*. 2017;61(3):607-615. doi:10.1007/s00125-017-4519-5
7. Gentile S, Strollo F, Viazzi F, et al. Five-Year Predictors of Insulin Initiation in People with Type 2 Diabetes under Real-Life Conditions. *J Diabetes Res*. 2018:1-11.
8. Mast R, Jansen APD, Walraven I, et al. Time to insulin initiation and long-term effects of initiating insulin in people with type 2 diabetes mellitus: the Hoorn Diabetes Care System Cohort Study. *Eur Soc Endocrinol*. 2016;174(5):563-571. doi:10.1530/EJE-15-1149

9. Ringborg A, Lindgren P, Yin DD, Martinell M, Stålhammar J. Time to insulin treatment and factors associated with insulin prescription in Swedish patients with type 2 diabetes. *Diabetes Metab*. 2010;36(3):198-203. doi:10.1016/j.diabet.2009.11.006

10. Zhou K, Donnelly LA, Morris AD, et al. Clinical and Genetic Determinants of Progression of Type 2 Diabetes: A DIRECT Study. *Diabetes Care*. 2014;37(3):718-724. doi:10.2337/dc13-1995

11. Pilla SJ, Yeh H, Juraschek SP, Clark JM, Maruthur NM. Predictors of Insulin Initiation in Patients with Type 2 Diabetes: An Analysis of the Look AHEAD Randomized Trial. *J Gen Intern Med*. 2018;33(6):839-846. doi:10.1007/s11606-017-4282-9

12. Mahajan A, Taliun D, Thurner M, Robertson NR, Torres JM. Fine-mapping of an expanded set of type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat Genet*. 2018;50(11):1505-1513.

13. Rose S. Machine Learning for Prediction in Electronic Health Data. *JAMA Netw Open*. 2018;1(4):e181404. doi:10.1097/ede.0b013e3181c30fb2

14. Miller D, Scheinker D, Bambos N. A Practical Approach to Machine Learning for Clinical Decision Support Projects at Lucile Packard Children' s Hospital Stanford in Partnership with Stanford Engineering. In: *Health Care Systems Engineering*. ; 2018:111-120.

15. Contreras I, Vehi J. Artificial Intelligence for Diabetes Management and Decision Support: Literature Review. *J Med Internet Res*. 2018;20(5):e10775. doi:10.2196/10775

16. Talaei-Khoei A, Wilson JM. Identifying people at risk of developing type 2 diabetes: A comparison of predictive analytics techniques and predictor variables. *Int J Med Inform*. 2018;119(July):22-38. doi:10.1016/j.ijmedinf.2018.08.008

17. Murphree D, Arabmakki E, Ngufor C, Storlie C, McCoy R. Stacked classifiers for individualized prediction of glycemic control following initiation of metformin therapy in

type 2 diabetes. *Comput Biol Med*. 2018;103:109-115. doi:10.1016/j.compbiomed.2018.10.017

18. Ngufor C, Van Houten H, Caffo BS, Shah ND, McCoy RG. Mixed effect machine learning: A framework for predicting longitudinal change in hemoglobin A1c. *J Biomed Inform*. 2019;89(May 2018):56-67. doi:10.1016/j.jbi.2018.09.001

19. Sudharsan B, Peeples M, Shomali M. Hypoglycemia prediction using machine learning models for patients with type 2 diabetes. *J Diabetes Sci Technol*. 2015;9(1):86-90. doi:10.1177/1932296814554260

20. Dagliati A, Marini S, Sacchi L, et al. Machine Learning Methods to Predict Diabetes Complications. *J Diabetes Sci Technol*. 2018;12(2):295-302. doi:10.1177/1932296817706375

21. Ting DSW, Cheung CYL, Lim G, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA - J Am Med Assoc*. 2017;318(22):2211-2223. doi:10.1001/jama.2017.18152

22. Hébert HL, Shepherd B, Milburn K, et al. Cohort profile: Genetics of Diabetes Audit and Research in Tayside Scotland (GoDARTS). *Int J Epidemiol*. 2018;47(2):380-381j. doi:10.1093/ije/dyx140

23. Nielsen AM, Nielsen RL, Donnelly L, et al. A Comparison of Methods for Disease Progression Prediction Through a GoDARTS Study. 2018:1-23.

24. Ripley B, Venables W. Package 'nnet.' *https://cran.r-project.org/web/packages/nnet/nnet.pdf*. 2016.

25. Kuhn M. Building Predictive Models in R Using the caret Package. *J Stat Softw*. 2008;28(5):1-26. doi:10.1053/j.sodo.2009.03.002

26. Friedman J, Hastie T, Tibshirani R, Narasimhan B, Simon N, Qian J. Package 'glmnet.' *https://cran.r-project.org/web/packages/glmnet/glmnet.pdf*. 2019.
27. Olden JD, Joy MK, Death RG. An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecol Modell*. 2004;178(3-4):389-397. doi:10.1016/j.ecolmodel.2004.03.013
28. Morris AP, Voight BF, Teslovich TM, et al. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat Genet*. 2012;44(9):981-990. doi:10.1038/ng.2383
29. Yarnell J, Langenberg C, Prokopenko I, et al. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nat Genet*. 2010;42:105-119. doi:10.1038/ng.520
30. Locke AE, Kahali B, Berndt SI, et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature*. 2015;518:197-206.
31. Lotta LA, Gulati P, Day FR, et al. Integrative genomic analysis implicates limited peripheral adipose storage capacity in the pathogenesis of human insulin resistance. *Nat Genet*. 2017;49(1):17-26. doi:10.1038/ng.3714
32. Toh S, García Rodríguez LA, Hernán MA. Analyzing partially missing confounder information in comparative effectiveness and safety research of therapeutics. *Pharmacoepidemiol Drug Saf*. 2012;21(SUPPL.2):13-20. doi:10.1002/pds.3248
33. Agniel D, Kohane IS, Weber GM. Biases in electronic health record data due to processes within the healthcare system: Retrospective observational study. *BMJ*. 2018;361. doi:10.1136/bmj.k1479

Tables

Table 1: Performance given as mean ± standard deviation of all five-fold test models using two-level cross-validation. M1 refers to the first trained models, and M2 refers to the models retrained using false positive and false negative from M1. For M1 an average is reported for across all years from where data is extracted. Due to duplicated patients across multiple years, only one time point was included for the retrained models (M2). MCC: Matthews correlation coefficient, J: Youden's index.

	ROC-AUC	***Sensitivity***	***Specificity***	***MCC***	***J***
M1 clinical (Fixed time point) ANN					
+1Y	0.83 ± 0.04	0.71 ± 0.09	0.79 ± 0.08	0.25 ± 0.06	0.50 ± 0.09
+2Y	0.73 ± 0.04	0.64 ± 0.08	0.72 ± 0.05	0.16 ± 0.04	0.35 ± 0.09
+3Y	0.69 ± 0.05	0.61 ± 0.10	0.68 ± 0.06	0.13 ± 0.04	0.30 ± 0.10
+4Y	0.66 ± 0.06	0.58 ± 0.10	0.68± 0.06	0.11 ± 0.04	0.26 ± 0.10
M1 clinical+GRS (Fixed time point) ANN	***ROC-AUC***	***Sensitivity***	***Specificity***	***MCC***	***J***
+1Y	0.81 ± 0.06	0.69 ± 0.09	0.77 ± 0.06	0.22 ± 0.06	0.46 ± 0.10
+2Y	0.69 ± 0.05	0.61 ± 0.07	0.67 ± 0.06	0.12 ± 0.04	0.28 ± 0.08
+3Y	0.64 ± 0.05	0.56 ± 0.10	0.65 ± 0.05	0.09 ± 0.05	0.21 ± 0.11
+4Y	0.64 ± 0.04	0.56 ± 0.09	0.64 ± 0.06	0.09 ± 0.03	0.21 ± 0.08
M1 clinical+forward selection SNPs associated with T2D (Fixed time point) ANN	***ROC-AUC***	***Sensitivity***	***Specificity***	***MCC***	***J***
+1Y	0.81 ± 0.06	0.69 ± 0.09	0.78 ± 0.07	0.23 ± 0.07	0.46 ± 0.10
+2Y	0.70 ± 0.06	0.60 ± 0.10	0.70 ± 0.05	0.14 ± 0.04	0.30 ± 0.09
+3Y	0.66 ± 0.07	0.59 ± 0.11	0.66 ± 0.05	0.11 ± 0.05	0.25 ± 0.11
+4Y	0.65 ± 0.06	0.57 ± 0.10	0.65 ± 0.05	0.10 ± 0.04	0.22 ± 0.10

M2 clinical (Fixed time point) ANN	*ROC-AUC*	*Sensitivity*	*Specificity*	*MCC*	*J*
+1Y	0.73 ± 0.02	0.73 ± 0.08	0.61 ± 0.03	0.18 ± 0.03	0.34 ± 0.06
+2Y	0.70 ± 0.02	0.64 ± 0.04	0.64 ± 0.02	0.17 ± 0.02	0.29 ± 0.03
+3Y	0.65 ± 0.03	0.63 ± 0.04	0.55 ± 0.04	0.11 ± 0.03	0.19 ± 0.06
+4Y	0.69 ± 0.02	0.66 ± 0.04	0.63 ± 0.02	0.18 ± 0.03	0.29 ± 0.06

Table 1: The mean±standard deviation for the sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) for the +1Y, +2Y, +3Y and +4Y time to insulin prediction models across different prediction score classification thresholds for M1 clinical (Fixed time point) ANN.

M1 TTI +1Y	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
Sensitivity	0.98±0.01	0.96±0.01	0.91±0.03	0.81±0.06	0.71±0.04	0.60±0.05	0.52±0.06	0.42±0.08	0.24±0.05
Specificity	0.17±0.08	0.35±0.06	0.50±0.07	0.66±0.10	0.79±0.07	0.88±0.03	0.92±0.01	0.96±0.01	0.98±0.01
PPV	0.05±0.01	0.06±0.02	0.07±0.01	0.10±0.02	0.14±0.04	0.18±0.05	0.23±0.06	0.29±0.07	0.33±0.10
NPV	1±0	1±0	0.99±0	0.99±0	0.99±0	0.98±0	0.98±0.1	0.97±0.01	0.97±0.01
M1 TTI +2Y	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
Sensitivity	0.99±0.01	0.96±0.02	0.90±0.02	0.79±0.04	0.64±0.03	0.51±0.07	0.38±0.06	0.22±0.06	0.08±0.03
Specificity	0.05±0.02	0.17±0.05	0.32±0.05	0.53±0.08	0.72±0.04	0.83±0.02	0.90±0.01	0.94±0.01	0.98±0.0
PPV	0.04±0.01	0.05±0.01	0.06±0.01	0.07±0.01	0.09±0.02	0.11±0.02	0.14±0.03	0.15±0.04	0.15±0.05
NPV	0.99±0.01	0.99±0.0	0.99±0.0	0.98±0.0	0.98±0.01	0.97±0.01	0.97±0.01	0.97±0.01	0.96±0.01
M1 TTI +3Y	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
Sensitivity	0.99±0.01	0.97±0.02	0.89±0.03	0.77±0.04	0.61±0.03	0.46±0.06	0.31±0.07	0.16±0.04	0.05±0.02
Specificity	0.03±0.02	0.12±0.04	0.29±0.04	0.48±0.08	0.68±0.04	0.80±0.01	0.87±0.01	0.94±0.01	0.98±0.01
PPV	0.04±0.01	0.05±0.01	0.05±0.01	0.06±0.01	0.08±0.01	0.09±0.01	0.10±0.01	0.11±0.01	0.10±0.03
NPV	0.98±0.02	0.99±0.0	0.98±0.0	0.98±0.0	0.98±0.01	0.97±0.01	0.97±0.01	0.96±0.01	0.96±0.01
M1 TTI +4Y	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
Sensitivity	0.99±0.01	0.96±0.01	0.89±0.04	0.75±0.04	0.58±0.03	0.42±0.04	0.26±0.06	0.13±0.05	0.03±0.02
Specificity	0.02±0.01	0.10±0.03	0.24±0.04	0.46±0.06	0.68±0.05	0.81±0.02	0.90±0.01	0.96±0.01	0.99±0.0
PPV	0.04±0.01	0.04±0.01	0.05±0.01	0.06±0.01	0.07±0.01	0.09±0.01	0.10±0.02	0.11±0.03	0.09±0.06
NPV	0.97±0.03	0.98±0.01	0.98±0.01	0.98±0.0	0.97±0.01	0.97±0.01	0.97±0.01	0.96±0.01	0.96±0.01

Figures

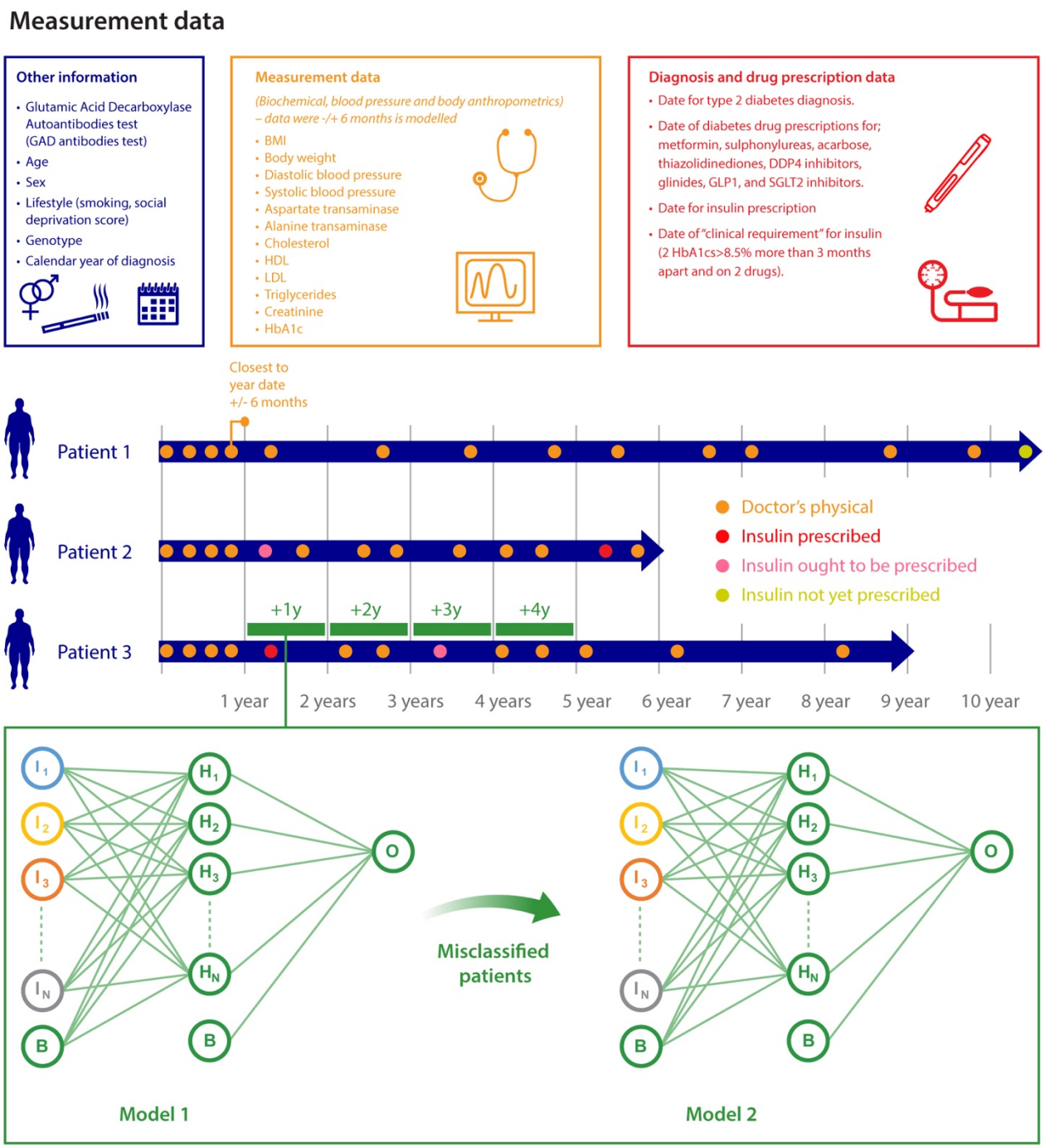


Figure 1: Type-2 diabetes management begins with lifestyle/diet intervention. Insufficient glucose management leads to drug initiation (often metformin) followed by a dual-therapy prescription and eventually insulin. In this study, an individual patient's journey towards insulin varies between 1 and

10 years or longer from initial diabetes diagnosis. Artificial neural networks (ANNs) modelled patients' treatment journey by predicting time to insulin +1, +2, +3, and +4 years based on information extracted from the electronical medical records from year 1 to year 10 after diagnosis with a one year prediction window (green bars as an example). Information collected one time around diagnosis was also included in the modelling. For year 1 to 10 the +1 year models were trained, for year 1 to 9 to +2year models were trained, for year 1 to 8 the +3year models were trained and for year 1 to 7 the +4 year models were trained. Mis-classified patients were extracted for retraining of an additional model.

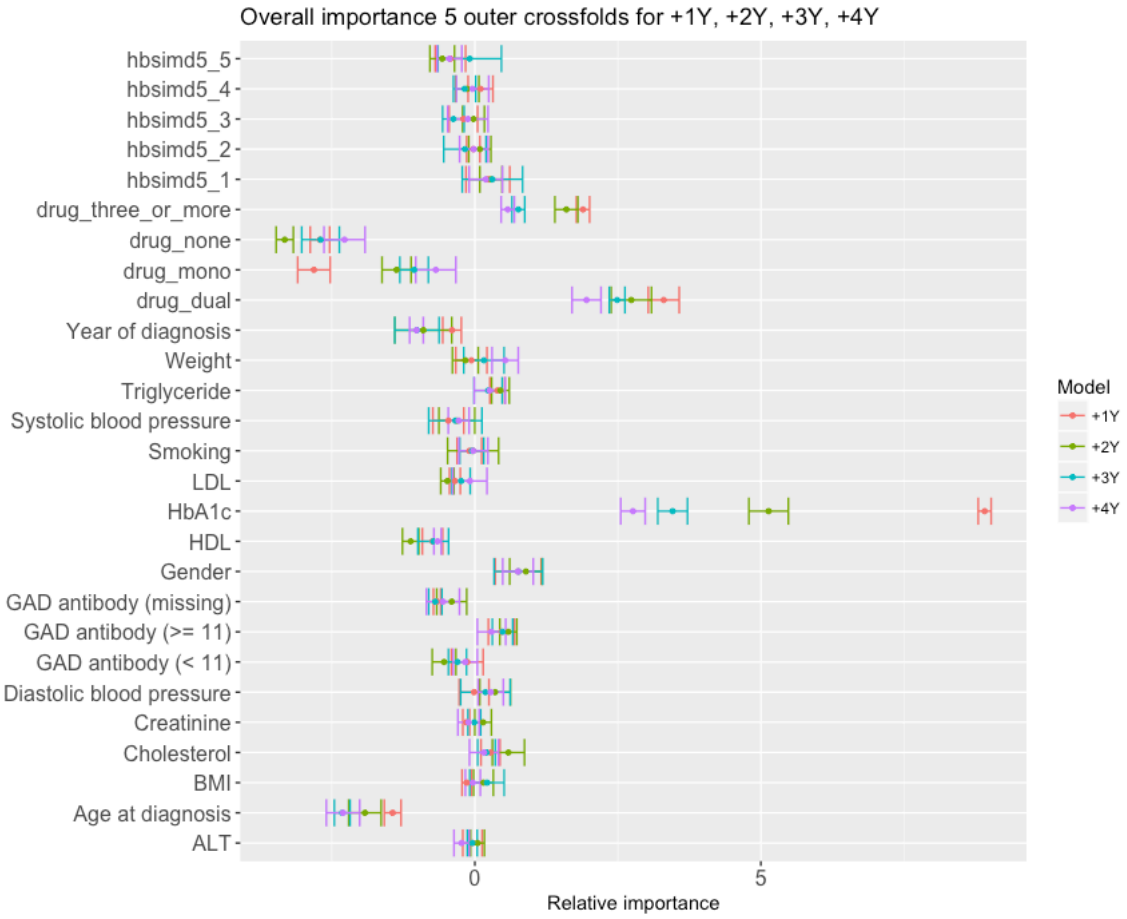


Figure 2: Relative feature importance given as mean and standard deviation of five outer cross-validation folds including the direction of clinical features used for training of the time to insulin. Positive values indicate association with progression to insulin requirement, while negative values indicate association with no progression on to insulin. The higher the value, the more important was the feature in the time to insulin models.

5.9 Epilogue: Genetics helps ... in some patients

The overall aim of this project has always been prediction of time to insulin. However, several different prediction outcomes and strategies for modelling longitudinal EMR data have been tested as we gained more experience from the data handling as well as clinical utility of different phenotypic outcomes.

One interesting setup worth mentioning in further details beyond the included manuscript is where a similar prediction outcome as presented in the manuscript was applied to predict the 1 to 4 years time horizon for insulin requirement. However, instead of using all patients available, the definition of controls (no insulin requirement) was restricted to only include patients that not yet had progressed on to dual therapy in the end of the one-year prediction window. This resulted in slightly more balanced prediction outcomes, but the definition is not an ideal clinical phenotype from a real world application perspective.

Longitudinal features were modelled similar to the data presented in the manuscript except for the drug prescriptions. Diabetes drug prescriptions were (at this stage of modelling) only considered at the time of diagnosis. The predictive performance was similar to the clinical baseline model reported in the manuscript (Table 5.1).

We also retrained a second model on all patients where the first clinical models' predictions were false (positive or negative) in order to capture information about misclassified patients. Similar to the presented manuscript, this again pointed to a group of patients in which high HbA1c is a major driver of progression and one where lower HbA1c and other clinical features are important for prediction of time to insulin. In the second clinical model, the patient profile was however less heterogeneous compared to the second clinical models in the presented manuscript. In this stratified second clinical model, other important features across +1Y, +2Y, +3Y and +4Y were earlier calendar year of diagnosis, higher HDL, older age at the time of diagnosis, and diabetes drug treatment at diagnosis. Furthermore, a linear effect of the social deprivation score was been in the +1Y and +2Y models, where least deprived patients (hbsimd5 5) had faster progression compared to most deprived patients (hbsimd5 1). Diabetes drug treatment at diagnosis was also important in the first clinical baseline models for +1Y, +2Y, +3Y and +4Y (Figures 5.3-5.6).

We included genotype information from the four GRS and genetic risk variants to the second clinical model as described in the manuscript. Integration of the GRS alone did not improve prediction of the clinical +1, +2, +3 and +4 year models (Table 5.1). However, using a forward selection strategy on the GRS and individual SNPs together with the clinical data, increased the performance significantly for the +3Y models (ROC-AUC increase from 0.67 to 0.82, $P = 0.001$) and the +4Y models (ROC-AUC increase from 0.70 to 0.82, $P = 0.02$). In order to assess stability of features driving the performance, we re-initialized training of these models across 50 random model initialization seeds

Table 5.1: Performance given as mean±standard deviation of all models using two-level cross-validation and down-sampling of training set. M1 refers to the first trained models, and M2 refers to the models retrained using FP and FN from M1. For M1 an average is reported for across all years from where data is extracted. Due to duplicated patients across multiple years, only one time point was included for the retrained models (M2). MCC: Matthews correlation coefficient, J: Youden's index.

	ROC-AUC	*Sensitivity*	*Specificity*	*MCC*	*J*
M1 clinical					
+1	0.82 ± 0.04	0.76 ± 0.10	0.71 ± 0.09	0.39 ± 0.11	0.47 ± 0.08
+2	0.71 ± 0.05	0.69 ± 0.09	0.62 ± 0.08	0.25 ± 0.10	0.30 ± 0.11
+3	0.67 ± 0.07	0.65 ± 0.08	0.60 ± 0.07	0.22 ± 0.10	0.25 ± 0.10
+4	0.68 ± 0.06	0.64 ± 0.09	0.60 ± 0.09	0.21 ± 0.10	0.24 ± 0.11
M2 clinical					
+1	0.72 ± 0.04	0.63 ± 0.07	0.72 ± 0.03	0.29 ± 0.05	0.36 ± 0.06
+2	0.77 ± 0.03	0.70 ± 0.06	0.71 ± 0.03	0.36 ± 0.06	0.41 ± 0.06
+3	0.68 ± 0.03	0.59 ± 0.05	0.66 ± 0.03	0.23 ± 0.05	0.25 ± 0.05
+4	0.68 ± 0.03	0.61 ± 0.05	0.65 ± 0.05	0.25 ± 0.07	0.26 ± 0.07
M2 Clinical + GRS					
+1	0.70 ± 0.01	0.61 ± 0.07	0.70 ± 0.03	0.25 ± 0.03	0.31 ± 0.05
+2	0.75 ± 0.05	0.69 ± 0.07	0.68 ± 0.04	0.33 ± 0.09	0.37 ± 0.10
+3	0.67 ± 0.01	0.62 ± 0.03	0.63 ± 0.04	0.24 ± 0.04	0.25 ± 0.04
+4	0.67 ± 0.02	0.64 ± 0.06	0.63 ± 0.05	0.25 ± 0.04	0.27 ± 0.04
M2 + forward selection SNPs and GRS					
+1	0.72 ± 0.04	0.65 ± 0.09	0.72 ± 0.05	0.30 ± 0.06	0.36 ± 0.08
+2	0.81 ± 0.02	0.71 ± 0.04	0.79 ± 0.05	0.47 ± 0.04	0.50 ± 0.03
+3	0.82 ± 0.02	0.70 ± 0.05	0.82 ± 0.04	0.52 ± 0.05	0.52 ± 0.05
+4	0.82 ± 0.04	0.74 ± 0.08	0.82 ± 0.04	0.55 ± 0.07	0.56 ± 0.08
M2 + forward selection SNPs and GRS 50 seeds, same CV					
+1	0.72 ± 0.01	0.62 ± 0.01	0.72 ± 0.01	0.28 ± 0.01	0.34 ± 0.02
+2	0.80 ± 0.01	0.72 ± 0.01	0.78 ± 0.01	0.46 ± 0.02	0.50 ± 0.02
+3	0.81 ± 0.01	0.71 ± 0.01	0.81 ± 0.01	0.50 ± 0.02	0.52 ± 0.02
+4	0.82 ± 0.005	0.72 ± 0.01	0.81 ± 0.01	0.52 ± 0.01	0.53 ± 0.01
M2 + forward selection SNPs 50 seeds, same CV					
+1	0.72 ± 0.01	0.62 ± 0.02	0.72 ± 0.01	0.28 ± 0.02	0.34 ± 0.02
+2	0.80 ± 0.01	0.72 ± 0.01	0.78 ± 0.01	0.46 ± 0.02	0.50 ± 0.02
+3	0.81 ± 0.01	0.71 ± 0.01	0.81 ± 0.01	0.50 ± 0.02	0.52 ± 0.02
+4	0.82 ± 0.005	0.72 ± 0.01	0.81 ± 0.01	0.52 ± 0.01	0.53 ± 0.01

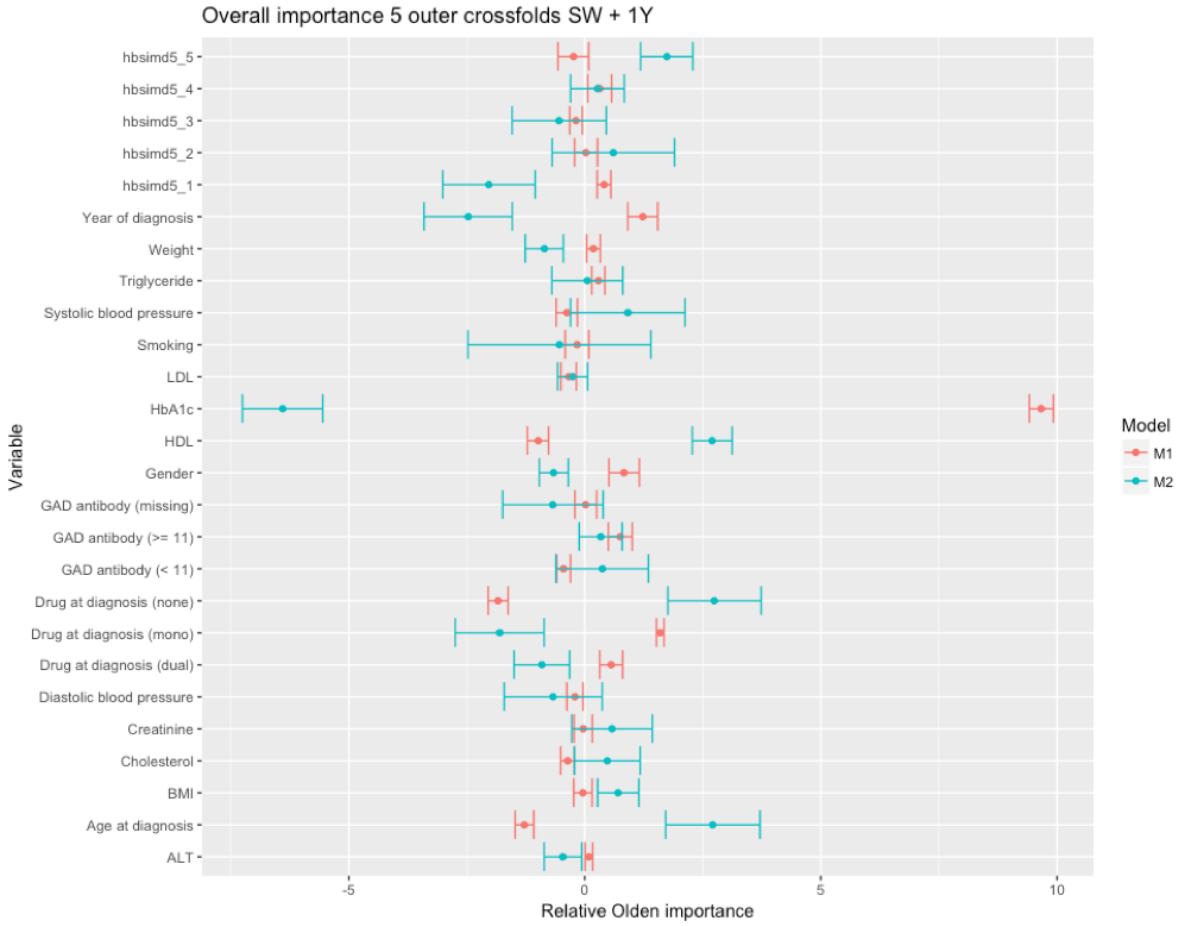


Figure 5.3: Relative feature importance given as mean±standard deviation of five outer cross-validation folds including the direction of clinical features used for training of model 1 and model 2 retained on unique patients that were incorrectly classified in M1 for +1Y predictions. M1 consisted of ten models trained at the time extracted for year 1 to 10 from after confirmed diagnosis of type 2 diabetes. The relative feature importance was summed across the ten models per outer cross-validation fold (different split of patients due to phenotype definition) and was divided with the number of models (10) before calculating the mean and standard deviation per feature.

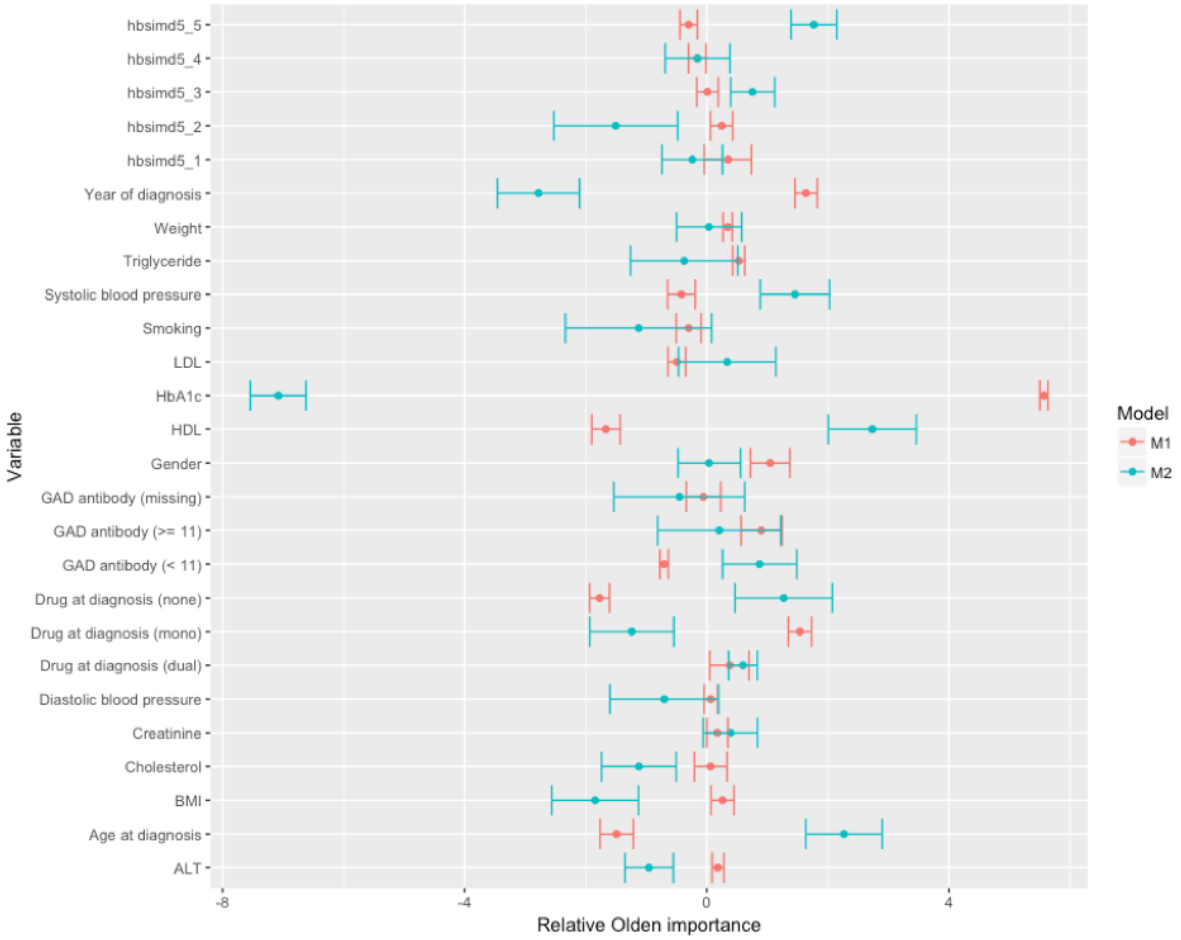


Figure 5.4: Relative feature importance given as mean±standard deviation of five outer cross-validation folds including the direction of clinical features used for training of model 1 and model 2 retained on unique patients that were incorrectly classified in M1 for +2Y predictions. M1 consisted of nine models trained at the time extracted for year 1 to 9 from after confirmed diagnosis of type 2 diabetes. The relative feature importance was summed across the nine models per outer cross-validation fold (different split of patients due to phenotype definition) and was divided with the number of models (9) before calculating the mean and standard deviation per feature.

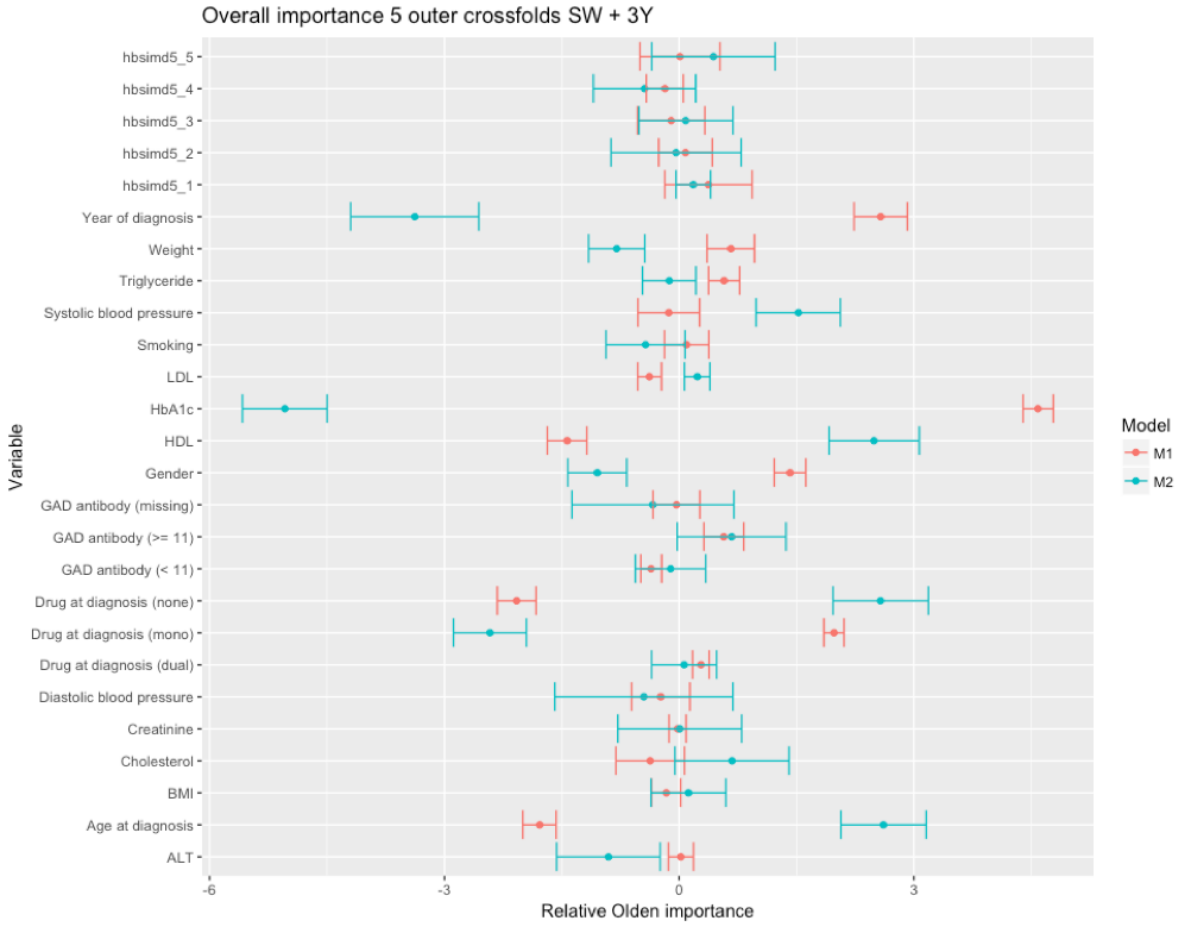


Figure 5.5: Relative feature importance given as mean±standard deviation of five outer cross-validation folds including the direction of clinical features used for training of model 1 and model 2 retained on unique patients that were incorrectly classified in M1 for +3Y predictions. M1 consisted of eight models trained at the time extracted for year 1 to 8 from after confirmed diagnosis of type 2 diabetes. The relative feature importance was summed across the eight models per outer cross-validation fold (different split of patients due to phenotype definition) and was divided with the number of models (8) before calculating the mean and standard deviation per feature.

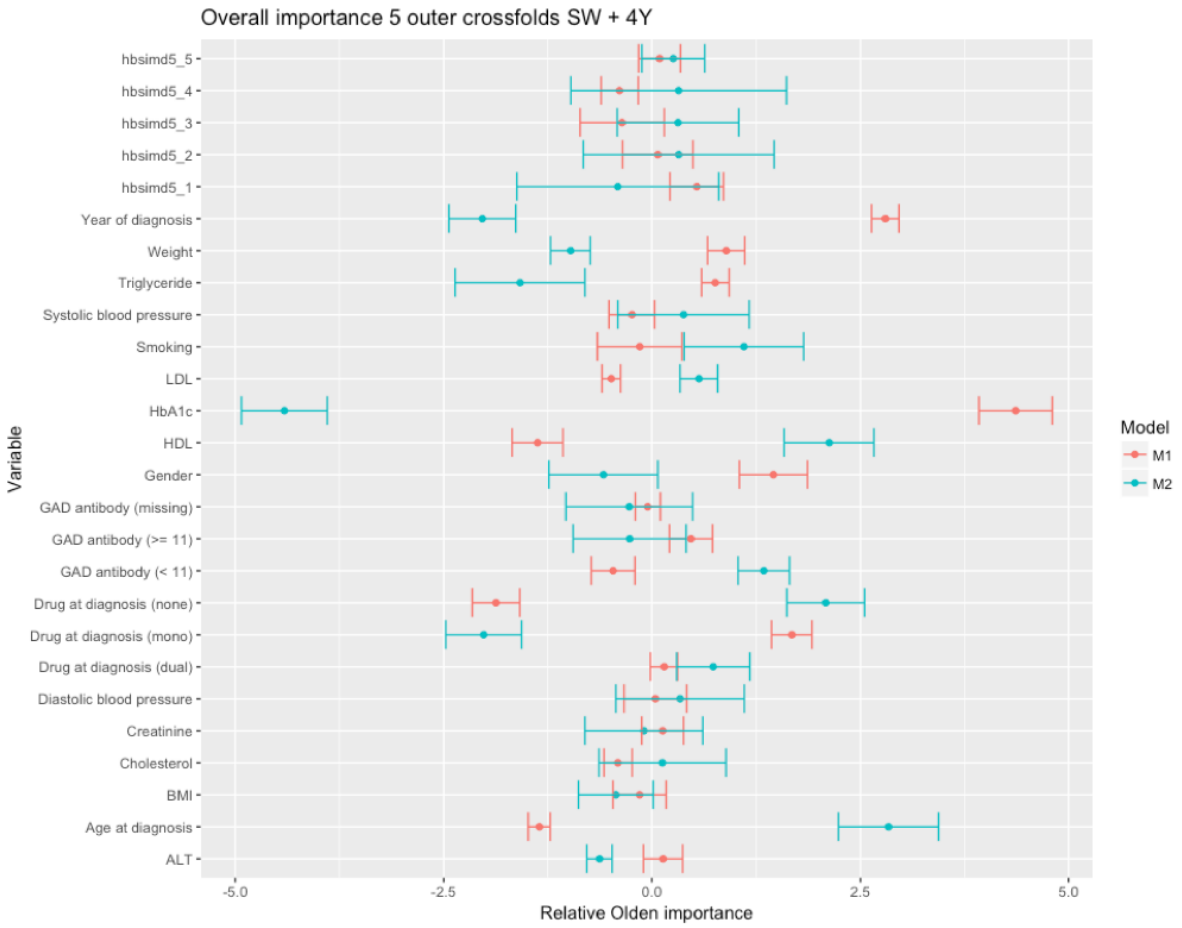


Figure 5.6: Relative feature importance given as mean±standard deviation of five outer cross-validation folds including the direction of clinical features used for training of model 1 and model 2 retained on unique patients that were incorrectly classified in M1 for +4Y predictions. M1 consisted of seven models trained at the time extracted for year 1 to 7 from after confirmed diagnosis of type 2 diabetes. The relative feature importance was summed across the seven models per outer cross-validation fold (different split of patients due to phenotype definition) and was divided with the number of models (7) before calculating the mean and standard deviation per feature.

resulting in ROC-AUC: 0.81 (±0.01) and 0.82 (±0.005) for the +3Y and +4Y models, respectively. The top 10 SNPs selected across the +3 and +4Y TTI:genetics models were rs1412234 (*LINGO2*, selected 48), rs11699802 (*CEBPB*, selected 37), rs1801212 (*WFS1*, selected 36), rs7987740 (*IRS2*, selected 31), rs10962 (*HNF1B*, selected 23), rs4946812 (*BEND3*, selected 23), rs12910825 (*PRC1*, selected 21), rs112498319 (*RREB1*, selected 20), rs11709077 (*PPARG*, selected 18) and rs28525376 (*THADA*, selected 17). SNPs selected by the forward selection during inner cross-validation model training across 50 different model initializations were investigated for pathway enrichment using GSEA[171], GOrilla[172], and STRING v11.0[173]. For GSEA, top 50 enriched pathways with a FDR q-value below 0.05 were reported. From the STRING database, only enriched pathways by Reactome were considered. For all pathway enrichment analyses, we compared the gene list from literature on T2D risk variants [159] to the genes annotated to any given SNP that were selected a minimum 7 times by the machine learning model. Enrichment by STRING and GSEA uncovered pathways involved in the insulin signalling and cancer pathways that were not enriched in the original gene list where the genetic risk variants were selected from for the +4Y predictions.

The impact of genetics on the rate of T2D would thus be interesting to investigate further. Here, we found the genetics especially improved the predictions in the long-term outcomes (+3 or +4 years ahead) indicating genetics may help in the prediction of time to insulin in some subgroups of patients.

CHAPTER 6

Clinical application II: Prediction of asparaginase-associated pancreatitis in childhood acute lymphoblastic leukemia

6.1 Childhood acute lymphoblastic leukemia

Childhood acute lymphoblastic leukemia (ALL) is a hematological cancer characterized by excess proliferation and malignant transformation of lymphoid precursor cells of B- or T-lymphocytes (lymphoblasts) in the bone marrow [174, 175]. Childhood ALL can be divided into immunophenotypes by the lymphoid linages, where approximately 85% develops from the B-cell linage and 15% develops from the T-cell linage [176]. Classification of childhood ALL is determined further by cytogenetics which determines chromosomal abnormalities [175–177]. Childhood ALL is the most common pediatric cancer and occurs in 25% of all cancers before the age of 15 [178]. However, it is a rare disease with the incidence rate being 4 in 100,000 children in the Nordic countries including Sweden, Denmark, Norway, Finland and Iceland [177]. The progression of the cancer is fast, hence referred to as acute, which is caused by rapid tumor cell proliferation. If left untreated, childhood ALL would be fatal within weeks. The peak of childhood ALL incidences are in children aged 2−5 years old [174]. Other risk factors of childhood ALL include a combination of male sex, genetic predisposition given ancestry and genetic syndromes, cytogenetic abnormalities, high socioeconomic status and environmental risk exposures [179, 180].

6.2 Treatment of childhood ALL

Childhood ALL is treated over an approximate two-year period. The treatment protocol typically involves three treatment phases [174]. Treatment is initialized by an intensive induction therapy (∼30 days) with the overall aim of eliminating malignant lymphoblasts and restoring normal blood cell formation in the bone marrow [174]. The second phase, consolidation therapy (or re-intensification, ∼60 days), aims to target remaining drug-resistant malignant lymphoblasts by introducing a different set of anti-leukemic drugs as treatment. Finally, the maintenance therapy (> 18 months) is introduced with the aim of preventing ALL relapse. If a patient responds poorly to initial treatment or are in a high risk (HR) group of childhood ALL, allogenic hemopoietic stem-cell transplantation is needed, which is the most intensive therapy for childhood ALL [174].

In the Nordic Society of Pediatric Hematology and Oncology (NOPHO) ALL-2008 treatment protocol, which is the treatment protocol used in Denmark, patients are stratified into a standard risk (SR) group, an intermediate risk (IR) group and a HR group that guides treatment. Patients are stratified three times; in the beginning of the induction phase, consolidation phase and maintenance phase. The first stratification relies on the immunophenotype and white blood cell count at diagnosis for induction stratification on prednisone or dexamethasone treatment. The patients are stratified into SR, IR and HR groups at the beginning of the consolidation therapy which is based on bone marrow minimal residual disease (MRD), central nervous system involvement and cytogenetics. The stratification in the beginning of the maintenance therapy is based on bone marrow MRD.

6.3 Survival versus treatment toxicity

Great improvements of childhood ALL treatment protocols and risk stratification have resulted in survival rates of more than 80% and up to 95% in low risk groups [178, 181]. However, the price for survival may be high, as survivors are burdened by treatment-associated toxicities that prolong hospital admissions, increase risk of long-term complications and are potentially life-threatening [182, 183]. It has been reported that almost 50% of all patients on the NOPHO ALL-2008 protocol experience serious adverse events during treatment [184]. Furthermore, the life expectancy in adult childhood cancer survivors is estimated to be compromised by treatment-related late effects several decades after diagnosis [185].

Given that survival rates of childhood ALL are high, focus is now shifting towards understanding and prediction of treatment-associated toxicities. This chapter focus on treatment with asparaginase and its related toxicity asparaginase-associated pancreatitis (AAP).

