

# VU Research Portal

## POLYGENIC RISK PREDICTION OF COMMON DISEASES

Martens, Forike Kirsten

2021

### **document version**

Publisher's PDF, also known as Version of record

[Link to publication in VU Research Portal](#)

### **citation for published version (APA)**

Martens, F. K. (2021). *POLYGENIC RISK PREDICTION OF COMMON DISEASES: Design, evaluation and interpretation of prediction studies*. [PhD-Thesis - Research and graduation internal, Vrije Universiteit Amsterdam]. s.n.

### **General rights**

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

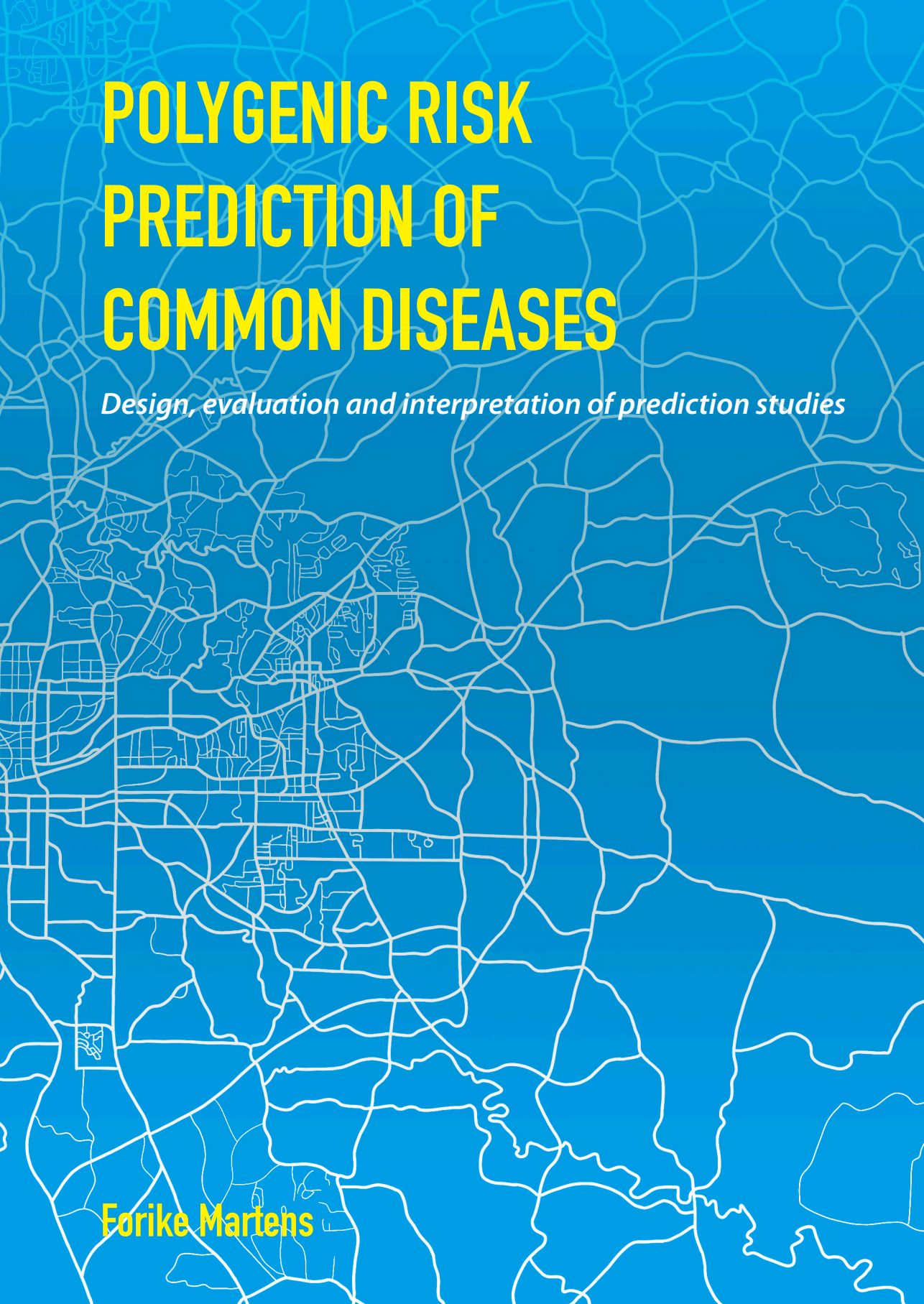
- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal ?

### **Take down policy**

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

### **E-mail address:**

[vuresearchportal.ub@vu.nl](mailto:vuresearchportal.ub@vu.nl)



# POLYGENIC RISK PREDICTION OF COMMON DISEASES

*Design, evaluation and interpretation of prediction studies*

**Forike Martens**



# **Polygenic Risk Prediction of Common Diseases**

Design, evaluation and interpretation of prediction studies

**Forike Kirsten Martens**

## **Colofon**

Polygenic risk prediction of common diseases: design, evaluation and interpretation of prediction studies

Martens, F.K.

Copyright © 2021 Forike K. Martens

All rights reserved. No part of this publication may be reproduced or transmitted in any form or by any means, without prior permission of the author or the copyright-owning journals for previously published chapters.

ISBN: 978-94-6421-563-2

Cover design by: Forike Martens | Layout by: Wendy Bour-van Telgen

Printed by: Ipskamp Printing, Enschede, the Netherlands

The printing of this thesis was kindly supported by the Department of Human Genetics at Amsterdam UMC.

VRIJE UNIVERSITEIT

# **Polygenic Risk Prediction of Common Diseases**

Design, evaluation and interpretation of prediction studies

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor aan  
de Vrije Universiteit Amsterdam,  
op gezag van de rector magnificus  
prof.dr. C.M. van Praag,  
in het openbaar te verdedigen  
ten overstaan van de promotiecommissie  
van de Faculteit der Geneeskunde  
op maandag 20 december 2021 om 9.45 uur  
in een bijeenkomst van de universiteit,  
De Boelelaan 1105

door

Forike Kirsten Martens

geboren te Nuenen, Gerwen en Nederwetten

promotoren: prof.dr. M.C. Cornel  
prof.dr. A.C.J.W. Janssens

copromotor: dr. E.C.M. Tonk

promotiecommissie: prof.dr. L. Henneman  
prof.dr. C.R. Bezzina  
prof.dr.ir. M.K. Schmidt  
prof.dr. E.W. Steyerberg  
prof.dr. D.R.M. Timmermans

# Contents

|                  |   |     |
|------------------|---|-----|
| <b>Chapter 1</b> | General introduction  | 9   |
| <b>Chapter 2</b> | How the intended use of polygenic risk scores guides the design and evaluation of prediction studies.<br><i>Curr Epidemiol Rep. 2019;6(2):184-90</i>  | 35  |
| <b>Chapter 3</b> | Reflection on modern methods: Revisiting the area under the ROC Curve.<br><i>Int J Epidemiol. 2020;49(4):1397-1403</i>  | 51  |
| <b>Chapter 4</b> | Small improvement in the area under the receiver operating characteristic curve indicated small changes in predicted risks.<br><i>J Clin Epidemiol. 2016;79:159-64</i>  | 65  |
| <b>