6.4 Asparaginase-associated pancreatitis

Asparaginase functions by depleting the non-essential amino acid asparagine (and glutamine) from circulation [186]. Healthy cells have asparagine synthetase and can produce asparagine, whereas malignant lymphoblasts only are capable of synthesizing asparagine in limited amounts. This compromises the cellular function of the tumor which eventually leads to cell death [186]. However, asparaginase also increases patients' risk of developing pancreatitis. AAP occurs in 2-18% of children given their treatment protocol and is associated with both the cumulative dose of asparaginase and the duration of asparaginase treatment across different treatment protocols [187]. Acute complications of AAP include respiratory or circulatory failure, insulin therapy, pancreatic pseudocysts or death, whereas more persistent complications involve abdominal pain and endocrine and exocrine dysfunctions of the pancreas [188]. Treatment protocols usually recommend truncation of asparaginase treatment if AAP develops, which increases the risk of relapse [181, 189]. Patients that are re-exposed to asparaginase after first AAP event have an almost 50% risk of developing a second AAP [188]. The chapter defines AAP based on consensus definitions determined by the Ponte di Legno Toxicity working group, where two out of three criteria had to be full-filled [182]:

- Abdominal pain
- Imaging consistent with pancreatitis
- Pancreatic enzymes (pancreatic lipase and pancreatic amylase) >3x upper normal limit

Risk factors associated with pancreatitis are older age at diagnosis, concomitant anti-leukemic treatment and genetic predisposition [187, 190–192]. However, AAP is currently a difficult toxicity to predict as established risk factors have modest effect sizes. Even harder to predict is second AAP following re-exposure to asparaginase after the first AAP event where currently no consensus guidelines exist on whether or not to re-expose a patient to asparaginase.

The genetic predisposition in AAP and non-asparaginase-associated pancreatitis have been suggested to be determined by genes involved in the trypsinogen activation pathway [192]. The acinar cells in the pancreas produce zymogens (inactive digestive enzymes) including trypsinogens. The inactive digestive enzymes are transported to the duodenum by a sodium bicarbonate-rich fluid produced by the duct cells where they are activated under normal physiological conditions [193]. Asparaginase is proposed to activate an influx of Ca^{2+} resulting in premature cleavage of trypsinogens (*PRSS1* and *PRSS2*) in the acinar cell resulting in pancreatitis [192] (Figure 6.1).

Activation of trypsinogens in the acinar cells results in an inflammatory response and several enzymes encoded by *SPINK1*, *CASR*, *CTRC* are protecting against pre-mature cleavage of trypsinogens. *CFTR*, encoding an epithelial cell anion channel involved in duct cell function, has also been suggested to be involved in pancreatitis [193].

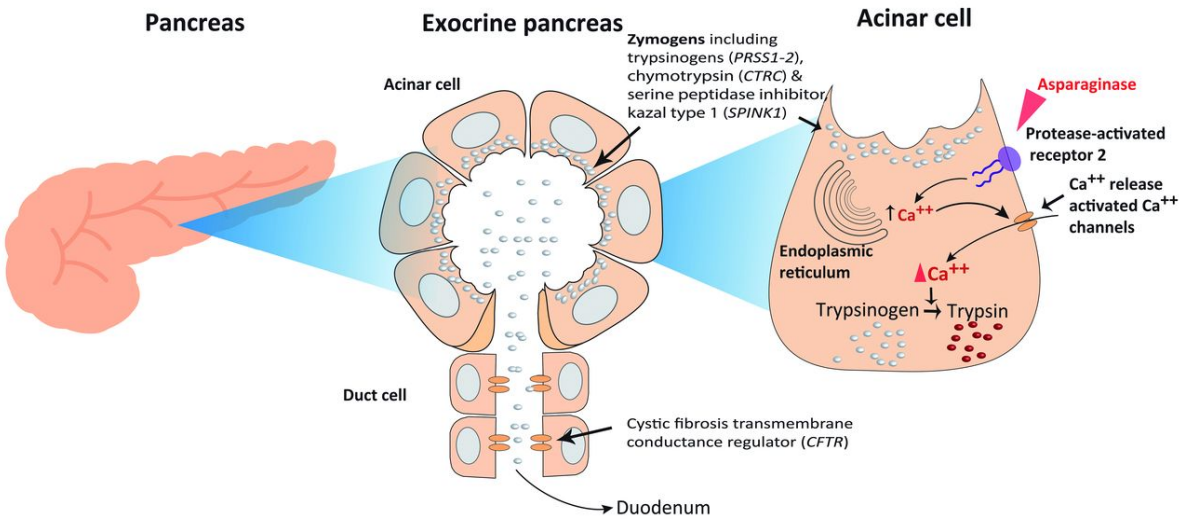


Figure 6.1: Effect of asparaginase (red triangle on acinar cell) on the trypsinogen activation pathway in the exocrine pancreas. Source: The figure was obtained from the Haematologica Journal website http://www.haematologica.org in Wolthers *et al*, 2019 [192] with permission granted from the Ferrata Storti Foundation, Pavia, Italy.

6.5 Study introduction

This chapter presents a machine learning-based strategy for prediction of patients at risk for AAP and risk for a second AAP following re-exposure to asparaginase after the first pancreatitis event. The prediction models were trained based on data collected by the Ponte di Legno toxicity working group, which resulted in the largest AAP cohort $N = 1564$ (244 AAP cases). The data included age, sex and SNP genotypes obtained across participating trial groups. This cohort was previously used for an AAP GWAS [192]. This study identified SNPs with insufficient effect sizes for clinical use. Since GWAS test each SNP independent and assume an additive effect, it may miss out on more complex interactions between SNPs as well as their genetic effect on phenotype. We thus used several different machine learning models, encoding of genetic features as well as SNP-based datasets to capture genotype-phenotype relationships and use this to predict individual risk of AAP. The most predictive models were used to describe predictive risk factors for AAP. Following validation in an independent cohort with similar characteristics, this model can function as an early screening tool that potentially can assist in guidance of treatment stratification of childhood ALL.

6.6 Bioinformatic challenges: exploring genotype-phenotype interactions

Risk factors of AAP have previously been identified through statistical methods, which have identfied risk factors with limited predictive ability for AAP. An important aspect to consider is that some phenotypes are influenced by multiple genetic variants in epistasis and the impact of SNPs discovered by GWAS are not necessarily predictive biomarkers for all patients' response to treatment. We thus hypothesized that machine learning would offer more insight for individual level predictions of AAP by allowing detection of more complex interactions in the tested SNPs. For integration of genotype data into machine learning models of AAP several considerations were made as described in the following sections.

6.6.1 Data imbalance

The cases (AAP) versus controls (no AAP) in the cohort were imbalanced. We found this was important for the predictive performance in a clinical baseline model only with age and sex. Without down-sampling, the models classified every sample as a control (specificity = 100% and sensitivity = 0%). Thus, we dealt with the imbalance using down-sampling. No overall difference appeared in the ROC-AUC when using only age and sex as features for the baseline models with or with-out down-sampling.

6.6.2 Population sub-structure

We initially trained models on the entire cohort $N = 1564$ (244 AAP cases). To account for population sub-structure, the first four principal components of the genotype were included as features in the model. However, the principal components of genetic ancestry appeared as the most predictive features, indicating the model possibly learned genetic ancestry profiles rather than separating AAP cases versus controls. Thus, only patients with CEU ancestry were included in the modelling. This reduced the number of patients included for analysis to $N = 1390$ (205 AAP cases) and possibly limited the clinical utility of the model. This should be explored through validation with different cohorts to understand the SNP effects.

6.6.3 Feature selection, reduction and encoding of SNPs

$\sim$1.4M SNPs from genotype arrays were available. A key challenge of this project was to increase the predictive performance of AAP from genetic variants. Several feature selection and reduction strategies were applied. Three main approaches were explored for

feature selection; i) SNPs from previous genetic studies of AAP, ii) data-driven feature selection, and iii) prior knowledge on adult pancreatitis (Figure 6.2).

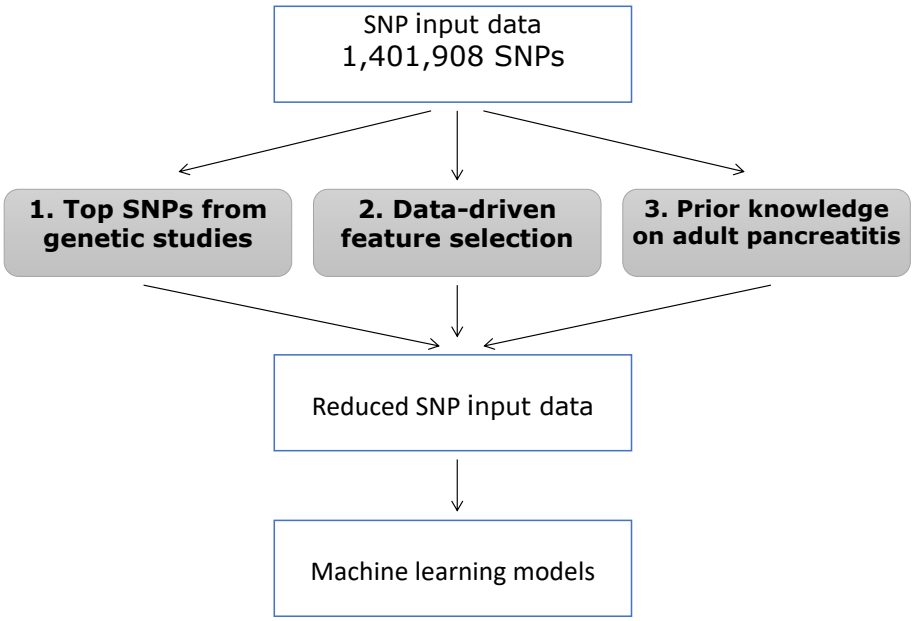


Figure 6.2: Feature selection strategies of SNPs for prediction of AAP.

The findings from three previous studies investigating the genotype-phenotype relationship of pancreatitis in ALL were used for selection of SNPs [187, 191, 192]. For the data-driven feature selection (*not described in the manuscript*), we performed a GWAS on the training dataset within the cross-validation framework and only included SNPs in the final model that passed a given threshold of the P-value or odds ratio. If a large set of SNPs were selected at the given threshold (>30 SNPs), the SNPs were further selected by a forward selection wrapper. A SNP was included if ROC-AUC increased with minimum 0.01 for maximum 30 iterations of the forward selection. As prior knowledge, we focused on eight candidate genes of pancreatitis (*PRSS1*, *PRSS2*, *SPINK1*, *CTRC*, *CASR*, *CFTR*, *CPA1* and *CLDN2*) [193]. SNPs that were annotated to these eight genes as well as associated eQTLs in pancreatic tissue were extracted as datasets to explore potential new SNPs in the prediction of AAP.

In general for all three selection strategies, if several SNPs (>30 SNPs) were selected for modelling, the datasets were further reduced. Different strategies for feature reduction included pruning of SNP-sets based on linkage disequilibrium to reduce SNP redundancy or through functional annotation and priority of the most severe variants. Thus, several defined sets of SNPs were used for modelling. These were further encoded by dominant, recessive, additive effect on the phenotypic trait or as binary by presence of the major or minor allele. The SNP datasets were also presented to the models by genetic principal components or a GRS.

6.6.4 Choice of machine learning algorithm

Different types of machine learning algorithms including logistic regression, random forest, AdaBoost and ANNs (both with one and two hidden layers) were trained to provide different decision boundary options. Testing different models would allow detection of any linear or non-linear relationship between genotype and phenotype. The ROC-AUC was however relatively stable across different types of models. The ROC-AUC was instead more dependent on the selected features and genotype feature encoding. Different sets of patients were correctly classified in different models by inspection of individual predictions. To capture the broadest biological view on AAP, some models were included for a personalized AI ensemble model to improve stability of the final prediction for AAP where clinical thresholds were suggested for detection of high risk patients.

6.6.5 Prediction of second AAP event after re-exposure

Prediction of a second AAP event was challenging given a very limited number of samples in the cohort being re-exposed to asparaginase ($N = 37$, 13 AAP cases). To train a machine learning model on this number of samples, we used leave-one-out cross-validation and limited the number of features. The model for second AAP following re-exposure should be re-trained to confirm initial performance and findings. More data is currently being collected on this phenotype by the Ponte di Legno Toxicity working group.

6.7 Manuscript

The following manuscript is submitted to *Haematologica*. The supplementary material is in Appendix C.

Prediction of asparaginase-associated pancreatitis cases in childhood acute lymphoblastic leukemia

Author list: Rikke L. Nielsen[1,2], Benjamin O. Wolthers[3], Marianne Helenius[1], Birgitte K. Albertsen[4], Line Clemmensen[5], Kasper Nielsen[6], Jukka Kanerva[7], Riitta Niinimäki[8], Thomas L Frandsen[3], Andishe Attarbaschi[9], Shlomit Barzilai[10], Antonella Colombini[11], Gabriele Escherich[12], Derya Aytan-Aktug[13], Hsi-Che Liu[14], Anja Möricke[15], Sujith Samarasinghe[16], Inge M van der Sluis[17], Martin Stanulla[18], Morten Tulstrup[3], Ester Zapotocka[19], Kjeld Schmiegelow[3,20] and Ramneek Gupta[1].

KS and RG have shared last authorship.

Affiliations

1. Department of Health Technology, Technical University of Denmark, Kgs. Lyngby, Denmark.
2. Sino-Danish Center for Education and Research, University of Chinese Academy of Sciences, Huairou, China.
3. Department of Pediatrics and Adolescent Medicine, University Hospital Rigshospitalet, Copenhagen, Denmark.
4. Department of Pediatric Oncology, Skejby Hospital, Aarhus, Denmark.
5. Department of Applied Mathematics and Computer Science, Kgs. Lyngby, Denmark.
6. Center for Biological Sequence Analysis, Technical University of Denmark, Kgs. Lyngby, Denmark.
7. Children's Hospital, Helsinki University Central Hospital, University of Helsinki, Helsinki, Finland.

8. Oulu University Hospital, Department of Children and Adolescents, and University of Oulu, PEDEGO Research Unit, Oulu, Finland.
9. Department of Pediatric Hematology and Oncology, St Anna Children's Hospital and Department of Pediatric and Adolescent Medicine, Medical University of Vienna, Austria.
10. Pediatric Hematology and Oncology, Schneider Children's Medical Center of Israel, Petah-Tikva, Israel and Sackler Faculty of Medicine, Tel Aviv University, Israel.
11. Department of Pediatrics, Ospedale San Gerardo, University of Milano-Bicocca, Fondazione MBBM, Monza, Italy.
12. University Medical Center Eppendorf, Clinic of Pediatric Hematology and Oncology, Hamburg, Germany.
13. Department of Bioinformatics, Technical University of Denmark, Kgs. Lyngby, Denmark.
14. Division of Pediatric Hematology-Oncology, Mackay Memorial Hospital, Taipei, Taiwan.
15. Christian-Albrechts-University Kiel and University Medical Center Schleswig-Holstein, Department of Pediatrics, Kiel, Germany.
16. Great Ormond Street Hospital for Children, London, UK.
17. Dutch Childhood Oncology Group, The Hague and Princess Máxima Center for Pediatric Oncology, Utrecht, the Netherlands.
18. Department of Pediatric Hematology and Oncology, Hannover Medical School, Germany.
19. University Hospital Motol, Department of Pediatric Hematology/Oncology, Prague, Czech Republic.
20. Institute of Clinical Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Denmark.

Running heads (34/50)

Prediction of AAP in childhood ALL

Contact information for correspondence

Kjeld Schmiegelow, Kjeld.Schmiegelow@regionh.dk

Ramneek Gupta, ramg@dtu.dk

Word count

Abstract: 250/250

Main text: 4080/4000

Tables/Figures in Main text: 8/8

Supplementary file(s); 1

References: 26/50

Acknowledgments

The authors would like to thank Olga Rigina for functional annotation and extraction of prioritized genetic variants from Ensembl. Moreover, we thank all the researchers who scrutinized patient files and completed phenotype questionnaires, and organizational support from the research staff at Bonkolab, at the University Hospital Rigshospitalet. Lastly, we thank the Bloodwise Childhood Leukaemia Cell Bank, UK, for providing samples and data for this research.

Funding

This study was funded by the Kirsten and Freddy Johansen Foundation, the Danish Childhood Cancer Foundation, the Swedish Childhood Cancer Foundation and the Danish Cancer Society, The Nordic

72 Cancer Union, The Otto Christensen Foundation, University Hospital Rigshospitalet, and The Novo
73 Nordic Foundation.

Abstract

Asparaginase-associated pancreatitis (AAP) frequently affects children treated for acute lymphoblastic leukemia (ALL) protocols and causes severe acute and persisting complications. Known risk factors such as asparaginase dosing, older age and germline variants have insufficient odds ratios to allow personalized asparaginase therapy. We integrated information on age, sex, and genotypes based on Illumina Omni2.5exome-8 BeadChip arrays of 1,390 childhood ALL patients aged 1.0−17.9 years, including 205 with AAP, into machine learning models to build robust predictive classifiers of AAP risk. We applied logistic regression, random forest, AdaBoost, and artificial neural networks. A clinical baseline model with age and sex had an area under the receiver operating characteristic curve (ROC-AUC) of 0.62, while inclusion single nucleotide polymorphisms (SNPs, N=30) identified through our the largest GWAS cohort on AAP boosted performance to 0.78−0.81, while six SNPs in candidate genes of adult pancreatitis boosted ROC-AUC to only 0.67. Most predictive SNPs in our models included rs1505495 (*GALNTL6*), rs4655107 (*EPHB2*), rs13228878 (*PRSS2, PRSS3P2*), and rs10436957 (*CTRC*). Several predictive models of AAP were combined into a personalized artificial intelligence ensemble model, where detection of patients at high risk of AAP had positive predictive value of 0.95 at a sensitivity of 0.37 in the validation dataset suggesting surveillance of these. Second AAP (N=13) following re-exposure to asparaginase (N=37) could be predicted with ROC-AUC:0.63 but requires validation. The machine learning models can enable individual-level high risk assessment of AAP for future intervention trials, and also decisions on asparaginase re-exposure after AAP, when risk is predicted to be low.

Introduction

Asparaginase is an essential drug in the treatment of childhood acute lymphoblastic leukemia (ALL) associated with increased survival rates[1]. By depleting circulating asparagine levels, malignant lymphoblasts are targeted for apoptosis, due to limited capacity for re-synthesis of asparagine[2]. Asparaginase use is, however, associated with significant treatment related toxicities[3] out of which pancreatitis (AAP) occurs in 2−18% of patients[4], mostly in older children and adults[5], which often leads to truncation of therapy, potentially increasing the risk of relapse[1,6]. On the other hand, re-exposure to asparaginase after AAP has been associated with an almost 50% risk for a second AAP, mostly after several doses of pegylated asparaginase (PegAsp)[7]. Previous studies have identified higher age at diagnosis[4], co-administration of other cancer drugs[8], and host genome variants[4,9,10] as risk factors for AAP. However, these risk factors are currently not used to individualize treatment with asparaginase, as they only have modest effect sizes for clinical decision support. To address this challenge, we integrated genetic data of single nucleotide polymorphisms (SNPs) from the largest childhood ALL AAP case-control cohort (N=1564, including 244 AAP cases) into machine learning models to classify patients at very high risk of AAP and explored the predictiveness of SNPs previously associated with AAP. We and others have previously applied machine learning modelling to identify childhood ALL patients at high risk of relapse[11,12]. Individual level predictions across several machine learning models can be compared to improve understanding of relevant risk factors associated with a high risk of AAP, and potential subgroups that are predictable by separate models. By identifying patients at high confidence for risk of AAP, this analysis may lead to increased monitoring or selecting patients who should undergo less intensive asparaginase therapy for childhood ALL. We also tested if the use of machine learning models performed equally well for the risk of developing a second AAP when re-exposed to asparaginase.

Methods

Patients

To map AAP phenotypes and identify significant host genome variant associated with AAP risk, the international Ponte di Legno (PdL) toxicity working (PTWG) group collected post remission blood samples from 1564 children (aged 1.0–17.9 years) with newly diagnosed t(9;22)-negative ALL between June 1, 1996, and January 1, 2016 as described previously[10]. All patients received asparaginase according to their respective treatment protocols. The applied diagnostic criteria for AAP stated that two of the three following international consensus criteria must be fulfilled: i) amylase, pancreatic amylase, or pancreatic lipase >3x UNL, ii) abdominal pain, or iii) imaging compatible with AAP[13]. To eliminate population sub-structure, we only included patients of European ancestry (CEU, N=1390 of the 1564) whereof 205 patients developed AAP (cases) and 1185 did not (controls). DNA was genotyped on Illumina Omni2.5exome-8 BeadChip arrays. After quality control as previously described[10], the data collected by PTWG consisted of 1,401,908 SNPs, age and sex. A subset of the PTWG cohort originated from the Nordic Society of Pediatric Hematology and Oncology (NOPHO) study group with included 815 controls who had received 15 doses of Pegylated asparaginase (1,000 IU/ml intramuscularly) and 77 cases of AAP. On these patients, additional clinical biomarkers such as anthropometrics, ALL risk group (standard (SR), intermediate (IR) or high risk (HR)), white blood cell count at ALL diagnosis, minimal residual disease (MRD), and the cumulative amount of asparaginase dosages were available (Supplementary material S1).

Machine learning training, feature importance and validation

For modelling of AAP, logistic regression, random forest, AdaBoost and artificial neural networks (incl. one and two hidden layers) models were fitted using python (version 3.6.8)[14] with Scikit-learn (version 0.21.3)[15]. Feature importance was evaluated by a ‘leave-one-out’ approach on the features.

143 Performance was evaluated in an independent test set of 100 samples including all the 37 patients re-
144 exposed to asparaginase after truncation and 63 further randomly selected samples from the PTWG
145 cohort, extracted prior to training of the machine learning models (Supplementary material S1). Each
146 model provides an individual prediction score (0−1) corresponding to the probability of AAP ranging
147 0−100%.
148
149 *Processing and annotation of genetic variants*
150 Data pre-processing of genotypes was done using R (version 3.2.5)[16] and PLINK (version
151 plink2/1.90beta3)[17]. Genetic variants were annotated to genes ±50kB of a gene boundary using
152 Variant Effect Predictor GCRCh37[18].
153
154 *Genetic feature representation and selection strategies*
155 To maximize learning from genetic data, SNPs were represented by additive, dominant and recessive
156 genetic encodings as well as a non-additive manner according to the presence of the major allele or
157 minor allele. SNPs were selected by different strategies to test their predictiveness of AAP (Figure
158 1). Feature selection included SNPs previously associated with AAP[4,9,10] and prior knowledge on
159 eight candidate genes in pancreatitis[19] and their expression quantitative trait loci (eQTLs) from the
160 GTEx biobank[20]. SNPs annotated to the eight candidate genes (with minor allele frequency >5%)
161 were reduced to three principal components (Supplementary material S1). Furthermore, a genetic risk
162 score (GRS) was calculated based on six SNPs rs17107315 (*SPINK1*), rs56296320 (*CFTR*),
163 rs12853674 (*CLDN2*), rs13228878 (*PRSS*), rs16832787 (*CASR*) and rs10436957 (*CTRC*) identified
164 as most significant in candidate genes of pancreatitis[19] in the PTWG genome-wide association studies
165 (GWAS)[10]. rs13228878 was previously validated in the AALL0232 cohort[10], while rs12853674 has

been identified in GWAS of adult pancreatitis[21], and rs17107315, rs13228878 and rs10436957 have been identified in GWAS of both childhood AAP[10] and adult pancreatitis[21].

Ensemble model

An ensemble of prediction models was created in order to capture multiple aspects of AAP biology without overly increasing the complexity of any individual model (Supplementary material S1). The ensemble model was scored by three different approaches i) average mean scoring, ii) majority voting or iii) average mean scoring of confident individual predictions i.e. the score should be ≤0.35 or ≥0.65 to make a count in the final prediction.

Results

Clinical baseline AAP models

A clinical baseline risk model of AAP was first established using only age and sex as features. The machine learning models were trained on a subset of the study cohort with European ancestry (N=1290, whereof 155 patients developed AAP). The age of patients with AAP was significantly higher than those without (cases: 8.7±4.8 controls: 6.2±4.5, P=9.26e-11), while no significant difference was found for sex (P= 0.68). The remaining subset (N=100 of the 1390) of patients in the study cohort was used for validation. Clinical baseline AAP prediction models were trained using logistic regression, random forest, AdaBoost, and artificial neural networks (including one and two hidden layers), which all resulted in ROC-AUC of 0.62±0.01 (Supplementary Table S2).

Integration of AAP-associated genetic variants

SNPs previously associated with AAP were integrated into the clinical baseline machine learning models. SNPs associated with AAP were obtained from three previous genetic studies by; Liu *et al*

2016[4], Abaji *et al* 2017[9] and Wolthers *et al* 2019[10], to test the predictive performance. SNPs identified by Liu *et al* 2016 and Abaji *et al* 2017 did not change ROC-AUC compared to the clinical baseline model (ROC-AUC: 0.61−0.63 for Liu *et al* 2016, ROC-AUC: 0.60−0.62 for Abaji *et al* 2017 (Supplementary Table S3)). Wolthers *et al* 2019 presents the most powerful GWAS given the largest cohort on AAP[10]. Using the top thirty AAP associated SNPs by Wolthers *et al* 2019 resulted in ROC-AUC: 0.78−0.81 (Table 1). The performance appeared to be independent of the type of machine learning model used, as well as of additive, dominant or sparse encoding of the genetic variants. However, using the recessive encoding of genetic features resulted in multiple near-zero variance predictors due to very few homozygous recessive alleles, and thus lower ROC-AUC: 0.67−0.70 (Table 1). Several prediction models of AAP trained on the thirty strongest associated AAP risk variants, age and sex resulted in similar ROC-AUC, thus the artificial neural network (one hidden layer) with sparsely encoded genetic risk variants (ROC-AUC: 0.80, P=1.16e-208 compared to the clinical baseline AAP model across 100 model initializations) was chosen for further investigation (ROC curve and permutation test in Supplementary Figures S4A and S4B). The most important features of this model were the minor alleles of rs1505495 (*GALNTL6*) and rs4655107 (*EPHB2*) (Figure 2). The feature importance plots further showed that individual SNPs have minor impact on predictive performance indicating a combination of several genetic components are needed to achieve ROC-AUC: 0.80 (Figure 2). The performance of this model was robust across 100 model initializations compared to a permuted outcome label of AAP with significantly higher ROC-AUC for the true AAP-labelled models (P=1.12e-113, Supplementary Figure 4B). This model was significantly more confident in AAP patients with higher age (≥6 years, mean [95% CI] of individual risk of AAP: 0.74 [0.69−0.78]) compared to children younger than 6 years (mean, 95%CI of individual risk of AAP: 0.58 [0.52−0.65], P = 0.0001).

Integration of genetic risk variants in pancreatitis pathways

Recent AAP GWAS discovered shared genetic predisposition between asparaginase-associated pancreatitis and non-asparaginase-associated pancreatitis through similar pathways[10]. We thus explored predictability using SNPs annotated to eight genes involved in development of pancreatitis, i.e. *PRSS1, PRSS2, SPINK1, CTRC, CASR, CFTR, CPA1,* and *CLDN2*[19]. Datasets for modelling were based on SNPs annotated to these eight genes by linkage disequilibrium-pruning selection of all SNPs annotated to the genes (eight genes dataset) or reducing the variance of all SNPs to three principal components (PCA dataset). Genetic variants were further expanded by eQTLs of these eight genes in pancreatic tissue (eight genes eQTL dataset). Furthermore, within the eight genes in the pancreatitis pathway, six candidate SNPs previously associated with AAP were selected: rs17107315 (*SPINK1*), rs56296320 (*CFTR*), rs12853674 (*CLDN2*), rs13228878 (*PRSS2* and *PRSS3P2*), rs16832787 (*CASR*) and rs10436957 (*CTRC*). These were modelled as separate SNPs (six candidate SNPs dataset) and by a weighted polygenic risk score (PRS dataset). Additionally, the top four genome-wide SNPs by Wolthers *et al* 2019[10] and SNPs annotated to the *PRSS1/PRSS2* locus were prioritized for modelling (2 genes, chromosome 20 dataset). The predictive performance for models with age, sex and the selected SNP datasets ranged with ROC-AUC: 0.47−0.67 (Supplementary Table S3). Overall, the best performance was achieved from the six candidate SNPs, age and sex model with ROC-AUC: 0.67 (ROC curve in Supplementary Figure 4C), which were significantly improved when comparing the artificial neural network model across 100 model initializations to the clinical baseline model (P= 2.72e-113). The most important features of this model were rs13228878 (*PRSS2* and *PRSS3P2*), the minor allele of rs10436957 (*CTRC*) and the age group of 1−7 years (Figure 3). All features in the models - when individually dropped - induced only minor changes to the ROC-AUC and thus the value was in their combination for the final models. The ROC-AUC of the six candidate SNPs models

was robust across 100 model initializations compared to a permuted outcome label of AAP with significantly higher ROC-AUC for the true AAP models (P = 9.04e-94, Supplementary Figure S4D).

Validation of models

A test dataset left out from the original training of the models was used for validation of the AAP risk models (N=100). The age of patients with AAP was significantly higher than those without (cases: 8.66±5.01, controls: 6.73±4.89, P=0.04) and no significant difference in the sex distribution (P=0.11). Performances were validated for the most successful machine models and are presented in Tables 1 and 2. The models based on the thirty SNPs previously associated with AAP, age and sex had ROC-AUC: 0.79−0.85 (Table 1). The models based on six SNPs that were most significant across eight candidate genes of adult pancreatitis[19] in the PTWG genome-wide association studies (GWAS), age and sex (six candidate SNPs models) had ROC-AUC: 0.62−0.66 (Table 2).

Personalized artificial intelligence ensemble model

The most predictive models compared to the clinical baseline model (ROC-AUC≥0.62) with different genetic encoding and features capturing different subsets of patients were integrated into an ensemble model. This ensemble model was composed of 18 models across six genetic feature subsets out of a total 600 possible (Table 3). For establishing a joint prediction score on each patient, the scores of the individual models within an ensemble are combined via a) averaging, b) majority voting and c) averaging only on confident scores (as described in Methods). Best performance was seen for the confident averaging (at a score threshold of ≤0.35 or ≥0.65) ROC-AUC: 0.84 on the cross-validation set (N=1290) and ROC-AUC: 0.83 on the independent test set (N=100) (Figures 4A and 4D). The confident scores were considered for a personalized AI ensemble. The ROC-AUC thus slightly improved from the best model with age, sex and thirty previously associated AAP SNPs

(ROC-AUC=0.78−0.81). For most of the individual predictions, models with the 30 SNPs associated with AAP were highly confident compared to other models in the 'personalized AI' ensemble. However, the combined prediction in the 'personalized AI' ensemble helped correct previously false predictions for five AAP cases and 47 controls. In situations where the predicted outcome was not influenced, the personalized AI ensemble provided more confidence to many of the correct predictions and reduce confidence levels for incorrect predictions by moving the combined prediction score closer to the class threshold indicating lower confidence. Figures 4B, 4C, 4E and 4F depict the prediction scores against their true class as well as estimations of sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) for the 'personalized AI' ensemble model and test data set performance.

In addition to the overall performance scores of the combined 'personalized AI' ensemble model, we also assessed how good the model was in the extremes of its score distribution. Patients at a high risk of AAP were identified by setting the prediction threshold of 0.8 in the trained model and test performance reported in Figure 4B and Figure 4C, respectively. In the cross-validation, the sensitivity was 42% with a PPV of 45%, where 143 patients (both case/controls) had prediction threshold ≥0.8, whereof 64 cases are correctly predicted in the trained models (Figure 4C). On the independent test dataset (N=100), a prediction threshold of 0.8 resulted in sensitivity of 37% and a PPV of 95% where 19 patients (both cases/controls) had a prediction score ≥0.8 whereof 18 AAP cases are correctly predicted (Figure 4F). It was not possible to validate the personalized AI ensemble model trained on the first AAP event cases to predict 2nd AAP following re-exposure to asparaginase (ROC-AUC: 0.53, N=37, 13 second AAP cases). The model falsely predicted the group of no-second AAP as cases (Figure 4H). This indicate the models only applies to the first event of AAP. It was only possible to validate the first AAP models on the independent hold-out set for second AAP (N=37) using

models including the six SNPs that were most significant across eight candidate genes of pancreatitis[19] in the PTWG genome-wide association studies (GWAS), age and sex (six candidate SNPs models) (Table 2).

Second AAP following re-exposure to asparaginase

As the personalized AI ensemble did not validate the patients that were re-exposed, we explored training of this subset of patients in a separate model as the phenotype between first AAP and second AAP patients may differ. Thirty-seven patients were re-exposed to asparaginase and thirteen patients developed a second AAP. We trained models separately on these 37 patients using leave-one-out cross validation. The input data was age, sex and the six candidate SNPs or thirty previously SNPs associated with AAP with sparse encoding of genetic features. The best performing model was the random forest with the six candidate SNPs with ROC-AUC: 0.63, sensitivity: 0.40 and specificity: 0.76 (Supplementary Table S.5). The most important feature was the minor allele of the variant rs13228878 (*PRSS2* and *PRSS3P2*) (Figure 5). Sex, rs12853674 (*CLDN2*) rs16832787 (*CASR*) and the age group 1−7 years were other important predictive features. The distribution of the true outcome AAP labels is significantly different from the random AAP outcome labels in the permutation test (P=1.04e-19, Supplementary Figure S.6).

Discussion

Asparaginase is an essential drug for ALL therapy, and truncation of therapy due to e.g. pancreatitis, hypersensitivity, serious thrombosis, or silent inactivation has in several studies been associated with an increased risk of relapse. Thus, there is a currently unmet need to identify patients at high risk of such adverse events or to guide clinicians on when re-exposure to asparaginase is likely to be safe.

In this study, we successfully integrated multiple SNPs and clinical features in machine learning models to provide a sufficiently strong model that could, when validated by other groups, be clinically applicable for identification of patients with very high risk of AAP (minimum 80%).

Several methodologies were employed in this study to model asparaginase-associated pancreatitis (AAP) in children diagnosed with ALL (aged 1−17.9). SNPs were added to the clinical baseline model of age and sex (ROC-AUC of 0.62±0.01) using a variety of feature selection strategies from a pool of ~1.4M SNPs. Improvements from six candidate SNPs in adult pancreatitis drove performance up from ROC-AUC: 0.62 to 0.67 and with thirty previously associated AAP SNPs to ROC-AUC: 0.81 indicating that germline genetic profiling can significantly assist in the prediction of some patients at risk of AAP, given a basic clinical model. The two single best performing models trained by six SNPs in pancreatitis candidate genes from the PTWG genome-wide association studies (GWAS)[10] or thirty previously associated AAP variants in a study by Wolthers *et al*[10] performed significantly better compared to a random permuted outcome label of AAP (P=1.12e-113 for model with previously associated AAP variants and P=9.04e-94 for model with six candidate SNPs in pancreatitis). The performance of both models was validated in an independent test set with ROC-AUC: 0.84 for the model with thirty previously associated AAP variants and with ROC-AUC: 0.64 for the model with six candidate SNPs in adult pancreatitis. Despite both these models of AAP being validated on a held-out test set, validation on an external data set must be used prior to adaption of such machine learning models in clinic. A limitation of the feature selection for the best performing model of the thirty most significantly SNPs associated with AAP, is that these SNPs were identified through a GWAS on the same dataset used for training and testing of the machine learning models in this study. Therefore, the feature selection is not completely independent. Of the thirty SNPs selected from the Wolthers *et al* study, only rs13228878 and rs10273639 were validated in another cohort previously[10]. It was not possible to obtain similar performance with other genetic variants identified

in prior studies by Liu *et al*[4] or Abaji *et al*[9] of AAP in childhood ALL and these models only obtained ROC-AUC of 0.60−0.63. This is possibly due to the cohorts of patients being very different, both in asparaginase exposure and diagnostic criteria of AAP, or more likely reflected false positive findings as none of the those SNPs reached GWAS significance[4,9]. The thirty SNPs identified in Wolthers *et al* represents findings from the largest AAP GWAS providing the strongest power to detect true findings.

Feature importance of these two genetic clinical models of AAP showed that a combination of all features was required to achieve clinically useful performance as each feature had minor impact when being left out on the ROC-AUC performance indicating an interaction between features. In the model based on thirty previously associated AAP variants, the most important features included the minor alleles of rs1505495 and rs4655107, which are annotated to *GALNTL6* and *EPHB2*, respectively. *GALNTL6* encodes polypeptide N-acetylgalactosaminyltransferase-like 6, which is a transferase-like enzyme involved in the posttranslational process of O-linked glycosylation responsible for transferring N-acetylgalactosamine to an exposed serine or threonine[22,23]. *EPHB2* encodes the transmembrane EPH receptor B2, which part of the largest family of tyrosine kinase receptors and capable of bidirectional signaling through binding with ephrin ligands on neighboring cells. This signaling is involved in developmental processes such as cell and axon growth, as well as involved in cancers[24,25]. In the model based on six SNPs in candidate genes of adult pancreatitis, the most important features involved younger age (1−7), rs13228878 and rs10436957. The variant rs13228878 (*PRSS2* and *PRSS3P2*) is located in the *PRSS1-PRSS2* locus which encode trypsinogens that can be cleaved into trypsin in order to activate digestive enzymes prematurely leading to cases of AAP[19]. *PRSS3P2* is a pseudogene and its function are difficult to determine. The minor allele of rs13228878 was previously found to reduce risk of AAP[10]. The minor allele rs10436957 is annotated to *CTRC*

355 which encodes the enzyme chymotrypsin C that helps regulate the activation and degradation of
356 trypsinogens[10,19].
357 AAP predictions was improved by an ensemble model approach where 18 of all trained models were
358 included, which helped correctly re-classify some wrongly predicted patients in the training and
359 testing datasets as well as increased the confidence of the true prediction class. In order to correctly
360 classify the cases of AAP with high confidence, a score threshold of t=0.8 was applied, resulting in
361 37% of the true cases to be correctly predicted, with false positives limited to only 5% on the test set.
362 These thus identified over one third of the AAP cases with very high confidence.
363 Even more important is the need for identifying patients who are likely to tolerate re-exposure to
364 PegAsp after their 1st AAP episode. Re-exposure with PegAsp after an episode of AAP is currently
365 one of the most critical questions associated with asparaginase therapy, since truncation of therapy
366 has been associated with an increased risk of relapse[1,6]. However, neither the AAP phenotype,
367 including severity of the first AAP, or the age of patients are sufficiently strong risk factors to guide
368 the decision. Similarly, no single SNP has sufficient power for predicting the risk of a 2nd AAP. It
369 was not possible to apply the ensemble to validate findings only of the set of patients that experienced
370 re-exposure to asparaginase after first AAP. Validation of the subset of patients was only possible in
371 models including the six SNPs that were most significant across eight candidate genes of adult
372 pancreatitis[19] in the PTWG genome-wide association studies (GWAS), age and sex. Thus, separate
373 models were trained on re-exposure patients. The present study has limited power for prediction of
374 second AAP, but the PdL group is currently collecting very detailed data on more than 100 patients
375 re-exposed with PegAsp after AAP of whom approximately 40% are expected to develop AAP. Since
376 a 2nd episode of AAP usually occurs after several doses of PegAsp, future developments of this tool
377 could increase the number of patients that will be re-exposed to asparaginase and thus avoid the risk
378 of relapse by identifying low-risk patients. Currently, consensus guidelines do not exist, and decisions

to re-expose a patient will thus reflect physicians' attitudes and gut feeling, and the balance between anticipated risks of a 2nd AAP versus leukemic relapse.

The potential clinical utility of the models should be evaluated in the light of predictive performance as well as their interpretability of features which is an important challenge to address for adaptation into clinic[26]. The machine learning models learn patterns from data which can be complex and non-linear and achieve good predictive performance, while the feature importance – especially with complex feature interaction – at the individual patient level can be harder to identify. On the path towards clinical translation of an AAP prediction model, several additional aspects need to be considered in relation to the current treatment guidelines. Since asparaginase is an essential drug in the treatment of childhood ALL, the model should primarily identify patients with a very high risk of developing AAP, which could guide surveillance as approximately 10% of patients need transferal to intensive care unit and mechanical ventilation and 2% of patients die from AAP[7]. In addition, the risk score can enrich the cohort of patients recruited for pre-emptive interventions.

Lowering the asparaginase dose will impair anti-leukemic efficacy and is not feasible. Furthermore, for clinical implementation it is also important to know the timing of when a patient will develop AAP. However, if the patients can tolerate more asparaginase before possibly developing AAP, it becomes more difficult to determine at what timing asparaginase therapy should be truncated. The NOPHO subset of patients (N=892, whereof 77 had AAP) had more clinical features available, including a randomization group of asparaginase therapy with either 2- or 6-weeks treatment intervals. The total number of doses given was the most important feature as previously shown[4] compared to age, sex and SNPs (*data not shown in paper*). A suggested follow-up study is integrating the number of asparaginase doses in a time-dependent machine learning model on AAP together with the identified predictive SNPs to determine the timing of a patient's risk of AAP and if this changes during treatment.

In conclusion, this study supports the very strong role of host-genome variants on risk of AAP and exemplifies strategies for applying such modelling on other severe acute toxicities to ALL therapy[13].

Author contributions

Conception or design of the work: RLN, BOW, KS, RG

Acquisition of data: BOW, BKA, JK, RN, TLF, AA, SB, AC, GE, HL, AM, SS, IMS, MS, EZ, KS

Analysis of the data: RLN, BOW, MH, KN, DAA, MT

Interpretation of the data: RLN, BOW, MH, LC, KS, RG

Drafted the manuscript: RLN

Revision of manuscript: All authors.

References

1. Pieters R, Hunger SP, Boos J, et al. L-asparaginase treatment in acute lymphoblastic leukemia: a focus on Erwinia asparaginase. Cancer 2011;117(2):238–249.
2. Müller HJ, Boos J. Use of L-asparaginase in childhood ALL. Crit Rev Oncol Hematol 1998;28(2):97–113.
3. Hijiya N, van der Sluis IM. Asparaginase-associated toxicity in children with acute lymphoblastic leukemia. Leuk Lymphoma 2016;57(4):748–757.
4. Liu C, Yang W, Devidas M, et al. Clinical and genetic risk factors for acute pancreatitis in patients with acute lymphoblastic leukemia. J Clin Oncol 2016;34(18):2133–2140.
5. Rank CU, Wolthers BO, Grell K, et al. Asparaginase-Associated Pancreatitis in Acute Lymphoblastic Leukemia: Results From the NOPHO ALL2008 Treatment of Patients 1-45 Years of Age. J Clin Oncol 2020;38(2):145–154.
6. Gupta S, Wang C, Raetz EA, et al. Impact of asparaginase discontinuation on outcome in childhood ALL: A report from the Children's Oncology Group (COG). J Clin Oncol 2019;37(15_suppl):10005.
7. Wolthers BO, Frandsen TL, Baruchel A, et al. Asparaginase-associated pancreatitis in childhood acute lymphoblastic leukaemia : an observational Ponte di Legno Toxicity Working Group study. Lancet Oncol 2017;18(9):1238–48.
8. Raja RA, Schmiegelow K, Frandsen TL. Asparaginase-associated pancreatitis in children. Br J Haematol 2012;159(1):18–27.
9. Abaji R, Gagné V, Xu CJ, et al. Whole-exome sequencing identified genetic risk factors for asparaginase-related complications in childhood ALL patients. Oncotarget 2017;8(27):43752–43767.
10. Wolthers BO, Frandsen TL, Patel CJ, et al. Trypsin-encoding PRSS1-PRSS2 variations

influence the risk of asparaginase-associated pancreatitis in children with acute lymphoblastic leukemia: A ponte di legno toxicity working group report. Haematologica 2019;104(3):556–563.

11. Wesołowska-Andersen A, Borst L, Dalgaard MD, et al. Genomic profiling of thousands of candidate polymorphisms predicts risk of relapse in 778 Danish and German childhood acute lymphoblastic leukemia patients. Leukemia 2015;29(2):297–303.

12. Pan L, Liu G, Lin F, et al. Machine learning applications for prediction of relapse in childhood acute lymphoblastic leukemia. Sci Rep 2017;7(1):1–9.

13. Schmiegelow K, Attarbaschi A, Barzilai S, et al. Consensus definitions of 14 severe acute toxic effects for childhood lymphoblastic leukaemia treatment: a Delphi consensus. Lancet Oncol 2016;17(6):e231–e239.

14. Python Software Foundation. Python, version 3.6.8. https://www.python.org/ (accessed January 6, 2020).

15. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res 2011;122825–2830.

16. R Core Team. R: A Language and Environment for Statistical Computing. https://www.r-project.org/ (2016, accessed January 6, 2020).

17. Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience;4(1):.

18. McLaren W, Gil L, Hunt SE, et al. The Ensembl Variant Effect Predictor. Genome Biol 2016;17(1):1–14.

19. Zator Z, Whitcomb DC. Insights into the genetic risk factors for the development of pancreatic disease. Therap Adv Gastroenterol 2017;10(3):323–336.

20. The GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. Nat Genet

2013;45(6):580–585.

21. Rosendahl J, Kirsten H, Hegyi E, et al. Genome-wide association study identifies inversion in the CTRB1-CTRB2 locus to modify risk for alcoholic and non-alcoholic chronic pancreatitis. Gut 2018;67(10):1855–1863.

22. Steen P Van Den, Rudd PM, Dwek RA, Opdenakker G. Concepts and Principles of O-Linked Glycosylation. Crit Rev Biochem Mol Biol 1998;33(3):151–208.

23. Uniprot. https://www.uniprot.org/uniprot/Q49A17. https://www.uniprot.org/uniprot/Q49A17 (accessed November 22, 2019).

24. Himanen J, Chumley MJ, Lackmann M, et al. Repelling class discrimination: ephrin-A5 binds to and activates EphB2 receptor signaling. Nat Neurosci 2004;7(5):501–509.

25. Uniprot. https://www.uniprot.org/uniprot/P29323. https://www.uniprot.org/uniprot/P29323 (accessed November 22, 2019).

26. Prosperi M, Min JS, Bian J, Modave F. Big data hurdles in precision medicine and precision public health. BMC Med Inform Decis Mak 2018;18(139):1–15.

Tables

Table 1: The ROC-AUC performances are reported as average test performance when training the models on N=1290 samples (with 155 AAP cases) across 100 model initializations in five-fold cross-validation based on sex, age and top 30 SNPs reported by Wolthers *et al* 2019[11]. All models are trained with down-sampling on the control group within the cross-validation folds. ANN = artificial neural networks. Test performance was reported holdout dataset with N=100 patients (50 cases and controls) and a subset of these N=37 that were re-exposed to asparaginase (13 AAP cases).

Data type	**Model**	**ROC-AUC (N=1290)**	**ROC-AUC Test (N=100)**	**ROC-AUC Test 2nd AAP (N=37)**
Top 30 P-value SNPs (Additive encoding of genetics)	Logistic regression	0.80 ± 0.01	0.84 ± 0.01	0.56 ± 0.03
	Random forest	0.79 ± 0.01	0.83 ± 0.01	0.54 ± 0.03
	AdaBoost	0.81 ± 0.01	0.84 ± 0.01	0.52 ± 0.03
	ANN (1 hidden layer)	0.81 ± 0.01	0.84 ± 0.01	0.54 ± 0.03
	ANN (2 hidden layers)	0.80 ± 0.01	0.83 ± 0.02	0.56 ± 0.04
Top 30 P-value SNPs (Dominant encoding of genetics)	Logistic regression	0.79 ± 0.01	0.81 ± 0.01	0.57 ± 0.03
	Random forest	0.78 ± 0.01	0.79 ± 0.01	0.55 ± 0.04
	AdaBoost	0.80 ± 0.01	0.79 ± 0.01	0.54 ± 0.03
	ANN (1 hidden layer)	0.80 ± 0.01	0.80 ± 0.01	0.56 ± 0.03
	ANN (2 hidden layers)	0.79 ± 0.01	0.80 ± 0.02	0.56 ± 0.03
Top 30 P-value SNPs (Recessive encoding of genetics)	Logistic regression	0.67 ± 0.01	0.75 ± 0.01	0.55 ± 0.02
	Random forest	0.68 ± 0.01	0.73 ± 0.01	0.55 ± 0.02
	AdaBoost	0.70 ± 0.01	0.76 ± 0.01	0.55 ± 0.02
	ANN (1 hidden layer)	0.69 ± 0.01	0.76 ± 0.01	0.52 ± 0.02
	ANN (2 hidden layers)	0.68 ± 0.01	0.76 ± 0.01	0.53 ± 0.02
Top 30 P-value SNPs (Sparse	Logistic regression	0.78 ± 0.01	0.85 ± 0.01	0.55 ± 0.04
	Random forest	0.79 ± 0.01	0.84 ± 0.01	0.55 ± 0.03
	AdaBoost	0.81 ± 0.01	0.85 ± 0.01	0.53 ± 0.03
	ANN (1 hidden layer)	0.80 ± 0.01	0.84 ± 0.01	0.55 ± 0.04

encoding of genetics)	ANN (2 hidden layers)	0.79 ± 0.01	0.83 ± 0.02	0.54 ± 0.04

Table 2: The ROC-AUC performances are reported as average performance when training the models on N=1290 samples (with 155 AAP cases) across 100 model initializations in five-fold cross-validation based on sex, age and six candidate SNPs. All models are trained with down-sampling on the control group within the cross-validation folds. ANN = artificial neural networks. Test performance was reported holdout dataset with N=100 patients (50 cases and controls) and a subset of these N=37 that were re-exposed to asparaginase (13 AAP cases).

Data type	**Model**	**ROC-AUC (N=1290)**	**ROC-AUC Test (N=100)**	**ROC-AUC Test 2nd AAP (N=37)**
Six candidate SNPs (Additive encoding of genetics)	Logistic regression	0.66 ± 0.01	0.66 ± 0.01	0.61 ± 0.02
	Random forest	0.64 ± 0.01	0.65 ± 0.02	0.60 ± 0.02
	AdaBoost	0.66 ± 0.01	0.65 ± 0.01	0.61 ± 0.02
	ANN (1 hidden layer)	0.67 ± 0.01	0.65 ± 0.01	0.61 ± 0.02
	ANN (2 hidden layers)	0.66 ± 0.01	0.65 ± 0.01	0.62 ± 0.02
Six candidate SNPs (Dominant encoding of genetics)	Logistic regression	0.66 ± 0.01	0.63 ± 0.01	0.64 ± 0.02
	Random forest	0.65 ± 0.01	0.63 ± 0.01	0.59 ± 0.03
	AdaBoost	0.67 ± 0.01	0.63 ± 0.01	0.62 ± 0.02
	ANN (1 hidden layer)	0.66 ± 0.01	0.62 ± 0.01	0.64 ± 0.02
	ANN (2 hidden layers)	0.66 ± 0.01	0.63 ± 0.01	0.64 ± 0.02
Six candidate SNPs (Recessive encoding of genetics)	Logistic regression	0.63 ± 0.01	0.63 ± 0.01	0.61 ± 0.02
	Random forest	0.63 ± 0.01	0.62 ± 0.01	0.61 ± 0.02
	AdaBoost	0.64 ± 0.01	0.63 ± 0.01	0.59 ± 0.02
	ANN (1 hidden layer)	0.63 ± 0.01	0.62 ± 0.01	0.58 ± 0.02
	ANN (2 hidden layers)	0.63 ± 0.01	0.63 ± 0.01	0.60 ± 0.02
Six candidate SNPs (Sparse encoding of genetics)	Logistic regression	0.66 ± 0.01	0.65 ± 0.01	0.61 ± 0.02
	Random forest	0.65 ± 0.01	0.63 ± 0.02	0.59 ± 0.03
	AdaBoost	0.67 ± 0.01	0.64 ± 0.01	0.60 ± 0.02

	ANN (1 hidden layer)	0.67 ± 0.01	0.64 ± 0.01	0.60 ± 0.01
	ANN (2 hidden layers)	0.67 ± 0.01	0.64 ± 0.01	0.59 ± 0.01

Table 3: The ROC-AUC performances of models included in the ensemble models. These are reported for training the models on N=1290 samples (with 155 AAP cases) for each of the included model initializations in five-fold cross-validation. All models are trained with down-sampling on the control group within the cross-validation folds. Test performance was reported holdout dataset with N=100 patients (50 cases and controls) and a subset of these N=37 that were re-exposed to asparaginase (13 AAP cases).
The model initializations to include were chosen as the top 3 initializations that performed above ROC-AUC: 0.63. ANN = artificial neural network. Init = model initialization seed.

Data type	**Model**	**ROC-AUC (N=1290)**	**ROC-AUC Test (N=100)**	**ROC-AUC Test 2nd AAP (N=37)**
Genetic risk scores	AdaBoost (init 5708456)	0.66	0.62	0.60
	AdaBoost (init 7078673)		0.62	0.60
	AdaBoost (init 3612365)		0.63	0.59
Eight genes GTEx eQTL (dominant encoding of genetics	AdaBoost (init 1714803)	0.64	0.57	0.48
	AdaBoost (init 5770619)		0.58	0.42
	AdaBoost (init 9686361)		0.56	0.44
Two genes, chromosome 20 (dominant encoding of genetics)	AdaBoost (init 1110460)	0.69	0.64	0.54
	AdaBoost (init 2396987)		0.62	0.58
	AdaBoost (init 9149732)		0.66	0.53
Eight genes (recessive	AdaBoost (init 4855124)	0.65	0.61	0.62
	AdaBoost (init 3612365)	0.65	0.62	0.55
	AdaBoost (init 499914)	0.64	0.62	0.63

encoding of genetics)				
Six candidate SNPs (sparse encoding of genetics)	ANN, 1 hid layer (init 9149732) ANN, 1 hid layer (init 1166941) ANN, 1 hid layer (init 7078673)	0.70 0.69 0.69	0.65 0.64 0.66	0.59 0.58 0.57
Wolthers *et al* 2019 (sparse encoding of genetics)	ANN, 1 hid layer (init 3698379) ANN, 1 hid layer (init 6730428) ANN, 1 hid layer (init 6022674)	0.83 0.82 0.82	0.84 0.83 0.81	0.47 0.47 0.62

Figure legends

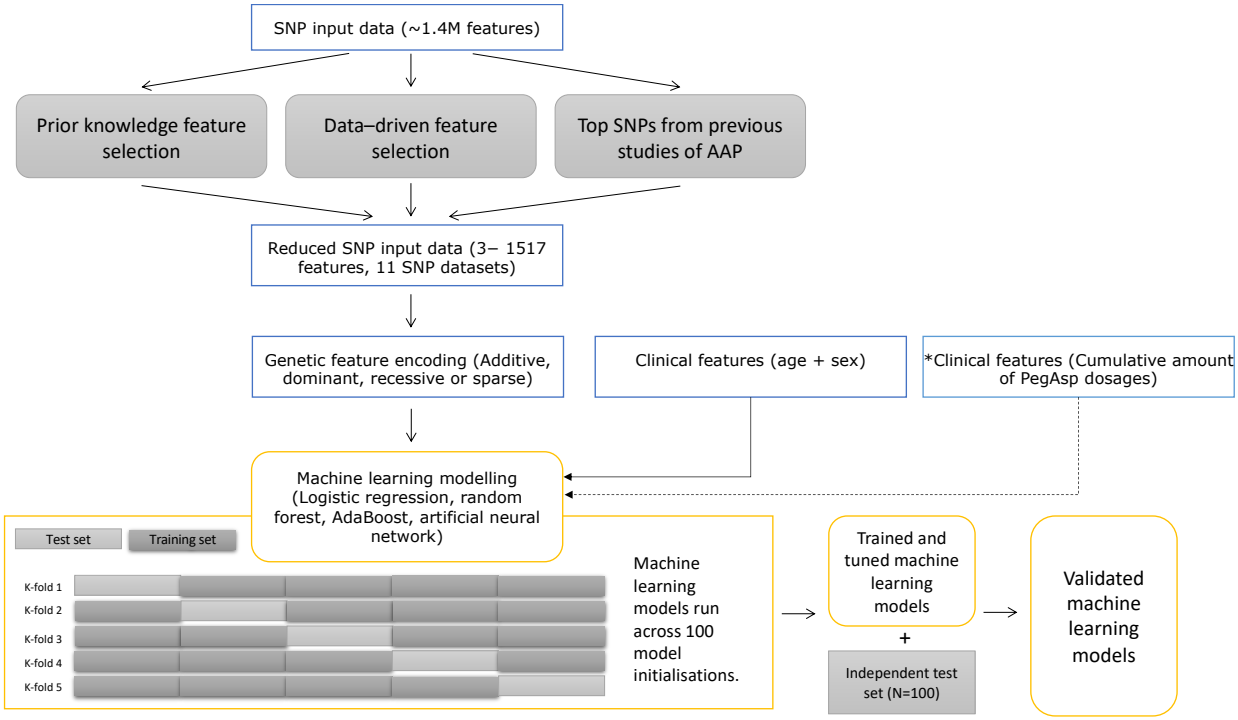


*Figure 1: Overview of the feature selection and machine learning strategies used in the study *A future model would benefit from inclusion of the cumulative dosage of PegAsp. In this study, it was only available on a subset of patients and was thus not fully explored. Age and sex were always included in modelling.*

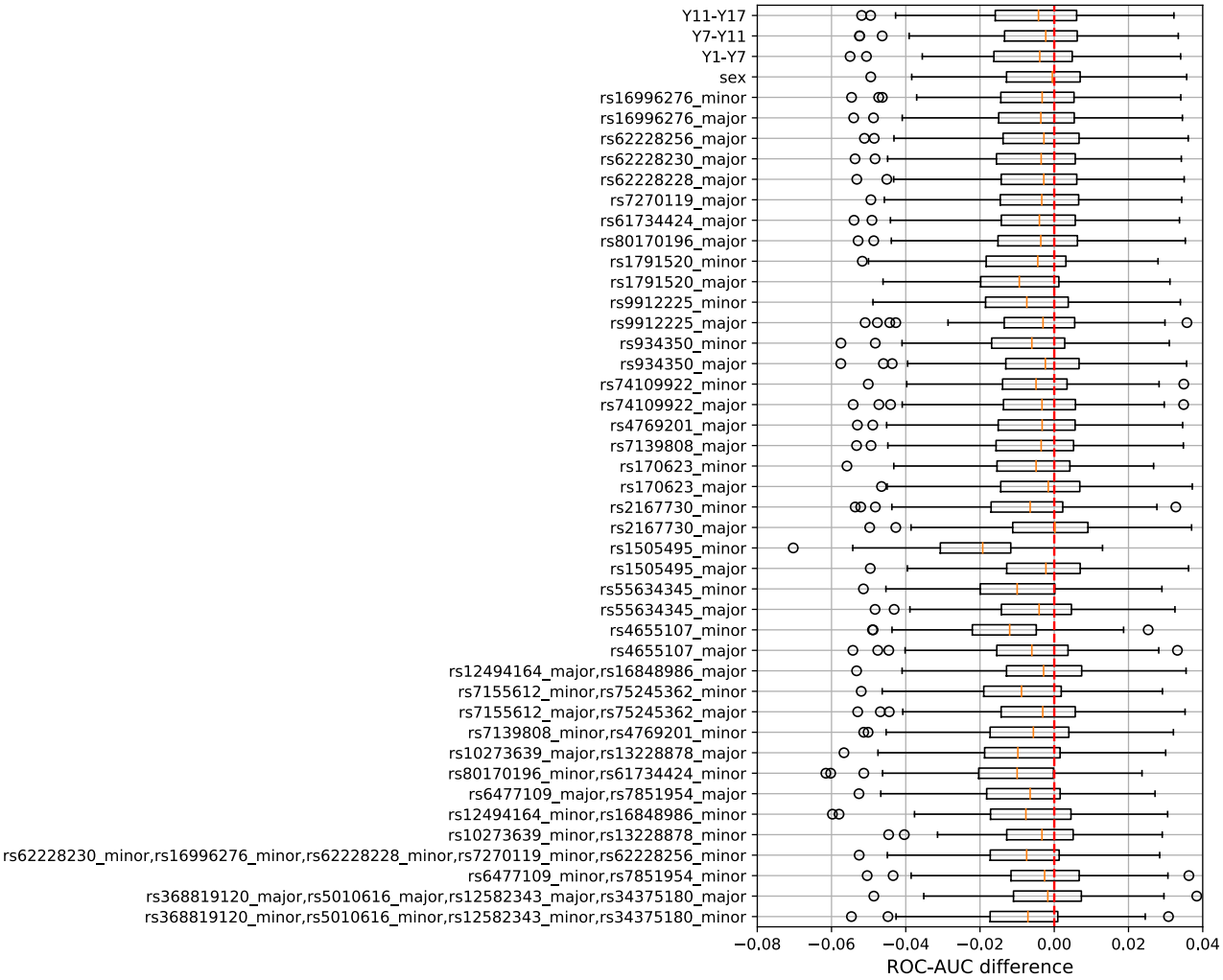


Figure 2: **Leave-one-out feature importance for AAP risk model** using artificial neural network with one hidden layer trained on age, sex and top 30 SNPs associated with AAP from Wolthers *et al* 2019 GWAS[11] with sparse encoding of genetics (N=1290).

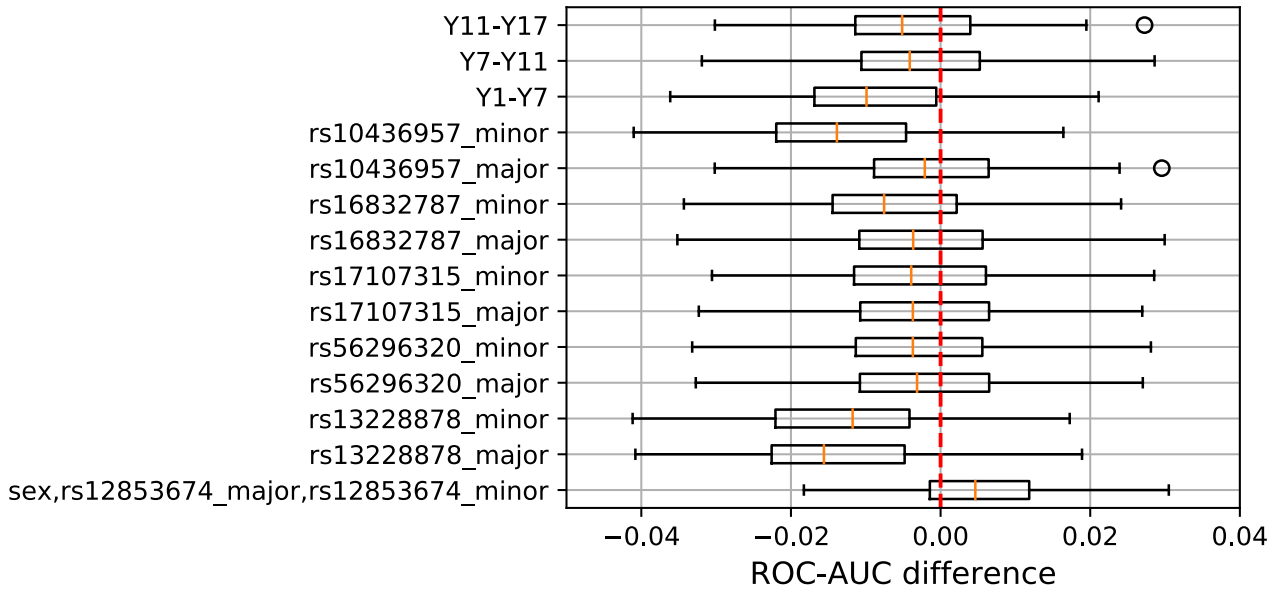
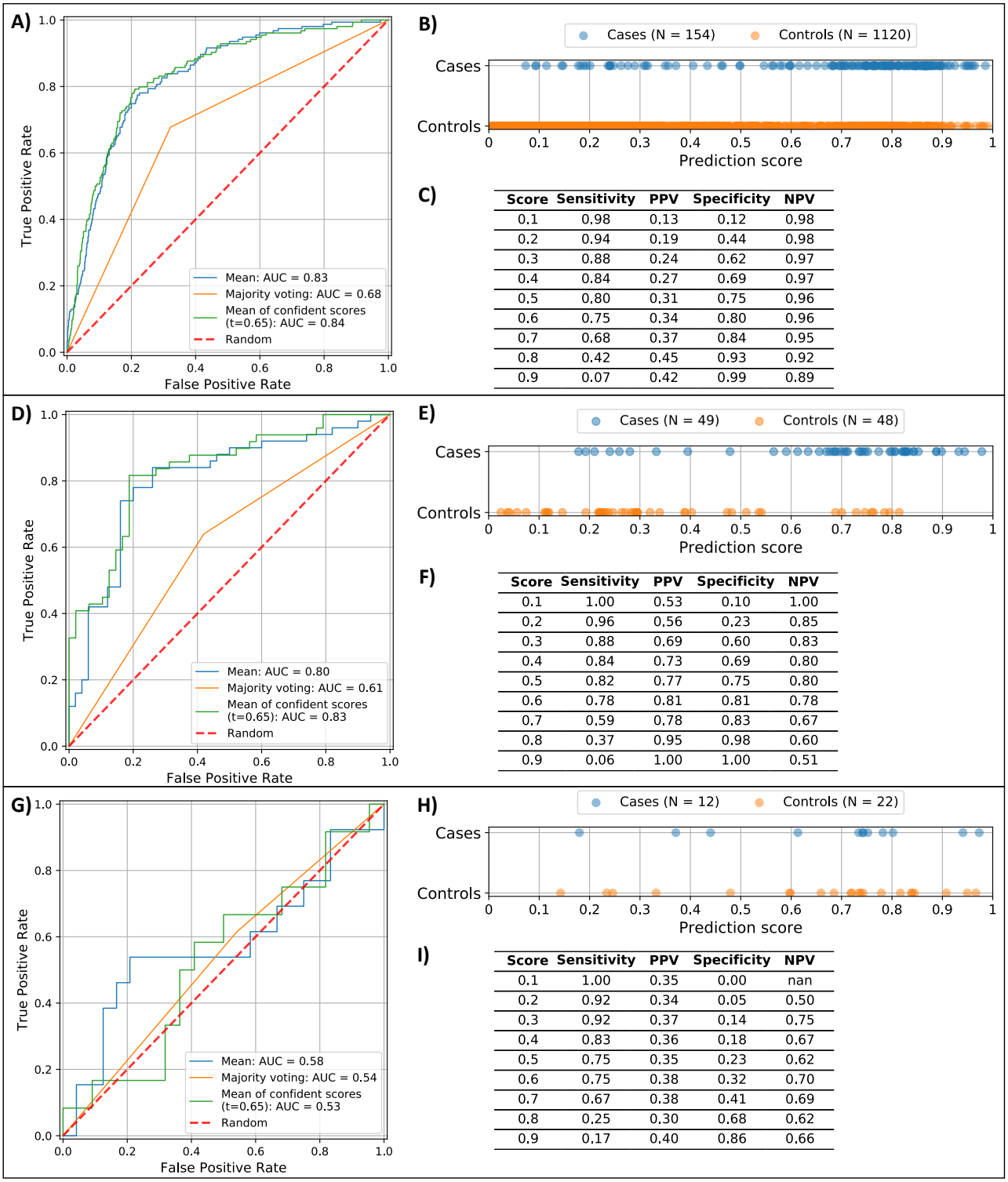


Figure 3: **Leave-one-out feature importance for AAP risk model** from the artificial neural network using age, sex and six candidate SNPs with sparse encoding of genetics as input features (N=1290).



Score	Sensitivity	PPV	Specificity	NPV
0.1	0.98	0.13	0.12	0.98
0.2	0.94	0.19	0.44	0.98
0.3	0.88	0.24	0.62	0.97
0.4	0.84	0.27	0.69	0.97
0.5	0.80	0.31	0.75	0.96
0.6	0.75	0.34	0.80	0.96
0.7	0.68	0.37	0.84	0.95
0.8	0.42	0.45	0.93	0.92
0.9	0.07	0.42	0.99	0.89

Score	Sensitivity	PPV	Specificity	NPV
0.1	1.00	0.53	0.10	1.00
0.2	0.96	0.56	0.23	0.85
0.3	0.88	0.69	0.60	0.83
0.4	0.84	0.73	0.69	0.80
0.5	0.82	0.77	0.75	0.80
0.6	0.78	0.81	0.81	0.78
0.7	0.59	0.78	0.83	0.67
0.8	0.37	0.95	0.98	0.60
0.9	0.06	1.00	1.00	0.51

Score	Sensitivity	PPV	Specificity	NPV
0.1	1.00	0.35	0.00	nan
0.2	0.92	0.34	0.05	0.50
0.3	0.92	0.37	0.14	0.75
0.4	0.83	0.36	0.18	0.67
0.5	0.75	0.35	0.23	0.62
0.6	0.75	0.38	0.32	0.70
0.7	0.67	0.38	0.41	0.69
0.8	0.25	0.30	0.68	0.62
0.9	0.17	0.40	0.86	0.66

Figure 4: **Personalized AI ensemble models** based on mean of scores, majority voting and mean of confident scores (t=0.65). Score = Applied prediction score threshold for classification (≥Score), PPV = positive predictive value, NPV = negative predictive value. **A)** ROC curve for the ensemble when predicting on the test set (N=1290). **B-C**) Plot of prediction scores vs true class and table of performance metrics for different score thresholds when scoring the predictions on the holdout set (N=1290) model ensemble with the

mean of confident scores (score threshold of ≤0.35 or ≥0.65) where one case and fifteen controls were left out of the scoring. **D)** ROC curve for the ensemble when predicting on the holdout set (N=100). **E-F)** Plot of prediction scores vs true class and table of performance metrics for different score thresholds when scoring the predictions on the holdout set (N=100) model ensemble with the mean of confident scores (score threshold of ≤0.35 or ≥0.65) where one case and two controls were left out of the scoring. **G)** ROC curve for the ensemble when predicting secondary AAP cases. **H-I)** Plot of prediction scores vs true class and table of performance metrics for different score thresholds when scoring the predictions on the second AAP phenotype (N=37) model ensemble with the mean of confident scores (score threshold of ≤0.35 or ≥0.65) where one case and two controls were left out of the scoring.

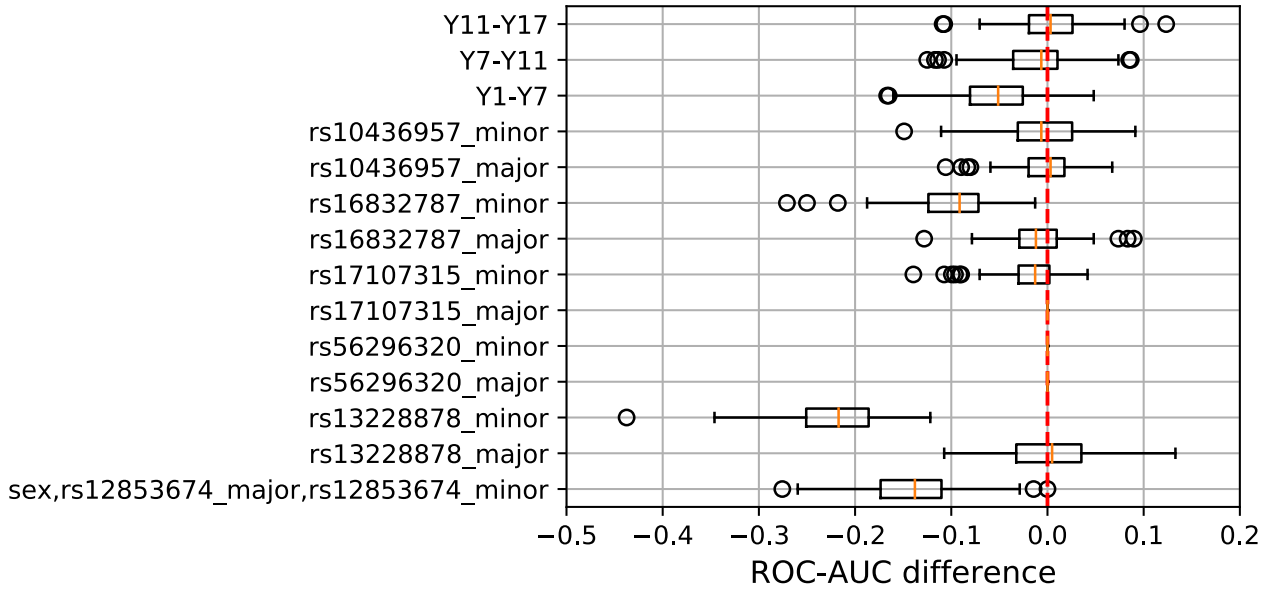


Figure 5: **Leave-one-out feature importance for AAP re-exposure model** using a random forest trained on N=37 to predict second cases of AAP during when re-exposed to asparaginase. The model used age, sex and six candidate SNPs with sparse encoding of genetics as input features and was trained with leave-one-out cross-validation.

CHAPTER 7

Clinical utility

The thesis showcase three potential applications of prediction models using machine learning methodologies. This chapter will briefly explore some learnings on what is needed of a model to provide clinical utility.

The real-life success of predictive machine learning models relies on a clinically useful outcome i.e. it is important its predictions can be used to modify and target strategies to prevent, delay or improve the given outcome. Outcomes can span various clinically relevant areas such as disease onset, disease progression, treatment efficacy or treatment toxicity. As an example, a successful machine learning model for breast cancer screening was recently published. This model outperformed radiologists in detection of breast cancer and could reduce radiologist's workload in the screenings [100]. In terms of treatment, RCTs demonstrating that machine learning-based decisions improve patient outcomes are the current accepted standard to evaluate their effect on patient outcomes. Some RCTs have proven machine learning models can provide guidance in precision medicine. This has been shown in patients with severe sepsis, where the duration of hospital stays were reduced and in-hospital mortality decreased when applying guidelines from the predictive model [194]. More recently, colonoscopy with computer-guided diagnosis assisted in finding more small polyps than a colonoscopy alone [195]. In the early stages of machine learning for health and clinical applications, this should by no means replace physicians, but provide a tool for additional decision guidance. It is thus important to guide the model development alongside experts in health care to ensure a relevant model that following implementation in its respective application area eventually would create value to benefit individual patients at the right time, or assist physicians and hospital setups to be more efficient in operation or decision making.

A key consideration through the projects in this thesis has been what accuracy machine learning models need to achieve in order to be clinically useful. In an attempt to address this question, it was useful to identify individuals at low or high risk of the clinical outcomes e.g. by evaluating the true-positive, true-negative, false-positive and false-negative rates at different prediction thresholds. This allowed evaluation of the model's ability to detect high-risk individuals at the cost of falsely detecting low-risk individuals and vice versa. It is imperative to keep in mind the risk of benefit and harm at the individual patient level in order to develop the most clinically relevant model. Some considerations are; Will increased monitoring of patients in high risk lead to additional financial expenses compared to now? In cancer, is it useful to truncate chemotherapy that potentially can result in toxicity given a slightly higher risk of treatment failure?

Or rather, is there a level of increased treatment failure risk that might be acceptable, given the potential for significant improvements in quality of life? And how will the decision of a machine learning model eventually influence a person's quality of life if the wrong decision is made? These are considerations that need to be decided given current clinical practice, and also spell the need for close iterative development of these algorithms with clinicians.

Several types of machine learning algorithms are available, and the choice upon which model to use is not always given. Some more popular models within bioinformatics include random forests, support vector machines, ANNs or deep learning. However, when applying models for predictions in clinical settings, there is currently some emphasis on interpretability. Another consideration would be deployment: models should be made simple, fast and minimally disruptive to current clinical practice. Furthermore, it is important that a machine learning model and its predictions can be applied across populations. A large diverse population during training will allow prediction models to generalize better and avoid selection bias on genetic and environmental factors that may hinder the generalization of a model.

Generalizability of any machine learning model requires robust cross-validation setups to counteract overfitting of the model. An essential requirement for translation of a predictive model is that it can be validated on an independent dataset [15]. In the validation step of the machine learning workflow, the performance of the predictive model is evaluated for its ability to generalize to new datasets. Despite robust precautions, there is always a risk that a machine learning model will not validate on an independent dataset even though the model is not overfitted [13, 15]. Reasons for this can be that the independent test dataset does not match the properties of the data used for training of the model, such as substantial biological differences e.g. a different background population, treatment protocol or technical differences in the data with respect to tissue collection, laboratory tests and quality control as well as data representation [13, 15, 77]. However, the machine learning model can in this case still be used to generate new hypotheses of biological impact features where another study may come up with new knowledge that can be applied to develop the model further.

Additional challenges regarding the success of machine learning in precision medicine relate to data collection and handling including data storage, data security and computing facilities, communication and education of stakeholders including legal experts, data analysts, clinicians and patients, as well as regulatory, social and ethical considerations [15, 30, 196, 197]. These topics are beyond the scope of this thesis and the reader is instead referred to the listed references.

In addition, there are even more factors to consider to clinical utility of machine learning models. These could form the basis of a Viewpoint in a clinically oriented journal.

CHAPTER 8

Learnings and future prospects

8.1 Learnings

In this thesis, machine learning was applied on Big data within different health and disease areas to develop predictive models with potential applications in precision medicine. The machine learning models provided approaches on how to integrate highly diverse types of data of large volumes obtained from different clinical study designs with varying number of samples across the three presented research projects. All projects aimed to elucidate biomarkers that predispose individuals towards a given response. In each of the three research projects presented, the specific study had its individual bioinformatic challenges that were identified and addressed. The findings are summarized in the following per study.

The first study (*Chapter 4*) explored prediction of weight loss in Danish individuals at an increased risk of metabolic complications undergoing eight weeks dietary intervention with changes in the dietary whole grain or gluten content. The study was challenged by small cohort sizes with a high number of heterogeneous features. Use of machine learning methodologies did provide predictive signals from small cohorts when the models were carefully guided by different strategies in feature transformation, reduction as well as selection. These signals were confirmed across several models when their robustness was assessed in different types of data. The most successful models capturing different biological aspects were integrated through an ensemble that only considered the most confident predictions to minimize the risk of false positive or false negative predictions. The ensemble has potential application as an early screening tool for selection of weight loss strategies by providing a score of how likely it is that a person would be to experience weight loss in combination with one of the intervention diets. Given that the model would be difficult to backtrace with regards to which dietary intervention a person had received for weight loss, the safest utility in deploying the model is in determining individuals that would not benefit from a dietary intervention as a weight loss strategy. This guidance could save frustrations by not having individuals going through a dietary intervention that most likely not will be efficient for weight loss.

The second study (*Chapter 5*) presents prediction models of the time to insulin, 1 to 4 years ahead of any time point during disease progression. Models were developed

using a Scottish cohort of T2D patients with data on genotype and longitudinal irregular sampled data from EMRs. Different strategies for feature extraction and representation of the longitudinal EMR data were explored given a fixed time point approach and a longitudinal approach. When modelling the time to insulin with ANNs given the two data extractions methods, the fixed time point approach had higher predictive performance when the extracted measurement was close to the insulin requirement event (up to two years ahead) compared to the longitudinal approach. This illustrated the importance of longitudinal follow-ups on patients in order to provide best guidance of individual risk at the right time. Furthermore, it was surprising that the models could predict the time to insulin with robust and similar performance at any time point up to 10 years after T2D diagnosis given highly variable progression rates onto insulin and possibly heterogeneity of T2D patients. Clinical inertia is currently a serious problem in T2D treatment, and we hypothesize that the clinical application of the presented model can motivate patient behavior to prolong the time to insulin or identify patients where increased glycemic monitoring is needed. The models can also assist in identifying patients whose glycemia is well-controlled which can help reduce interventions and their associated cost and burden in health care systems, which only becomes more important given a continuously increasing number of T2D patients. Another interesting finding of this project was that despite genetics are a strong hereditary component for T2D risk, no genetic variants or GRS improved the predictive performance following integration of these into the clinical models when the models were developed towards real-world application. The impact of genetics of the predictability of time to insulin was only observed when using a stricter definition of people that did not require insulin (controls). SNPs enriched in insulin signaling and cancer pathways improved the predictive performance for a subgroup of patients indicating that genetics may be useful for prediction of the time to insulin in some T2D patients.

The third project (*Chapter 6*) presents a study on prediction of asparaginase-associated pancreatitis (AAP) as well as second AAP following re-exposure to asparaginase after truncation of treatment in childhood ALL. We explored genotype-phenotype predictions in the largest AAP cohort of childhood ALL patients to develop a screening tool at the time of diagnosis. Bioinformatic challenges included identification of the most predictive SNPs of the tested outcomes. Thus, different feature encodings, reduction and selection strategies as well as different types of machine learning models were explored. The model utility was throughout the project compared to treatment guidelines to keep a clinically relevant focus on the models. Models of AAP and second AAP risk have potential clinical applications by identifying high-risk patients of AAP that need more increased surveillance or less intensive asparaginase therapy (compared to the 2−18% risk of AAP) or by identifying low-risk patients of a second AAP where asparaginase is safe to re-introduce after the first AAP (compared to 50% risk of second AAP and physician gut feeling). By identifying these subgroup of patients, asparaginase treatment stratification may be updated in future childhood ALL protocols.

The three research projects illustrate that integration of Big data into machine learning models can provide predictions on future events that potentially can influence clinical

decision making. All presented models in this thesis require independent validation prior to any real-life applications. We believe that validation of the two suggested clinical applications on time to insulin in T2D patients and AAP in childhood ALL could soon provide a way in for implementing these models in clinic.

The learnings made from the projects can be repeated across several other disease areas with similar data. During my work as a PhD student, I have been excited in learning how to deal with data handling and pre-processing, modelling and considerations of AI for clinical translation through interdisciplinary cooperation and discussions with stakeholders. Far from all challenges in these areas have been dealt with in this PhD thesis, but presents the start of my research journey. One aspect that I found especially interesting across the three research projects, was the process of identifying utility of the machine learning models and how this could potentially translate into clinical applications. The model utility was not clear in the beginning of the projects. Instead, this developed through close interdisciplinary collaborations and have provided me with several valuable learnings on how bioinformatic methods and machine learning can play a role in future clinical practice. Developing and reaching this common understanding in interdisciplinary cooperation is something I personally feel has been one of the major achievements of this PhD.

8.2 Future prospects

In the first project, weight loss was reported as a classification outcome. Thus, there is a risk of misclassification of weight loss responders and non-responders given day-to-day fluctuations in body weight that are not considered in the study. It would be useful to inform the model with prior knowledge on everyday body weight fluctuations for an individual to determine how large a short-term weight loss should be in order to consider a study participant a responder. This natural randomness in biological data is also true for several other biomarkers used in the presented projects where this information on fluctuations potentially could be useful.

In the T2D project, a predictive model of time to insulin were developed. Most predictive features from the EMRs were previously associated with progression. Thus, it would be interesting to explore if the model also could be used as a general screening tool for prediction of the time to any treatment intensification throughout T2D progression. The study presented two clinical models. The first model was based on predictions that were clinical intuitive whereas the second model was trained on a more heterogeneous group of patients, where HbA1c was not the main predictive driver of time to insulin. Furthermore, we found that genetics may assist predictions for some patients. It would thus be interesting to understand potential subgroups of T2D patients further beyond the most generalizable model, as T2D is a heterogenous disorder. From a methodological point of view, it would also be interesting to explore how to represent EMR data to machine learning and deep learning models further. Deep learning models have shown prospect in handling temporal longitudinal data from EHRs to predict clinically relevant

outcomes [123, 197, 198]. In addition, deep learning may provide a better chance of learning patterns in T2D heterogeneity.

The AAP project predicted treatment toxicity from asparaginase in childhood ALL patients. The prediction models were focused on exploring the predictability using only genotype. Thus, an essential feature, asparaginase, was not included in the prediction models. Our initial findings on a subset of patients from the NOPHO ALL-2008 treatment protocol ($N = 892$, 77 AAP cases) where the cumulative dose of asparaginase was available identified asparaginase as the most predictive feature compared to genetic variants. Since asparaginase is an essential drug in childhood ALL therapy, it would be interesting to predict 'the time to AAP' and include features of how many doses a patient can tolerate (including concentration) as well as the time in between asparaginase therapy. By further developing the established prediction models, these features are likely to provide insight on the timing when AAP occurs possibly allowing even better identification of when intensified monitoring or less intensive asparaginase therapy should be warranted. It would also be interesting to explore the AAP phenotype beyond the case-control definition and identify if the models are better in prediction of AAP that occurs early or late during the ALL treatment period and if it provides mild or severe symptoms. The training of the AAP models was only possible on patients with CEU ancestry. However, it is to be explored further if the models will be valid for prediction of AAP in other populations.

A common challenge in all projects included feature selection on high-dimensional datasets where the models needed guidance from prior knowledge to reduce the number of features prior to data-driven selection. Continuous work on feature representation and selection may assist in improved selection strategies to identify the most predictive features. Furthermore, biological data is often sparse which challenges learning if the signal is only predictive of outcome in a few individuals. Representation of such data could potentially gain more power from utilizing information on pathways or other functional concepts to e.g. improve the genetic signal from host genotype or microbiome data.

Another ongoing challenge is interpretation of feature importance at the individual level. Continuous efforts will eventually provide a more individual level understanding of features. As an example, the impact of feature importance at the individual level has recently been elucidated using Shapley values in patients at risk of gestational diabetes [78]. The need for understanding the complex learnings in machine learning models may change in the future when machine learning methods will gain more trust. The benefit of using of machine learning methods in clinical settings to guide precision medicine is still to be seen in RCTs and this could happen in a matter of years. Machine learning models will most likely not change current clinical guidelines which are based on several years of clinical experience but can assist in identification of patients who would benefit from more personalized treatment strategies as well as optimize clinical decision making or operations.

APPENDIX A

Appendix A

Supplementary Material

Supplementary Material 1: Flowchart of machine learning framework and data integration strategy

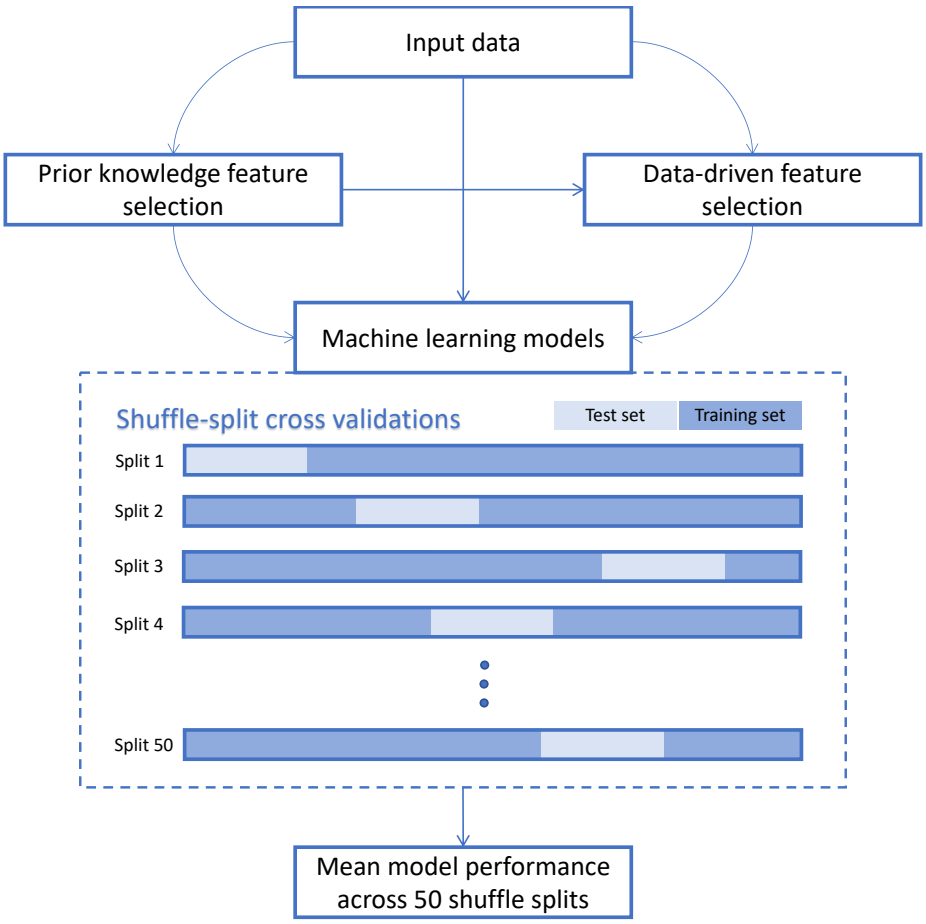


Figure S.1: **Feature selection and machine learning setup.**

Supplementary Material 2: Genes from literature pathways, butyrate-producing species from literature study, top metagenomic species and SNPs used in genetic risk scores

Table S.2a): **Biological aspect, pathways, genes and literature references used to select 703 SNPs for the LithPath dataset**. The SNPs for modelling were found by annotation to SNPs to genes by Variant Effect Predictor (VEP build 37).

Biological aspect	Pathways	Genes	Reference
B12 and folate pathways	B12 and folate pathway	ABCD4, CD320, CLYBL, CUBN, FOLR3, FUT2, FUT6, MMAA, MMACHC, MTHFR, MUT, TCN1 and TCN2	Grarup, Niels; Sulem, Patrick; Sandholt, Camilla H; et al. Genetic architecture of vitamin B12 and folate levels uncovered applying deeply sequenced large datasets. P L O S Genetics — 2013, Volume 9, Issue 6
Cancer, fibrin/collagen formation and anti-xenobiotic effects	Various that affect cancer, fibrin and collagen formation and anti-xenobiotic effects.	ACTA2, ACTG2, AKR1B10, CAPN13, CAPZA2, CES1, COL6A1, CYP1A1, CYP2D6, CYP3A7, CYP4A22, CYP3A66, cytochrome P450, DHDH, FXYD3, GOLPH2, GSN, GSTM4, GSTM5, KRT18, MT1E, MT1F, NAD, PTTG1, p24, RETNLB, SULT2A1, S100P, UGT2B10, UGT2B11 and ZBTB16	Sestak K, Conroy L, Aye PP, Mehra S, Doxiadis GG, et al. (2011) Improved Xenobiotic Metabolism and Reduced Susceptibility to Cancer in Gluten-sensitive Macaques upon Introduction of a Gluten-Free Diet. PLoS ONE 6(4):e18648
Fatty acid metabolism and cytokine signalling	PPAR signalling pathway (activation of Fatty acid degradation in liver) and cytokine signalling (FAS/TNFRSF6, MCP-1, TNFalpha and IL-6)	CPT1, FAS, IL-6, L-FABP, MCP-1, PPAR-a, SCD1, TNF-α and UCP2	Park, Mi-Young; Jang, Hwan-Hee; Kim, Jung Bong; et al. Hog millet (Panicum miliaceum L.)-supplemented diet ameliorates hyperlipidemia and hepatic lipid accumulation in C57BL/6J-ob/ob mice. NUTRITION RESEARCH AND PRACTICE — 2011, Volume 5, Issue 6, pp. 511-519
Gut microbiota modulation	ABO blood group	FUT2	Rausch et al., Colonic mucosa-associated microbiota is influenced by an interaction of Crohn disease and FUT2 (Secretor) genotype, PNAS, vol. 108, no.

			47, 19030–19035 (2011), -Mäkivuokko et al., Association between the ABO blood group and the human intestinal microbiota composition, BMC microbiology, 12:94, (2012), Wacklin et al., Secretor Genotype (FUT2 gene) is Strongly Associated with the Composition of Bifidobacteria in the Human Intestine, Plos One, vol 6, 5, e20113 (2011), and more
	G-protein coupled receptor	FFAR2, FFAR3, GPR109a, GPR41 and GPR43	Aw W and Fukuda S. Toward the comprehensive understanding of the gut ecosystem via metabolomics-based integrated omics approach. Semin Immunopathol 2014. DOI 10.1007/s00281-014-0456-2
	HDL complex	APOA1	Zhang, C. et al. Interactions between gut microbiota, host genetics and diet relevant to development of metabolic syndromes in mice. ISME J. 4, 232–241 (2010).
	IFN signalling	IFN-α, IFN-ß, IFN-γ, IFNG, IFR9 and STAT4	Thompson, C. L., Hofer, M. J., Campbell, I. L. & Holmes, A. J. Community dynamics in the mouse gut microbiota: a possible role for IRF9-regulated genes in community homeostasis. PLoS ONE 5, e10335 (2010)
	Tol-Like receptor signalling	TLR2 and TLR5	Albert, E. J., Sommerfeld, K., Gophna, S., Marshall, J. S. & Gophna, U. The gut microbiota of toll-like receptor 2-deficient mice exhibits lineage-specific modifications. Environ. Microbiol. Rep. 1, 65–70 (2009). Vijay-Kumar, M. et al. Metabolic syndrome and altered gut microbiota in mice lacking Toll-like receptor 5. Science 328, 228–231 (2010).
	Various	ANG, IFN, IGF1, IL-1B, IL-17B, IL-23, IRAK2, JUN, OXT,	van Baarlen, Peter, ; Troost, Freddy; van der Meer, Cindy; et al. Human mucosal in

		POMC, PYY, STAT4, TLR3, TLR9 and UCN	vivo transcriptome responses to three lactobacilli indicate how probiotics may modulate human cellular pathways. PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA — 2011, Volume 108, Issue Suppl. 1, 1, pp. 4562-4569
Immune response	Cytokine signaling, G protein signaling, T cell activation	CCR3, IL12A, IL18RAP, RGS1, SH2B3 and TAGAP	K. A. Hunt, A. Zhernakova, G. Turner et al., "Newly identified genetic risk variants for celiac disease related to the immune response," Nature Genetics, vol. 40, no. 4, pp. 395–402, 2008.
	Production of PAI-1, NADPHox, iNOS, TLR4, THF-a, (gut permability)	iNOS, NADPHox, PAI-1, THF-a and TLR4	Cani, P. D., ; Possemiers, S.; Van de Wiele, T.; et al. Changes in gut microbiota control inflammation in obese mice through a mechanism involving GLP-2-driven improvement of gut permeability. GUT — 2009, Volume 58, Issue 8, pp. 1091-1103.
Immune response, lipogenic/adipogenic regulation, and cell proliferation and differentiation	Transcription factors in immune, lipid and cell regulation	NF-kB, NEMO, akt, PPARg and SREBP1	Radonjic, Marijana; de Haan, Jorn R.; van Erk, Marjan J.; et al. Genome-Wide mRNA Expression Analysis of Hepatic Adaptation to High-Fat Diets Reveals Switch from an Inflammatory to Steatotic Transcriptional Program. PLOS ONE — 2009, Volume 4, Issue 8, pp. -
Inflammation	Cytokine signalling (IL-6 response by CRP and TNF-R2)	CRP and TNF-R2	Mantzoros, Christos,Franz, Mary Van Dam, Rob M et al. Wholeo-grain, bran, and cereal fiber intakes and markers of systemic inflammation in diabetic women. Diabetes Care 29:207–211, 2006.
Insulin secretion	Wnt signalling pathway	TCF7L2	Fisher E, Boeing H, Fritsche A, et al. Whole-grain consumption and transcription factor-7-like 2 (TCF7L2)

			rs7903146: gene–diet interaction in modulating type 2 diabetes risk. Br J Nutr 2009; 101:478–481.
Potential gluten degrading bacteria	NOD-like receptor signalling pathway	MEFV	Khachatryan, Z. A. et al. Predominant role of host genetics in controlling the composition of gut microbiota. PLoS ONE 3, e3064 (2008)
Signal transduction, metabolism (other than energy metabolism), cell cycle control, transcription and translation, control of cellualar organization and transport facilitation	Various that influence Signal transduction, metabolism (other than energy metabolism), cell cycle control, transcription and translation, control of cellualar organization and transport facilitation	ALDH7A1, ATP6V0B, BAP1, CDNA, CLY6G5B, CTNNB1, DGKD, DHCR24, DIO2, D123, EFNB2, EGFR, FLJ43113, FTCD, HIBADH, MRPL4, MTCP1, PKIA, PIK3R1, RIT1, RLF, RNTRE, ROK1, SFA1, TM4SF1, WAVE1 and ZNF161	K. Juuti-Uusitalo, M. Mäki, H. Kainulainen, J. Isola and K. Kaukinen. Gluten affects epithelial differentiation-associated genes in small intestinal mucosa of coeliac patients. 2007 British Society for Immunology, Clinical and Experimental Immunology, 150: 294–305
Transit time	Incretin response	GCG	Anita Wichmann, Ava Allahyar, Thomas U. Greiner, Hubert Plovier, Gunnel Östergren Lundén, Thomas Larsson, Daniel J. Drucker, Nathalie M. Delzenne, Patrice D. Cani, Fredrik Bäckhed, Microbial Modulation of Energy Availability in the Colon Regulates Intestinal Transit, Cell Host & Microbe, Volume 14, Issue 5, 13 November 2013, Pages 582-590
Genes related weight, obesity or metabolic syndromes		FAIM2, FBXO22, FTO, NRG4, PAK7, PLCB1, SEC16B, SH2B1 and UBE2Q2	De Giorgio MR, Yoshioka M, St-amand J: Feeding induced changes in the hypothalamic transcriptome. *Clinica Chimica Acta* 2009, 406:103–107. Frazier-Wood, A. C. & Wang, Z. Genetics of Obesity. in *Metabolic Syndrome* 123–140 (Springer International Publishing, 2016). doi:10.1007/978-3-319-11251-0_10

			Mehta NK, Mehta KD: Protein kinase C-beta: an emerging connection between nutrient excess and obesity. *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids* 2014, 1841:1491–1497 Park, S. H., Lee, J. Y. & Kim, S. A methodology for multivariate phenotype-based genome-wide association studies to mine pleiotropic genes. *BMC Syst. Biol.* **5**, S13 (2011). Wang G, Zhao X, Meng Z, Kern M, Dietrich A, Chen Z, Cozacov Z, Zhou D, Okunade AL, Su X, Li S, Blüher M, Lin JD: The brown fat – enriched secreted factor Nrg4 preserves metabolic homeostasis through attenuation of hepatic lipogenesis. *Nature medicine* 2014, 20:1436–1443.

Table S.2b): **Butyrate-producing species selected by literature study**, and the MGmapper catalogue they are present in. Some species were found in more of the references listed, but only the first it was found in is listed.

Species	**MGmapper catalogue**	**Reference**
Anaerostipes caccae	Bacteria draft and Human Microbiome	Petra Louis, Harry J. Flint, "Diversity, metabolism and microbial ecology of butyrate-producing bacteria from the human large intestine", *FEMS Microbiology Letters*, Volume 294, Issue 1, May 2009, Pages 1–8, doi: 10.1111/j.1574-6968.2009.01514.x
Anaerotruncus colihominis	Bacteria draft and Human Microbiome	
Butyrivibrio fibrisolvens	Bacteria and Bacteria draft	
Eubacterium hallii	Bacteria draft	
Eubacterium ramulus	Bacteria draft	
Faecalibacterium prausnitzii	Bacteria, Bacteria draft and Human Microbiome	

Roseburia faecis	Bacteria draft	
Roseburia hominis	Bacteria	
Roseburia intestinalis	Bacteria, Bacteria draft and Human Microbiome	
Roseburia inulinivorans	Bacteria draft and Human Microbiome	
Subdoligranulum variabile	Bacteria draft and Human Microbiome	
Anaerostipes hadrus	Bacteria and Human Micrbiome	Sato, T., Kusuhara, S., Yokoi, W., Ito, M. & Miyazaki, K. Prebiotic potential of L-sorbose and xylitol in promoting the growth and metabolic activity of specific butyrate-producing bacteria in human fecal culture. *FEMS Microbiol. Ecol.* **93**, fiw227 (2017). doi: 10.1093/femsec/fiw227
Bacteroides uniformis	Bacteria draft and Human Microbiome	K. Takahashi *et al.*, "Reduced Abundance of Butyrate-Producing Bacteria Species in the Fecal Microbial Community in Crohn's Disease," *DIG*, vol. 93, no. 1, pp. 59–65, 2016. doi: 10.1159/000441768
Clostridium butyricum	Bacteria	G. Cai, B. Jin, C. Saint, and P. Monis, "Genetic manipulation of butyrate formation pathways in Clostridium butyricum," *J. Biotechnol.*, vol. 155, no. 3, pp. 269–274, Sep. 2011. doi: 10.1016/j.jbiotec.2011.07.004
Clostridium kluyveri	Bacteria	H. Seedorf *et al.*, "The genome of Clostridium kluyveri, a strict anaerobe with unique metabolic features," *PNAS*, vol. 105, no. 6, pp. 2128–2133, Feb. 2008. doi: 10.1073/pnas.0711093105
Eubacterium limosum	Bacteria	S. Park *et al.*, "Acetate-assisted increase of butyrate production by Eubacterium limosum KIST612 during carbon monoxide

		fermentation," *Bioresour. Technol.*, vol. 245, no. Pt A, pp. 560–566, Dec. 2017. doi: 10.1016/j.biortech.2017.08.132
Fusobacterium nucleatum	Bacteria and Human Microbiome	M. Vital, A. C. Howe, and J. M. Tiedje, "Revealing the Bacterial Butyrate Synthesis Pathways by Analyzing (Meta)genomic Data," *mBio*, vol. 5, no. 2, pp. e00889-14, May 2014. doi: 10.1128/mBio.00889-14

14

15

16 ***Table S.2c):*** **Top 14 most altered metagenomic species from the whole grain and gluten studies[25,26].**

Metagenomic species	**Taxonomic annotation**	**Adjusted (FDR) p-value**	**Reference**
MGS:igc210	Lachnospiraceae (family)	5.42E-09	Skov, L. B. et al. A low-gluten diet induces changes in the intestinal microbiome of healthy Danish adults. 28, (2019). Supplementary Data 2.
MGS:igc939	*Bifidobacterium angulatum*	2.98E-07	
MGS:igc413	*Bifidobacterium longum*	1.73E-06	
MGS:igc529	*Bifidobacterium adolescentis*	8.38E-06	
MGS:igc356	*Bifidobacterium pseudocatenulatum*	1.10E-04	
MGS:igc121	Lachnospiraceae (family)	1.07E-03	
MGS:igc846	*Dorea* (genus)	2.13E-03	
MGS:igc221	Unclassified	2.67E-03	
MGS:igc491	*Dorea longicatena*	7.02E-03	
MGS:igc47	*Blautia wexlerae*	7.34E-03	
MGS:igc1021	Unclassified	9.77E-03	
MGS:igc492	Clostridiales (order)	9.80E-03	
MGS:igc78	*[Eubacterium] hallii*	1.72E-02	
MGS:igc169	*Anaerostipes hadrus*	4.94E-02	
MGS:igc654	Clostridiales (order)	0.16	Roager, H. M. et al. Whole grain-rich diet reduces body weight and systemic low-grade inflammation without
MGS:igc102	*Erysipelatoclostridium ramosum*	0.16	
MGS:igc460	Clostridiales (order)	0.16	

MGS:igc633	Clostridiales (order)	0.16	inducing major changes of the gut microbiome: a randomised cross-over trial. Gut 68, 83–93 (2019). Online Supplementary Material 2, Table S7.
MGS:igc139	*Ruminococcus* (genus)	0.18	
MGS:igc291	*Faecalibacterium prausnitzii*	0.22	
MGS:igc359	Clostridiales (order)	0.30	
MGS:igc309	*Ruminococcus lactaris*	0.31	
MGS:igc734	*Streptococcus thermophilus*	0.33	
MGS:igc584	*Holdemanella biformis*	0.38	
MGS:igc517	*Faecalibacterium prausnitzii*	0.38	
MGS:igc171	*Faecalibacterium prausnitzii*	0.38	
MGS:igc213	Coprococcus	0.38	
MGS:igc9	*Bacteroides thetaiotaomicron*	0.38	

Since the OR from other studies was used to weight the SNPs, the GRS was calculated per study. GRS's were calculated when more than one SNP was present in our data for a study and the particular SNP had a p-value lower than 10^{-4}.

Table S.2d): Genetic risk scores. This table lists the SNPs used for the GRS's, as well as their p-values, odds ratios and the reference from which each was found.

SNP	**p-value**	**Odds ratio**	**Phenotype**	**Paper / GWAS**
rs2030323	3e-22	1.12	Obesity (BMI)	Berndt, S. I. et al., "Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture," *Nature Genetics*, vol. *45 no. 5*, pp. 501–512, 2013. doi: 10.1038/ng.2606
rs13130484	4e-28	1.08		
rs4735692	3.51e-10	1.04		
rs7138803	1e-20	1.09		
rs4833407	1e-6	1.11	Obesity (BMI)	Bradfield, J. P. et al. (Early Growth Genetics Consortium),
rs4864201	2e-7	1.12		

rs9299	4e-9	1.14		"A genome-wide association meta-analysis identifies new childhood obesity loci," *Nature Genetics*, vol. *44 no.* 5, pp. 526–531, 2012. doi: 10.1038/ng.2247
rs9568856	2e-9	1.22		
rs7138803	6.4e-8	1.24	Extremely overweight young adults (BMI)	Paternoster, L. et al., "Genome-Wide Population-Based Association Study of Extremely Overweight Young Adults – The GOYA Study", *PLoS ONE*, vol. *6 no.* 9: e24303, 2011. doi: 10.1371/journal.pone.0024303
rs9936385	1.4e-13	1.35		
rs13130484	1.9e-5	0.85		
rs699363	2e-5	0.21		
rs543874	6.5e-5	1.2		
rs734597	2e-5	1.25		
rs6077585	3.52e-7	-1.28	Body weight [kg]	GWAS of weight changes in whole grain study
rs1039547	1.59e-6	-2.04		
rs2648435	1.59e-6	-2.04		
rs8036952	1.59e-6	-2.04		
rs7169122	4.68e-6	-1.82		
rs6591079	7.48e-6	-1.27		
rs2239985	9.18e-6	-0.92		
rs13379337	9.79e-6	-0.91		
rs12972098	1.34e-5	-0.91		
rs1015092	1.39e-5	-2.89		
rs17382342	5.74e-6	1.39	Sagittal abdominal diameter [cm]	GWAS of sagittal abdominal diameter changes in whole grain study
rs12156272	6.12e-6	-1.77		
rs10957603	1.47e-6	2.1		
rs536995	1.65e-5	-1.47		
rs17466747	1.7e-5	1.39		
rs4782784	2.17e-5	-1.51		

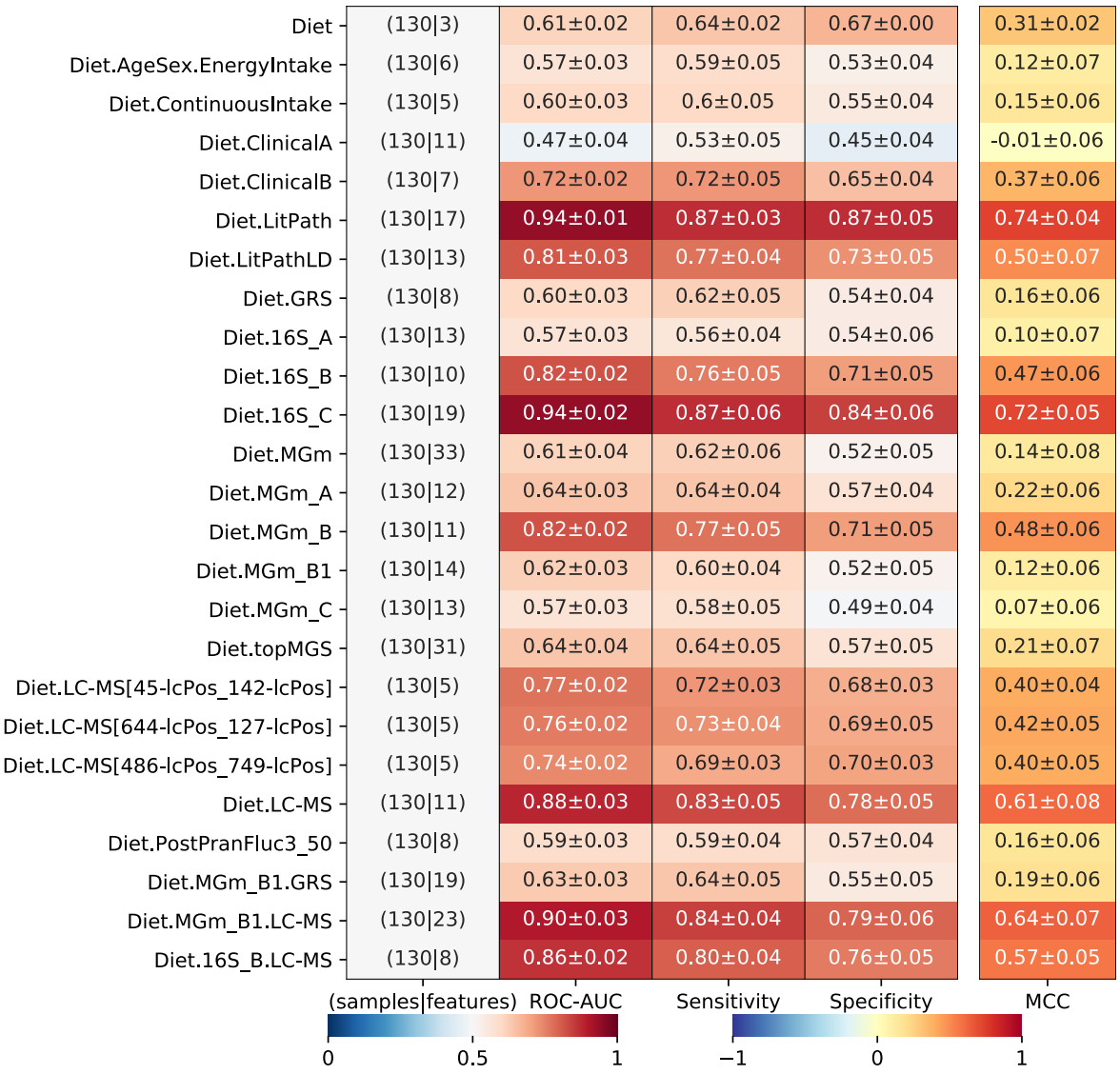
rs11725412	2.17e-5	-2.26		
exm1537540	2.21e-5	-2.77		
exm1537642	2.21e-5	-2.77		
rs12654643	2.7e-5	-1.8		

24

Supplementary Material 3: Comparison of models

Table S.3 below is an extended version of Table 2. Here, all models trained and tested on the common set of 130 samples are shown, meaning that models which were subsequently not selected for further analysis because of lower performance or suspicion of overfit are included. The metagenomic data, which was mapped against the MGmapper database, gave five datasets from mapping against catalogues Bacteria, Bacteria draft and Human Microbiome: *Diet.MGm* (contains pre-selected butyrate-producing species for all selected catalogues), *Diet.MGm_A* (contains pre-selected butyrate-producing species for the Bacteria catalogue), *Diet.MGm_B* (contains all species mapped to the Bacteria draft catalogue), *Diet.MGm_B1* (contains pre-selected butyrate-producing species for the Bacteria draft catalogue) and *Diet.MGm_C* (contains pre-selected butyrate-producing species for the Human Microbiome catalogue). The dataset for Bacteria draft catalogue with butyrate-producing species (MGm_B1) was selected for subsequent models, since it performed much better when considering all available data, not just 130 samples, as seen in Table S.6. Furthermore, the model including forward selected species from the Bacteria draft catalogue (MGm_B) showed less separation when considering the difference between models trained on a true and a permuted target.

Table S.3: **Test performances for models run on a common set of 130 samples across 50 random shuffles of the data**. The blue-red colorbar is for area under the receiver operating characteristic curve (ROC-AUC), sensitivity and specificity, while the blue.yellow-red colorbar is for Matthews correlation coefficient (MCC). Abbreviations for model combinations are explained in Table 1.


Diet (130|3) 0.61±0.02 0.64±0.02 0.67±0.00 0.31±0.02
Diet.AgeSex.EnergyIntake (130|6) 0.57±0.03 0.59±0.05 0.53±0.04 0.12±0.07
Diet.ContinuousIntake (130|5) 0.60±0.03 0.6±0.05 0.55±0.04 0.15±0.06
Diet.ClinicalA (130|11) 0.47±0.04 0.53±0.05 0.45±0.04 -0.01±0.06
Diet.ClinicalB (130|7) 0.72±0.02 0.72±0.05 0.65±0.04 0.37±0.06
Diet.LitPath (130|17) 0.94±0.01 0.87±0.03 0.87±0.05 0.74±0.04
Diet.LitPathLD (130|13) 0.81±0.03 0.77±0.04 0.73±0.05 0.50±0.07
Diet.GRS (130|8) 0.60±0.03 0.62±0.05 0.54±0.04 0.16±0.06
Diet.16S_A (130|13) 0.57±0.03 0.56±0.04 0.54±0.06 0.10±0.07
Diet.16S_B (130|10) 0.82±0.02 0.76±0.05 0.71±0.05 0.47±0.06
Diet.16S_C (130|19) 0.94±0.02 0.87±0.06 0.84±0.06 0.72±0.05
Diet.MGm (130|33) 0.61±0.04 0.62±0.06 0.52±0.05 0.14±0.08
Diet.MGm_A (130|12) 0.64±0.03 0.64±0.04 0.57±0.04 0.22±0.06
Diet.MGm_B (130|11) 0.82±0.02 0.77±0.05 0.71±0.05 0.48±0.06
Diet.MGm_B1 (130|14) 0.62±0.03 0.60±0.04 0.52±0.05 0.12±0.06
Diet.MGm_C (130|13) 0.57±0.03 0.58±0.05 0.49±0.04 0.07±0.06
Diet.topMGS (130|31) 0.64±0.04 0.64±0.05 0.57±0.05 0.21±0.07
Diet.LC-MS[45-lcPos_142-lcPos] (130|5) 0.77±0.02 0.72±0.03 0.68±0.03 0.40±0.04
Diet.LC-MS[644-lcPos_127-lcPos] (130|5) 0.76±0.02 0.73±0.04 0.69±0.05 0.42±0.05
Diet.LC-MS[486-lcPos_749-lcPos] (130|5) 0.74±0.02 0.69±0.03 0.70±0.03 0.40±0.05
Diet.LC-MS (130|11) 0.88±0.03 0.83±0.05 0.78±0.05 0.61±0.08
Diet.PostPranFluc3_50 (130|8) 0.59±0.03 0.59±0.04 0.57±0.04 0.16±0.06
Diet.MGm_B1.GRS (130|19) 0.63±0.03 0.64±0.05 0.55±0.05 0.19±0.06
Diet.MGm_B1.LC-MS (130|23) 0.90±0.03 0.84±0.04 0.79±0.06 0.64±0.07
Diet.16S_B.LC-MS (130|8) 0.86±0.02 0.80±0.04 0.76±0.05 0.57±0.05
(samples|features)
ROC-AUC
Sensitivity
Specificity
MCC
0
0.5
1
−1
0
1

45

Supplementary Material 4: Permuted target distributions of best models with diet, 16S-based OTUs or MGmapped gut microbiome taxa with urine metabolites identified by LC-MS

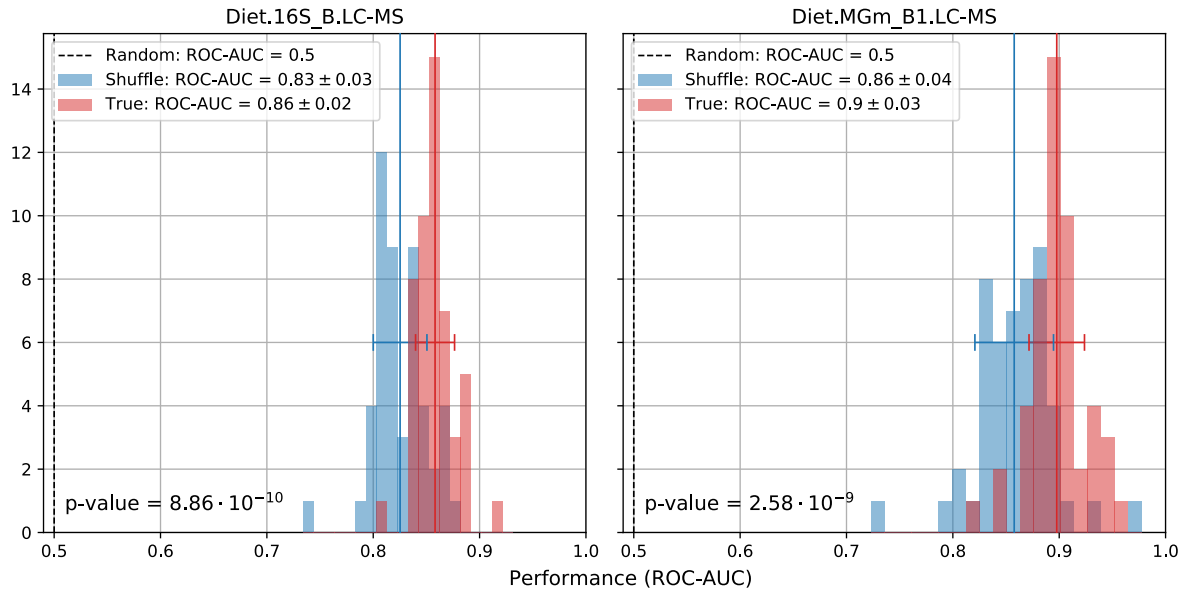


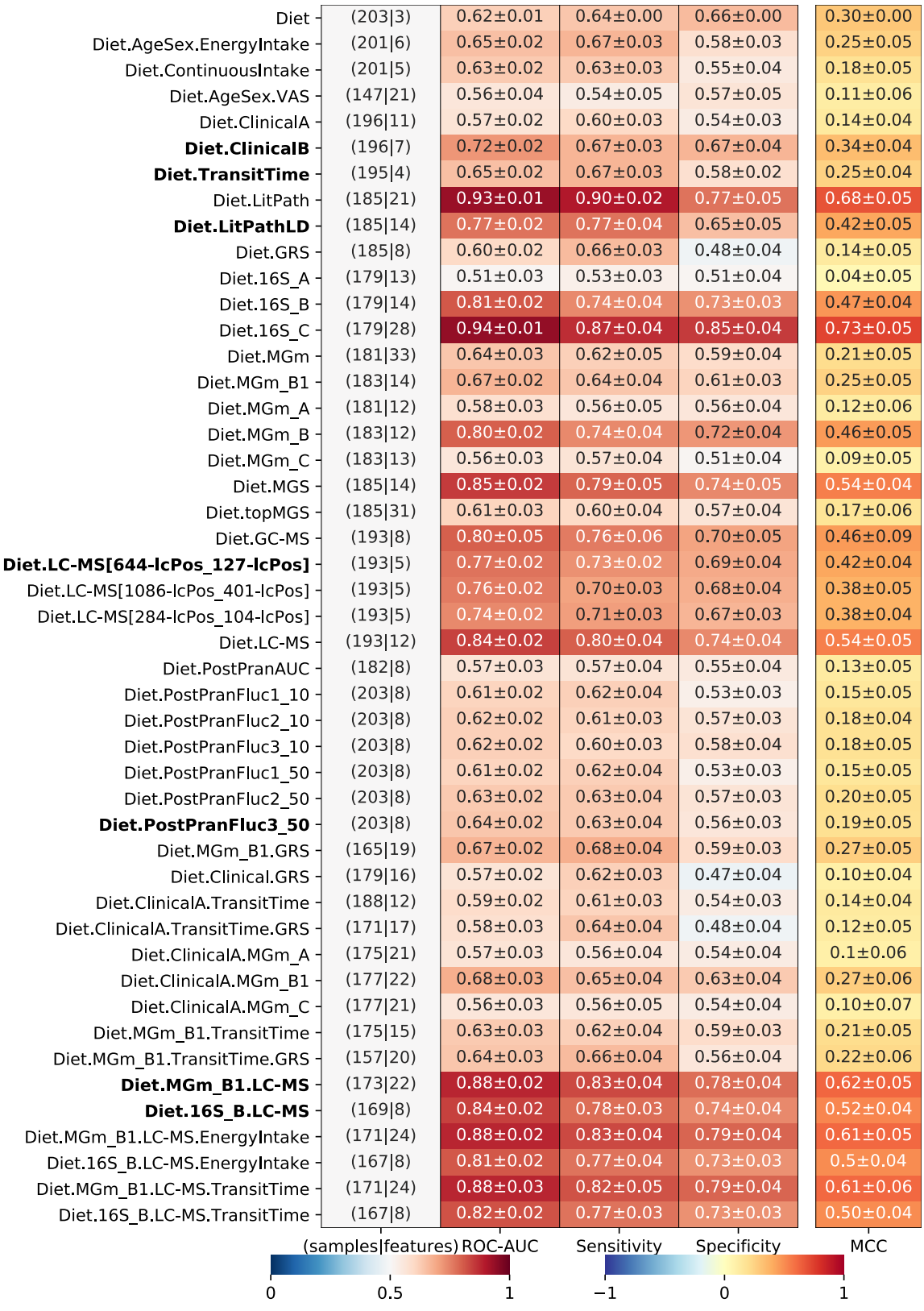
Figure S.5: Permutation tests. ROC-AUC distributions for two best models with data combinations diet, forward selected 16S-based OTUs (left, Diet.16S_B.LC-MS) or butyrate-producing species (right, Diet.MGm_B1.LC-MS) and forward selected urine metabolites identified by LC-MS trained on 130 common samples on the true target (red) and on permuted targets (blue). The black dashed lines denote random ROC-AUC performance of 0.5, while the red and blue lines are the performance means and standard deviations of the model performances.

Supplementary Material 5: Performance of models trained with all data available

Table S.4 below is an extended version of Table 3 and shows the model for each data combination run on the full intersection of samples, i.e. all samples available for a given feature combination. The bold text marks data combinations included in the ensemble model.

Multiple representations for the postprandial features were made, which can be seen in comparison in Table S.4 below, where models have been run with each representation along with the diet features. As seen in the figure, the new representation using the clustered sum of consecutive ones (Diet.PostPranFluc3_50) performs slightly better, which is why this representation was used in subsequent model combinations.

Table S.4: **Test performances for all feature combinations, where models are run on all data available for the combined data types**. The model names in bold are included in the model ensemble. The blue-red colorbar is for area under the receiver operating characteristic curve (ROC-AUC), sensitivity and specificity, while the blue.yellow-red colorbar is for Matthews correlation coefficient (MCC). Abbreviations for model combinations are explained in Table 1.


Diet (203|3) 0.62±0.01 0.64±0.00 0.66±0.00 0.30±0.00
Diet.AgeSex.EnergyIntake (201|6) 0.65±0.02 0.67±0.03 0.58±0.03 0.25±0.05
Diet.ContinuousIntake (201|5) 0.63±0.02 0.63±0.03 0.55±0.04 0.18±0.05
Diet.AgeSex.VAS (147|21) 0.56±0.04 0.54±0.05 0.57±0.05 0.11±0.06
Diet.ClinicalA (196|11) 0.57±0.02 0.60±0.03 0.54±0.03 0.14±0.04
Diet.ClinicalB (196|7) 0.72±0.02 0.67±0.03 0.67±0.04 0.34±0.04
Diet.TransitTime (195|4) 0.65±0.02 0.67±0.03 0.58±0.02 0.25±0.04
Diet.LitPath (185|21) 0.93±0.01 0.90±0.02 0.77±0.05 0.68±0.05
Diet.LitPathLD (185|14) 0.77±0.02 0.77±0.04 0.65±0.05 0.42±0.05
Diet.GRS (185|8) 0.60±0.02 0.66±0.03 0.48±0.04 0.14±0.05
Diet.16S_A (179|13) 0.51±0.03 0.53±0.03 0.51±0.04 0.04±0.05
Diet.16S_B (179|14) 0.81±0.02 0.74±0.04 0.73±0.03 0.47±0.04
Diet.16S_C (179|28) 0.94±0.01 0.87±0.04 0.85±0.04 0.73±0.05
Diet.MGm (181|33) 0.64±0.03 0.62±0.05 0.59±0.04 0.21±0.05
Diet.MGm_B1 (183|14) 0.67±0.02 0.64±0.04 0.61±0.03 0.25±0.05
Diet.MGm_A (181|12) 0.58±0.03 0.56±0.05 0.56±0.04 0.12±0.06
Diet.MGm_B (183|12) 0.80±0.02 0.74±0.04 0.72±0.04 0.46±0.05
Diet.MGm_C (183|13) 0.56±0.03 0.57±0.04 0.51±0.04 0.09±0.05
Diet.MGS (185|14) 0.85±0.02 0.79±0.05 0.74±0.05 0.54±0.04
Diet.topMGS (185|31) 0.61±0.03 0.60±0.04 0.57±0.04 0.17±0.06
Diet.GC-MS (193|8) 0.80±0.05 0.76±0.06 0.70±0.05 0.46±0.09
Diet.LC-MS[644-lcPos_127-lcPos] (193|5) 0.77±0.02 0.73±0.02 0.69±0.04 0.42±0.04
Diet.LC-MS[1086-lcPos_401-lcPos] (193|5) 0.76±0.02 0.70±0.03 0.68±0.04 0.38±0.05
Diet.LC-MS[284-lcPos_104-lcPos] (193|5) 0.74±0.02 0.71±0.03 0.67±0.03 0.38±0.04
Diet.LC-MS (193|12) 0.84±0.02 0.80±0.04 0.74±0.04 0.54±0.05
Diet.PostPranAUC (182|8) 0.57±0.03 0.57±0.04 0.55±0.04 0.13±0.05
Diet.PostPranFluc1_10 (203|8) 0.61±0.02 0.62±0.04 0.53±0.03 0.15±0.05
Diet.PostPranFluc2_10 (203|8) 0.62±0.02 0.61±0.03 0.57±0.03 0.18±0.04
Diet.PostPranFluc3_10 (203|8) 0.62±0.02 0.60±0.03 0.58±0.04 0.18±0.05
Diet.PostPranFluc1_50 (203|8) 0.61±0.02 0.62±0.04 0.53±0.03 0.15±0.05
Diet.PostPranFluc2_50 (203|8) 0.63±0.02 0.63±0.04 0.57±0.03 0.20±0.05
Diet.PostPranFluc3_50 (203|8) 0.64±0.02 0.63±0.04 0.56±0.03 0.19±0.05
Diet.MGm_B1.GRS (165|19) 0.67±0.02 0.68±0.04 0.59±0.03 0.27±0.05
Diet.Clinical.GRS (179|16) 0.57±0.02 0.62±0.03 0.47±0.04 0.10±0.04
Diet.ClinicalA.TransitTime (188|12) 0.59±0.02 0.61±0.03 0.54±0.03 0.14±0.04
Diet.ClinicalA.TransitTime.GRS (171|17) 0.58±0.03 0.64±0.04 0.48±0.04 0.12±0.05
Diet.ClinicalA.MGm_A (175|21) 0.57±0.03 0.56±0.04 0.54±0.04 0.1±0.06
Diet.ClinicalA.MGm_B1 (177|22) 0.68±0.03 0.65±0.04 0.63±0.04 0.27±0.06
Diet.ClinicalA.MGm_C (177|21) 0.56±0.03 0.56±0.05 0.54±0.04 0.10±0.07
Diet.MGm_B1.TransitTime (175|15) 0.63±0.03 0.62±0.04 0.59±0.03 0.21±0.05
Diet.MGm_B1.TransitTime.GRS (157|20) 0.64±0.03 0.66±0.04 0.56±0.04 0.22±0.06
Diet.MGm_B1.LC-MS (173|22) 0.88±0.02 0.83±0.04 0.78±0.04 0.62±0.05
Diet.16S_B.LC-MS (169|8) 0.84±0.02 0.78±0.03 0.74±0.04 0.52±0.04
Diet.MGm_B1.LC-MS.EnergyIntake (171|24) 0.88±0.02 0.83±0.04 0.79±0.04 0.61±0.05
Diet.16S_B.LC-MS.EnergyIntake (167|8) 0.81±0.02 0.77±0.04 0.73±0.03 0.5±0.04
Diet.MGm_B1.LC-MS.TransitTime (171|24) 0.88±0.03 0.82±0.05 0.79±0.04 0.61±0.06
Diet.16S_B.LC-MS.TransitTime (167|8) 0.82±0.02 0.77±0.03 0.73±0.03 0.50±0.04
(samples|features)
ROC-AUC
Sensitivity
Specificity
MCC
0
0.5
1
−1
0
1

Supplementary Material 6: Feature levels by class for microbiome and metabolome

Table S.5: **Feature levels by class for microbiome and metabolomic** features (mean ± standard deviation) if considered very important by the relative Gini coefficient for the number of selected features in Figure 3. The microbiome features are abundances and the metabolites are noted as the logarithm of the relative abundance. NA: not available.

Features	m/z	Retention time	Prevalence (microbiome) or annotation (metabolite)	Prevalence Responders	Prevalence Non responders	Responders	Non responders	p_value of single feature association
3-lcPos	82.02	32.55	NA	NA	NA	-3.82 ± 0.18	-3.85 ± 0.19	0.77
701-lcPos	265.14	323.54	gamma-carboxyethyl hydroxychroman	NA	NA	-3.62 ± 0.16	-3.59 ± 0.19	0.17
241-lcPos	169.10	61.54	NA	NA	NA	-3.93 ± 0.1	-3.95 ± 0.12	0.39
142-lcPos	145.10	325.59	food additive (S1)-Methoxy-3-haptanethiol	NA	NA	-3.78 ± 0.12	-3.8 ± 0.11	0.38
624-lcPos	249.07	40.12	Dipeptide	NA	NA	-3.92 ± 0.09	-3.93 ± 0.09	0.09
921-lcPos	326.04	216.69	Urothion	NA	NA	-3.63 ± 0.12	-3.63 ± 0.11	0.96
1180-lcPos	567.18	246.85	NA	NA	NA	-3.91 ± 0.13	-3.94 ± 0.16	0.37
131-lcNeg	216.92	37.63	NA	NA	NA	-2.85 ± 0.22	-2.83 ± 0.21	0.85
539-lcPos	230.19	81.67	NA	NA	NA	-3.85 ± 0.12	-3.86 ± 0.14	0.44
470-lcPos	219.03	36.54	Polyol OR sugar metabolite	NA	NA	-3.62 ± 0.22	-3.65 ± 0.15	0.29
763-lcPos	280.05	216.36	3-(3,5-dihydroxyphenyl)-1-propanoic acid sulphate OR Dihydrocaffeic acid 3-sulfate OR similar	NA	NA	-4.08 ± 0.22	-4.05 ± 0.22	0.44
557-lcPos	235.09	41.67	Dipeptide	NA	NA	-3.68 ± 0.1	-3.71 ± 0.11	0.1
918-lcPo	325.03	37.20	NA	NA	NA	-3.71 ± 0.31	-3.65 ± 0.21	0.55
680-lcPos	262.04	94.03	2-Methoxyacetaminophen sulfate OR p-Coumaric acid sulfate	NA	NA	-2.49 ± 0.09	-2.48 ± 0.08	0.46
749-lcPos	276.99	451.41	NA	NA	NA	-4.18 ± 0.1	-4.19 ± 0.09	0.37
1182-lcPos	577.13	240.86	NA	NA	NA	-4.05 ± 0.17	-4.07 ± 0.15	0.42
486-lcPos	221.08	38.04	dipeptide L-β-aspartypl-L-serine/Aspartyl-Serine	NA	NA	-4 ± 0.18	-4 ± 0.15	0.5
Ruminococcaceae.	NA	NA	97/169	47/89	50/80	4.74e-03±9.31e-03	3.74e-03±6.03e-03	0.32
Streptococcus. sp	NA	NA	60/169	36/89	24/80	3.50e-03±1.22e-02	4.74e-03±1.81e-02	0.46
F. prausnitzii	NA	NA	172/173	88/89	84/84	9.46e-03±5.33e-03	1.32e-02±9.07e-03	0.01
E. ramulus	NA	NA	173/173	89/89	84/84	7.34e-02±6.28e-02	7.21e-02±5.30e-02	0.41
R. faecis	NA	NA	173/173	89/89	84/84	8.70e-01±1.14e+00	6.60e-01±7.70e-01	0.25

R. intestinalis	NA	NA	124/173	68/89	56/84	7.98e-04±4.81e-04	6.79e-04±4.95e-04	0.12
R. inulinivorans	NA	NA	149/173	79/89	70/84	9.55e-04±4.24e-04	9.52e-04±5.58e-04	0.82
B. fibrisolvens	NA	NA	15/173	5/89	10/84	1.12e-04±5.53e-04	1.90e-04±6.67e-04	0.15
S. variabile	NA	NA	144/173	76/89	68/84	1.24e-03±8.26e-04	1.19e-03±8.98e-04	0.53
A. colihominis	NA	NA	91/173	48/89	43/84	5.62e-04±5.43e-04	5.24e-04±5.26e-04	0.67
B. uniformis	NA	NA	98/173	53/89	45/84	1.07e-03±1.46e-03	1.36e-03±2.32e-03	0.88
E. hallii	NA	NA	13/173	8/89	5/84	8.99e-05±2.88e-04	7.14e-05±3.02e-04	0.47

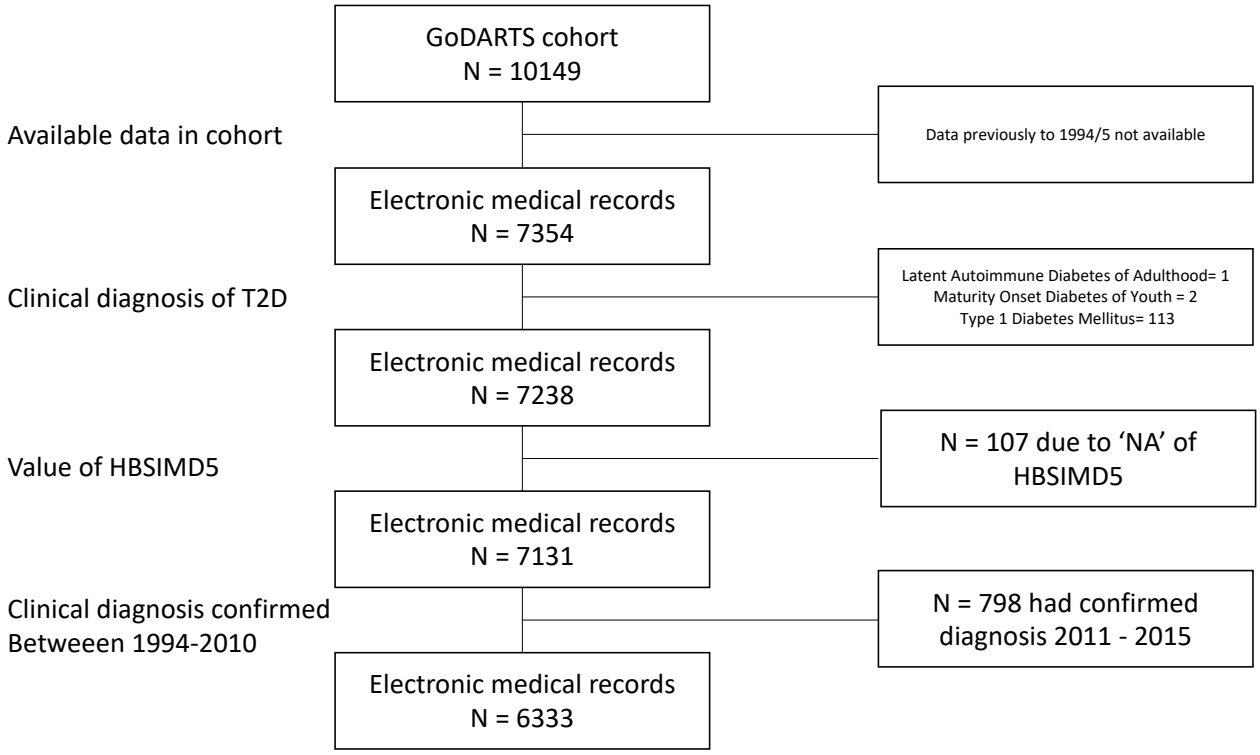
73

APPENDIX B

Appendix B

Supplementary Material

Supplementary Material 1; Inclusion criteria for study



Supplementary Figure 1:Pre-filtering of patients available in the GoDARTS cohort included in the study.

Supplementary Material 2; GoDARTS cohort overview

Supplementary Table 1: Overview of the GoDARTS study cohort after filtering of data (N = 6333). Data is extracted ±6 months from time of diagnosis date confirmed by first HbA1c≥6.5% or first drug prescription.

	Unit	*Mean ± SD, or number of patients*	*%missing*	*Encoding in machine learning models*
Anthropometrics				
Age	Years	61.367 ± 11.419	0%	Numeric
Gender	M/F	M: 3518 / F: 2815	0%	0/1
Weight	Kg	88.544 ± 19.541	18.648%	Numeric
BMI	Kg/m^2	31.644 ± 6.282	18.648%	Numeric
Blood pressure				
Systolic		143.138 ± 20.125	9.648%	Numeric
Diastolic		82.219 ± 11.022	9.648%	Numeric
Life-style				
Smoking status	Y/N	Y: 4730 / N: 1603	0%	1/0
Social deprivation	1 most deprived	1: 1587	0%	One-out-of-K 0/1 encoding
	2	2: 1385		
	3	3: 1074		
	4	4: 1115		
	5 least deprived	5: 1172		
Biochemical blood markers				
HbA1c	%	8.340 ± 1.996	13.106%	Numeric
Cholesterol		5.393 ± 1.389	42.713%	Numeric
LDL		2.928 ± 1.052	77.072%	Numeric
HDL		1.191 ± 0.337	45.666%	Numeric
Triglycerides		3.364 ± 3.926	72.177%	Numeric
Creatinine		79.397 ± 24.060	32.907%	Numeric
ALT		37.636 ± 30.181	66.509%	Numeric
AST		27.743 ± 23.864	98.216%	Numeric
GAD antibody		38.943 ± 236.937	23.133%	One-out-of-K 0/1 encoding for 'missing', '< 11', '>= 11'.
Year of confirmed diagnosis by HbA1c or first drug	Calendar year	2002.073 ± 3.987	0%	Numeric

Drug prescriptions were reported by all prescription encashments. For modelling, dates were compared to baseline timepoint and for ±2months all prescriptions were summed per different diabetes drug. One-out-of-K 0/1 encoding for 'no therapy', 'mono-therapy', 'dual-therapy' or 'triple or more therapy' at the time of interest between each year following 1 year after confirmed type 2 diabetes diagnosis upto 10 years after.

Genetic variants were available from SNP arrays with 77.106.320 from summit study and 87.434.780 from wtccc2 study. For modelling, the genetic risk variants were represented by sparse encoding of the presence of the major or minor allele.

Supplementary Material 3 – %missingness in continuous anthropometry and biochemical variables after data extraction (fixed time point and autoregressive modelled features)

These tables report %missingness for each variable and is reported separately for cases and controls in the models that were trained in the first clinical model with fixed time point data and autoregressive-modelled data. Models were trained using data from year 1 to 10 after diagnosis for the +1Y model, data from year 1 to 9 after diagnosis for the +2Y model, data from year 1 to 8 after diagnosis for the +3Y model, and data from year 1 to 7 after diagnosis for the +4Y model where the %-missingness is given separately. The %-missingness is reported as mean and standard deviation across all datasets included for modelling.

Information of gender, age at confirmed diagnosis, phenotype, year of confirmed diagnosis, social deprivation (hbsimd5), GAD antibody (missing, < 11 or >= 11), drug prescribed at the time of interest (none, mono therapy, dual therapy or triple or more therapy) was complete for all patients.

Supplementary Table 2: This table reports %missingness for each longitudinal extracted variable and is reported separately for cases (patients requiring insulin) in the models that were trained in the first clinical model with fixed time point data.

%missingness (mean ± SD) cases M1	+1Y	+2Y	+3Y	+4Y
BMI	28.73 ± 6.01	30.03 ± 5.11	28.98 ± 5.62	28.27 ± 7.04
DBP	28.13 ± 6.19	27.89 ± 5.40	27.10 ± 5.73	27.30 ± 6.19
SBP	28.13 ± 6.19	27.89 ± 5.40	27.10 ± 5.73	27.30 ± 6.19
WEIGHT	28.73 ± 6.01	30.03 ± 5.11	28.98 ± 5.62	28.27 ± 7.04
ALT	38.82 ± 10.55	40.41 ± 8.52	41.12 ± 9.90	42.85 ± 10.34
AST	99.32 ± 0.69	99.39 ± 0.52	99.07 ± 0.78	98.85 ± 0.78
CHOL	29.83 ± 6.62	30.48 ± 5.70	30.28 ± 6.23	29.75 ± 6.64
CREAT	30.39 ± 8.31	30.79 ± 6.36	31.09 ± 6.50	30.28 ± 7.40
HBA1C_DCCT	26.83 ± 6.44	28.19 ± 4.82	27.31 ± 5.69	27.08 ± 6.02
HDL	30.13 ± 6.86	31.14 ± 6.28	31.01 ± 7.01	30.78 ± 7.95
LDL	57.67 ± 6.70	59.54 ± 7.07	58.15 ± 7.06	57.21 ± 5.64
TRIGS	51.26 ± 4.40	54.24 ± 5.16	53.25 ± 6.75	51.00 ± 6.52

Supplementary Table 3: This table reports %missingness for each longitudinal extracted variable and is reported separately for controls (patients not requiring insulin) in the models that were trained in the first clinical model with fixed time point data.

%missingness (mean ± SD) control M1	+1Y	+2Y	+3Y	+4Y
BMI	39.19 ± 8.44	37.51 ± 6.51	36.08 ± 4.73	34.95 ± 3.23
DBP	37.83 ± 8.48	36.16 ± 6.49	34.75 ± 4.68	33.62 ± 3.14
SBP	37.83 ± 8.48	36.16 ± 6.49	34.75 ± 4.68	33.62 ± 3.14
WEIGHT	39.19 ± 8.44	37.51 ± 6.51	36.08 ± 4.73	34.95 ± 3.23
ALT	47.02 ± 5.36	46.01 ± 4.12	45.28 ± 3.45	44.91 ± 3.36
AST	99.55 ± 0.17	99.54 ± 0.16	99.55 ± 0.14	99.56 ± 0.12
CHOL	39.95 ± 8.13	38.36 ± 6.27	36.95 ± 4.46	35.87 ± 2.94
CREAT	39.57 ± 7.30	38.10 ± 5.52	36.82 ± 3.84	35.91 ± 2.54
HBA1C_DCCT	37.91 ± 8.53	36.24 ± 6.59	34.80 ± 4.73	33.68 ± 3.27
HDL	40.31 ± 7.91	38.75 ± 6.07	37.38 ± 4.30	36.33 ± 2.84
LDL	64.32 ± 4.60	63.37 ± 3.57	62.62 ± 2.63	62.18 ± 2.12
TRIGS	62.26 ± 5.31	61.31 ± 4.12	60.51 ± 2.95	60.05 ± 2.07

*Supplementary Table 4: This table reports %missingness for each longitudinal extracted variable and is reported separately for cases (patients requiring insulin) in the models that were trained in the first clinical model with autoregressive-modelling data. *_tau describes the autoregressive aspect, *_par1 describes the general level and *_par2 describes the linear trend, where * is the variable.*

%missingness (mean ± SD) cases M1 autoregressive data	+1Y	+2Y	+3Y	+4Y
ALT_tau, ALT_par1, ALT_par2	17.94 ± 11.62	20.26 ± 11.42	22.37 ± 13.47	23.33 ± 13.58
AST_tau, AST_par1, AST_par2	97.22 ± 0.54	97.27 ± 0.50	97.50 ± 0.58	97.87 ± 0.55
CHOL_tau, CHOL_par1, CHOL_par2	8.06 ± 5.64	8.30 ± 4.61	9.34 ± 6.49	11.00 ± 8.81
CREAT_tau	6.48 ± 2.96	7.16 ± 3.14	7.99 ± 4.07	8.58 ± 5.19
HBA1C_DCCT_tau, HBA1C_DCCT_par1, HBA1C_DCCT_par2	6.33 ± 2.43	6.34 ± 2.34	6.84 ± 3.16	7.21 ± 4.16
HDL_tau, HDL_par1, HDL_par2	9.75 ± 7.63	10.09 ± 6.61	11.60 ± 7.47	13.33 ± 10.19
LDL_tau, LDL_par1, LDL_par2	33.09 ± 22.44	34.47 ± 20.59	36.65 ± 20.51	40.21 ± 20.49
TRIGS_tau, TRIGS_par1, TRIGS_par2	24.69 ± 17.94	26.74 ± 16.78	28.91 ± 16.93	32.89 ± 18.97

*Supplementary Table 5: This table reports %missingness for each longitudinal extracted variable and is reported separately for controls (patients not requiring insulin) in the models that were trained in the first clinical model with autoregressive-modelling data. *_tau describes the autoregressive aspect, *_par1 describes the general level and *_par2 describes the linear trend, where * is the variable.*

%missingness (mean ± SD) control M1 autoregressive data	+1Y	+2Y	+3Y	+4Y
ALT_tau, ALT_par1, ALT_par2	15.60 ± 11.27	16.51 ± 11.40	17.57 ± 11.38	18.89 ± 11.27
AST_tau, AST_par1, AST_par2	97.84 ± 0.38	97.91 ± 0.40	97.99 ± 0.41	98.07 ± 0.41
CHOL_tau, CHOL_par1, CHOL_par2	7.76 ± 4.85	8.04 ± 5.07	8.36 ± 5.25	8.70 ± 5.30
CREAT_tau	6.34 ± 2.69	6.47 ± 2.81	6.61 ± 2.88	6.77 ± 2.91
HBA1C_DCCT_tau, HBA1C_DCCT_par1, HBA1C_DCCT_par2	5.77 ± 1.74	5.85 ± 1.82	5.92 ± 1.87	6.01 ± 1.87
HDL_tau, HDL_par1, HDL_par2	8.85 ± 5.89	9.21 ± 6.08	9.62 ± 6.25	10.09 ± 6.26
LDL_tau, LDL_par1, LDL_par2	32.00 ± 18.20	33.76 ± 18.25	35.80 ± 18.23	38.14 ± 18.09

TRIGS_tau, TRIGS_par1, TRIGS_par2	27.46 ± 16.18	29.10 ± 16.37	30.99 ± 16.51	33.13 ± 16.47

Supplementary Material 4; Diabetes drug prescriptions

BNF_code '6.1.2.1' = sulphonylureas.
BNF_code '6.1.2.2'= metformin.
BNF_code '6.1.2.3'= all other diabetes drugs.

All other diabetes drugs were counted by the following categories:

3. 'PIOGLITAZONE' 'ROSIGLITAZONE' = TZDs
4. 'ACARBOSE' = Acarbose
5. 'ALOGLIPTIN' 'LINAGLIPTIN' 'SAXAGLIPTIN' 'SITAGLIPTIN' 'VILDAGLIPTIN' = DDP4 Inhibitors
6. 'NATEGLINIDE' 'REPAGLINIDE' = glinides
7. 'EXENATIDE' 'LIRAGLUTIDE' 'LIXISENATIDE'= GLP1
8. 'CANAGLIFLOZIN' 'DAPAGLIFLOZIN' 'EMPAGLIFLOZIN'= SGLT2

OTHER ANTIDIABETICS=These were ignored for modelling.

Supplementary Material 5; Number of patients in trained models

Supplementary Table 6: Prediction model overview and the number of patients as case/controls according to the time to insulin phenotype for the given year.

Dataset extraction for year from confirmed diagnosis	+1 year	+ 1 year (N)	+ 2 year	+ 2 year (N)	+ 3 year	+ 3 year (N)	+ 4 year	+ 4 year (N)
1	1-2	209/5878	2-3	241/5637	3-4	270/5367	4-5	300/5066
2	2-3	241/5637	3-4	270/5367	4-5	300/5066	5-6	235/4832
3	3-4	270/5367	4-5	300/5066	5-6	235/4832	6-7	218/4614
4	4-5	300/5066	5-6	235/4832	6-7	218/4614	7-8	180/4432
5	5-6	235/4832	6-7	218/4614	7-8	180/4432	8-9	181/4252
6	6-7	218/4614	7-8	180/4432	8-9	181/4252	9-10	152/4100
7	7-8	180/4432	8-9	181/4252	9-10	152/4100	10-11	113/3988
8	8-9	181/4252	9-10	152/4100	10-11	113/3988	-	-
9	9-10	152/4100	10-11	113/3988	-	-	-	-
10	10-11	113/3988	-	-	-	-	-	-

Supplementary Material 6; %death in machine learning models for patients that require insulin in a +1Y, +2Y, +3Y and +4Y time horizon

Supplementary Table 7: Reported %death events of cases used for training the M1 models across +1Y, +2Y, +3Y and +4Y.

% death amongst cases M1 within prediction window	+ 1Y	+ 2Y	+ 3Y	+ 4Y
Year 1	1.43	0.41	1.09	1.30
Year 2	0.41	1.09	1.30	1.27
Year 3	1.09	1.30	1.27	0.45
Year 4	1.30	1.27	0.45	0.55
Year 5	1.27	0.45	0.55	2.72
Year 6	0.45	0.55	2.72	1.31
Year 7	0.55	2.72	1.31	2.63
Year 8	2.72	1.31	2.63	-
Year 9	1.31	2.63	-	-
Year 10	2.63	-	-	-

Supplementary Material 7; Machine learning modelling and performance

If an input feature contained > 80% missing data prior to cross-validation, it was removed prior to training of the model (AST was removed for all models). Continuous variables in training and test sets were standardized within the outer and inner cross-validation by a z-score from the training sets. Missing values were substituted by extreme single value imputation (value: -999). Due to the class imbalance, the largest group was randomly down-sampled to achieve balanced training. ANNs were trained where optimization of the number of hidden nodes (Number of nodes in hidden layer optimized in artificial neural network; 5,10,15,20) was guided by ROC-AUC with a weight decay of 1 to avoid overfitting and early stopping on a maximum 1000 epochs. For the LASSO regression, the models were optimized for lambda (Lambda optimization in LASSO regression; 0.0001,0.0002, 0.0003, 0.0004, 0.0005, 0.0006, 0.0007, 0.0008, 0.0009, 0.001, 0.002, 0.003, 0.004, 0.005, 0.006, 0.007, 0.008, 0.009, 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9). The logistic regression was trained in a similar setup as the ANN and LASSO regression.

Performance of the ANN and LASSO models was reported using the average of the 5-fold test performance from the outer level of the cross-validation. The following performance measurements are reported in the study (TP = true positive, FP = false positive, TN = true negative, FN = false negative):

$$ROC - AUC$$

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

$$J = sensitivity + specificity - 1$$

$$Positive\ predictive\ value\ (PPV) = \frac{TP}{TP + FP}$$

$$Negative\ predictive\ value\ (NPV) = \frac{TN}{TN + FN}$$

The feature importance was reported for models across multiple years were summed per outer cross-validation fold and was normalized by division from the number of models included.

Supplementary Material 8; Genetic risk scores
Calculation of weighted and unweighted genetic risk scores (GRS);

$$GRS_{unweighted} = \sum_{SNP}(number\ of\ risk\ allele)$$

$$GRS_{weighted} = \sum_{SNP}(OR * number\ of\ risk\ allele)$$

Supplementary Material 9; Results of +1, +2, +3 +4 years prediction by artificial neural networks using datasets modelled by the fixed-time point approach or autoregressive approach
Overall performance summarized for the feature engineering of the fixed time-point extraction (close data by ±6 months) and autoregressive modelling (historical modelling of all previous measurements).

Supplementary Table 8: Average ROC-AUC performance for models trained from year 1-10 using the fixed time-point engineered dataset, the autoregressive modelled dataset and the fixed time-point and autoregressive datasets combined. On average, use of the fixed time-point data resulted in the highest ROC-AUC for +1Y, +2Y, +3Y, +4Y. The table shows mean ± standard deviation.

	Fixed time point data **ROC-AUC**	Autoregressive data **ROC-AUC**	Fixed time point data And autoregressive data **ROC-AUC**
+1Y	0.83 ± 0.04	0.78 ± 0.04	0.78 ± 0.04
+2Y	0.73 ± 0.04	0.71 ± 0.05	0.67 ± 0.05
+3Y	0.69 ± 0.05	0.68 ± 0.04	0.63 ± 0.04
+4Y	0.66 ± 0.06	0.66 ± 0.05	0.63 ± 0.05

Supplementary Table 9: Detailed performance of +1, +2, +3, +4 modelled with clinical data by the fixed time-point datasets extracted from year 1 to 10, clinical data by the autoregressive approach datasets extracted from year 1 to 10 and clinical data by the fixed time-point approach and autoregressive combined datasets extracted from year 1 to 10 using artificial neural networks. Performance is given by ROC-AUC, sensitivity, specificity, MCC: Matthews correlation coefficient, J: Youden's index on the test sets by mean ± standard deviation from five-fold 2-level cross-validation.

+ 1Y prediction Fixed time-point	**ROC-AUC**	**Sensitivity**	**Specificity**	**MCC**	**J**
Year 1	0.76 ± 0.04	0.69 ± 0.10	0.66 ± 0.05	0.14 ± 0.04	0.36 ± 0.10
Year 2	0.82 ± 0.05	0.80 ± 0.13	0.71 ± 0.04	0.22 ± 0.04	0.51 ± 0.10
Year 3	0.84 ± 0.03	0.74 ± 0.06	0.77 ± 0.04	0.25 ± 0.03	0.51 ± 0.05
Year 4	0.84 ± 0.01	0.76 ± 0.06	0.74 ± 0.04	0.26 ± 0.03	0.51 ± 0.06
Year 5	0.82 ± 0.05	0.68 ± 0.14	0.75 ± 0.06	0.21 ± 0.05	0.43 ± 0.11
Year 6	0.84 ± 0.04	0.67 ± 0.06	0.86 ± 0.03	0.31 ± 0.06	0.54 ± 0.08
Year 7	0.81 ± 0.05	0.67 ± 0.08	0.86 ± 0.02	0.27 ± 0.04	0.52 ± 0.07
Year 8	0.84 ± 0.03	0.69 ± 0.09	0.88 ± 0.01	0.32 ± 0.05	0.56 ± 0.09

Year 9	0.85 ± 0.01	0.70 ± 0.03	0.85 ± 0.01	0.28 ± 0.02	0.56 ± 0.03
Year 10	0.85 ± 0.04	0.71 ± 0.09	0.84 ± 0.02	0.23 ± 0.02	0.54 ± 0.07
+ 2Y prediction Fixed time-point	**ROC-AUC**	**Sensitivity**	**Specificity**	**MCC**	**J**
Year 1	0.70 ± 0.03	0.59 ± 0.03	0.68 ± 0.03	0.12 ± 0.02	0.28 ± 0.04
Year 2	0.74 ± 0.03	0.64 ± 0.04	0.71 ± 0.04	0.16 ± 0.02	0.35 ± 0.04
Year 3	0.73 ± 0.04	0.67 ± 0.14	0.66 ± 0.05	0.16 ± 0.04	0.33 ± 0.10
Year 4	0.71 ± 0.03	0.64 ± 0.09	0.65 ± 0.02	0.13 ± 0.04	0.29 ± 0.09
Year 5	0.75 ± 0.06	0.66 ± 0.12	0.73 ± 0.04	0.18 ± 0.05	0.39 ± 0.10
Year 6	0.72 ± 0.02	0.60 ± 0.11	0.72 ± 0.05	0.14 ± 0.04	0.32 ± 0.09
Year 7	0.76 ± 0.04	0.64 ± 0.08	0.79 ± 0.03	0.20 ± 0.02	0.43 ± 0.05
Year 8	0.71 ± 0.06	0.62 ± 0.07	0.76 ± 0.03	0.16 ± 0.03	0.38 ± 0.07
Year 9	0.76 ± 0.02	0.69 ± 0.04	0.73 ± 0.03	0.16 ± 0.03	0.42 ± 0.06
+ 3Y prediction Fixed time-point	**ROC-AUC**	**Sensitivity**	**Specificity**	**MCC**	**J**
Year 1	0.67 ± 0.05	0.58 ± 0.13	0.68 ± 0.08	0.12 ± 0.05	0.26 ± 0.12
Year 2	0.67 ± 0.02	0.60 ± 0.04	0.62 ± 0.03	0.11 ± 0.01	0.22 ± 0.03
Year 3	0.67 ± 0.04	0.65 ± 0.06	0.62 ± 0.03	0.12 ± 0.03	0.27 ± 0.07
Year 4	0.70 ± 0.01	0.60 ± 0.08	0.69 ± 0.04	0.13 ± 0.03	0.28 ± 0.06
Year 5	0.68 ± 0.04	0.59 ± 0.07	0.67 ± 0.03	0.11 ± 0.03	0.27 ± 0.07
Year 6	0.70 ± 0.06	0.60 ± 0.10	0.73 ± 0.04	0.14 ± 0.05	0.32 ± 0.11
Year 7	0.71 ± 0.07	0.62 ± 0.11	0.73 ± 0.03	0.14 ± 0.04	0.35 ± 0.10
Year 8	0.73 ± 0.09	0.67 ± 0.16	0.72 ± 0.05	0.14 ± 0.04	0.39 ± 0.13
+ 4Y prediction Fixed time-point	**ROC-AUC**	**Sensitivity**	**Specificity**	**MCC**	**J**
Year 1	0.66 ± 0.07	0.62 ± 0.09	0.63 ± 0.02	0.12 ± 0.05	0.25 ± 0.11
Year 2	0.64 ± 0.07	0.61 ± 0.12	0.61 ± 0.02	0.09 ± 0.05	0.22 ± 0.11
Year 3	0.67 ± 0.03	0.56 ± 0.08	0.70 ± 0.03	0.12 ± 0.03	0.26 ± 0.07
Year 4	0.65 ± 0.04	0.54 ± 0.14	0.65 ± 0.08	0.08 ± 0.04	0.19 ± 0.09
Year 5	0.67 ± 0.03	0.55 ± 0.06	0.72 ± 0.05	0.12 ± 0.02	0.27 ± 0.04
Year 6	0.69 ± 0.03	0.57 ± 0.06	0.72 ± 0.03	0.12 ± 0.02	0.28 ± 0.06
Year 7	0.67 ± 0.12	0.61 ± 0.14	0.73 ± 0.03	0.13 ± 0.07	0.34 ± 0.17
+ 1Y prediction Autoregressive	**ROC-AUC**	**Sensitivity**	**Specificity**	**MCC**	**J**
Year 1	0.80 ± 0.03	0.73 ± 0.09	0.75 ± 0.03	0.19 ± 0.03	0.47 ± 0.09
Year 2	0.81 ± 0.02	0.71 ± 0.05	0.75 ± 0.01	0.20 ± 0.03	0.46 ± 0.06
Year 3	0.77 ± 0.02	0.67 ± 0.07	0.72 ± 0.02	0.18 ± 0.03	0.39 ± 0.06
Year 4	0.75 ± 0.02	0.69 ± 0.05	0.72 ± 0.02	0.20 ± 0.01	0.41 ± 0.03
Year 5	0.76 ± 0.03	0.69 ± 0.09	0.69 ± 0.01	0.17 ± 0.03	0.37 ± 0.08

Year 6	0.78 ± 0.03	0.76 ± 0.05	0.70 ± 0.02	0.21 ± 0.02	0.46 ± 0.04
Year 7	0.75 ± 0.04	0.71 ± 0.06	0.69 ± 0.02	0.16 ± 0.02	0.39 ± 0.04
Year 8	0.76 ± 0.05	0.71 ± 0.11	0.67 ± 0.02	0.16 ± 0.04	0.38 ± 0.11
Year 9	0.80 ± 0.05	0.82 ± 0.08	0.67 ± 0.03	0.19 ± 0.04	0.50 ± 0.09
Year 10	0.81 ± 0.04	0.79 ± 0.08	0.71 ± 0.02	0.18 ± 0.03	0.50 ± 0.09
+ 2Y prediction Autoregressive	**ROC-AUC**	**Sensitivity**	**Specificity**	**MCC**	**J**
Year 1	0.66 ± 0.04	0.63 ± 0.05	0.61 ± 0.03	0.10 ± 0.03	0.24 ± 0.07
Year 2	0.72 ± 0.02	0.67 ± 0.05	0.66 ± 0.04	0.15 ± 0.01	0.33 ± 0.03
Year 3	0.66 ± 0.04	0.62 ± 0.10	0.63 ± 0.03	0.11 ± 0.03	0.24 ± 0.07
Year 4	0.67 ± 0.02	0.59 ± 0.03	0.62 ± 0.02	0.09 ± 0.02	0.22 ± 0.04
Year 5	0.74 ± 0.02	0.72 ± 0.05	0.65 ± 0.02	0.16 ± 0.02	0.37 ± 0.05
Year 6	0.71 ± 0.03	0.65 ± 0.05	0.65 ± 0.01	0.12 ± 0.02	0.30 ± 0.05
Year 7	0.70 ± 0.04	0.66 ± 0.07	0.65 ± 0.03	0.13 ± 0.02	0.32 ± 0.05
Year 8	0.70 ± 0.03	0.65 ± 0.04	0.63 ± 0.05	0.11 ± 0.01	0.28 ± 0.03
Year 9	0.79 ± 0.03	0.78 ± 0.12	0.68 ± 0.03	0.16 ± 0.04	0.46 ± 0.13
+ 3Y prediction Autoregressive	**ROC-AUC**	**Sensitivity**	**Specificity**	**MCC**	**J**
Year 1	0.68 ± 0.05	0.63 ± 0.09	0.64 ± 0.01	0.12 ± 0.03	0.27 ± 0.08
Year 2	0.65 ± 0.04	0.64 ± 0.07	0.59 ± 0.02	0.11 ± 0.03	0.23 ± 0.06
Year 3	0.66 ± 0.04	0.61 ± 0.05	0.63 ± 0.03	0.11 ± 0.02	0.25 ± 0.05
Year 4	0.68 ± 0.05	0.63 ± 0.06	0.63 ± 0.02	0.11 ± 0.03	0.26 ± 0.06
Year 5	0.68 ± 0.02	0.62 ± 0.04	0.66 ± 0.03	0.12 ± 0.01	0.29 ± 0.02
Year 6	0.65 ± 0.04	0.62 ± 0.06	0.61 ± 0.03	0.09 ± 0.03	0.23 ± 0.07
Year 7	0.68 ± 0.04	0.63 ± 0.06	0.62 ± 0.03	0.10 ± 0.01	0.26 ± 0.04
Year 8	0.72 ± 0.04	0.67 ± 0.07	0.64 ± 0.01	0.11 ± 0.02	0.32 ± 0.07
+ 4Y prediction Autoregressive	**ROC-AUC**	**Sensitivity**	**Specificity**	**MCC**	**J**
Year 1	0.64 ± 0.04	0.64 ± 0.06	0.59 ± 0.03	0.11 ± 0.03	0.23 ± 0.07
Year 2	0.63 ± 0.04	0.54 ± 0.13	0.63 ± 0.09	0.08 ± 0.03	0.17 ± 0.06
Year 3	0.64 ± 0.04	0.59 ± 0.07	0.62 ± 0.03	0.09 ± 0.03	0.21 ± 0.07
Year 4	0.67 ± 0.02	0.63 ± 0.08	0.61 ± 0.02	0.09 ± 0.02	0.24 ± 0.06
Year 5	0.68 ± 0.06	0.65 ± 0.13	0.60 ± 0.03	0.10 ± 0.05	0.25 ± 0.12
Year 6	0.68 ± 0.04	0.67 ± 0.10	0.62 ± 0.04	0.11 ± 0.04	0.29 ± 0.11
Year 7	0.69 ± 0.06	0.65 ± 0.10	0.63 ± 0.02	0.09 ± 0.03	0.28 ± 0.09
+ 1Y prediction Autoregressive and Fixed time-point	**ROC-AUC**	**Sensitivity**	**Specificity**	**MCC**	**J**
Year 1	0.62 ± 0.08	0.67 ± 0.15	0.54 ± 0.07	0.08 ± 0.05	0.21 ± 0.14
Year 2	0.76 ± 0.05	0.73 ± 0.07	0.69 ± 0.03	0.18 ± 0.04	0.41 ± 0.08
Year 3	0.73 ± 0.04	0.61 ± 0.08	0.75 ± 0.05	0.17 ± 0.03	0.36 ± 0.06

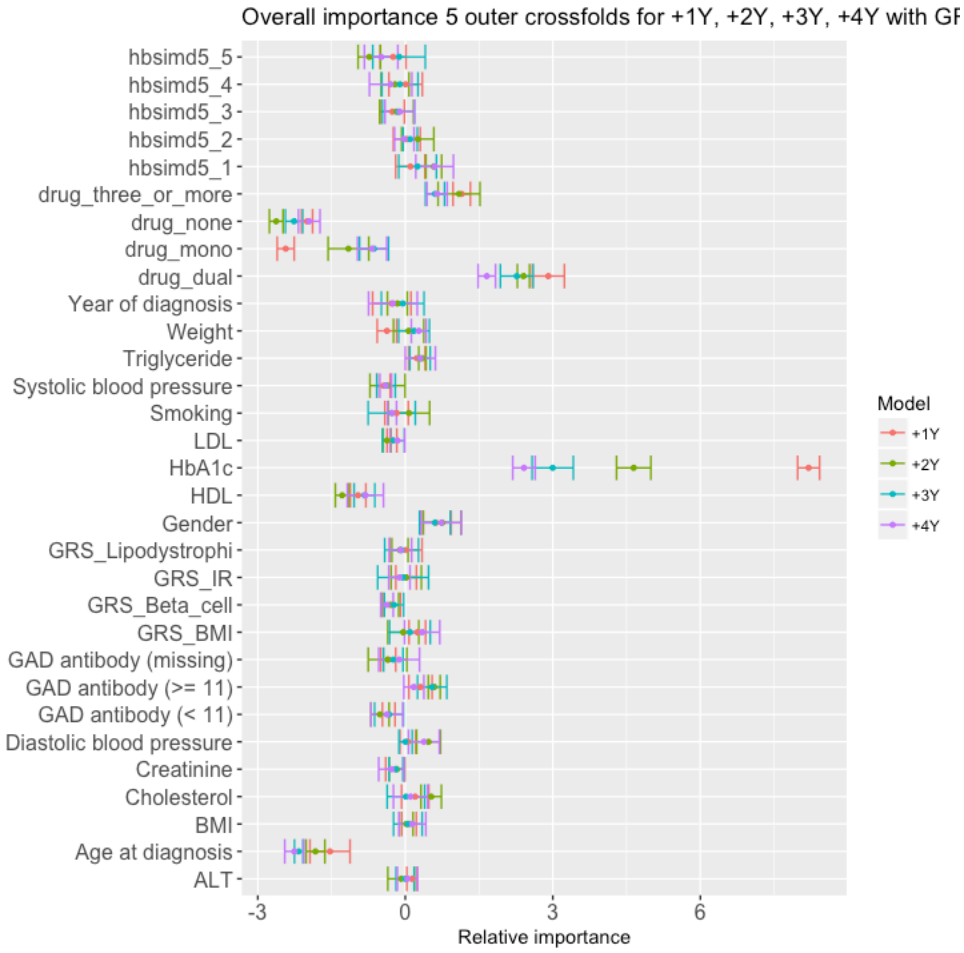
Year 4	0.79 ± 0.03	0.74 ± 0.09	0.73 ± 0.03	0.23 ± 0.04	0.46 ± 0.08
Year 5	0.80 ± 0.03	0.73 ± 0.01	0.72 ± 0.04	0.21 ± 0.02	0.45 ± 0.03
Year 6	0.81 ± 0.04	0.70 ± 0.07	0.79 ± 0.04	0.24 ± 0.02	0.48 ± 0.04
Year 7	0.81 ± 0.04	0.70 ± 0.11	0.75 ± 0.04	0.20 ± 0.03	0.45 ± 0.08
Year 8	0.82 ± 0.02	0.75 ± 0.06	0.76 ± 0.02	0.23 ± 0.02	0.50 ± 0.06
Year 9	0.83 ± 0.07	0.73 ± 0.15	0.77 ± 0.02	0.22 ± 0.05	0.50 0.13
Year 10	0.84 ± 0.03	0.69 ± 0.04	0.78 ± 0.05	0.18 ± 0.03	0.47 0.04
+ 2Y prediction Autoregressive and Fixed time-point	**ROC-AUC**	**Sensitivity**	**Specificity**	**MCC**	**J**
Year 1	0.56 ± 0.05	0.49 ± 0.09	0.63 ± 0.08	0.05 ± 0.02	0.12 ± 0.05
Year 2	0.65 ± 0.03	0.59 ± 0.09	0.65 ± 0.05	0.10 ± 0.03	0.24 ± 0.07
Year 3	0.66 ± 0.03	0.62 ± 0.07	0.61 ± 0.06	0.11 ± 0.03	0.23 ± 0.06
Year 4	0.62 ± 0.04	0.55 ± 0.08	0.60 ± 0.03	0.06 ± 0.03	0.15 ± 0.07
Year 5	0.70 ± 0.04	0.62 ± 0.10	0.67 ± 0.02	0.12 ± 0.04	0.28 ± 0.09
Year 6	0.72 ± 0.07	0.69 ± 0.12	0.66 ± 0.03	0.14 ± 0.05	0.35 ± 0.12
Year 7	0.69 ± 0.05	0.61 ± 0.12	0.67 ± 0.03	0.12 ± 0.05	0.28 ± 0.11
Year 8	0.73 ± 0.05	0.70 ± 0.10	0.67 ± 0.04	0.15 ± 0.04	0.37 ± 0.10
Year 9	0.75 ± 0.03	0.73 ± 0.07	0.68 ± 0.02	0.14 ± 0.03	0.41 ± 0.08
+ 3Y prediction Autoregressive and Fixed time-point	**ROC-AUC**	**Sensitivity**	**Specificity**	**MCC**	**J**
Year 1	0.58 ± 0.05	0.53 ± 0.14	0.62 ± 0.05	0.06 ± 0.04	0.15 ± 0.10
Year 2	0.61 ± 0.02	0.57 ± 0.06	0.57 ± 0.05	0.07 ± 0.02	0.14 ± 0.05
Year 3	0.59 ± 0.06	0.49 ± 0.09	0.62 ± 0.04	0.05 ± 0.03	0.11 ± 0.08
Year 4	0.65 ± 0.06	0.59 ± 0.10	0.65 ± 0.02	0.10 ± 0.03	0.24 ± 0.08
Year 5	0.64 ± 0.03	0.61 ± 0.07	0.62 ± 0.03	0.09 ± 0.03	0.23 ± 0.06
Year 6	0.64 ± 0.03	0.57 ± 0.12	0.63 ± 0.03	0.08 ± 0.04	0.20 ± 0.10
Year 7	0.71 ± 0.04	0.63 ± 0.10	0.67 ± 0.02	0.11 ± 0.03	0.29 ± 0.08
Year 8	0.65 ± 0.10	0.64 ± 0.15	0.61 ± 0.03	0.08 ± 0.05	0.25 ± 0.16
+ 4Y prediction Autoregressive and Fixed time-point	**ROC-AUC**	**Sensitivity**	**Specificity**	**MCC**	**J**
Year 1	0.57 ± 0.07	0.59 ± 0.14	0.48 ± 0.13	0.03 ± 0.03	0.07 ± 0.07
Year 2	0.57 ± 0.06	0.52 ± 0.09	0.61 ± 0.05	0.06 ± 0.04	0.13 ± 0.09
Year 3	0.64 ± 0.04	0.59 ± 0.05	0.61 ± 0.06	0.09 ± 0.03	0.21 ± 0.07
Year 4	0.62 ± 0.03	0.61 ± 0.07	0.58 ± 0.02	0.07 ± 0.02	0.19 ± 0.05
Year 5	0.63 ± 0.03	0.59 ± 0.13	0.61 ± 0.03	0.08 ± 0.05	0.20 ± 0.12
Year 6	0.67 ± 0.04	0.60 ± 0.08	0.64 ± 0.03	0.09 ± 0.03	0.24 ± 0.08
Year 7	0.69 ± 0.09	0.64 ± 0.16	0.64 ± 0.05	0.10 ± 0.04	0.28 ± 0.13

Supplementary Material 10; TTI:clinical model with genetic subset of patients (M1)
Since 5922 of the 6333 patients had genotype information available, we retrained the TTI:clinical model on these patients only which showed similar performance to the original data set with N = 6333.

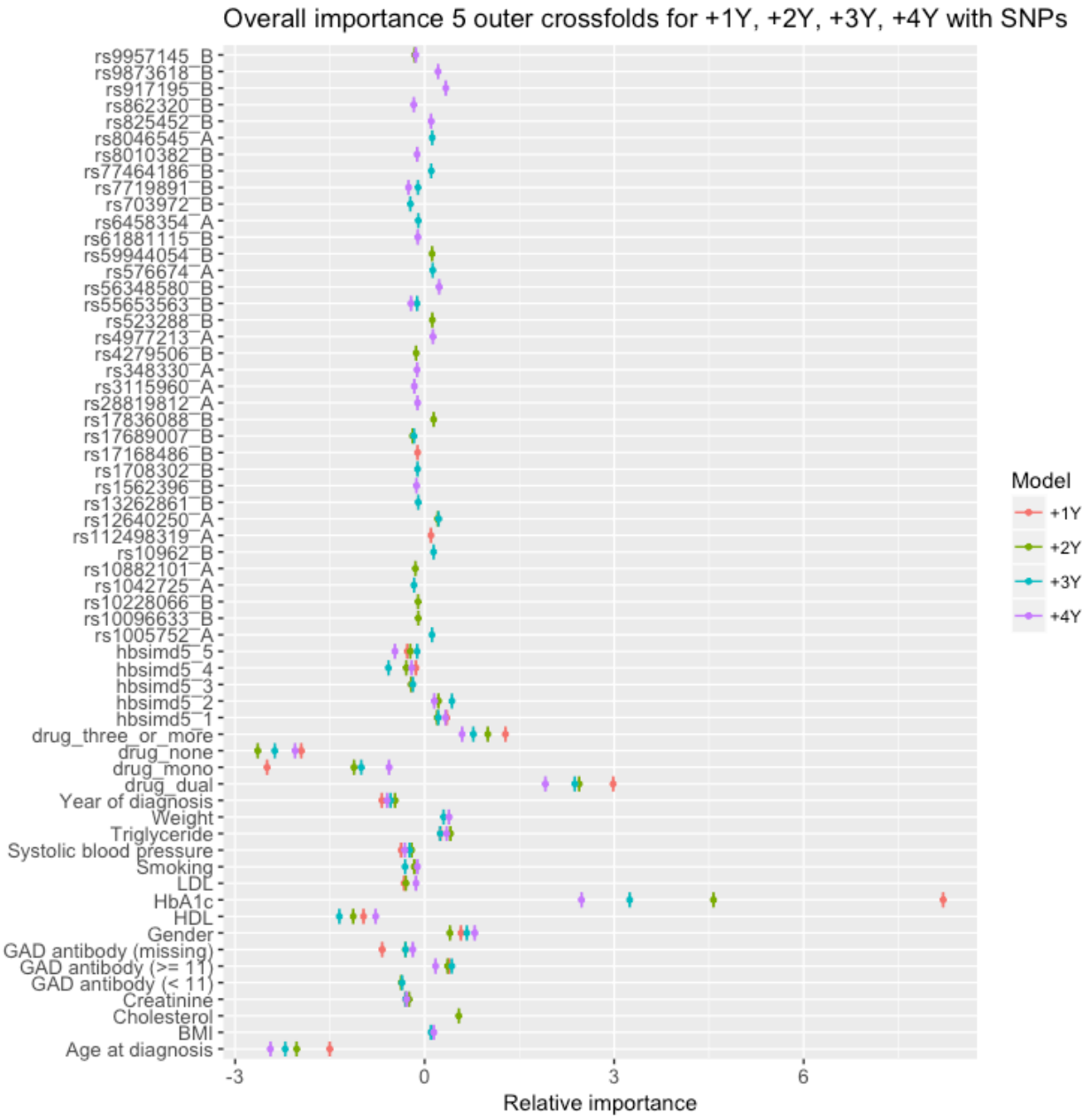
Supplementary Table 10: M1 trained on clinical data fixed time point data with artificial neural networks on patients where genotype is available. Performance is given by ROC-AUC, sensitivity, specificity, MCC: Matthews correlation coefficient, J: Youden's index on the test sets (mean ± standard deviation).

M1 clinical (Fixed time point) on genetic subset	*ROC-AUC*	*Sensitivity*	*Specificity*	*MCC*	*J*
+1Y	0.82 ± 0.05	0.70 ± 0.11	0.79 ± 0.08	0.24 ± 0.06	0.48 ± 0.09
+2Y	0.70 ± 0.05	0.61 ± 0.09	0.69 ± 0.05	0.14 ± 0.04	0.30 ± 0.07
+3Y	0.66 ± 0.05	0.58 ± 0.11	0.67 ± 0.06	0.11 ± 0.04	0.24 ± 0.10
+4Y	0.64 ± 0.05	0.55 ± 0.09	0.65 ± 0.06	0.09 ± 0.04	0.20 ± 0.09

Supplementary Material 11; Feature importance for models +1Y, +2Y, +3Y and +4Y for artificial neural networks with fixed time point data and genetic risk scores and fixed time point data and genetic risk variants



Supplementary Figure 2: Relative feature importance (mean ±standard deviation) for artificial neural network modelled with fixed time point data (M1) with genetic risk scores of lipodystrophy, BMI, insulin resistance (IR) and beta cell function (Beta_cell).



Supplementary Figure 3: Relative feature importance (mean ±standard deviation) for artificial neural network modelled with fixed time point data (M1) with forward selected SNPs. The relative importance is normalised by the number of outer cross-validation folds and number of years for datasets extracted since the time of diagnosis. Only features where relative importance >=0.1 or <= -0.1 was included in plot. '_A' = allele A, '_B' = allele B. TTI_1 forward SNP, N(feature) = 244, TTI_2 forward SNP, N(feature) = 273, TTI_3 forward SNP, N(feature) = 262, TTI_4 forward SNP, N(feature) = 243.

Supplementary Material 12; Performance of LASSO regression and Logistic regression models for +1, +2, +3 and +4 years TTI models.

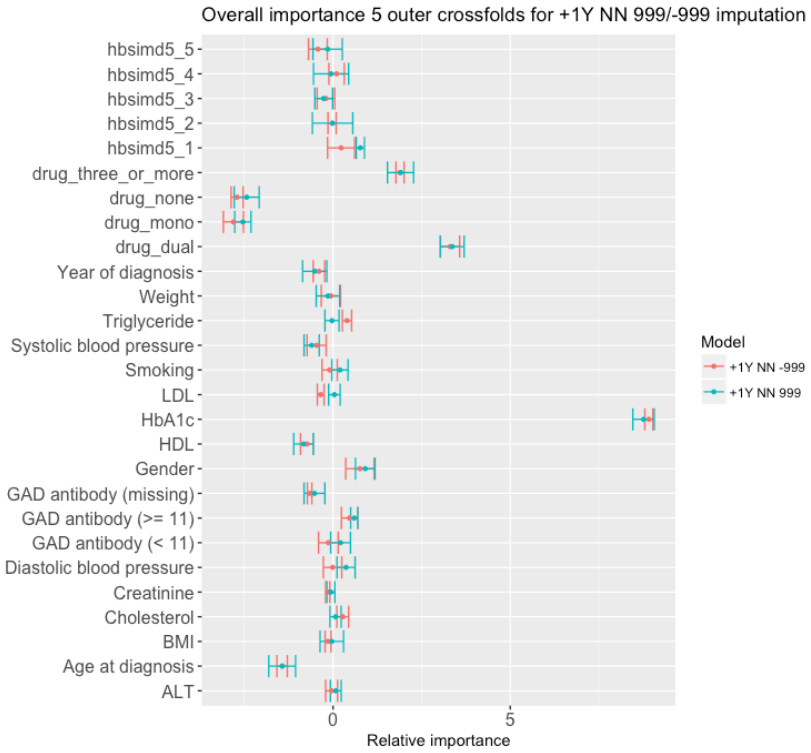
The performance of the logistic regression +1Y model is not reported, as this model was overfitted and thus +1Y could only be trained by introduced regularization of the linear coefficients.

Supplementary Table 11: Performance is given by ROC-AUC, sensitivity, specificity, MCC: Matthews correlation coefficient, J: Youden's index on the test set for LASSO regression and logistic regression.

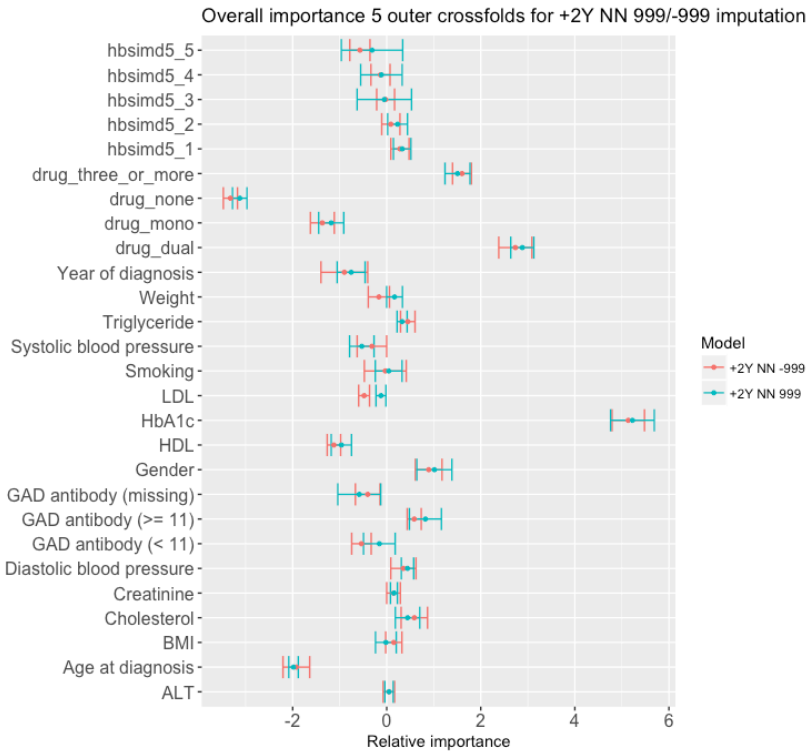
M1 clinical (Fixed time point) LASSO	***ROC-AUC***	***Sensitivity***	***Specificity***	***MCC***	***J***
+1Y	0.66 ± 0.05	0.63 ± 0.16	0.68 ± 0.11	0.14 ± 0.04	0.31 ± 0.11
+2Y	0.64 ± 0.05	0.61 ± 0.10	0.66 ± 0.06	0.11 ± 0.04	0.27 ± 0.10
+3Y	0.62 ± 0.05	0.57 ± 0.14	0.66 ± 0.07	0.10 ± 0.04	0.23 ± 0.10
+4Y	0.62 ± 0.05	0.58 ± 0.14	0.66 ± 0.07	0.10 ± 0.04	0.24 ± 0.10
M1 clinical (Fixed time point) for Drug at diagnosis (none/mono/dual/triple or more) + calendar year of diagnosis + HbA1c + BMI + age at diagnosis, Logistic regression	***ROC-AUC***	***Sensitivity***	***Specificity***	***MCC***	***J***
+2Y	0.66 ± 0.04	0.62 ± 0.09	0.69 ± 0.06	0.13 ± 0.03	0.31 ± 0.09
+3Y	0.64 ± 0.04	0.59 ± 0.07	0.69 ± 0.04	0.12 ± 0.03	0.28 ± 0.08
+4Y	0.64 ± 0.05	0.62 ± 0.08	0.66 ± 0.05	0.12 ± 0.04	0.28 ± 0.09

Supplementary Material 13; Comparison of -999/999 imputation on feature importance.

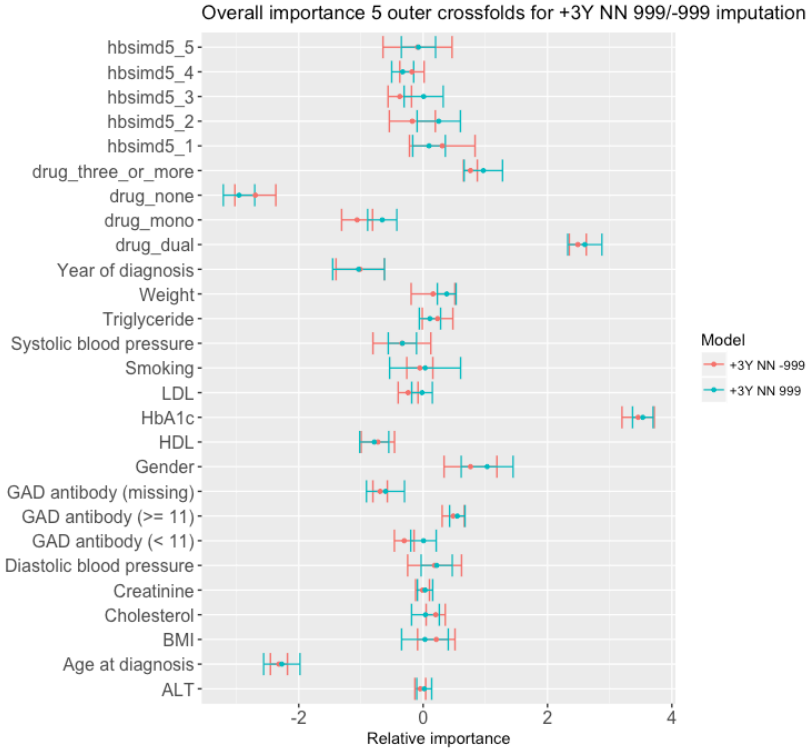
Artificial neural networks



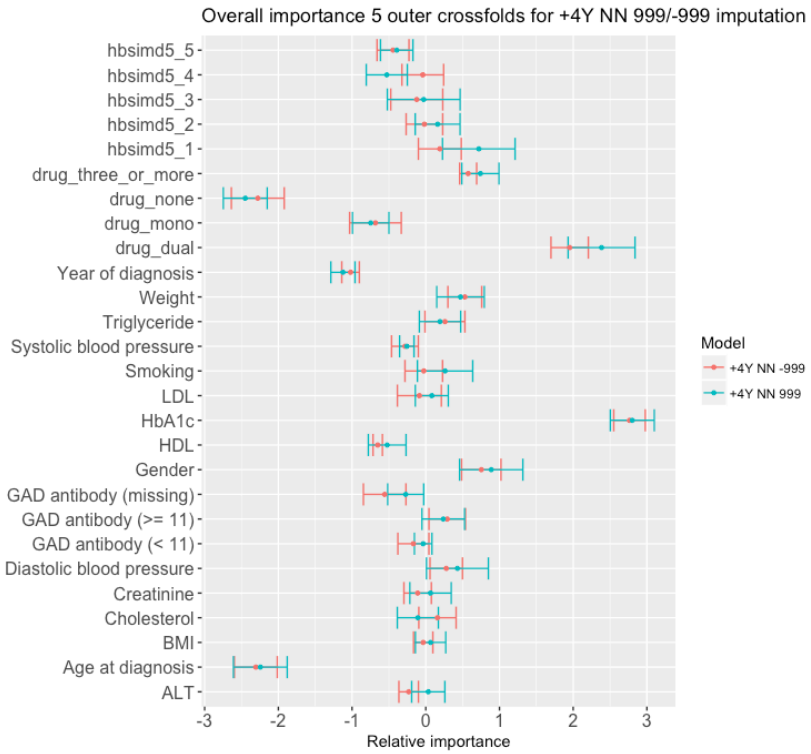
Supplementary Figure 4: Artificial neural networks trained on -999 vs 999 data for fixed time point data +1Y. No major impact of imputation.



Supplementary Figure 5: Artificial neural networks trained on -999 vs 999 data for fixed time point data +2Y. No major impact of imputation.

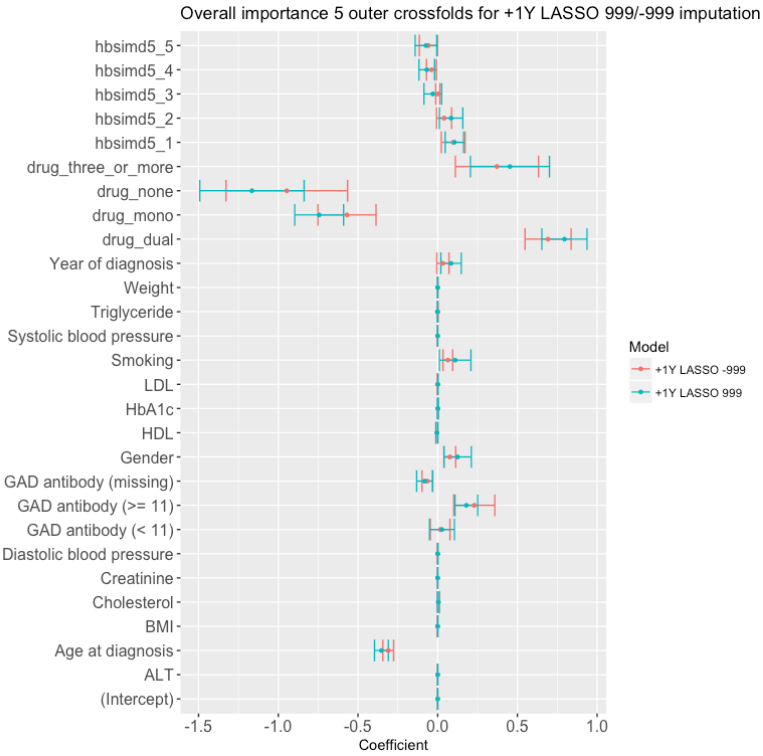


Supplementary Figure 6: Artificial neural networks trained on -999 vs 999 data for fixed time point data +3Y. No major impact of imputation.

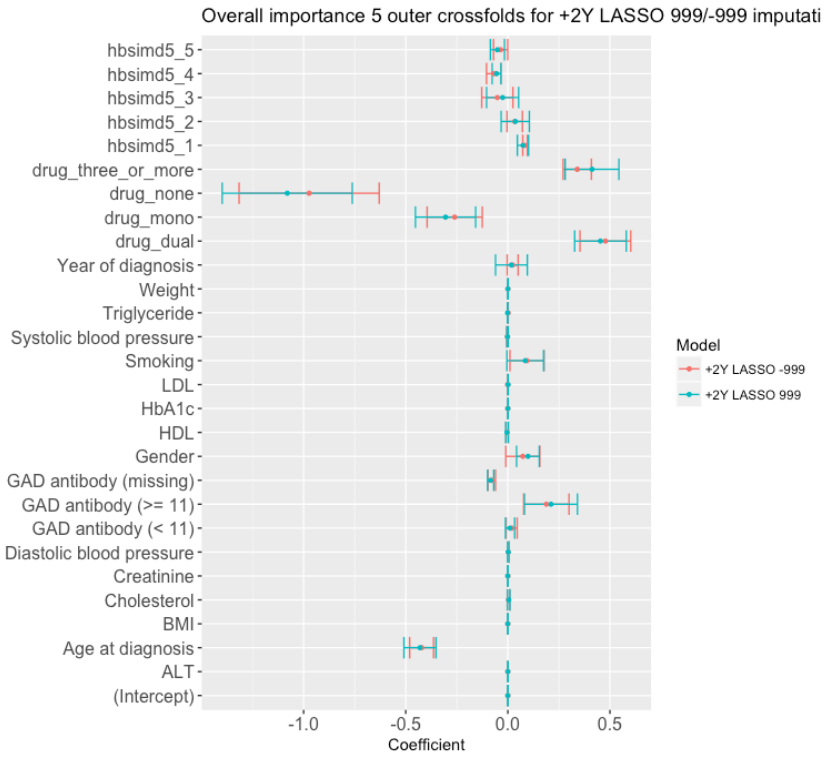


Supplementary Figure 7: Artificial neural networks trained on -999 vs 999 data for fixed time point data +4Y. No major impact of imputation.

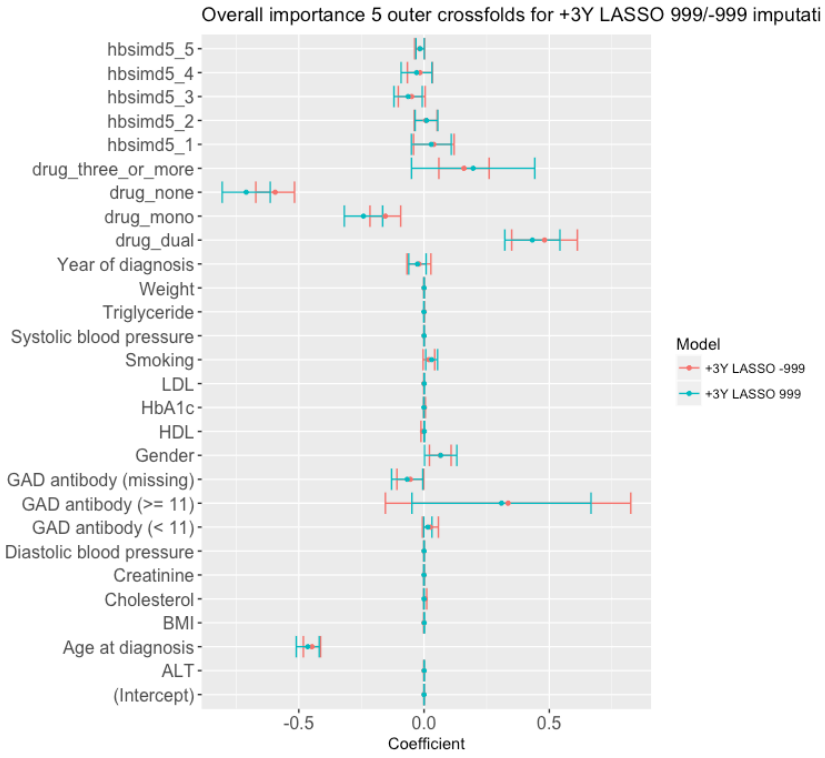
LASSO regression



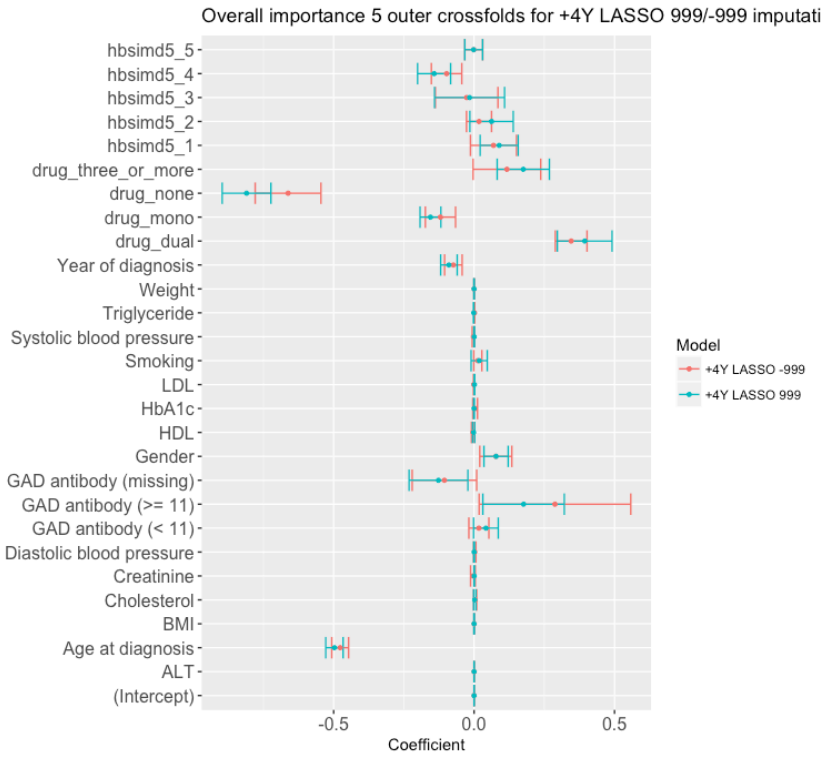
Supplementary Figure 8: LASSO regression coefficients trained on -999 vs 999 data for fixed time point data +1Y. No major impact of imputation.



Supplementary Figure 9: LASSO regression coefficients trained on -999 vs 999 data for fixed time point data +2Y. No major impact of imputation.

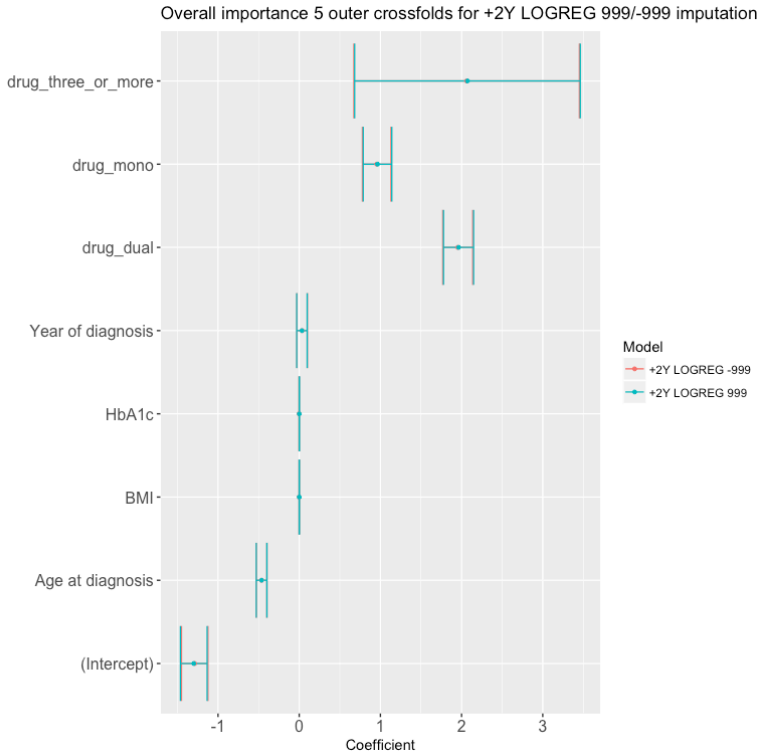


Supplementary Figure 10: LASSO regression coefficients trained on -999 vs 999 data for fixed time point data +3Y. No major impact of imputation.

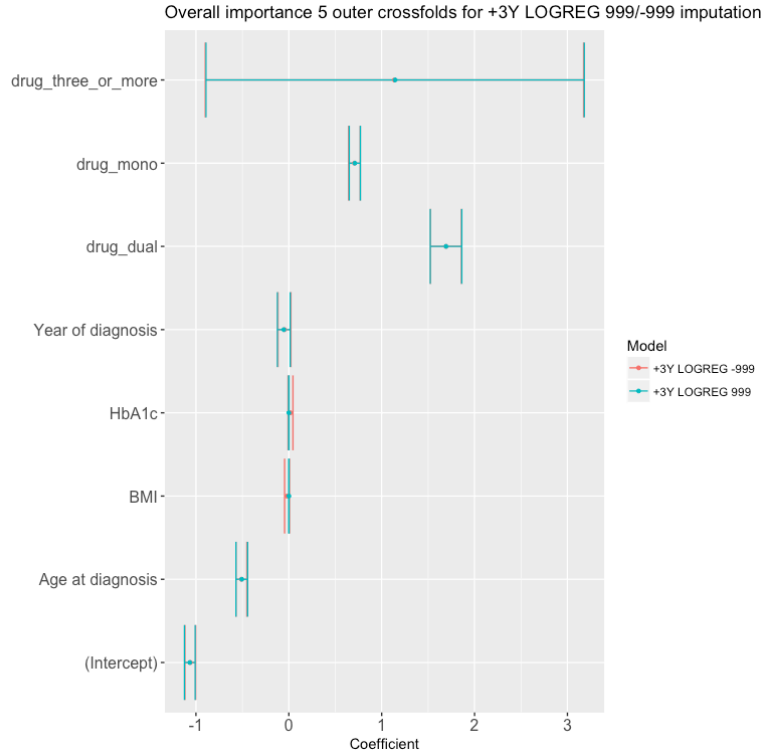


Supplementary Figure 11: LASSO regression coefficients trained on -999 vs 999 data for fixed time point data +4Y. No major impact of imputation.

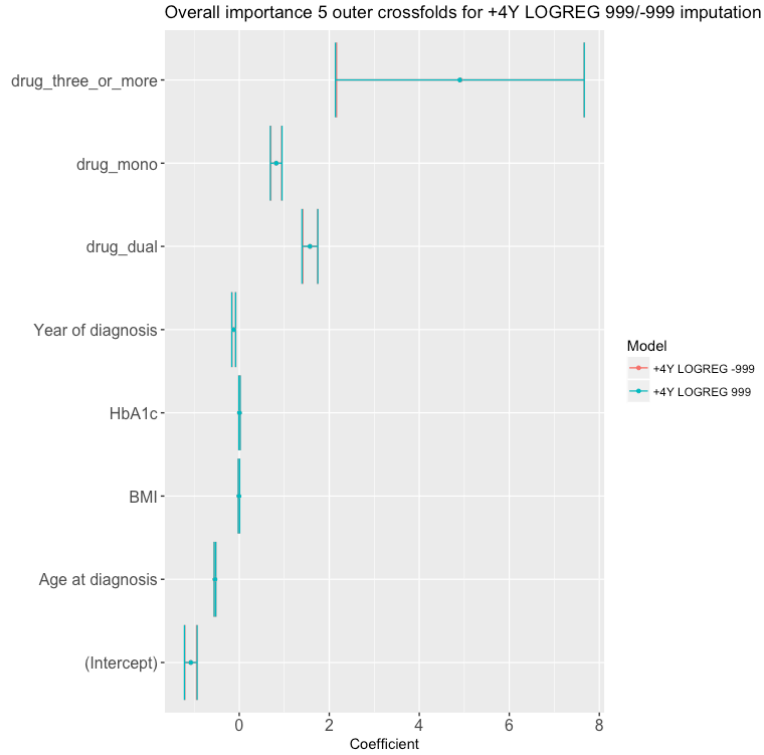
Logistic regression



Supplementary Figure 12: Logistic regression coefficients trained on -999 vs 999 data for fixed time point data +2Y. No major impact of imputation.



Supplementary Figure 13: Logistic regression coefficients trained on -999 vs 999 data for fixed time point data +3Y. No major impact of imputation.



Supplementary Figure 14: Logistic regression coefficients trained on -999 vs 999 data for fixed time point data +4Y. No major impact of imputation.

Supplementary Material 14; Different model initialisation of the artificial neural networks and permutation of prediction labels

For time to insulin models predicting +1, +2, +3, +4 year ahead of time with clinical features on fixed time point data.

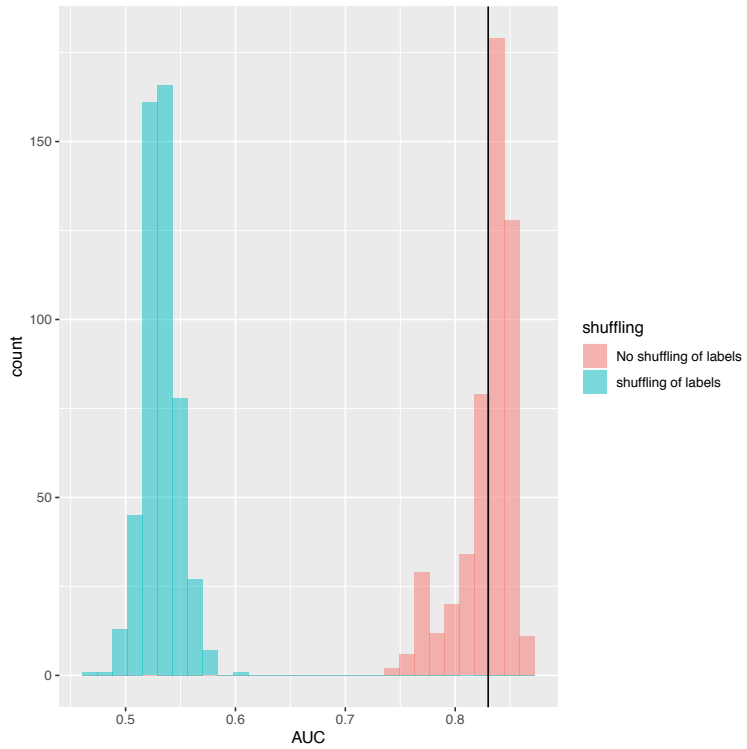
The performance of the reported models was trained using seed 432.

The fifty seeds for retraining of models with and without permutation included the following seeds: 5633, 2855, 3984, 4373, 8208, 7351, 4329, 7109, 1382, 6487, 3610, 9133, 5242,2, 5424, 4430, 1506, 7431,7373, 86, 3584, 9645, 9599, 5768, 1623, 5886, 2765, 5222, 4542, 6516,6024, 5863, 5156, 299, 2992, 9367, 4408, 8830, 9407, 6087, 2549, 6967, 7799, 7558, 6450, 1565, 2949, 2736, 5405, 2178.

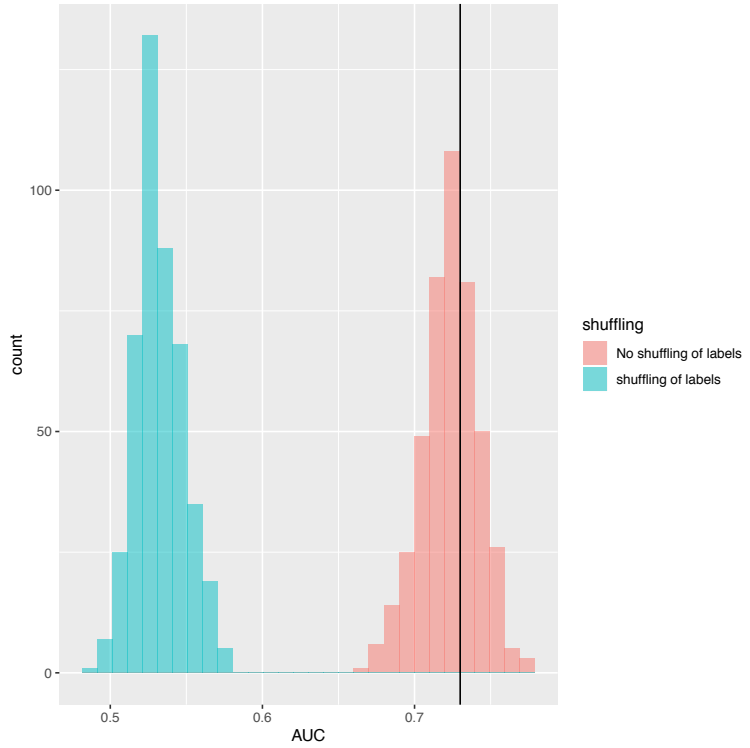
The black vertical lines indicate the average ROC-AUC performance across the 5 folds from the model reported in the paper with seed 432 as is the ROC-AUC that is reported in the paper.

The average ROC-AUC performance across the 5 folds was plotted for each of the 50 model initialisation seeds.

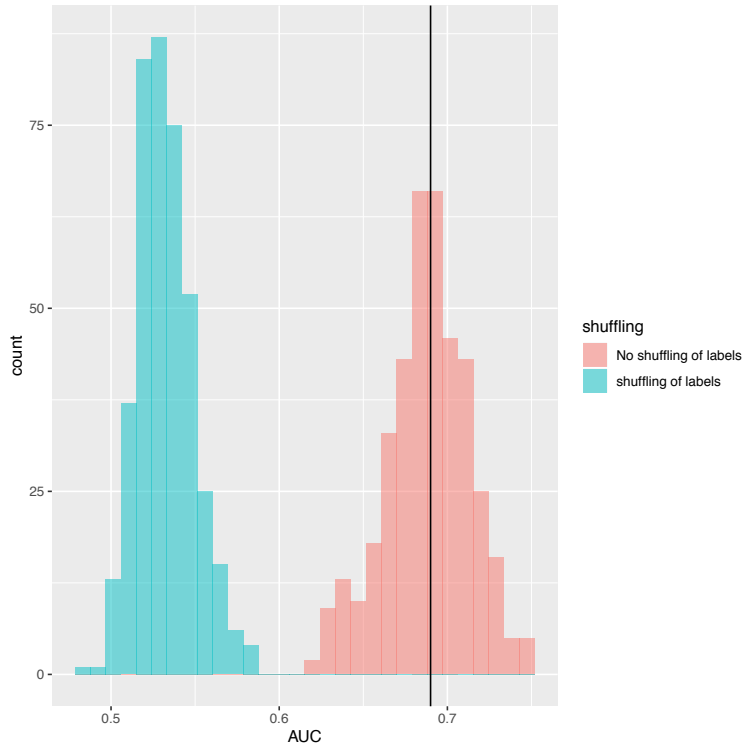
The same cross-validated splits were maintained across the non-shuffled and shuffled labels. The shuffled labels were shuffled for each training and test dataset in order to maintain the original cross-validation splits from seed 432.



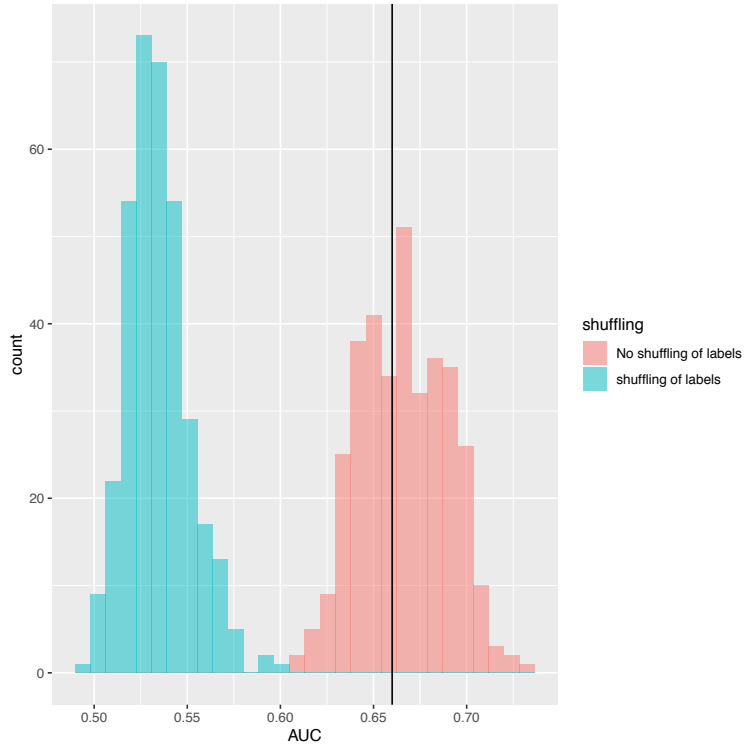
Supplementary Figure 15: Artificial neural network +1Y with fixed time point data (M1) predictions averaged across year 1 - 10 with shuffled and non-shuffled labels. The black line represents average performance for reported performance in paper.



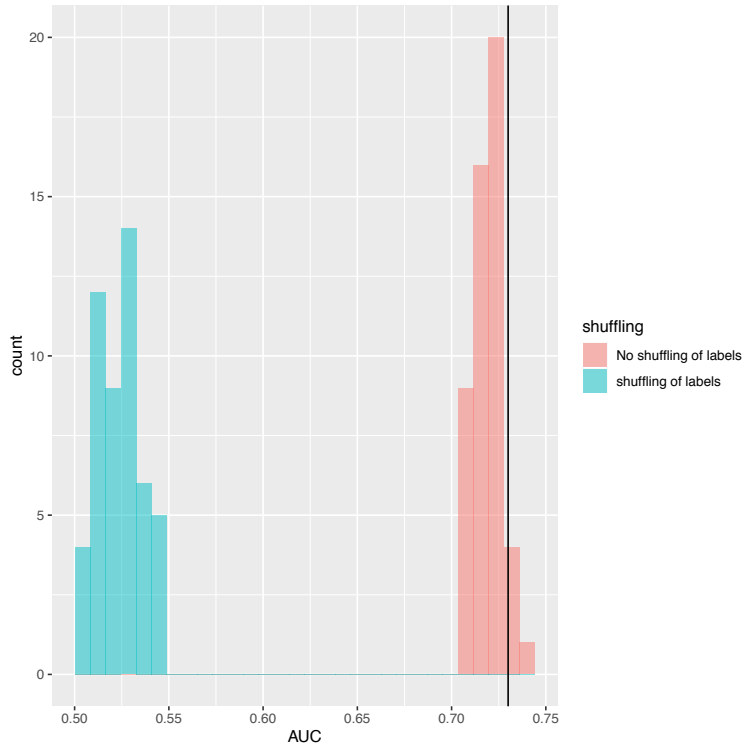
Supplementary Figure 16: Artificial neural network +2Y with fixed time point data (M1) predictions averaged across year 1 - 9 with shuffled and non-shuffled labels. The black line represents average performance for reported performance in paper.



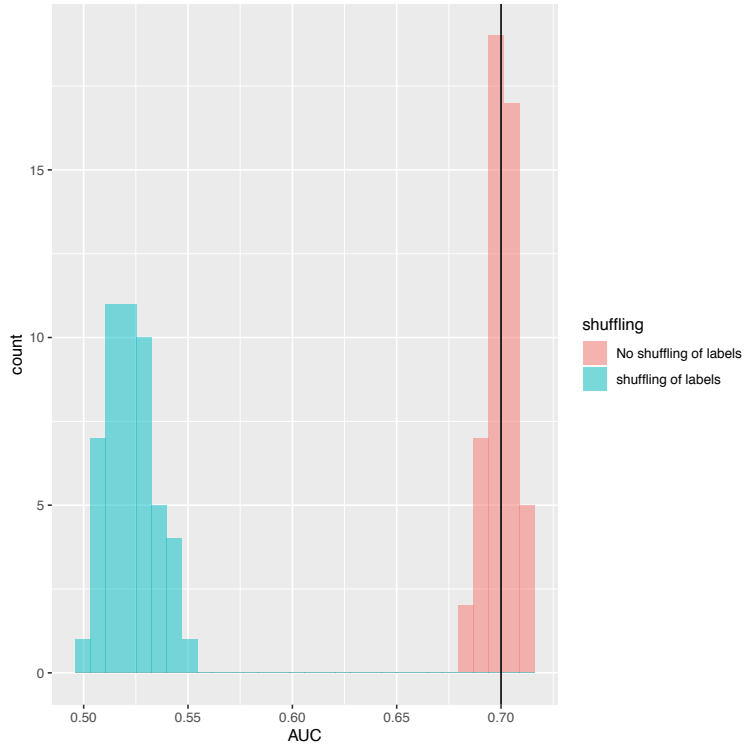
Supplementary Figure 17: Artificial neural network +3Y with fixed time point data (M1) predictions averaged across year 1 - 8 with shuffled and non-shuffled labels. The black line represents average performance for reported performance in paper.



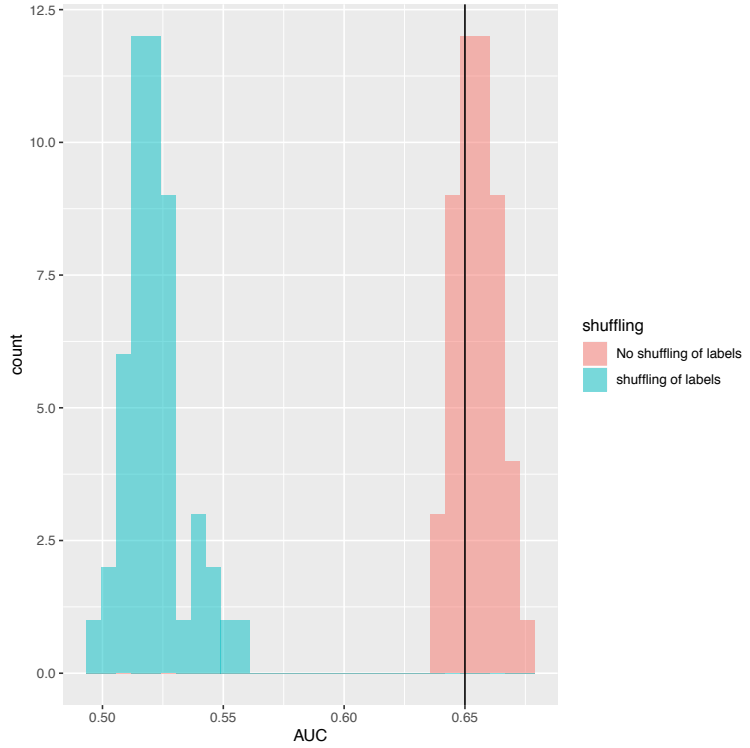
Supplementary Figure 18: Artificial neural network +4Y with fixed time point data (M1) predictions averaged across year 1 - 7 with shuffled and non-shuffled labels. The black line represents average performance for reported performance in paper.



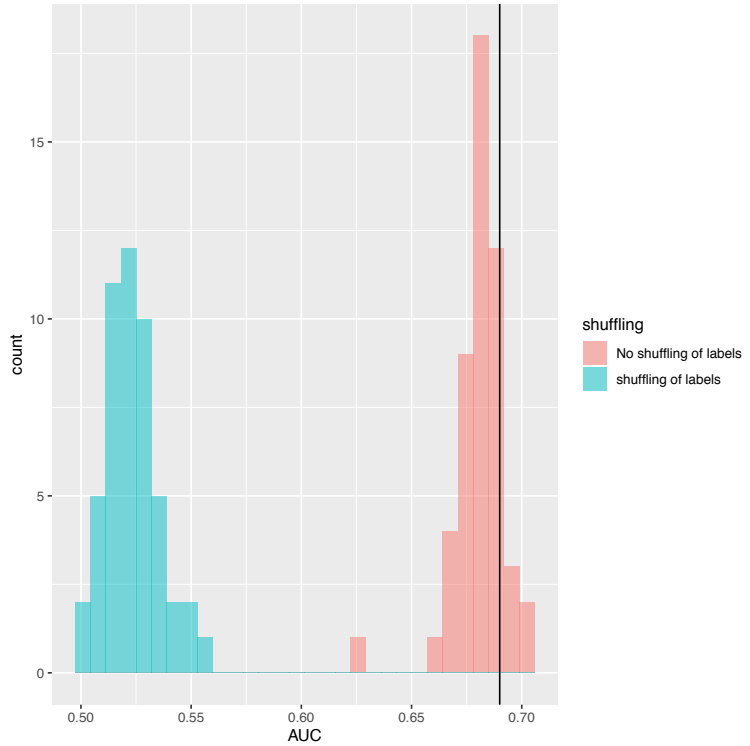
Supplementary Figure 19: M2 artificial neural network clinical and genetics +1Y predictions with shuffled and non-shuffled labels. The black line represents reported performance in paper.



Supplementary Figure 20: M2 artificial neural network clinical and genetics +2Y predictions with shuffled and non-shuffled labels. The black line represents reported performance in paper.



Supplementary Figure 21: M2 artificial neural network clinical and genetics +3Y predictions with shuffled and non-shuffled labels. The black line represents reported performance in paper.



Supplementary Figure 22: M2 artificial neural network clinical and genetics +4Y predictions with shuffled and non-shuffled labels. The black line represents reported performance in paper.

Supplementary Material 15; Distribution of HbA1c and the type of drugs in prediction score tail distributions (<0.3 or >0.7)

Supplementary Table 12: The mean± standard deviation for HbA1c and the type of drug in patients classified with low(<0.3) or high (<0.7) confidence of requiring insulin for the +1Y, +2Y, +3Y and +4Y time to insulin prediction models.

	Patients not requiring insulin (prediction score <0.3)	Patients requiring insulin (prediction score >0.7)
+1Y		
HbA1c (mean±SD [min-max])	6.75±0.68 [3.7-12.2]	8.75±1.66 [4.4-16.00]
Diabetes drug	None: 10335 Mono: 10841 Dual: 2463 Triple or more: 271	None: 855 Mono: 1556 Dual: 2050 Triple or more: 284
+2Y		
HbA1c (mean±SD [min-max])	6.60±0.66 [3.7-12.6]	8.21±1.40
Diabetes drug	None: 7139 Mono: 5768 Dual: 836 Triple or more: 77	None: 768 Mono: 2080 Dual: 1901 Triple or more: 242
+3Y		
HbA1c (mean±SD [min-max])	6.60±0.66 [4.3-13.3]	7.90±1.33 [4.2-16.0]
Diabetes drug	None: 6182 Mono: 4027 Dual: 424	None: 731 Mono: 2034 Dual: 1728

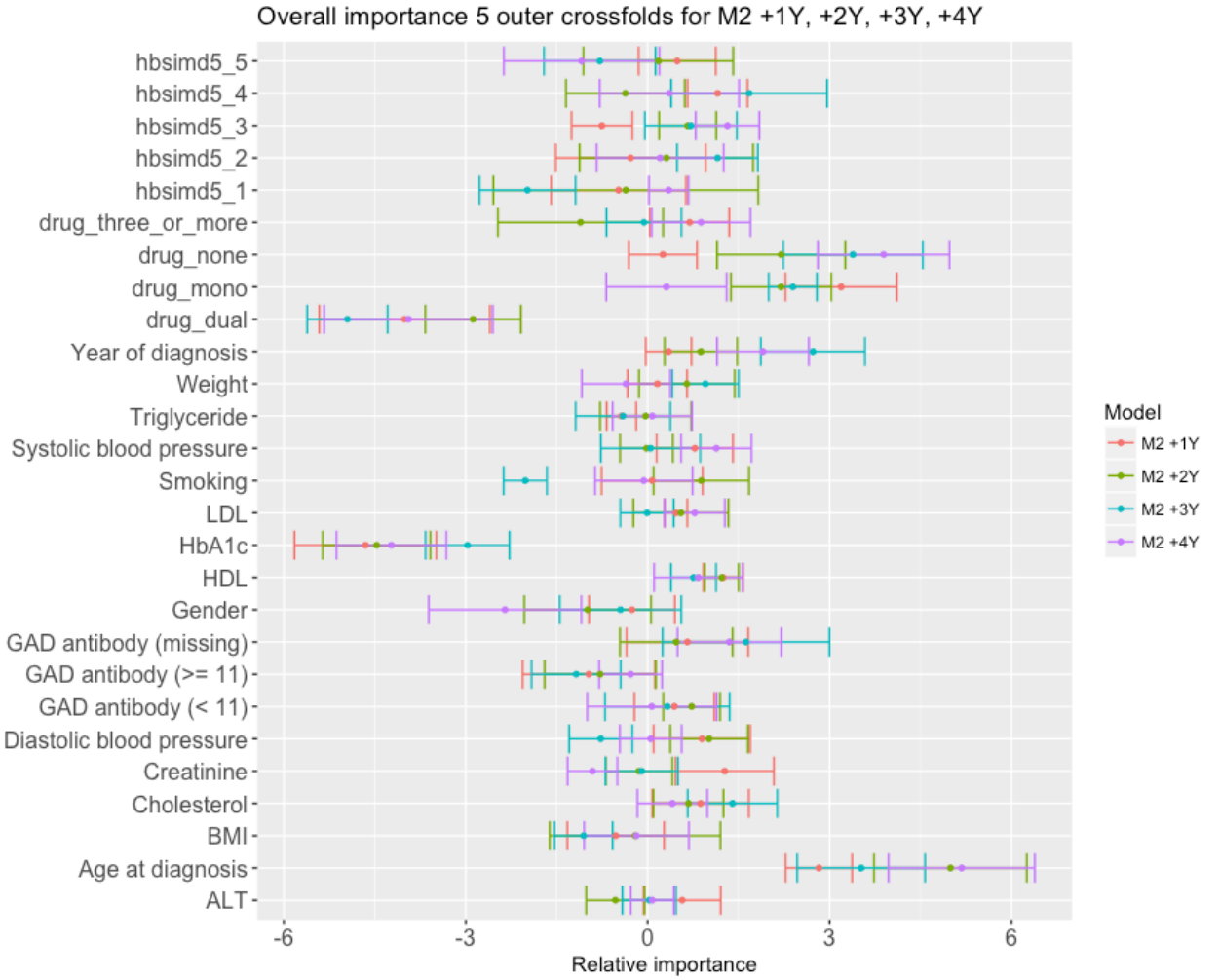
	Triple or more: 32	Triple or more: 138
+4Y		
HbA1c (mean±SD [min-max])	6.61±0.68 [3.7-12.1]	7.79±1.31 [4.4-16.0]
Diabetes drug	None: 4495 Mono: 2979 Dual: 317 Triple or more: 27	None: 715 Mono: 1694 Dual: 1126 Triple or more: 69

Supplementary Material 16; Performance of second time to insulin models (M2) trained on false positive and false negative from the first clinical model

Supplementary Table 13: Performance of M2 with and without genetics integrated with the fixed time point clinical data. Performance given as mean ± standard deviation of all five-fold test models using 2-level cross-validation MCC: Matthews correlation coefficient, J: Youden's index.

M2 clinical+GRS (Fixed time point) ANN	***ROC-AUC***	***Sensitivity***	***Specificity***	***MCC***	***J***
+1Y	0.71 ± 0.05	0.70 ± 0.07	0.62 ± 0.04	0.17 ± 0.04	0.32 ± 0.08
+2Y	0.68 ± 0.03	0.66 ± 0.06	0.62 ± 0.03	0.17 ± 0.04	0.28 ± 0.07
+3Y	0.67 ± 0.04	0.64 ± 0.04	0.57 ± 0.03	0.13 ± 0.04	0.21 ± 0.06
+4Y	0.67 ± 0.02	0.69 ± 0.04	0.60 ± 0.03	0.18 ± 0.03	0.29 ± 0.05
M2 clinical+ forward selection SNPs associated with T2D (Fixed time point) ANN	***ROC-AUC***	***Sensitivity***	***Specificity***	***MCC***	***J***
+1Y	0.72 ± 0.03	0.72 ± 0.04	0.61 ± 0.01	0.18 ± 0.03	0.33 ± 0.05
+2Y	0.71 ± 0.03	0.67 ± 0.07	0.64 ± 0.03	0.19 ± 0.05	0.31 ± 0.07
+3Y	0.66 ± 0.04	0.65 ± 0.07	0.60 ± 0.02	0.15 ± 0.04	0.24 ± 0.07
+4Y	0.70 ± 0.01	0.65 ± 0.03	0.65 ± 0.01	0.19 ± 0.03	0.30 ± 0.04
M2 clinical (Fixed time point) on genetic subset	***ROC-AUC***	***Sensitivity***	***Specificity***	***MCC***	***J***
+1Y	0.72 ± 0.03	0.72 ± 0.04	0.61 ± 0.02	0.17 ± 0.02	0.33 ± 0.04
+2Y	0.69 ± 0.02	0.67 ± 0.07	0.60 ± 0.07	0.17 ± 0.04	0.27 ± 0.07
+3Y	0.65 ± 0.04	0.59 ± 0.05	0.58 ± 0.03	0.11 ± 0.03	0.18 ± 0.05
+4Y	0.68 ± 0.02	0.67 ± 0.03	0.61 ± 0.04	0.17 ± 0.02	0.27 ± 0.04

Supplementary Material 17; Feature importance plots for M2 +1Y, +3Y and +4Y models



Supplementary Figure 23: Relative feature importance given as mean±standard deviation of five outer cross-validation folds including the direction of clinical features used for training of model 2 retained on unique patients that were incorrectly classified in M1 for +1Y, +2Y, +3Y and +4Y predictions respectively. M2 included various data extracted from different years (baseline time points).

Supplementary Material 18; Differences clinical features between the two clinical models
Mean and standard deviation (SD) for all continuous biomarkers and number of patients with a given value in model 1 and model 2 with cases and controls combined and stratified by cases and controls.

For M1, patients were duplicated across multiple years and thus a random subset was selected so a patient only was represented at one time point.

Supplementary Table 14: mean ± SD of continuous clinical variables and number of patients with a given category in the encoding of features across cases and controls for +1Y prediction TTI:clinical models. Cases (N = 209 in M1, N = 326 in M2) and Controls (N = 5878 in M1 and N = 4242 in M2).

Continuous variables	**Model 1 Cases**	**Model 1 Controls**	**Model 2 Cases**	**Model 2 Controls**

BMI	32.79 ± 7.57	31.80 ± 5.85	31.61 ± 6.25	32.15 ± 6.18
Systolic blood pressure	137.20 ± 16.38	138.76 ± 16.73	138.40 ± 18.66	137.12 ± 16.52
Diastolic blood pressure	80.60 ± 10.01	79.83 ± 10.09	79.45 ± 11.76	77.74 ± 10.17
Weight	92.33 ± 24.14	89.52 ± 18.58	88.95 ± 19.25	90.84 ± 19.62
ALT	37.20 ± 21.49	33.00 ± 24.06	34.37 ± 25.11	34.42 ± 25.14
Cholesterol	4.98 ± 1.25	4.81 ± 1.08	4.53 ± 1.15	4.50 ± 1.09
Creatinine	76.69 ± 48.38	76.21 ± 22.47	84.22 ± 49.89	77.72 ± 30.32
HbA1c	9.32 ± 1.64	7.11 ± 1.15	8.14 ± 1.17	8.11 ± 1.42
HDL	1.18 ± 0.32	1.24 ± 0.32	1.27 ± 0.37	1.16 ± 0.30
LDL	2.43 ± 0.95	2.56 ± 0.94	2.27 ± 0.97	2.24 ± 0.89
Triglycerides	3.71 ± 3.57	2.65 ± 2.27	2.39 ± 1.85	2.90 ± 2.31
Age diagnosis	60.39 ± 14.27	64.55 ± 13.53	70.25 ± 14.31	63.04 ± 13.61
Calendar year of confirmed diagnosis	2004.66 ± 5.41	2005.04 ± 5.20	2006.76 ± 5.71	2004.78 ± 5.41
Categorical variables				
Diabetes drug	None: 82 Mono: 57 Dual: 66 Triple or more: 4	None: 3670 Mono: 1946 Dual: 259 Triple or more: 3	None: 191 Mono: 90 Dual: 37 Triple or more: 8	None: 1954 Mono: 1168 Dual: 1019 Triple or more: 101
GAD	Missing: 45 GAD < 11: 139 GAD >= 11: 25	Missing: 1353 GAD < 11: 4378 GAD >= 11: 147	Missing: 67 GAD < 11: 251 GAD >= 11: 8	Missing: 956 GAD < 11: 3166 GAD >= 11: 120
Gender (Female/Male)	F: 103 M: 106	F: 2589 M: 3289	F: 138 M: 188	F: 1887 M: 2355
hbsimd5 (social deprivation)	hbsimd5_1: 69 hbsimd5_2: 43 hbsimd5_3: 32 hbsimd5_4: 34 hbsimd5_5: 31	hbsimd5_1: 1457 hbsimd5_2: 1286 hbsimd5_3: 998 hbsimd5_4: 1034 hbsimd5_5: 1103	hbsimd5_1: 75 hbsimd5_2: 70 hbsimd5_3: 44 hbsimd5_4: 71 hbsimd5_5: 66	hbsimd5_1: 1115 hbsimd5_2: 968 hbsimd5_3: 707 hbsimd5_4: 721 hbsimd5_5: 731
Smoking (ever/never)	Ever: 47 Never: 162	Ever: 4384 Never: 1494	Ever: 243 Never: 83	Ever: 3187 Never: 1055

Supplementary Table 15: mean ± SD of continuous clinical variables and number of patients with a given category in the encoding of features across cases and controls for +2Y prediction TTI:clinical models. Cases (N =241 in M1, N =443 in M2) and Controls (N = 5637 in M1 and N = 4180 in M2).

Continuous variables	**Model 1 Cases**	**Model 1 Controls**	**Model 2 Cases**	**Model 2 Controls**
BMI	33.36 ± 7.19	31.74 ± 5.79	32.45 ± 56.72	32.12 ± 56.09
Systolic blood pressure	137.50 ± 17.47	138.81 ± 16.70	137.84 ± 517.13	137.73 ± 516.39
Diastolic blood pressure	82.73 ± 11.27	79.71 ± 10.02	79.22 ± 512.06	77.95 ± 510.13
Weight	93.57 ± 20.21	89.36 ± 18.49	91.75 ± 521.41	90.56 ± 519.16
ALT	44.13 ± 63.13	32.54 ± 20.83	32.30 ± 521.54	33.48 ± 522.56
Cholesterol	5.12 ± 1.16	4.80 ± 1.08	4.60 ± 51.09	4.45 ± 51.08
Creatinine	71.78 ± 16.64	76.40 ± 22.66	80.31 ± 525.23	77.58 ± 525.31
HbA1c	8.54 ± 1.56	7.05 ± 1.087	7.56 ± 51.04	7.71 ± 51.21
HDL	1.13 ± 0.27	1.24 ± 0.320	1.24 ± 50.32	1.18 ± 50.30
LDL	2.64 ± 0.98	2.559 ± 0.93	2.33 ± 51.01	2.19 ± 50.89
Triglycerides	3.75 ± 2.68	2.60 ± 2.23	2.50 ± 51.40	2.72 ± 52.22
Age diagnosis	59.27 ± 13.70	64.77 ± 13.48	69.15 ± 513.84	62.16 ± 512.65
Calendar year of confirmed diagnosis	2004.54 ± 5.34	2005.06 ± 5.20	2006.53 ± 55.59	2004.40 ± 55.16
Categorical variables				
Diabetes drug	None: 104 Mono: 106 Dual: 31 Triple or more: 0	None: 3566 Mono: 1840 Dual: 228 Triple or more: 3	None: 253 Mono: 139 Dual: 49 Triple or more: 2	None: 1687 Mono: 1537 Dual: 871 Triple or more: 85
GAD	Missing: 54 GAD < 11: 172 GAD >= 11: 15	Missing: 1299 GAD < 11: 4206 GAD >= 11: 132	Missing: 90 GAD < 11: 340 GAD >= 11: 13	Missing: 965 GAD < 11: 3109 GAD >= 11: 106
Gender (Female/Male)	F: 111 M: 130	F: 2478 M: 3159	F: 192 M: 251	F: 1895 M: 2285
hbsimd5 (social deprivation)	hbsimd5_1: 62 hbsimd5_2: 65 hbsimd5_3: 42 hbsimd5_4: 46 hbsimd5_5: 26	hbsimd5_1: 1395 hbsimd5_2: 1221 hbsimd5_3: 956 hbsimd5_4: 988 hbsimd5_5: 1077	hbsimd5_1: 96 hbsimd5_2: 109 hbsimd5_3: 80 hbsimd5_4: 81 hbsimd5_5: 77	hbsimd5_1: 1100 hbsimd5_2: 945 hbsimd5_3: 697 hbsimd5_4: 714 hbsimd5_5: 724
Smoking (ever/never)	Ever: 188 Never: 53	Ever: 4196 Never: 1441	Ever: 339 Never: 104	Ever: 3096 Never: 1084

Supplementary Table 16: mean ± SD of continuous clinical variables and number of patients with a given category in the encoding of features across cases and controls for +3Y prediction TTI:clinical models. Cases (N = 270 in M1, N = 420 in M2) and Controls (N = 5367 in M1 and N = 4053 in M2).

Continuous variables	**Model 1 Cases**	**Model 1 Controls**	**Model 2 Cases**	**Model 2 Controls**
BMI	33.12 ± 6.00	31.66 ± 5.77	31.83 ± 6.11	32.09 ± 5.95
Systolic blood pressure	137.05 ± 16.03	138.90 ± 16.73	136.19 ± 17.40	137.58 ± 16.00
Diastolic blood pressure	81.14 ± 9.30	79.64 ± 10.05	77.92 ± 10.87	78.19 ± 9.99
Weight	93.53 ± 20.45	89.14 ± 18.36	90.03 ± 19.64	90.44 ± 18.67
ALT	36.36 ± 18.52	32.35 ± 20.92	31.53 ± 17.90	33.27 ± 29.03
Cholesterol	4.96 ± 1.23	4.79 ± 1.07	4.61 ± 1.00	4.45 ± 1.03
Creatinine	70.02 ± 16.27	76.73 ± 22.90	78.09 ± 22.50	77.67 ± 25.85
HbA1c	7.90 ± 1.35	7.01 ± 1.05	7.35 ± 1.01	7.53 ± 1.15
HDL	1.13 ± 0.31	1.25 ± 0.32	1.21 ± 0.31	1.20 ± 0.31
LDL	2.52 ± 0.91	2.56 ± 0.94	2.35 ± 0.92	2.22 ± 0.86
Triglycerides	3.45 ± 4.06	2.55 ± 2.08	2.52 ± 1.95	2.54 ± 2.08
Age diagnosis	58.63 ± 13.11	65.08 ± 13.42	67.64 ± 14.44	62.36 ± 12.48
Calendar year of confirmed diagnosis	2004.13 ± 5.31	2005.11 ± 5.19	2006.12 ± 5.49	2004.45 ± 5.16
Categorical variables				
Diabetes drug	None: 109 Mono: 134 Dual: 27 Triple or more: 0	None: 3457 Mono: 1706 Dual: 201 Triple or more: 3	None: 242 Mono: 141 Dual: 34 Triple or more: 3	None: 1704 Mono: 1534 Dual: 761 Triple or more: 54
GAD	Missing: 52 GAD < 11: 204 GAD >= 11: 14	Missing: 1247 GAD < 11: 4002 GAD >= 11: 118	Missing: 95 GAD < 11: 103 GAD >= 11: 8	Missing: 902 GAD < 11: 3054 GAD >= 11: 97
Gender (Female/Male)	F: 117 M: 153	F: 2361 M: 3006	F: 176 M: 244	F: 1840 M: 2213
hbsimd5 (social deprivation)	hbsimd5_1: 83 hbsimd5_2: 52 hbsimd5_3: 51 hbsimd5_4: 41 hbsimd5_5: 43	hbsimd5_1: 1312 hbsimd5_2: 1169 hbsimd5_3: 905 hbsimd5_4: 947 hbsimd5_5: 1034	hbsimd5_1: 94 hbsimd5_2: 112 hbsimd5_3: 59 hbsimd5_4: 82 hbsimd5_5: 73	hbsimd5_1: 1046 hbsimd5_2: 873 hbsimd5_3: 664 hbsimd5_4: 690 hbsimd5_5: 780
Smoking (ever/never)	Ever: 212 Never: 58	Ever: 3984 Never: 1383	Ever: 313 Never: 107	Ever: 3013 Never: 1040

Supplementary Table 17: mean ± SD of continuous clinical variables and number of patients with a given category in the encoding of features across cases and controls for +4Y prediction TTI:clinical models. Cases (N = 301 in M1, N = 388 in M2) and Controls (N = 5066 in M1 and N = 3656 in M2).

Continuous variables	**Model 1 Cases**	**Model 1 Controls**	**Model 2 Cases**	**Model 2 Controls**
BMI	33.29 ± 6.45	31.58 ± 5.72	31.02 ± 5.85	32.16 ± 6.00
Systolic blood pressure	138.58 ± 17.21	138.92 ± 16.71	138.63 ± 15.95	137.92 ± 16.19
Diastolic blood pressure	82.76 ± 9.80	79.47 ± 10.04	77.60 ± 9.92	78.36 ± 10.24
Weight	95.46 ± 20.58	88.81 ± 18.18	87.52 ± 19.76	90.83 ± 18.82
ALT	34.25 ± 16.85	32.26 ± 21.10	30.34 ± 17.62	32.48 ± 19.16
Cholesterol	4.94 ± 1.05	4.78 ± 1.07	4.58 ± 1.08	4.47 ± 1.01
Creatinine	72.78 ± 16.74	76.94 ± 23.17	80.43 ± 29.32	77.56 ± 26.81
HbA1c	7.68 ± 1.28	6.97 ± 1.03	7.14 ± 0.91	7.43 ± 1.15
HDL	1.13 ± 0.26	1.25 ± 0.32	1.25 ± 0.33	1.21 ± 0.31
LDL	2.65 ± 0.89	2.56 ± 0.94	2.38 ± 1.04	2.21 ± 0.89
Triglycerides	2.70 ± 1.49	2.54 ± 2.11	2.42 ± 1.57	2.53 ± 1.84
Age diagnosis	59.84 ± 14.38	65.39 ± 13.30	68.16 ± 13.48	62.30 ± 12.47
Calendar year of confirmed diagnosis	2004.67 ± 5.79	2005.13 ± 5.15	2005.60 ± 5.33	2004.54 ± 5.30
Categorical variables				
Diabetes drug	None: 165 Mono: 107 Dual: 29 Triple or more: 0	None: 3292 Mono: 1599 Dual: 172 Triple or more: 3	None: 214 Mono: 130 Dual: 40 Triple or more: 4	None: 1558 Mono: 1451 Dual: 612 Triple or more: 35
GAD	Missing: 64 GAD < 11: 226 GAD >= 11: 11	Missing: 1183 GAD < 11: 3776 GAD >= 11: 107	Missing: 84 GAD < 11: 293 GAD >= 11: 11	Missing: 794 GAD < 11: 2780 GAD >= 11: 82
Gender (Female/Male)	F: 134 M: 167	F: 2227 M: 2839	F: 169 M: 219	F: 1624 M: 2032
hbsimd5 (social deprivation)	hbsimd5_1: 84 hbsimd5_2: 65 hbsimd5_3: 49 hbsimd5_4: 52 hbsimd5_5: 51	hbsimd5_1: 1228 hbsimd5_2: 1104 hbsimd5_3: 856 hbsimd5_4: 895 hbsimd5_5: 983	hbsimd5_1: 91 hbsimd5_2: 92 hbsimd5_3: 74 hbsimd5_4: 60 hbsimd5_5: 71	hbsimd5_1: 912 hbsimd5_2: 820 hbsimd5_3: 613 hbsimd5_4: 650 hbsimd5_5: 661
Smoking (ever/never)	Ever: 228 Never: 73	Ever: 3756 Never: 1310	Ever: 291 Never: 97	Ever: 2740 Never: 916

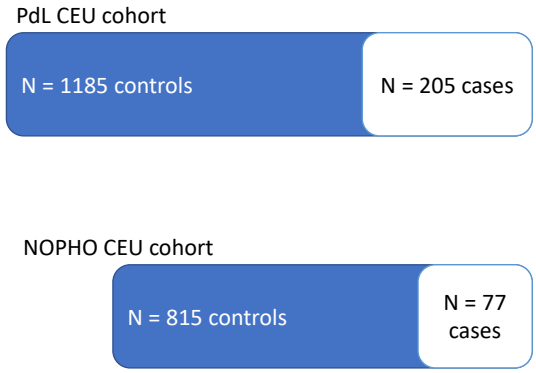
APPENDIX C

Appendix C

Supplementary Methods S1

Patients

Figure A gives an overview of the patients available for the study and which features were available in all samples for modelling.



Common features in PdL and NOPHO cohort	Additional features in NOPHO cohort only
• Age • Gender • 1,401,908 SNPs	• Country • Weight • Height • White blood cell count at ALL diagnosis • MRD d29 flow • MRD d29 PCR • Risk group (SR, IR or HR) • Randomization of asparaginase group (2 or 6 weeks interval) • Total dosages of asparaginase

Figure A: Overview of number patient cohorts *of AAP cases and controls and data available from the PdL and NOPHO study with European ethnicity.*

Machine learning setup and model parameters

Prior to training and testing of AAP prediction models, we divided data from the PdL cohort into two datasets; one used for training and validation (N=1290) and one dataset used for independent test of the model (N=100). The hold out test dataset included 50 cases of AAP and 50 non-APP cases (controls) where the 37 patients that had re-exposure to asparaginase (hereof 13 with second AAP) were included. The remaining patients were randomly selected for the test set.

A five-fold cross-validation with splits stratified by AAP case classes repeated across 100 model initializations was applied on the training set. The classes based on AAP cases were balanced using down-sampling of the controls (no AAP patients) in the training set. Continuous input features for the models were standardized by subtracting the mean and dividing with the standard deviation from

the training set. Categorical variables are divided into one-out-of-K encoded binary variables, and binary variables remained binary. Age was binned into features of different age groups; Y1−Y7, Y7−Y11, Y11−Y17.

The logistic regression models were made using the Scikit-learn class LogisticRegression() with the algorithm extension SAGA of the Stochastic Average Gradient descent algorithm for optimization of the log loss function ('solver'='saga') and no penalty ('penalty'='none').

The random forests were created with the Scikit-learn class RandomForestClassifier(), where the Gini function was used as the measure of impurity (criterion='gini') and minimum decrease in impurity after a split was set to 0.001, i.e. the minimum decrease in Gini impurity for a split to be made (min_impurity_decrease=0.001). Furthermore, the maximum number of features to consider when making each split was set to the square root of the total number of input features (max_features='sqrt'), and the minimum number of samples required to make a split was set to 2.5% of the input samples (min_samples_split=0.025). The random forests were optimized for number of trees in the forest (n_estimators=[5, 10, 15, 20, 25, 50, 100]).

AdaBoost[1] models were trained with the Scikit-learn class AdaBoostClassifier() to get the models to learn the 'hard-to-classify' observations. This model was optimized for number of random forest base estimators (n_estimators=[2, 3, 4, 5, 10, 50]). The algorithm for boosting was the SAMME.R real boosting algorithm (algorithm='SAMME.R'), which is able to boost the base estimators using the class probabilities instead of the binary classification outcome.

The Scikit-learn class MLPClassifier() was used to build the artificial neural networks (ANN). Two types of ANNs were made with one and two hidden layers, respectively, where the networks with a single hidden layer were optimized for the number of hidden neurons (hidden_layer_sizes=[(5,), (10,), (15,), (20,), (25,), (50,)]) and the ANNs with two hidden layers always had 25 neurons in each hidden layer (hidden_layer_sizes=(25,25)). The ANNs with a single hidden layer used mostly default

parameters from the scikit-learn function, but with the logistic activation function (activation='logistic'), stochastic gradient descent as optimization method (solver='sgd'), batch size of 20 samples (batch_size=20) and the maximum number of iterations to train the model was 500 (max_iter=500). The ANNs with two hidden layers also used default parameters except for logistic activation function (activation='logistic'), batch size of 20 samples (batch_size=20), a learning rate of 0.001 and a maximum number of iterations to train the model was 500 (max_iter=500).

Statistical analyses

Statistical analyses were performed using a t-test or Wilcoxon rank sum test for continuous variables given a Gaussian or non-gaussian distribution. Categorical variables were analyzed by Pearson's Chi-squared test in R (version 3.2.5).

Functional consequence ranking of genetic variants

The functional consequences of genetic variants were annotated from Variant Effect Predictor (build 37) in order to create a ranking of the magnitude of effect that the presence of a variant would cause. The consequences were ranked from 1-4 where 1 is the most severe, such as a variant resulting in a frameshift or a gained stop codon (Table S.A). The variants were matched using the chromosome and position, as well as the "rs" ID when available. Only variants with 1 or 2 in severity was prioritized.

Table S.A: Consequences ranked by severity.

Consequence	**Severity**
frameshift_variant	1
splice_acceptor_variant	1
splice_donor_variant	1

splice_region_variant	1
start_lost	1
stop_gained	1
stop_lost	1
inframe_deletion	2
inframe_insertion	2
missense_variant	2
protein_altering_variant	2
TF_binding_site_variant	2
3_prime_UTR_variant	3
5_prime_UTR_variant	3
mature_miRNA_variant	3
NMD_transcript_variant	3
start_retained_variant	3
stop_retained_variant	3
synonymous_variant	3
downstream_gene_variant	4
intergenic_variant	4
intron_variant	4
non_coding_transcript_exon_variant	4
non_coding_transcript_variant	4
upstream_gene_variant	4

Genetic feature selection

We upfront selected which genetic features to be included in the models. The following sections describes the name of the dataset and how this was generated.

Three datasets of SNPs previously associated with AAP

Wolthers *et al* 2019

In 2019, Wolthers *et al* identified SNPs associated with asparaginase-associated pancreatitis in the PdL cohort, where no genetic background was used to subset the data (n=244 cases of AAP, n=1320 controls)[2]. The top thirty P-value SNPs associated with AAP were included for modelling; rs10273639, rs12494164, rs12582343, rs13228878, rs1505495, rs16848986, rs16996276, rs170623, rs1791520, rs2167730, rs34375180, rs368819120, rs4655107, rs4769201, rs5010616, rs5563434, rs61734424, rs62228228, rs62228230, rs62228256, rs6477109, rs7139808, rs7155612, rs7270119, rs74109922, rs75245362, rs7851954, rs80170196, rs934350, rs9912225. Of these 30 SNPs, rs13228878 and rs10273639 were significantly associated with the risk of developing pancreatitis in the AALL0232 cohort[2].

Liu *et al* 2016

In 2016, Liu *et al* identified SNPs associated with acute pancreatitis in patients diagnosed with ALL in 117 cases and 5068 controls. 49 SNPs were reported with an rsID by Liu *et al* were available in the PdL genotype data. These included; rs1023840, rs1034347, rs10831758, rs10979693, rs112141546, rs113207856, rs11606424, rs116191233, rs11961305, rs12589386, rs1341861, rs139875564, rs141452852, rs141708090, rs142728074, rs143348702, rs144556038, rs146053779, rs146796996, rs147501499, rs15176341, rs1564947, rs16892294, rs16907254, rs16945382, rs16953067, rs16955095, rs17053225, rs17117423, rs17377657, rs1768056, rs18247836, rs19187532, rs199695765, rs200495769, rs2010463, rs201734206, rs202124287, rs41463245, rs45461499, rs67047829, rs6721376, rs6724701, rs7544262, rs7826058, rs78283108, rs79537388, rs9849262, rs9859201. This set was further reduced by pruning of SNPs, resulting in 10 SNPs

included for modelling: rs1023840, rs10831758, rs1564947, rs16892294, rs16907254, rs16945382, rs17053225, rs17117423, rs41463245, rs67047829.

Abaji *et al* 2017

In 2017, Abaji *et al* identified genetic risk factors by whole exome sequencing associated with AAP in the Quebec childhood ALL cohort[3]. Five SNPs were significant associated with AAP; rs11556218, rs34708521, rs3809849, rs72755233, rs9908032. in our dataset four SNPs where available: rs11556218, rs34708521, rs72755233 and rs9908032.

Eight datasets of SNPs in the pancreatitis pathway

Two genes, chromosome 20

In the dataset, we selected the top 4 SNPs located on chromosome 20 reported in the Wolthers *et al* AAP GWAS; rs16996276, rs62228230, rs62228256, rs7270119[2]. In addition, we annotated variants to the PRSS1-PRSS2 locus. This resulted in a total of 160 SNPs, which further where pruned resulting in 38 SNPs used for modelling; rs1008248, rs114531821, rs117220554, rs12537777, rs12703471, rs144142335, rs149960437, rs1524140, rs1573618, rs1800907, rs199697208, rs2213192, rs2734060, rs2734117, rs2855918, rs2855924, rs34587783, rs4427082, rs4498456, rs55709813, rs55930828, rs62228256, rs62473589, rs6464494, rs66701035, rs6950275, rs6974425, rs71545384, rs731521, rs73170640, rs76039184, rs8175963, rs8176044, rs8176058, rs8176063, rs8177146, rs8177171, rs975494.

Eight candidate genes

Zator et al described genes involved in pancreatic disease; *PRSS1, PRSS2, SPINK1, CTRC, CASR, CFTR, CPA1* and *CLDN2*[4]. We annotated SNPs to these genes in our dataset resulting in 1517 SNPs.

These were pruned and resulted in 377 SNPs for modelling; exm1651845, exm1651859, exm2262902, kgp22732487, kgp22736019, kgp22804168, kgp22819526, kgp22829942, rs1008248, rs10226236, rs10230435, rs10239213, rs10247427, rs10256541, rs10272052, rs10273043, rs10488188, rs10511409, rs10803377, rs10927713, rs10927749, rs10927751, rs10927797, rs10954269, rs111278301, rs111793536, rs1119800, rs112627346, rs112861203, rs112896741, rs112977751, rs1136995, rs113741816, rs114175678, rs114433626, rs114531821, rs114571236, rs114658781, rs114697189, rs114796420, rs114959849, rs115020348, rs115153864, rs115182091, rs115243436, rs115251799, rs1153084, rs115324868, rs115395979, rs1154729, rs116027795, rs116190983, rs116576422, rs116688903, rs116900477, rs117052523, rs117080455, rs117099609, rs117141465, rs117220554, rs11765940, rs117700687, rs117913533, rs118021467, rs11924218, rs12014762, rs12038868, rs12058287, rs12070915, rs12106790, rs12123804, rs12136606, rs12486849, rs12534580, rs12562216, rs12562412, rs12669592, rs12672166, rs12672889, rs12673576, rs12706936, rs12853674, rs1285712, rs12857932, rs13154671, rs13154930, rs1316277, rs13221882, rs13222308, rs13226446, rs13229221, rs13230593, rs13239073, rs13244661, rs13320117, rs1368412, rs1393198, rs1422990, rs1432823, rs1432975, rs143955465, rs144142335, rs1467513, rs149960437, rs1524140, rs16833165, rs16851387, rs16851463, rs16851662, rs16851665, rs16851946, rs17106906, rs17106994, rs17106995, rs17107315, rs17133172, rs17139584, rs17139774, rs17139904, rs17140425, rs17164720, rs17164729, rs17203488, rs17253746, rs17266628, rs17281995, rs17330912, rs17388190, rs17405463, rs17415991, rs17494960, rs1751998, rs17538716, rs17547485, rs17561784, rs17703848, rs17718041, rs17774502, rs1780595, rs1800076, rs1800095, rs1800907, rs1862329, rs1883757, rs189486222, rs189800017, rs200747551, rs201477143, rs201691901, rs202012741, rs2067080, rs2116766, rs213938, rs2141329, rs2193264, rs2242337, rs2253372, rs2285544, rs2287371, rs2291457, rs2308941, rs2308950, rs2367618, rs2400439, rs2402268, rs2436340, rs2496313, rs2496328,

rs2496331, rs2681401, rs2681420, rs2681425, rs2715259, rs2734060, rs28743073, rs2880013, rs33914662, rs34042920, rs34094595, rs34122809, rs34137539, rs34178491, rs34206873, rs34345120, rs34474469, rs34548758, rs34587783, rs35014830, rs35018893, rs35320768, rs35453239, rs35549389, rs35562243, rs35766781, rs35795032, rs35815704, rs35986679, rs36072711, rs36105551, rs361487, rs367603516, rs367791466, rs367920631, rs3734122, rs3753314, rs3765370, rs3800562, rs3800563, rs3806925, rs3807340, rs3816830, rs3817599, rs3823582, rs38895, rs38896, rs38901, rs38903, rs38904, rs38913, rs4059, rs4100170, rs41269443, rs4252372, rs4252522, rs4363128, rs4470444, rs4498456, rs4661330, rs4661599, rs4731674, rs4731678, rs4974398, rs504775, rs558251, rs55847691, rs55875822, rs55899454, rs55930828, rs56219492, rs56227115, rs56273492, rs57339232, rs5962770, rs60310027, rs60394478, rs62469442, rs62473589, rs62491331, rs62617115, rs6438692, rs6438707, rs6466615, rs6622104, rs6658612, rs66697073, rs6680738, rs66839817, rs66850376, rs6693417, rs67003441, rs67023840, rs6769765, rs6803240, rs6871771, rs6884703, rs6889194, rs6942670, rs6947134, rs6972479, rs6975391, rs71579229, rs71579242, rs72643647, rs72643677, rs72835117, rs72835141, rs72835180, rs72875728, rs73170640, rs73181898, rs73184032, rs73186075, rs73186082, rs73211993, rs73213827, rs73215983, rs73269110, rs73274040, rs7355209, rs73724326, rs739693, rs74668887, rs74974914, rs74993149, rs75118026, rs7512971, rs75222399, rs7525279, rs7533567, rs7536482, rs7549759, rs75598696, rs75700249, rs75793153, rs75911963, rs75945259, rs75980868, rs76039184, rs76124003, rs7618747, rs76249487, rs76368460, rs7637874, rs7647405, rs76509354, rs76516385, rs76613504, rs76659336, rs76707403, rs76788357, rs76860590, rs76908136, rs76919727, rs77041199, rs77045691, rs7704889, rs7713010, rs7725292, rs7729260, rs77385877, rs77448606, rs77688344, rs77808099, rs77859819, rs7786419, rs7789559, rs7796359, rs78000398, rs7800765, rs78074583, rs78113236, rs78135358, rs78641803, rs78660569, rs78681108, rs79042346, rs79156986, rs79190650, rs79243435, rs79280654, rs79463568, rs79736961,

rs79845433, rs80009321, rs80099942, rs80213789, rs80221012, rs80298935, rs8175963, rs8176044, rs8176058, rs8176063, rs8177107, rs8177146, rs8177171, rs874742, rs901799, rs9282641, rs9289191, rs937629, rs949425, rs975494, rs9817571, rs9826770, rs9829181, rs9839875, rs9867460.

Eight candidate genes (prioritized)

We prioritized genetic variants with the most severe consequence; therefore, we selected the variants with consequences ranked at 1 and 2 of all SNPs annotated to the eight genes involved in adult pancreatitis (See Supplementary Methods). This resulted in a prioritized set of 60 SNPs; rs1010294, rs1029396, rs1042077, rs1042636, rs10489962, rs1052571, rs1129055, rs1132312, rs113966492, rs11580170, rs11583306, rs11761888, rs11979330, rs12126178, rs12706927, rs17107315, rs17164867, rs17208, rs17248, rs17260, rs17267, rs17388190, rs17589, rs17854248, rs1800076, rs1800095, rs1801725, rs1801726, rs2020902, rs213950, rs2171492, rs2250145, rs2308941, rs2308950, rs2681417, rs34173813, rs34474469, rs34587586, rs35795032, rs361359, rs361439, rs3765356, rs3766163, rs4252372, rs4252499, rs4661330, rs4728190, rs4731668 rs3915061, rs486557, rs4987667, rs4987682, rs55709813, rs62617115, rs6429757, rs763821, rs7722926, rs8176058, rs8177146, rs9282641.

Eight candidate genes - GTEx eQTL

The eight genes associated involved in pancreatic disease pathways were used a basis to expand the eQTL variants associated with these genes in pancreatic tissue from GTEx data[5] [https://gtexportal.org/home/] downloaded by ensemble tools [https://rest.ensembl.org/]. Genotype-Tissue Expression (GTEx) annotation was selected from pancreatitis tissue and $\log(p) > 0.75$. This resulted 7176 SNPs, where 1197 were available in our genotype data. These were further LD-pruned

to 449 SNPs; rs10064194, rs10068916, rs1008248, rs10084650, rs10215972, rs10216068, rs10216140, rs10224011, rs10245094, rs10253715, rs1025489, rs10256541, rs10256879, rs10258170, rs10265693, rs10266621, rs10266895, rs10270056, rs10270974, rs10273043, rs10273639, rs10276606, rs1042720, rs10477360, rs10487372, rs10487399, rs10489962, rs10491403, rs1049334, rs10511414, rs10515588, rs10754866, rs10755876, rs1079221, rs10803309, rs10927452, rs10927530, rs10927543, rs10927571, rs10927578, rs10927632, rs10927670, rs10927676, rs10927765, rs10934612, rs1114086, rs11167957, rs11168048, rs111850617, rs1119800, rs1124213, rs112445433, rs11260738, rs1127343, rs1136995, rs114340813, rs114433626, rs1144986, rs1153084, rs115324868, rs115395979, rs1155458, rs115618685, rs11585810, rs116752872, rs11709496, rs11760434, rs11760909, rs11764066, rs11766819, rs11770721, rs11771082, rs117780183, rs118046556, rs11921048, rs11950634, rs11958839, rs11973084, rs11973869, rs11975899, rs11978185, rs11982223, rs11982376, rs11983414, rs11983741, rs12057512, rs12085004, rs12117872, rs12127407, rs12129618, rs12130370, rs12141069, rs12409399, rs12486849, rs12489855, rs12494271, rs12521378, rs12532165, rs12533442, rs12537777, rs12539323, rs12654852, rs12655663, rs12660014, rs12667732, rs12671578, rs12673576, rs12695429, rs12738424, rs12757909, rs1285965, rs13064281, rs13073767, rs13074706, rs13099843, rs13154930, rs13157587, rs13159091, rs1316257, rs13186402, rs13222576, rs13223756, rs13224443, rs13229581, rs13231609, rs13234660, rs13357143, rs1346945, rs1354162, rs1395249, rs1404061, rs1422636, rs1424414, rs1432654, rs1432691, rs1432823, rs1480163, rs149107492, rs157933, rs1588770, rs160971, rs1665105, rs16850907, rs16851358, rs16851463, rs16873935, rs16873941, rs17106850, rs17139364, rs17139904, rs17140425, rs17160984, rs17162946, rs17165134, rs17209173, rs1721831, rs17262, rs17266628, rs17333054, rs17495305, rs1752021, rs17538716, rs17589, rs1763611, rs1763612, rs1763632, rs17639329, rs17639735, rs17688079, rs17691111, rs17704764, rs17718041, rs1864944,

rs1867530, rs1877377, rs1881995, rs1909808, rs1919554, rs1966438, rs1976714, rs199688150, rs2040369, rs2049819, rs205725, rs205763, rs2074136, rs2082395, rs2111209, rs2137639, rs213969, rs2140904, rs2152458, rs2178158, rs2178313, rs221035, rs2253372, rs2271545, rs2272256, rs2280680, rs2285544, rs2293748, rs2305327, rs2312535, rs2332240, rs2402993, rs2436410, rs2473357, rs2480057, rs2486774, rs2570407, rs2670494, rs2803397, rs2807555, rs28491601, rs28497047, rs2855896, rs2862154, rs28743073, rs287624, rs2913326, rs2949766, rs2960760, rs2963075, rs2966701, rs30312, rs31031, rs31033, rs319143, rs3213858, rs34206873, rs34407651, rs34654409, rs34785444, rs34861031, rs34941082, rs35023707, rs350638, rs35182759, rs35196193, rs35230862, rs35233872, rs35441766, rs35653352, rs35903225, rs36003690, rs3792390, rs3800567, rs3800990, rs3807994, rs3823483, rs3823577, rs3823582, rs3845598, rs38825, rs3909553, rs40238, rs41782, rs4252372, rs4269476, rs4338012, rs4341084, rs4363128, rs4385705, rs4441917, rs4443601, rs4445168, rs4454229, rs4469397, rs4574533, rs4607527, rs4661293, rs4661543, rs4661667, rs4661702, rs4676683, rs4676686, rs4678013, rs4705036, rs4705045, rs4725617, rs4726642, rs4731771, rs486958, rs514370, rs542008, rs55699789, rs55837101, rs55899454, rs56219258, rs56234747, rs56284241, rs56289897, rs56358766, rs56409029, rs58212949, rs586362, rs58720091, rs58740854, rs60207331, rs60300478, rs60395301, rs60465931, rs60875013, rs61772194, rs61772283, rs61773640, rs61782448, rs62261554, rs62263867, rs62387740, rs62471584, rs62471973, rs62473545, rs62477619, rs62489337, rs62491280, rs6429661, rs6464544, rs6467293, rs6467327, rs6467347, rs6580582, rs6681120, rs66850376, rs6764971, rs6786208, rs68056147, rs68056161, rs6863426, rs6880853, rs6881655, rs6943541, rs6945268, rs6949709, rs6963381, rs6965643, rs6967528, rs6971551, rs6971899, rs6973013, rs6975771, rs702019, rs71545384, rs71594546, rs72640776, rs72643677, rs72833326, rs72834757, rs72862082, rs73146781, rs73152868, rs73157670, rs73159891, rs73159900, rs73170664, rs73179904, rs73181898, rs73195652, rs73211994, rs73213827, rs732470, rs73452827,

rs73452898, rs73523902, rs73722742, rs73855492, rs739557, rs742362, rs742363, rs742662, rs74558393, rs74564230, rs74801905, rs74827221, rs75098848, rs7514440, rs75147459, rs7515681, rs75258377, rs7525851, rs7529561, rs7530236, rs7531035, rs7534010, rs7539884, rs7553794, rs75570697, rs75631290, rs75932939, rs7628062, rs7637874, rs7645033, rs76706972, rs7707452, rs7714069, rs77153860, rs7715716, rs7726085, rs7731196, rs7734352, rs7783273, rs7794533, rs7799105, rs7800811, rs7802940, rs7803075, rs7805545, rs7806322, rs7807308, rs7812207, rs78336569, rs78403305, rs78425975, rs78592190, rs78618574, rs78696877, rs79853540, rs80117059, rs80156131, rs80221012, rs80227654, rs80298935, rs804132, rs8175963, rs863219, rs867522, rs869839, rs8713, rs877741, rs879003, rs879211, rs887574, rs8935, rs895072, rs895074, rs896153, rs910104, rs916725, rs9325026, rs9325057, rs9429230, rs9641562, rs970472, rs970952, rs974558, rs9812472, rs9839593, rs9839656, rs9848900.

Eight candidate genes - GTEx eQTL (prioritized)

We prioritized the most variants with the most severe consequence; therefore, we selected the variants with consequences ranked at 1 and 2 of all eQTLs to the eight genes involved in adult pancreatitis (See Supplementary Methods). This resulted in 27 SNPs; rs10043775, rs1010294, rs10489962, rs1076726, rs10803354, rs12186491, rs12669721, rs12706927, rs1464890, rs17208, rs17589, rs17849995, rs1801726, rs2070179, rs2171492, rs2234001, rs2234002, rs35196193, rs35903225, rs3777134, rs4252372, rs4725617, rs4808, rs6948695, rs7645033, rs8940, rs9968193.

Six candidate SNPs

Finally, we selected six candidate SNPs involved in the pancreatitis pathway rs10436957 (*CTRC*), rs12853674 (*CLDN2*), rs13228878 (*PRSS*), rs16832787 (*CASR*), rs17107315 (*SPINK1*) and rs56296320 (*CFTR*).

Six candidate SNPs – a polygenic risk score

A continuous genetic risk score (GRS) was calculated based on six candidate SNPs rs10436957 (*CTRC*), rs12853674 (*CLDN2*), rs13228878 (*PRSS*), rs16832787 (*CASR*), rs17107315 (*SPINK1*) and rs56296320 (*CFTR*); GRS = sum(Odds ratio * Number of minor alleles). The odds ratio was obtained from Wolthers *et al* GWAS on AAP[2].

Three principal components of eight pancreatitis SNP-annotated genes

In order to reduce the high-dimensional data set containing all SNPs annotated to the eight genes associated with AAP but still capture the variation of these, a principle component analysis (PCA) was performed. We implemented the PCA in the models in the machine learning setup, such that the PCA was performed on the training and test sets separately within the five-fold cross-validation. A filter was set to capture only SNPs with minor allele frequencies of 5% (--MAF 0.05). The PCA was implemented using the plink "--pca" flag that by default creates 20 components of which we extracted the first three.

Feature representation of genetic variants

Genetic variants were encoded following the additive, dominant and recessive genetic models[6], and by a binary non-additive manner according to the presence of the major allele or minor allele (Tables S.B and S.C). The minor allele was assumed to be the allele of effect in the genetic model representations.

Table S.B: Encoding representations for genetic models.

Encoding of genetics	**Additive**	**Dominant**	**Recessive**

Genotype			
Homozygotes major allele	0	0	0
Heterozygotes	1	1	0
Homozygotes minor allele	2	1	1

Table S.C: Sparse encoding of genetic features.

Encoding of genetics\ Genotype	**SNP minor allele**	**SNP major allele**
Homozygotes major allele	0	1
Heterozygotes	1	1
Homozygotes minor allele	1	0

The additive genetic encoding was {0,1,2} according to the number of minor alleles. For the dominant and recessive genetic encodings, variants where encoded with {0, 1} according to genotype. Using dominant encoding, homozygotes for the major allele received 0, and heterozygotes or homozygotes for the minor allele received 1. Using recessive encoding, homozygotes for the major allele or heterozygotes received 0, and homozygotes for the minor allele received 1. Missing values for genotype features were imputed with -1.

Model performance evaluation

Results of the machine learning models were evaluated using ROC-AUC, sensitivity and specificity, positive predictive value (PPV) and negative predictive value (NPV). Models were determined to overfit if the training ROC-AUC performance reached 1 during parameter optimization. To evaluate the tested models, the prediction outcome label (AAP case/control) was permuted 100 times to get distribution of random ROC-AUC performances.

Feature importance evaluation

Features were correlated using Spearman's correlation coefficient, R^2. A cut-off of $-0.5 > R^2 < 0.5$ was applied to cluster correlated features. The feature importance for selected models was evaluated based on a 'leave-one-out' approach for each highly correlated group or single variable. The groups were set to zero and the model trained with the optimized parameters, which gave a new partial ROC-AUC test performance. The feature importance was reported as the change in ROC-AUC when leaving out each group against the performance of the model trained on the full set of features:

$$\Delta = \text{ROC-AUC}_{\text{partial}} — \text{ROC-AUC}_{\text{full}}$$

Therefore, $\Delta < 0$ corresponds to an important feature that improves the model, while $\Delta \geq 0$ is a feature of no importance or a feature that worsens the model.

Personalized AI ensemble model strategy

For each model and type of genetic encoding type, the top three performing model initializations were selected across different SNP datasets and machine learning models were selected to gain as much predictive power as possible. The ensemble was made with a combined scoring approach using the predictions of the individual models in the ensemble. Three types of scoring the predictions were applied, in order to calculate a combined prediction per sample, which could then be used to estimate a collective performance of the ensemble model. 1) The *mean* of scores was a simple mean per sample of the predictions made by each model in the ensemble. 2) The *majority voting* approach rounded each prediction to either 0 or 1 at a threshold of 0.5, and then determined the prediction for each sample as the one selected most often. 3) The *mean of confident scores* uses only the predictions that are confident enough to calculate the mean per sample. The confidence of a prediction was set by a threshold, e.g. t=0.65, meaning that scores ≤ 0.35 and scores ≥ 0.65 are included when estimating the mean prediction score per sample.

Supplementary Table S.2: ROC-AUC of clinical baseline models with age and sex

Table S.2: Performance of logistic regression, random forest, AdaBoost and artificial neural network with and without down-sampling of the case/control ratio on the PdL training cohort (N=1290). The performance metrics are reported as mean ± standard deviation by ROC-AUC, sensitivity, specificity and Matthew's correlation coefficient (MCC).

Model type	**ROC-AUC (N=1290)**	**Sensitivity**	**Specificity**	**MCC**
Logistic regression No down-sampling	0.61 ± 0	0 ± 0	1 ± 0	0 ± 0
Random forest No down-sampling	0.63 ± 0	0 ± 0	1 ± 0	0 ± 0
AdaBoost No down-sampling	0.62 ± 0	0 ± 0	1 ± 0	0 ± 0
Artificial neural network (1 hidden layer) No down-sampling	0.62 ± 0	0 ± 0	1 ± 0	0 ± 0
Artificial neural network (2 hidden layers) No down-sampling	0.62 ± 0	0 ± 0	1 ± 0	0 ± 0
Logistic regression down-sampling	0.62 ± 0.01	0.57 ± 0.01	0.68 ± 0.01	0.18 ± 0
Random forest down-sampling	0.62 ± 0.01	0.56 ± 0.02	0.69 ± 0.01	0.18 ± 0.01
AdaBoost down-sampling	0.62 ± 0.01	0.57 ± 0.02	0.68 ± 0.01	0.18 ± 0
Artificial neural network (1 hidden layer) down-sampling	0.62 ± 0.01	0.57 ± 0.01	0.68 ± 0	0.18 ± 0
Artificial neural network (2 hidden layers) down-sampling	0.62 ± 0.01	0.57 ± 0.01	0.68 ± 0	0.18 ± 0

ROC-AUC for all trained clinical baseline models with sex and age ranged: 0.61−0.63, however, the models falsely classified every patient to be never develop AAP as specificity is 1 and the MCC indicate random performance (MCC = 0) in all cross-validated models. This tables shows that down-sampling of the majority class is needed (non-AAP controls are needed for training of the models). Implementation of down-sampling on the training sets show that all models obtain similar ROC-AUC performances, but the models do now not misclassify all patients as never developing AAP.

Supplementary Table S3: Performances of all trained models

Table S.3: ROC-AUC reported as mean ± standard deviation for the following machine learning models; logistic regression, random forest, AdaBoost, artificial neural networks (ANN, 1 and 2 hidden layers respectively) with age, sex and different sets of genetic variants encoded by dimensionality reduction methods incl. PCA or a weighted genetic risk score, or by single effect of SNPs by additive, dominant, recessive, or sparse genetic models.

ROC-AUC (N=1290)	**Logistic regression**	**Random forest**	**AdaBoost**	**ANN (1 hid)**	**ANN (2 hid)**
Dimensionality reduction					
3 principal components of 8 pancreatitis SNP-annotated genes	0.47 ± 0.01	0.51 ± 0.02	0.51 ± 0.02	0.51 ± 0.01	0.47 ± 0.01
Genetic risk scores for 6 candidate SNPs	0.63 ± 0.01	0.63 ± 0.01	0.64 ± 0.01	0.63 ± 0.01	0.63 ± 0.01
Additive encoding of genetics					
6 candidate SNPs	0.66 ± 0.01	0.64 ± 0.01	0.66 ± 0.01	0.67 ± 0.01	0.66 ± 0.01
8 genes	0.58 ± 0.02	0.57 ± 0.02	0.60 ± 0.02	0.58 ± 0.02	0.58 ± 0.02
8 genes (prioritized)	0.55 ± 0.02	0.59 ± 0.01	0.59 ± 0.02	0.56 ± 0.02	0.54 ± 0.02
8 genes GTEx eQTL	0.56 ± 0.02	0.55 ± 0.03	0.58 ± 0.02	0.57 ± 0.02	0.57 ± 0.02
8 genes GTEx eQTL (prioritized)	0.57 ± 0.02	0.61 ± 0.01	0.61 ± 0.01	0.58 ± 0.01	0.56 ± 0.02
2 genes, chromosome 20	0.63 ± 0.01	0.65 ± 0.01	0.66 ± 0.01	0.64 ± 0.01	0.63 ± 0.02
Wolthers *et al* 2019	0.80 ± 0.01	0.79 ± 0.01	0.81 ± 0.01	0.81 ± 0.01	0.80 ± 0.01
Liu *et al* 2016	0.61 ± 0.01	0.61 ± 0.01	0.63 ± 0.01	0.63 ± 0.01	0.62 ± 0.01
Abaji *et al* 2017	0.61 ± 0.01	0.61 ± 0.01	0.62 ± 0.01	0.61 ± 0.01	0.61 ± 0.01
Dominant encoding of genetics					
6 candidate SNPs	0.66 ± 0.01	0.65 ± 0.01	0.67 ± 0.01	0.66 ± 0.01	0.66 ± 0.01
8 genes	0.58 ± 0.02	0.57 ± 0.02	0.60 ± 0.02	0.58 ± 0.02	0.58 ± 0.02
8 genes (prioritized)	0.57 ± 0.02	0.61 ± 0.02	0.61 ± 0.01	0.58 ± 0.02	0.56 ± 0.02

8 genes GTEx eQTL	0.57 ± 0.02	0.56 ± 0.03	0.60 ± 0.02	0.58 ± 0.02	0.57 ± 0.02
8 genes GTEx eQTL (prioritized)	0.58 ± 0.01	0.61 ± 0.01	0.61 ± 0.01	0.59 ± 0.01	0.57 ± 0.02
2 genes, chromosome 20	0.63 ± 0.01	0.66 ± 0.01	0.66 ± 0.01	0.65 ± 0.01	0.63 ± 0.01
Wolthers *et al* 2019	0.79 ± 0.01	0.78 ± 0.01	0.80 ± 0.01	0.80 ± 0.01	0.79 ± 0.01
Liu *et al* 2016	0.61 ± 0.01	0.61 ± 0.01	0.63 ± 0.01	0.63 ± 0.01	0.62 ± 0.01
Abaji *et al* 2017	0.61 ± 0.01	0.61 ± 0.01	0.62 ± 0.01	0.62 ± 0.01	0.62 ± 0.01
Recessive encoding of genetics					
6 candidate SNPs	0.63 ± 0.01	0.63 ± 0.01	0.64 ± 0.01	0.63 ± 0.01	0.63 ± 0.01
8 genes	0.57 ± 0.02	0.61 ± 0.02	0.62 ± 0.02	0.57 ± 0.02	0.56 ± 0.02
8 genes (prioritized)	0.55 ± 0.02	0.57 ± 0.02	0.59 ± 0.01	0.56 ± 0.02	0.54 ± 0.02
8 genes GTEx eQTL	0.53 ± 0.02	0.58 ± 0.02	0.59 ± 0.02	0.53 ± 0.02	0.53 ± 0.02
8 genes GTEx eQTL (prioritized)	0.58 ± 0.02	0.61 ± 0.01	0.62 ± 0.01	0.60 ± 0.01	0.59 ± 0.02
2 genes, chromosome 20	0.59 ± 0.02	0.61 ± 0.01	0.62 ± 0.01	0.62 ± 0.01	0.60 ± 0.01
Wolthers *et al* 2019	0.67 ± 0.01	0.68 ± 0.01	0.70 ± 0.01	0.69 ± 0.01	0.68 ± 0.01
Liu *et al* 2016	0.62 ± 0.01	0.61 ± 0.01	0.62 ± 0.01	0.62 ± 0.01	0.62 ± 0.01
Abaji *et al* 2017	0.60 ± 0.01	0.61 ± 0.01	0.62 ± 0.01	0.62 ± 0.01	0.61 ± 0.01
Sparse encoding of genetics					
6 candidate SNPs	0.66 ± 0.01	0.65 ± 0.01	0.67 ± 0.01	0.67 ± 0.01	0.67 ± 0.01
8 genes	0.59 ± 0.02	0.56 ± 0.02	0.58 ± 0.02	0.58 ± 0.02	0.59 ± 0.02
8 genes (prioritized)	0.57 ± 0.02	0.59 ± 0.02	0.59 ± 0.01	0.58 ± 0.02	0.56 ± 0.02
8 genes GTEx eQTL	0.57 ± 0.02	0.55 ± 0.02	0.58 ± 0.02	0.57 ± 0.02	0.57 ± 0.02
8 genes GTEx eQTL (prioritized)	0.57 ± 0.01	0.61 ± 0.01	0.61 ± 0.01	0.58 ± 0.02	0.57 ± 0.02
2 genes, chromosome 20	0.62 ± 0.02	0.65 ± 0.02	0.65 ± 0.01	0.64 ± 0.01	0.63 ± 0.02
Wolthers e*t al* 2019	0.78 ± 0.01	0.79 ± 0.01	0.81 ± 0.01	0.80 ± 0.01	0.79 ± 0.01
Liu *et al* 2016	0.61 ± 0.01	0.61 ± 0.01	0.63 ± 0.01	0.63 ± 0.01	0.62 ± 0.01

Abaji *et al* 2017	0.61 ± 0.01	0.61 ± 0.01	0.62 ± 0.01	0.62 ± 0.01	0.61 ± 0.01

343

Supplementary Figure S4: ROC curves and permutation tests across 100 model initializations.

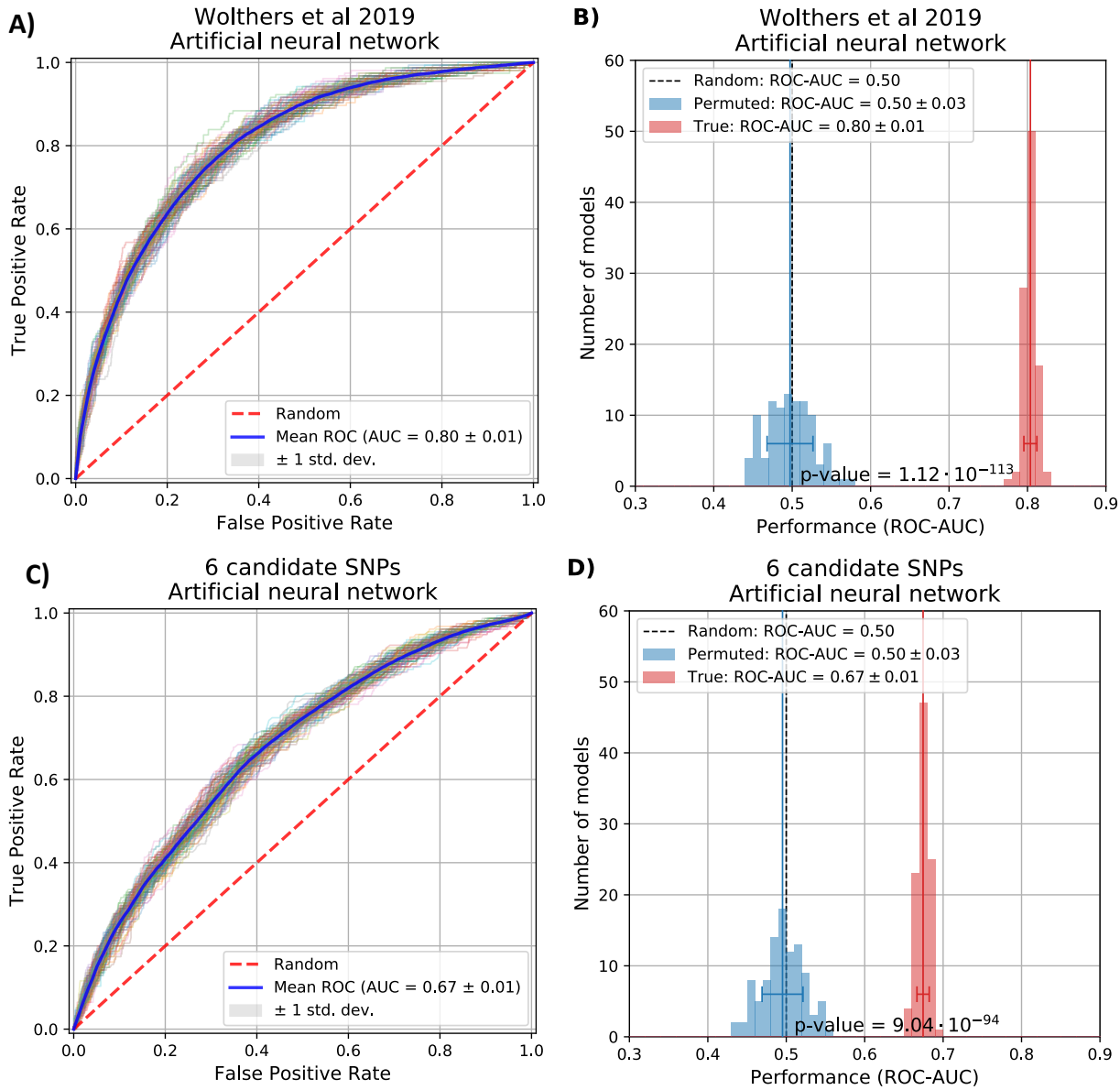


Figure S.4: A) ROC curves of artificial neural networks (1-layer) across 100 model initializations (N=1290) with down-sampling and used sparse encoding of genetics. Features: age, sex, and top 30 SNPs associated with AAP from Wolthers *et al*, 2019[2]. B) Permutation test for ROC-AUC of artificial neural networks (1-layer) across 100 model initializations (N=1290) with down-sampling and used sparse encoding of genetics. Features: age, sex, and top 30 SNPs associated with AAP from Wolthers *et al*, 2019[2]. C) ROC curves of artificial neural networks (1-layer) across 100 model initializations (N=1290) with down-sampling and used sparse encoding of genetics. Features: age, sex and six candidate SNPs. D) Permutation test for ROC-AUC of artificial neural networks (1-layer) across

100 model initializations (N=1290) with down-sampling and used sparse encoding of genetics. Features: age, sex and six candidate SNPs.

Supplementary Table S.5: Test performances for models predicting cases of AAP when re-exposed to asparaginase therapy

Table S.5: Performance of random forest and neural network model for a second AAP following re-exposure to asparaginase.

Model type	**ROC-AUC**	**Sensitivity**	**Specificity**
Logistic regression	0.49 ± 0	0.31 ± 0	0.67 ± 0
Random forest	0.63 ± 0.03	0.40 ± 0.09	0.76 ± 0.07
Artificial neural network (1 hidden layer)	0.56 ± 0	0.23 ± 0.02	0.78 ± 0.03
Logistic regression, Wolthers *et al* 2019	0.59 ± 0	0.38 ± 0	0.75 ± 0
Random forest, Wolthers *et al* 2019	0.58 ± 0.08	0.29 ± 0.12	0.75 ± 0.09
Artificial neural network (1 hidden layer), Wolthers *et al* 2019	0.58 ± 0.03	0.29 ± 0.05	0.70 ± 0.03

Supplementary Figure S.6: Permutation test (100 model initializations) for models predicting cases of AAP when re-exposed to asparaginase therapy.

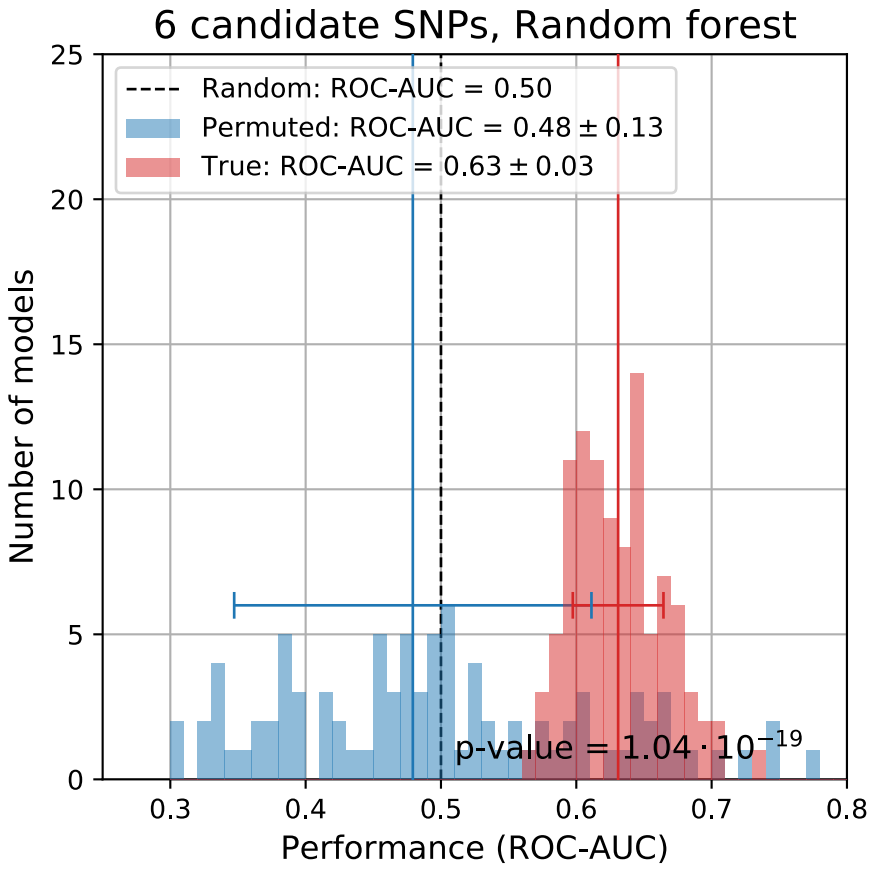


Figure S.6: Permutation test (100 model initializations) for models predicting cases of AAP when re-exposed to asparaginase therapy.

References

1. Freund Y, Schapire RE. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting*. J Comput Syst Sci 1997;55 119–139.

2. Wolthers BO, Frandsen TL, Patel CJ, et al. Trypsin-encoding PRSS1-PRSS2 variations influence the risk of asparaginase-associated pancreatitis in children with acute lymphoblastic leukemia: A ponte di legno toxicity working group report. Haematologica 2019;104(3):556–563.

3. Abaji R, Gagné V, Xu CJ, et al. Whole-exome sequencing identified genetic risk factors for asparaginase-related complications in childhood ALL patients. Oncotarget 2017;8(27):43752–43767.
4. Zator Z, Whitcomb DC. Insights into the genetic risk factors for the development of pancreatic disease. Therap Adv Gastroenterol 2017;10(3):323–336.
5. The GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. Nat Genet 2013;45(6):580–585.
6. Laird NM, Lange C. The Fundamentals of Modern Statistical Genetics. 2011. 15–28 p.

Bibliography

1. Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A. & Dokyoon, K. Methods of integrating data to uncover genotype–phenotype interactions. *Nature Reviews Genetics* **16,** 85–97 (2015).
2. Teschendorff, A. E. Avoiding common pitfalls in machine learning omic data science. *Nature Materials* **18,** 422–427 (2019).
3. Xu, C. & Jackson, S. A. Machine learning and complex biological data. *Genome Biology* **20,** 76 (2019).
4. McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. & Hirschhorn, J. N. Genome-wide association studies for complex traits: Consensus, uncertainty and challenges. *Nature Reviews Genetics* **9,** 356–369 (2008).
5. Alyass, A., Turcotte, M. & Meyre, D. From big data analysis to personalized medicine for all: Challenges and opportunities. *BMC Medical Genomics* **8,** 33 (2015).
6. Dempfle, A., Scherag, A., Hein, R., Beckmann, L., Chang-Claude, J. & Schäfer, H. Gene-environment interactions for complex traits: Definitions, methodological requirements and challenges. *European Journal of Human Genetics* **16,** 1164–1172 (2008).
7. Liu, R., Hong, J., Xu, X., Feng, Q., Zhang, D., Gu, Y., Shi, J., Zhao, S., Liu, W., Wang, X., Xia, H., Liu, Z., Cui, B., Liang, P., Xi, L., Jin, J., Ying, X., Wang, X., Zhao, X., Li, W., Jia, H., Lan, Z., Li, F., Wang, R., Sun, Y., Yang, M., Shen, Y., Jie, Z., Li, J., Chen, X., Zhong, H., Xie, H., Zhang, Y., Gu, W., Deng, X., Shen, B., Xu, X., Yang, H., Xu, G., Bi, Y., Lai, S., Wang, J., Qi, L., Madsen, L., Wang, J., Ning, G., Kristiansen, K. & Wang, W. Gut microbiome and serum metabolome alterations in obesity and after weight-loss intervention. *Nature Medicine* **23,** 859–868 (2017).
8. Goodarzi, M. O. Genetics of obesity: what genetic association studies have taught us about the biology of obesity and its complications. *The Lancet Diabetes & Endocrinology* **6,** 223–236 (2018).

9. Piening, B. D., Zhou, W., Contrepois, K., Röst, H., Gu Urban, G. J., Mishra, T., Hanson, B. M., Bautista, E. J., Leopold, S., Yeh, C. Y., Spakowicz, D., Banerjee, I., Chen, C., Kukurba, K., Perelman, D., Craig, C., Colbert, E., Salins, D., Rego, S., Lee, S., Zhang, C., Wheeler, J., Sailani, M. R., Liang, L., Abbott, C., Gerstein, M., Mardinoglu, A., Smith, U., Rubin, D. L., Pitteri, S., Sodergren, E., McLaughlin, T. L., Weinstock, G. M. & Snyder, M. P. Integrative Personal Omics Profiles during Periods of Weight Gain and Loss. *Cell Systems* **6,** 157–170 (2018).

10. Schüssler-Fiorenza Rose, S. M., Contrepois, K., Moneghetti, K. J., Zhou, W., Mishra, T., Mataraso, S., Dagan-Rosenfeld, O., Ganz, A. B., Dunn, J., Hornburg, D., Rego, S., Perelman, D., Ahadi, S., Sailani, M. R., Zhou, Y., Leopold, S. R., Chen, J., Ashland, M., Christle, J. W., Avina, M., Limcaoco, P., Ruiz, C., Tan, M., Butte, A. J., Weinstock, G. M., Slavich, G. M., Sodergren, E., McLaughlin, T. L., Haddad, F. & Snyder, M. P. A longitudinal big data approach for precision health. *Nature Medicine* **25,** 792–804 (2019).

11. Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G. & Meyre, D. Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics* **20,** 467–484 (2019).

12. Zhou, W., Sailani, M. R., Contrepois, K., Zhou, Y., Ahadi, S., Leopold, S. R., Zhang, M. J., Rao, V., Avina, M., Mishra, T., Johnson, J., Lee-McMullen, B., Chen, S., Metwally, A. A., Tran, T. D. B., Nguyen, H., Zhou, X., Albright, B., Hong, B. Y., Petersen, L., Bautista, E., Hanson, B., Chen, L., Spakowicz, D., Bahmani, A., Salins, D., Leopold, B., Ashland, M., Dagan-Rosenfeld, O., Rego, S., Limcaoco, P., Colbert, E., Allister, C., Perelman, D., Craig, C., Wei, E., Chaib, H., Hornburg, D., Dunn, J., Liang, L., Rose, S. M. S. F., Kukurba, K., Piening, B., Rost, H., Tse, D., McLaughlin, T., Sodergren, E., Weinstock, G. M. & Snyder, M. Longitudinal multi-omics of host–microbe dynamics in prediabetes. *Nature* **569,** 663–671 (2019).

13. Camacho, D. M., Collins, K. M., Powers, R. K., Costello, J. C. & Collins, J. J. Next-Generation Machine Learning for Biological Networks. *Cell* **173,** 1581–1592 (2018).

14. Gilbert, J. A., Blaser, M. J., Caporaso, J. G., Jansson, J. K., Lynch, S. V. & Knight, R. Current understanding of the human microbiome. *Nature Medicine* **24,** 392–400 (2018).

15. Hulsen, T., Jamuar, S. S., Moody, A. R., Karnes, J. H., Varga, O., Hedensted, S., Spreafico, R., Hafler, D. A. & Mckinney, E. F. From Big Data to Precision Medicine. *Frontiers in Medicine* **6,** 34 (2019).

16. Pinu, F. R., Beale, D. J., Paten, A. M., Kouremenos, K., Swarup, S., Schirra, H. J. & Wishart, D. Systems Biology and Multi-Omics Integration: Viewpoints from the Metabolomics Research Community. *Metabolites* **9,** 76 (2019).

17. Leff, D. R. & Yang, G.-Z. Big Data for Precision Medicine. *Engineering* **1,** 277–279 (2015).

18. Musy, S. N. & Simon, M. in *Big Data-enabled Nursing* (eds Delaney, C. W., Weaver, C. A., Warren, J. J., Clancy, T. R. & Simpson, R. L.) 79–101 (Springer International Publishing, 2017).

19. Cirillo, D. & Valencia, A. Big data analytics for personalized medicine. *Current Opinion in Biotechnology* **58,** 161–167 (2019).

20. Prosperi, M., Min, J. S., Bian, J. & Modave, F. Big data hurdles in precision medicine and precision public health. *BMC Medical Informatics and Decision Making* **18,** 139 (2018).

21. Dash, S., Shakyawar, S. K., Sharma, M. & Kaushik, S. Big data in healthcare: management, analysis and future prospects. *Journal of Big Data* **6,** 54 (2019).

22. Gligorijević, V., Malod-Dognin, N. & Pržulj, N. Integrative methods for analyzing big data in precision medicine. *Proteomics* **16,** 741–758 (2016).

23. Okser, S., Pahikkala, T. & Aittokallio, T. Genetic variants and their interactions in disease risk prediction - Machine learning and network perspectives. *BioData Mining* **6,** 5 (2013).

24. Ho, D. S. W., Schierding, W., Wake, M., Saffery, R. & O'Sullivan, J. Machine learning SNP based prediction for precision medicine. *Frontiers in Genetics* **10,** 267 (2019).

25. Hébert, H. L., Shepherd, B., Milburn, K., Veluchamy, A., Meng, W., Carr, F., Donnelly, L. A., Tavendale, R., Leese, G., Colhoun, H. M., Dow, E., Morris, A. D., Doney, A. S., Lang, C. C., Pearson, E. R., Smith, B. H. & Palmer, C. N. Cohort profile: Genetics of Diabetes Audit and Research in Tayside Scotland (GoDARTS). *International Journal of Epidemiology* **47,** 380–381 (2018).

26. Siderowf, A. D. Evidence from Clinical Trials: Can We Do Better? *NeuroRx* **1,** 363–371 (2004).

27. Spieth, P. M., Kubasch, A. S., Penzlin, A. I., Illigens, B. M.-W., Barlinn, K. & Siepmann, T. Randomized controlled trials - a matter of design. *Neuropsychiatric disease and treatment* **12,** 1341–1349 (2016).

28. Frieden, T. R. Evidence for Health Decision Making — Beyond Randomized, Controlled Trials. *New England Journal of Medicine* **377,** 465–475 (2017).

29. Agyeman, A. A. & Ofori-Asenso, R. Perspective: Does personalized medicine hold the future for medicine? *Journal of Pharmacy and Bioallied Sciences* **7,** 239–244 (2015).

30. Khoury, M. J., Iademarco, M. F. & Riley, W. T. Precision Public Health for the Era of Precision Medicine. *American Journal of Preventive Medicine* **50,** 398–401 (2016).

31. Ramaswami, R., Bayer, R. & Galea, S. Precision Medicine from a Public Health Perspective. *Annual Review of Public Health* **39,** 153–168 (2018).

32. Price, N. D., Magis, A. T., Earls, J. C., Glusman, G., Levy, R., Lausted, C., McDonald, D. T., Kusebauch, U., Moss, C. L., Zhou, Y., Qin, S., Moritz, R. L., Brogaard, K., Omenn, G. S., Lovejoy, J. C. & Hood, L. A wellness study of 108 individuals using personal, dense, dynamic data clouds. *Nature Biotechnology* **35,** 747–756 (2017).

33. Collins, F. S. & Varmus, H. A New Initiative on Precision Medicine. *The new england journal of medicine* **372,** 793–795 (2015).

34. Frost & Sullivan. *White Paper, "Drowning in Big Data? Reducing Information Technology Complexities and Costs For Healthcare Organizations"* 2012.

35. Andreu-Perez, J., Poon, C. C. Y., Merrifield, R. D., Wong, S. T. C. & Yang, G. Z. Big Data for Health. *IEEE Journal of Biomedical and Health Informatics* **19,** 1193–1208 (2015).

36. Cannataro, M. in *Encyclopedia of Big Data Technologies* (eds Sakr, S. & Zomaya, A.) (Springer, Cham., 2019).

37. Wooden, B., Goossens, N., Hoshida, Y. & Friedman, S. L. Using Big Data to Discover Diagnostics and Therapeutics for Gastrointestinal and Liver Diseases. *Gastroenterology* **152,** 53–67.e3 (2017).

38. Wu, H., Tremaroli, V. & Bäckhed, F. Linking Microbiota to Human Diseases: A Systems Biology Perspective. *Trends in Endocrinology & Metabolism* **26,** 758–770 (2015).

39. Koshy, A. N., Sajeev, J. K., Nerlekar, N., Brown, A. J., Rajakariar, K., Zureik, M., Wong, M. C., Roberts, L., Street, M., Cooke, J. & Teh, A. W. Smart watches for heart rate assessment in atrial arrhythmias. *International Journal of Cardiology* **266,** 124–127 (2018).

40. *UK Biobank* `https://www.ukbiobank.ac.uk/` (18 January 2020).

41. *the database of Genotypes and Phenotypes (dbGaP)* `https://www.ncbi.nlm.nih.gov/gap/` (18 January 2020).

42. *The Cancer Genome Atlas Program (TCGA)* `https://www.cancer.gov/tcga` (18 January 2020).

43. Schneider, M. V. & Orchard, S. in *Bioinformatics for Omics data* (ed Mayer, B.) 3–30 (Humana Press, 2011).

44. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431,** 931–945 (2004).

45. Shendure, J., Findlay, G. M. & Snyder, M. W. Genomic Medicine – Progress, Pitfalls, and Promise. *Cell* **177,** 45–57 (2019).

46. Editorial. Genome variation in precision medicine. *Nature Genetics* **48,** 701 (2016).

47. Collins, F. S. & Mansoura, M. K. The Human Genome Project: revealing the shared inheritance of all humankind. *Cancer* **91,** 221–225 (2001).

48. Makaowski, W. The human genome structure and organization. *Acta Biochimica Polonica* **48,** 587–598 (2001).

49. Lappalainen, T., Scott, A. J., Brandt, M. & Hall, I. M. Genomic Analysis in the Age of Human Genome Sequencing. *Cell* **177,** 70–84 (2019).

50. Crawford, D. C. & Nickerson, D. A. Definition and Clinical Importance of Haplotypes. *Annual Review of Medicine* **56,** 303–320 (February 2005).

51. Ke, X., Taylor, M. S. & Cardon, L. R. Singleton SNPs in the human genome and implications for genome-wide association studies. *European Journal of Human Genetics* **16,** 506–515 (2008).

52. Illumina. *The Omni Family of Microarrays* technical report (2010), Pub. No. 370–2009–020. `https://www.illumina.com/documents/products/datasheets/datasheet_gwas_roadmap.pdf`.

53. Nakagawa, H. & Fujita, M. Whole genome sequencing analysis for cancer genomics and precision medicine. *Cancer Science* **109,** 513–522 (2018).

54. Blum, H. E. The human microbiome. *Advances in Medical Sciences* **62,** 414–420 (2017).

55. Turnbaugh, P. J., Hamady, M., Yatsunenko, T., Cantarel, B. L., Duncan, A., Ley, R. E., Sogin, M. L., Jones, W. J., Roe, B. A., Affourtit, J. P., Egholm, M., Henrissat, B., Heath, A. C., Knight, R. & Gordon, J. I. A core gut microbiome in obese and lean twins. *Nature* **457,** 480–484 (2009).

56. Karlsson, F., Tremaroli, V., Nielsen, J. & Bäckhed, F. Assessing the human gut microbiota in metabolic diseases. *Diabetes* **62,** 3341–3349 (2013).

57. Thursby, E. & Juge, N. Introduction to the human gut microbiota. *The Biochemical journal* **474,** 1823–1836 (2017).

58. Holmes, E., Li, J. V., Marchesi, J. R. & Nicholson, J. K. Gut microbiota composition and activity in relation to host metabolic phenotype and disease risk. *Cell Metabolism* **16,** 559–564 (2012).

59. Rodríguez, J. M., Murphy, K., Stanton, C., Ross, R. P., Kober, O. I., Juge, N., Avershina, E., Rudi, K., Narbad, A., Jenmalm, M. C., Marchesi, J. R. & Collado, M. C. The composition of the gut microbiota throughout life, with an emphasis on early life. *Microbial Ecology in Health and Disease* **26,** 26050 (2015).

60. Hansen, L. B. S., Roager, H. M., Søndertoft, N. B., Gøbel, R. J., Kristensen, M., Vallès-Colomer, M., Vieira-Silva, S., Ibrügger, S., Lind, M. V., Mærkedahl, R. B., Bahl, M. I., Madsen, M. L., Havelund, J., Falony, G., Tetens, I., Nielsen, T., Allin, K. H., Frandsen, H. L., Hartmann, B., Holst, J. J., Sparholt, M. H., Holck, J., Blennow, A., Moll, J. M., Meyer, A. S., Hoppe, C., Poulsen, J. H., Carvalho, V., Sagnelli, D., Dalgaard, M. D., Christensen, A. F., Lydolph, M. C., Ross, A. B., Villas-Bôas, S., Brix, S., Sicheritz-Pontén, T., Buschard, K., Linneberg, A., Rumessen, J. J., Ekstrøm, C. T., Ritz, C., Kristiansen, K., Nielsen, H. B., Vestergaard, H., Færgeman, N. J., Raes, J., Frøkiær, H., Hansen, T., Lauritzen, L.,

Gupta, R., Licht, T. R. & Pedersen, O. A low-gluten diet induces changes in the intestinal microbiome of healthy Danish adults. *Nature Communications* **9** (2018).

61. Rausch, P., Rühlemann, M., Hermes, B. M., Doms, S., Dagan, T., Dierking, K., Domin, H., Fraune, S., von Frieling, J., Hentschel, U., Heinsen, F.-A., Höppner, M., Jahn, M. T., Jaspers, C., Kissoyan, K. A. B., Langfeldt, D., Rehman, A., Reusch, T. B. H., Roeder, T., Schmitz, R. A., Schulenburg, H., Soluch, R., Sommer, F., Stukenbrock, E., Weiland-Bräuer, N., Rosenstiel, P., Franke, A., Bosch, T. & Baines, J. F. Comparative analysis of amplicon and metagenomic sequencing methods reveals key features in the evolution of animal metaorganisms. *Microbiome* **7,** 133 (2019).
62. German, J. B., Hammock, B. D. & Watkins, S. M. Metabolomics: building on a century of biochemistry to guide human health. *Metabolomics* **1,** 3–9 (2005).
63. Kirpich, A. S., Ibarra, M., Moskalenko, O., Fear, J. M., Gerken, J., Mi, X., Ashrafi, A., Morse, A. M. & Mcintyre, L. M. SECIMTools: a suite of metabolomics data analysis tools. *BMC Bioinformatics* **19,** 151 (2018).
64. Smoleńska, Z. & Zdrojewski, Z. Metabolomics and its potential in diagnosis, prognosis and treatment of rheumatic diseases. *Reumatologia* **53,** 152–156 (2015).
65. Thiese, M. S. Observational and interventional study design types; an overview. *Biochemia Medica* **24,** 199–210 (2014).
66. Ranganathan, P. & Aggarwal, R. Study designs: Part 1 - An overview and classification. *Perspectives in Clinical Research* **9,** 184–186 (2018).
67. Woodward, M. *Epidemiology: Study Design and Data Analysis* 3rd, 898 (Chapman and Hall/CRC, 2014).
68. Ranganathan, P. & Aggarwal, R. Study designs: Part 3 - Analytical observational studies. *Perspectives in Clinical Research* **10,** 91–94 (2019).
69. Kreuzthaler, M., Schulz, S. & Berghold, A. Secondary use of electronic health records for building cohort studies through top-down information extraction. *Journal of Biomedical Informatics* **53,** 188–195 (2015).
70. Aggarwal, R. & Ranganathan, P. Study designs: Part 4 - Interventional studies. *Perspectives in Clinical Research* **10,** 137–139 (2019).
71. Guyatt, G. H., Haynes, R. B., Jaeschke, R. Z., Cook, D. J., Green, L., Naylor, C. D., Wilson, M. C. & Richardson, W. S. Users' guides to the medical literature: XXV. Evidence-based medicine: Principles for applying the users' guides to patient care. *JAMA: Journal of the American Medical Association* **284,** 1290–1296 (2000).
72. Duan, N., Kravitz, R. L. & Schmid, C. H. Single-patient (n-of-1) trials: A pragmatic clinical decision methodology for patient-centered comparative effectiveness research. *Journal of Clinical Epidemiology* **66,** S21–S28 (2013).
73. Agniel, D., Kohane, I. S. & Weber, G. M. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *BMJ* **361,** k1479 (2018).

74. USF HEALTH. *Differences Between EHR and EMR* https://www.usfhealthonline.com/resources/key-concepts/ehr-vs-emr/ (29 December 2019).

75. Sun, W., Cai, Z., Li, Y., Liu, F., Fang, S. & Wang, G. Data Processing and Text Mining Technologies on Electronic Medical Records: A Review. *Journal of Healthcare Engineering* **2018,** 4302425 (2018).

76. Denny, J. C., Ritchie, M. D., Basford, M. A., Pulley, J. M., Bastarache, L., Brown-Gentry, K., Wang, D., Masys, D. R., Roden, D. M. & Crawford, D. C. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. *Bioinformatics* **26,** 1205–1210 (2010).

77. Rose, S. Machine Learning for Prediction in Electronic Health Data. *JAMA Network Open* **1,** e181404 (2018).

78. Artzi, N. S., Shilo, S., Hadar, E., Rossman, H., Barbash-Hazan, S., Ben-Haroush, A., Balicer, R. D., Feldman, B., Wiznitzer, A. & Segal, E. Prediction of gestational diabetes based on nationwide electronic health records. *Nature Medicine* **26,** 71–76 (2020).

79. Winn, A. N. & Neuner, J. M. Making Sure We Don't Forget the Basics When Using Machine Learning. *JNCI: Journal of the National Cancer Institute* **111,** 529–530 (2019).

80. Vogt, H., Green, S., Ekstrøm, C. T. & Brodersen, J. How precision medicine and screening with big data could increase overdiagnosis. *BMJ* **366,** l5270 (2019).

81. Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L. & Parkinson, H. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research* **42,** 1001–1006 (2014).

82. Jannot, A.-s., Ehret, G. & Perneger, T. P < 5 x 10-8 has emerged as a standard of statistical significance for genome-wide association studies. *Journal of Clinical Epidemiology* **68,** 8783 (2015).

83. McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R., Thormann, A., Flicek, P. & Cunningham, F. The Ensembl Variant Effect Predictor. *Genome Biology* **17,** 1–14 (2016).

84. Buniello, A., Macarthur, J. A., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., Suveges, D., Vrousgou, O., Whetzel, P. L., Amode, R., Guillen, J. A., Riat, H. S., Trevanion, S. J., Hall, P., Junkins, H., Flicek, P., Burdett, T., Hindorff, L. A., Cunningham, F. & Parkinson, H. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research* **47,** D1005–D1012 (2019).

85. *GWAS Catalog. The NHGRI-EBI Catalog of published genome-wide association studies* https://www.ebi.ac.uk/gwas/ (18 January 2020).

86. Arking, D. & Rommens, J. Editorial overview: Molecular and genetic bases of disease: Enter the post-GWAS era. *Current Opinion in Genetics & Development* **33,** 77–79 (2015).

87. Sud, A., Kinnersley, B. & Houlston, R. S. Genome-wide association studies of cancer: Current insights and future perspectives. *Nature Reviews Cancer* **17,** 692–704 (2017).

88. Claussnitzer, M., Cho, J. H., Collins, R., Cox, N. J., Dermitzakis, E. T., Hurles, M. E., Kathiresan, S., Kenny, E. E., Lindgren, C. M., MacArthur, D. G., North, K. N., Plon, S. E., Rehm, H. L., Risch, N., Rotimi, C. N., Shendure, J., Soranzo, N. & McCarthy, M. I. A brief history of human disease genetics. *Nature* **577,** 179–189 (2020).

89. Marees, A. T., de Kluiver, H., Stringer, S., Vorspan, F., Curis, E., Marie-Claire, C. & Derks, E. M. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *International Journal of Methods in Psychiatric Research* **27,** e1608 (2018).

90. Singh, A., Nadkarni, G., Gottesman, O., Ellis, S. B., Bottinger, E. P. & Guttag, J. V. Incorporating temporal EHR data in predictive models for risk stratification of renal function deterioration. *Journal of Biomedical Informatics* **53,** 220–228 (2015).

91. Pedersen, H. K., Gudmundsdottir, V., Pedersen, M. K., Brorsson, C., Brunak, S. & Gupta, R. Ranking factors involved in diabetes remission after bariatric surgery using machine-learning integrating clinical and genomic biomarkers. *npj Genomic Medicine* **1,** 16035 (2016).

92. Caruana, E. J., Roman, M., Hernández-Sánchez, J. & Solli, P. Longitudinal studies. *Journal of Thoracic Disease* **7,** E537–E540 (2015).

93. Garrett-Bakelman, F. E., Darshi, M., Green, S. J., Gur, R. C., Lin, L., Macias, B. R., McKenna, M. J., Meydan, C., Mishra, T., Nasrini, J., Piening, B. D., Rizzardi, L. F., Sharma, K., Siamwala, J. H., Taylor, L., Vitaterna, M. H., Afkarian, M., Afshinnekoo, E., Ahadi, S., Ambati, A., Arya, M., Bezdan, D., Callahan, C. M., Chen, S., Choi, A. M., Chlipala, G. E., Contrepois, K., Covington, M., Crucian, B. E., De Vivo, I., Dinges, D. F., Ebert, D. J., Feinberg, J. I., Gandara, J. A., George, K. A., Goutsias, J., Grills, G. S., Hargens, A. R., Heer, M., Hillary, R. P., Hoofnagle, A. N., Hook, V. Y., Jenkinson, G., Jiang, P., Keshavarzian, A., Laurie, S. S., Lee-McMullen, B., Lumpkins, S. B., MacKay, M., Maienschein-Cline, M. G., Melnick, A. M., Moore, T. M., Nakahira, K., Patel, H. H., Pietrzyk, R., Rao, V., Saito, R., Salins, D. N., Schilling, J. M., Sears, D. D., Sheridan, C. K., Stenger, M. B., Tryggvadottir, R., Urban, A. E., Vaisar, T., Van Espen, B., Zhang, J., Ziegler, M. G., Zwart, S. R., Charles, J. B., Kundrot, C. E., Scott, G. B., Bailey, S. M., Basner, M., Feinberg, A. P., Lee, S. M., Mason, C. E., Mignot, E., Rana, B. K., Smith, S. M., Snyder, M. P. & Turek, F. W. The NASA Twins Study: A multidimensional analysis of a year-long human spaceflight. *Science* **364,** eaau8650 (2019).

94. European Medicines Agency. *EMA Regulatory Science to 2025* technical report (2018), 1–60. https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/ema-regulatory-science-2025-strategic-reflection_en.pdf.

95. US Food and Drug Administration. *GUIDANCE DOCUMENT: Policy for Device Software Functions and Mobile Medical Applications* technical report (2019). https://www.fda.gov/media/80958/download.

96. US Food and Drug Administration. *Clinical Decision Support Software Draft Guidance for Industry and Food and Drug Administration Staff* technical report (2019), 1–27. https://www.fda.gov/media/109618/download.

97. Kuan, R. *Adopting AI in Health Care Will Be Slow and Difficult* https://hbr.org/2019/10/adopting-ai-in-health-care-will-be-slow-and-difficult (20 January 2020).

98. Herlau, T., Schmidt, M. N. & Mørup, M. *Introduction to Machine Learning and Data Mining* Course Notes 2017 (2017).

99. Dagliati, A., Marini, S., Sacchi, L., Cogni, G., Teliti, M., Tibollo, V., Cata, P. D., Chiovato, L. & Bellazzi, R. Machine Learning Methods to Predict Diabetes Complications. *Journal of Diabetes Science and Technology* **12,** 295–302 (2018).

100. McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. C., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F. J., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C. J., King, D., Ledsam, J. R., Melnick, D., Mostofi, H., Peng, L., Reicher, J. J., Romera-Paredes, B., Sidebottom, R., Suleyman, M., Tse, D., Young, K. C., De Fauw, J. & Shetty, S. International evaluation of an AI system for breast cancer screening. *Nature* **577,** 89–94 (2020).

101. Alaa, A. M., Bolton, T., Angelantonio, E. D., Rudd, J. H. F. & Van Der Schaar, M. Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants. *PLoS ONE* **14,** e0213653 (2019).

102. Kim, K.-J., Kim, M., Adamopoulos, I. E. & Tagkopoulos, I. Compendium of synovial signatures identifies pathologic characteristics for predicting treatment response in rheumatoid arthritis patients. *Clinical Immunology* **202,** 1–10 (2019).

103. Zhao, J., Papapetrou, P., Asker, L. & Boström, H. Learning from heterogeneous temporal data in electronic health records. *Journal of Biomedical Informatics* **65,** 105–119 (2017).

104. Zhao, J., Feng, Q. P., Wu, P., Lupu, R. A., Wilke, R. A., Wells, Q. S., Denny, J. C. & Wei, W.-Q. Learning from Longitudinal Data in Electronic Health Record and Genetic Data to Improve Cardiovascular Event Prediction. *Scientific Reports* **9,** 717 (2019).

105. Carpenter, J. R. & Kenward, M. G. *Multiple Imputation and its Application* (WILEY, 2013).

106. Kang, H. The prevention and handling of the missing data. *Korean journal of anesthesiology* **64,** 402–406 (2013).

107. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9,** 559 (2008).

108. Saeys, Y., Inza, I. & Larrañaga, P. A review of feature selection techniques in bioinformatics. *Bioinformatics* **23,** 2507–2517 (2007).

109. Ritari, J., Hyvärinen, K., Koskela, S., Itälä-Remes, M., Niittyvuopio, R., Nihtinen, A., Salmenniemi, U., Putkonen, M., Volin, L., Kwan, T., Pastinen, T. & Partanen, J. Genomic prediction of relapse in recipients of allogeneic haematopoietic stem cell transplantation. *Leukemia* **33,** 240–248 (2019).

110. Misra, B. B., Langefeld, C., Olivier, M. & Cox, L. A. Integrated omics: Tools, advances and future approaches. *Journal of Molecular Endocrinology* **62,** R21–R45 (2019).

111. Rohart, F., Gautier, B., Singh, A. & Lê Cao, K. A. mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS Computational Biology* **13,** e1005752 (2017).

112. Wu, C., Zhou, F., Ren, J., Li, X., Jiang, Y. & Ma, S. A selective review of multi-level omics data integration using variable selection. *High-Throughput* **8,** 4 (2019).

113. Müller, A. C. & Guido, S. *Introduction to machine learning with Python* 1st (O'Reilly Media, 2017).

114. Couronné, R., Probst, P. & Boulesteix, A.-L. Random forest versus logistic regression: a large-scale benchmark experiment. *BMC Bioinformatics* **19,** 270 (2018).

115. James, G., Witten, D., Hastie, T. & Tibshirani, R. in *An Introduction to Statistical Learning: with Applications in R* (eds Casella, G., Fienberg, S. & Olkin, I.) 127–173 (Springer New York, 2013).

116. Hastie, T., Tibshirani, R. & Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* 2nd (Springer, 2009).

117. Kuhn, M. & Johnson, K. *Applied Predictive Modeling* (Springer, 2016).

118. Goldstein, B. A., Polley, E. C. & Briggs, F. B. S. Random forests for genetic association studies. *Statistical Applications in Genetics and Molecular Biology* **10,** 32 (2011).

119. Liaw, A. & Wiener, M. Classification and Regression by RandomForest. *Forest* **23.** `https://www.researchgate.net/publication/228451484_Classification_and_Regression_by_RandomForest` (2001).

120. McCulloch, W. S. & Pitts, W. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics* **5,** 115–133 (1943).

121. Li, Y., Huang, C., Ding, L., Li, Z., Pan, Y. & Gao, X. Deep learning in bioinformatics: Introduction, application, and perspective in the big data era. *Methods* **166,** 4–21 (2019).

122. Zhang, Z., Beck, M. W., Winkler, D. A., Huang, B., Sibanda, W. & Goyal, H. Opening the black box of neural networks: methods for interpreting neural network models in clinical applications. *Annals of Translational Medicine* **6,** 216 (2018).

123. Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F. & Sun, J. *Doctor AI: Predicting Clinical Events via Recurrent Neural Networks* in *Proceedings of the 1st Machine Learning for Healthcare Conference* **56** (2016), 301–318.

124. Parikh, R., Mathai, A., Parikh, S., Sekhar, G. C. & Thomas, R. Understanding and using sensitivity, specificity and predictive values. *Indian Journal of Ophthalmology* **56,** 45–50 (2008).

125. Kuhn, M. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software* **28,** 1–26 (2008).

126. Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure* **405,** 442–451 (1975).

127. Davis, J. & Goadrich, M. The Relationship Between Precision-Recall and ROC Curves. *Proceedings of the 23rd International Conference on Machine Learning* **2006,** 233–240 (2006).

128. Jiao, Y. & Du, P. Performance measures in evaluating machine learning based bioinformatics predictors for classifications. *Quantitative Biology* **4,** 320–330 (2016).

129. Mukherjee, S., Golland, P. & Panchenko, D. Permutation tests for classification. *AI Memo 2003-019.* `https://people.csail.mit.edu/polina/papers/AIM-2003-019.pdf` (30 January 2020) (2003).

130. Olden, J. D., Joy, M. K. & Death, R. G. An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecological Modelling* **178,** 389–397 (2004).

131. Agha, M. & Agha, R. The rising prevalence of obesity: part A: impact on public health. *International journal of surgery. Oncology* **2,** e17 (2017).

132. World Health Organization- WHO Media Centre. *Obesity and overweight: fact sheet (No. 311) 2015* `https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight` (22 January 2020).

133. Qi, L. Gene-diet interaction and weight loss. *Current Opinion in Lipidology* **25,** 27–34 (2014).

134. Djalalinia, S., Qorbani, M., Peykari, N. & Kelishadi, R. Health impacts of obesity. *Pakistan Journal of Medical Sciences* **31,** 239–242 (2015).

135. Popp, C. J., St-Jules, D. E., Hu, L., Ganguzza, L., Illiano, P., Curran, M., Li, H., Schoenthaler, A., Bergman, M., Schmidt, A. M., Segal, E., Godneva, A. & Sevick, M. A. The rationale and design of the personal diet study, a randomized clinical trial evaluating a personalized approach to weight loss in individuals with pre-diabetes and early-stage type 2 diabetes. *Contemporary Clinical Trials* **79,** 80–88 (2019).

136. Gibson, A. A. & Sainsbury, A. Strategies to Improve Adherence to Dietary Weight Loss Interventions in Research and Real-World Settings. *Behavioral Sciences* **7,** 7030044 (2017).

137. Malik, V. S. & Hu, F. B. Popular weight-loss diets: From evidence to practice. *Nature Clinical Practice Cardiovascular Medicine* **4,** 34–41 (2007).

138. Roager, H. M., Vogt, J. K., Kristensen, M., Hansen, L. B. S., Ibrügger, S., Mærkedahl, R. B., Bahl, M. I., Lind, M. V., Nielsen, R. L., Frøkiær, H., Gøbel, R. J., Landberg, R., Ross, A. B., Brix, S., Holck, J., Meyer, A. S., Sparholt, M. H., Christensen, A. F., Carvalho, V., Hartmann, B., Holst, J. J., Rumessen, J. J., Linneberg, A., Sicheritz-Pontén, T., Dalgaard, M. D., Blennow, A., Frandsen, H. L., Villas-Bôas, S., Kristiansen, K., Vestergaard, H., Hansen, T., Ekstrøm, C. T., Ritz, C., Nielsen, H. B., Pedersen, O. B., Gupta, R., Lauritzen, L. & Licht, T. R. Whole grain-rich diet reduces body weight and systemic low-grade inflammation without inducing major changes of the gut microbiome: a randomised cross-over trial. *Gut* **68,** 83–93 (2019).

139. Hjorth, M. F., Zohar, Y., Hill, J. O. & Astrup, A. Personalized Dietary Management of Overweight and Obesity Based on Measures of Insulin and Glucose. *Annual Review of Nutrition* **38,** 245–272 (2018).

140. Zeevi, D., Korem, T., Zmora, N., Israeli, D., Rothschild, D., Weinberger, A., Ben-Yacov, O., Lador, D., Avnit-Sagi, T., Lotan-Pompan, M., Suez, J., Mahdi, J. A., Matot, E., Malka, G., Kosower, N., Rein, M., Zilberman-Schapira, G., Dohnalová, L., Pevsner-Fischer, M., Bikovsky, R., Halpern, Z., Elinav, E. & Segal, E. Personalized Nutrition by Prediction of Glycemic Responses. *Cell* **163,** 1079–1094 (2015).

141. Mendes-Soares, H., Raveh-Sadka, T., Azulay, S., Edens, K., Ben-Shlomo, Y., Cohen, Y., Ofek, T., Bachrach, D., Stevens, J., Colibaseanu, D., Segal, L., Kashyap, P. & Nelson, H. Assessment of a Personalized Approach to Predicting Postprandial Glycemic Responses to Food Among Individuals Without Diabetes. *JAMA Network Open* **2,** e188102 (2019).

142. Mendes-Soares, H., Raveh-Sadka, T., Azulay, S., Ben-Shlomo, Y., Cohen, Y., Ofek, T., Stevens, J., Bachrach, D., Kashyap, P., Segal, L. & Nelson, H. Model of personalized postprandial glycemic response to food developed for an Israeli cohort predicts responses in Midwestern American individuals. *American Journal of Clinical Nutrition* **110,** 63–75 (2019).

143. Ibrügger, S., Gøbel, R. J., Vestergaard, H., Licht, T. R., Frøkiær, H., Linneberg, A., Hansen, T., Gupta, R., Pedersen, O. B., Kristensen, M. & Lauritzen, L. Two randomized cross-over trials assessing the impact of dietary gluten or wholegrain on the gut microbiome and host metabolic health. *Journal of Clinical Trials* **4** (2014).

144. Beretta, L. & Santaniello, A. Implementing ReliefF filters to extract meaningful features from genetic lifetime datasets. *Journal of Biomedical Informatics* **44,** 361–369 (2011).

145. Olson, R. S. *ReliefF* `http://pydoc.net/ReliefF/0.1.2/ReliefF/` (16 January 2020).

146. Urbanowicz, R. J., Meeker, M., La Cava, W., Olson, R. S. & Moore, J. H. Relief-based feature selection: Introduction and review. *Journal of Biomedical Informatics* **85,** 189–203 (2018).

147. Zimmet, P. Z. Diabetes and its drivers: the largest epidemic in human history? *Clinical Diabetes and Endocrinology* **3,** 1 (2017).

148. International Diabetes Federation. *International Diabetes Federation, IDF Diabetes Atlas 9th edition. Online version of IDF Diabetes Atlas: www.diabetesatlas.org* (11 January 2020).

149. Prasad, R. B. & Groop, L. Genetics of Type 2 Diabetes — Pitfalls and Possibilities. *Genes* **6,** 87–123 (2015).

150. Steenkamp, D. W., Alexanian, S. M. & Sternthal, E. Approach to the patient with atypical diabetes. *Canadian Medical Association Journal (CMAJ)* **186,** 678–684 (2014).

151. Zhou, K., Donnelly, L. A., Morris, A. D., Franks, P. W., Jennison, C., Palmer, C. N. A. & Pearson, E. R. Clinical and Genetic Determinants of Progression of Type 2 Diabetes: A DIRECT Study. *Diabetes Care* **37,** 718–724 (2014).

152. Jennison, C., Donnelly, L. A., Doney, A. S. F., Zhou, K., Pearson, E. R. & Franks, P. W. Rates of glycaemic deterioration in a real-world population with type 2 diabetes. *Diabetologia* **61,** 607–615 (2018).

153. Brunner, Y., Schvartz, D., Priego-capote, F., Couté, Y. & Sanchez, J.-c. Glucotoxicity and pancreatic proteomics. *Journal of Proteomics* **71,** 576–591 (2009).

154. Haeusler, R. A., McGraw, T. E. & Accili, D. Biochemical and cellular properties of insulin receptor signalling. *Nature Reviews Molecular Cell Biology* **19,** 31–44 (2018).

155. Cade, W. T. Diabetes-Related Microvascular and Macrovascular Diseases in the Physical Therapy Setting. *Physical Therapy* **88,** 1322–1335 (2008).

156. Agardh, E., Allebeck, P., Hallqvist, J., Moradi, T. & Sidorchuk, A. Type 2 diabetes incidence and socio-economic position: A systematic review and meta-analysis. *International Journal of Epidemiology* **40,** 804–818 (2011).

157. Grarup, N., Sandholt, C. H., Hansen, T. & Pedersen, O. Genetic susceptibility to type 2 diabetes and obesity: From genome-wide association studies to rare variants and beyond. *Diabetologia* **57,** 1528–1541 (2014).

158. Weng, J. P. & Hu, G. Diabetes: Leveraging the Tipping Point of the Diabetes Pandemic. *Diabetes* **66,** 1461–1463 (2017).

159. Mahajan, A., Taliun, D., Thurner, M., Robertson, N. R., Torres, J. M., Rayner, N. W., Payne, A. J., Steinthorsdottir, V., Scott, R. A., Grarup, N., Cook, J. P., Schmidt, E. M., Wuttke, M., Sarnowski, C., Mägi, R., Nano, J., Gieger, C., Trompet, S., Lecoeur, C., Preuss, M. H., Prins, B. P., Guo, X., Bielak, L. F., Below, J. E., Bowden, D. W., Chambers, J. C., Kim, Y. J., Ng, M. C., Petty, L. E., Sim, X., Zhang, W., Bennett, A. J., Bork-Jensen, J., Brummett, C. M., Canouil, M., Ec kardt, K. U., Fischer, K., Kardia, S. L., Kronenberg, F., Läll, K., Liu, C. T., Locke, A. E., Luan, J., Ntalla, I., Nylander, V., Schönherr, S., Schurmann, C., Yengo, L., Bottinger, E. P., Brandslund, I., Christensen, C., Dedoussis, G., Florez, J. C., Ford, I., Franco, O. H., Frayling, T. M., Giedraitis, V., Hackinger, S., Hattersley, A. T., Herder, C., Ikram, M. A., Ingelsson, M., Jørgensen, M. E., Jørgensen, T., Kriebel, J., Kuusisto, J., Ligthart, S., Lindgren, C. M., Linneberg, A., Lyssenko, V., Mamakou, V., Meitinger, T., Mohlke, K. L., Morris, A. D., Nadkarni, G., Pankow, J. S., Peters, A., Sattar, N., Stančáková, A., Strauch, K., Taylor, K. D., Thorand, B., Thorleifsson, G., Thorsteinsdottir, U., Tuomilehto, J., Witte, D. R., Dupuis, J., Peyser, P. A., Zeggini, E., Loos, R. J., Froguel, P., Ingelsson, E., Lind, L., Groop, L., Laakso, M., Collins, F. S., Jukema, J. W., Palmer, C. N., Grallert, H., Metspalu, A., Dehghan, A., Köttgen, A., Abecasis, G. R., Meigs, J. B., Rotter, J. I., Marchini, J., Pedersen, O., Hansen, T., Langenberg, C., Wareham, N. J., Stefansson, K., Gloyn, A. L., Morris, A. P., Boehnke, M. & McCarthy, M. I. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nature Genetics* **50,** 1505–1513 (2018).

160. Allin, K. H., Nielsen, T. & Pedersen, O. Gut microbiota in patients with type 2 diabetes mellitus. *European Journal of Endocrinology* **172,** R167–R177 (2015).

161. Udler, M. S., Kim, J., von Grotthuss, M., Bonàs-Guarch, S., Cole, J. B., Chiou, J., Boehnke, M., Laakso, M., Atzmon, G., Glaser, B., Mercader, J. M., Gaulton, K., Flannick, J., Getz, G. & Florez, J. C. Type 2 diabetes genetic loci informed by multi-trait associations point to disease mechanisms and subtypes: A soft clustering analysis. *PLOS Medicine* **15** (ed Langenberg, C.) e1002654 (2018).

162. Ahlqvist, E., Storm, P., Käräjämäki, A., Martinell, M., Dorkhan, M., Carlsson, A., Vikman, P., Prasad, R. B., Aly, D. M., Almgren, P., Wessman, Y., Shaat, N., Spégel, P., Mulder, H., Lindholm, E., Melander, O., Hansson, O., Malmqvist, U., Lernmark, Å., Lahti, K., Forsén, T., Tuomi, T., Rosengren, A. H. & Groop, L. Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. *The Lancet Diabetes and Endocrinology* **6,** 361–369 (2018).

163. American Diabetes Association. Classification and diagnosis of diabetes. Sec. 2. In Standards of Medical Care in Diabetes. *Diabetes Care* **38,** S8–S16 (2015).

164. Davies, M. J., D'Alessio, D. A., Fradkin, J., Kernan, W. N., Mathieu, C., Mingrone, G., Rossing, P., Tsapas, A., Wexler, D. J. & Buse, J. B. Management of hyperglycaemia in type 2 diabetes, 2018. A consensus report by the American Dia-

betes Association (ADA) and the European Association for the Study of Diabetes (EASD). *Diabetologia* **61,** 2461–2498 (2018).

165. American Diabetes Association. 6. Glycemic Targets: Standards of Medical Care in Diabetes—2019. *Diabetes Care* **42,** S61 LP –S70. http://care.diabetesjournals.org/content/42/Supplement%7B%5C_%7D1/S61 (2019).

166. Reusch, J. E. B. & Manson, J. E. Management of Type 2 Diabetes in 2017: Getting to Goal. *JAMA: Journal of the American Medical Association* **317,** 1015–1016 (2017).

167. Inzucchi, S. E., Tunceli, K., Qiu, Y., Rajpathak, S., Brodovicz, K. G., Engel, S. S., Mavros, P., Radican, L., Brudi, P., Li, Z., Fan, C. P. S., Hanna, B., Tang, J. & Blonde, L. Progression to insulin therapy among patients with type 2 diabetes treated with sitagliptin or sulphonylurea plus metformin dual therapy. *Diabetes, Obesity and Metabolism* **17,** 956–964 (2015).

168. Khunti, K., Wolden, M. L., Thorsted, B. L., Andersen, M. & Davies, M. J. Clinical inertia in people with type 2 diabetes: A retrospective cohort study of more than 80,000 people. *Diabetes Care* **36,** 3411–3417 (2013).

169. Khunti, K. & Millar-Jones, D. Clinical inertia to insulin initiation and intensification in the UK: A focused literature review. *Primary Care Diabetes* **11,** 3–12 (2017).

170. Nielsen, A. M., Nielsen, R. L., Donnelly, L., Zhou, K., Dahl, A. B., Gupta, R., Ersbøll, B. K., Pearson, E. & Clemmensen, L. *A Comparison of Methods for Disease Progression Prediction Through a GoDARTS Study* PhD thesis (Technical University of Denmark (DTU), 2018), 1–23.

171. Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S. & Ebert, B. L. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)* **102,** 15545–15550 (2005).

172. Eden, E., Navon, R., Steinfeld, I., Lipson, D. & Yakhini, Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* **10,** 1–7 (2009).

173. Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N. T., Morris, J. H., Bork, P., Jensen, L. J. & Mering, C. V. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research* **47,** D607–D613 (2019).

174. Pui, C. H., Robison, L. L. & Look, A. T. Acute lymphoblastic leukaemia. *Lancet* **371,** 1030–1043 (2008).

175. Terwilliger, T. & Abdul-Hay, M. Acute lymphoblastic leukemia: a comprehensive review and 2017 update. *Blood Cancer Journal* **7,** e577 (2017).

176. Chiaretti, S., Zini, G. & Bassan, R. Diagnosis and Subclassification of Acute Lymphoblastic Leukemia. *Mediterr J Hematol Infect Dis* **6,** e2014073 (2014).

177. Hjalgrim, L. L., Rostgaard, K., Schmiegelow, K., Söderhäll, S., Kolmannskog, S., Vettenranta, K., Kristinsson, J., Clausen, N., Melbye, M., Hjalgrim, H. & Gustafsson, G. Age- and Sex-Specific Incidence of Childhood Leukemia by Immunophenotype in the Nordic Countries. *JCNI: Journal of the National Cancer Institute* **95,** 1539–1544 (2003).

178. Hunger, S. P., Loh, M. L., Whitlock, J. A., Winick, N. J., Carroll, W. L., Devidas, M. & Raetz, E. A. Children's Oncology Group's 2013 blueprint for research: acute lymphoblastic leukemia. *Pediatr Blood Cancer* **60,** 957–963 (2013).

179. Spector, L. G., Ross, J. A., Robison, L. L. & Bhatia, S. in *Childhood Leukemias* (ed Pui, C.-H.) 48–66 (Cambridge University Press, 2006).

180. Vijayakrishnan, J., Studd, J., Broderick, P., Kinnersley, B., Holroyd, A., Law, P. J., Kumar, R., Allan, J. M., Harrison, C. J., Moorman, A. V., Vora, A., Roman, E., Rachakonda, S., Kinsey, S. E., Sheridan, E., Thompson, P. D., Irving, J. A., Koehler, R., Hoffmann, P., Nöthen, M. M., Heilmann-Heimbach, S., Jöckel, K.-H., Easton, D. F., Pharaoh, P. D. P., Dunning, A. M., Peto, J., Canzian, F., Swerdlow, A., Eeles, R. A., Kote-Jarai, Z., Muir, K., Pashayan, N., Henderson, B. E., Haiman, C. A., Benlloch, S., Schumacher, F. R., Olama, A. A. A., Berndt, S. I., Conti, D. V., Wiklund, F., Chanock, S., Stevens, V. L., Tangen, C. M., Batra, J., Clements, J., Gronberg, H., Schleutker, J., Albanes, D., Weinstein, S., Wolk, A., West, C., Mucci, L., Cancel-Tassin, G., Koutros, S., Sorensen, K. D., Maehle, L., Neal, D. E., Travis, R. C., Hamilton, R. J., Ingles, S. A., Rosenstein, B., Lu, Y.-J., Giles, G. G., Kibel, A. S., Vega, A., Kogevinas, M., Penney, K. L., Park, J. Y., Stanford, J. L., Cybulski, C., Nordestgaard, B. G., Brenner, H., Maier, C., Kim, J., John, E. M., Teixeira, M. R., Neuhausen, S. L., De Ruyck, K., Razack, A., Newcomb, L. F., Lessel, D., Kaneva, R., Usmani, N., Claessens, F., Townsend, P. A., Gago-Dominguez, M., Roobol, M. J., Menegaux, F., Greaves, M., Zimmerman, M., Bartram, C. R., Schrappe, M., Stanulla, M., Hemminki, K., Houlston, R. S. & Consortium, T. P. Genome-wide association study identifies susceptibility loci for B-cell childhood acute lymphoblastic leukemia. *Nature Communications* **9,** 1340 (2018).

181. Pieters, R., Hunger, S. P., Boos, J., Rizzari, C., Silverman, L., Baruchel, A., Goekbuget, N., Schrappe, M. & Pui, C. H. L-asparaginase treatment in acute lymphoblastic leukemia: a focus on Erwinia asparaginase. *Cancer* **117,** 238–249 (2011).

182. Schmiegelow, K., Attarbaschi, A., Barzilai, S., Escherich, G., Frandsen, T. L., Halsey, C., Hough, R., Jeha, S., Kato, M., Liang, D.-C., Mikkelsen, T. S., Möricke, A., Niinimäki, R., Piette, C., Putti, M. C., Raetz, E., Silverman, L. B., Skinner, R., Tuckuviene, R., van der Sluis, I. & Zapotocka, E. Consensus definitions of 14 severe acute toxic effects for childhood lymphoblastic leukaemia treatment: a Delphi consensus. *The Lancet oncology* **17,** e231–e239 (2016).

183. Schmiegelow, K., Müller, K., Mogensen, S. S., Mogensen, P. R., Wolthers, B. O., Stoltze, U. K., Tuckuviene, R. & Frandsen, T. Non-infectious chemotherapy-associated acute toxicities during childhood acute lymphoblastic leukemia therapy. *F1000Research* **6,** 444 (2017).

184. Frandsen, T. L., Heyman, M., Abrahamsson, J., Vettenranta, K., Åsberg, A., Vaitkeviciene, G., Pruunsild, K., Toft, N., Birgens, H., Hallböök, H., Quist-Paulsen, P., Griškevičius, L., Helt, L., Hansen, B. V. & Schmiegelow, K. Complying with the European Clinical Trials directive while surviving the administrative pressure – An alternative approach to toxicity registration in a cancer trial. *European Journal of Cancer* **50,** 251–259 (2014).

185. Yeh, J. M., Ward, Z. J., Chaudhry, A., Liu, Q., Yasui, Y., Armstrong, G. T., Gibson, T. M., Howell, R., Hudson, M. M., Krull, K. R., Leisenring, W. M., Oeffinger, K. C. & Diller, L. Life Expectancy of Adult Survivors of Childhood Cancer Over 3 Decades. *JAMA Oncology.* eprint: `https://jamanetwork.com/journals/jamaoncology/articlepdf/2757844/jamaoncology\_yeh\_2020\_oi\_190102.pdf` (2020).

186. Müller, H. J. & Boos, J. Use of L-asparaginase in childhood ALL. *Critical Reviews in Oncology/Hematology* **28,** 97–113 (1998).

187. Liu, C., Yang, W., Devidas, M., Cheng, C., Pei, D., Smith, C., Carroll, W. L., Raetz, E. A., Bowman, P. W., Larsen, E. C., Maloney, K. W., Martin, P. L., Mattano, L. A., Winick, N. J., Mardis, E. R., Fulton, R. S., Bhojwani, D., Howard, S. C., Jeha, S., Pui, C. H., Hunger, S. P., Evans, W. E., Loh, M. L. & Relling, M. V. Clinical and genetic risk factors for acute pancreatitis in patients with acute lymphoblastic leukemia. *Journal of Clinical Oncology* **34,** 2133–2140 (2016).

188. Wolthers, B. O., Frandsen, T. L., Baruchel, A., Attarbaschi, A., Barzilai, S., Colombini, A. & Escherich, G. Asparaginase-associated pancreatitis in childhood acute lymphoblastic leukaemia: an observational Ponte di Legno Toxicity Working Group study. *Lancet Oncology* **18,** 1238–1248 (2017).

189. Gupta, S., Wang, C., Raetz, E. A., Schore, R. J., Salzer, W. L., Larsen, E., Maloney, K. W., Mattano, L. A., Carroll, W. L., Winick, N. J., Hunger, S., Loh, M. L. & Devidas, M. Impact of asparaginase discontinuation on outcome in childhood ALL: A report from the Children's Oncology Group (COG). *Journal of Clinical Oncology* **37** (2019).

190. Raja, R. A., Schmiegelow, K. & Frandsen, T. L. Asparaginase-associated pancreatitis in children. *British Journal of Haematology* **159,** 18–27 (2012).

191. Abaji, R., Gagné, V., Xu, C. J., Spinella, J.-F., Ceppi, F., Laverdière, C., Leclerc, J. M., Sallan, S. E., Neuberg, D., Kutok, J. L., Silverman, L. B., Sinnett, D. & Krajinovic, M. Whole-exome sequencing identified genetic risk factors for asparaginase-related complications in childhood ALL patients. *Oncotarget* **8,** 43752–43767 (2017).

192. Wolthers, B. O., Frandsen, T. L., Patel, C. J., Abaji, R., Attarbaschi, A., Barzilai, S., Colombini, A., Escherich, G., Grosjean, M., Krajinovic, M., Larsen, E., Liang, D., Möricke, A., Rasmussen, K. K., Samarasinghe, S., Silverman, L. B., van der Sluis, I. M., Stanulla, M., Tulstrup, M., Yadav, R., Yang, W., Zapotocka, E., Gupta, R. & Schmiegelow, K. Trypsin-encoding PRSS1-PRSS2 variations influence the risk of asparaginase-associated pancreatitis in children with acute lymphoblastic leukemia: a Ponte di Legno toxicity working group reporr. *Haematologica* **104,** 556–563 (2019).

193. Zator, Z. & Whitcomb, D. C. Insights into the genetic risk factors for the development of pancreatic disease. *Therapeutic Advances in Gastroenterology* **10,** 323–336 (2017).

194. Shimabukuro, D. W., Barton, C. W., Feldman, M. D., Mataraso, S. J. & Das, R. Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial. *BMJ Open Respiratory Research* **4,** e000234 (2017).

195. Wang, P., Berzin, T. M., Glissen Brown, J. R., Bharadwaj, S., Becq, A., Xiao, X., Liu, P., Li, L., Song, Y., Zhang, D., Li, Y., Xu, G., Tu, M. & Liu, X. Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. *Gut* **68,** 1813–1819 (2019).

196. Ginsburg, G. S. & Phillips, K. A. Precision Medicine: From Science To Value. *Health Aff (Millwood)* **37,** 694–701 (2018).

197. Dias, R. & Torkamani, A. Artificial intelligence in clinical and genomic diagnostics. *Genome Medicine* **11,** 70 (2019).

198. Razavian, N., Marcus, J. & Sontag, D. *Multi-task Prediction of Disease Onsets from Longitudinal Laboratory Tests* in *Proceedings of the 1st Machine Learning for Healthcare Conference, PMLR* (PMLR, 2016), 73–100.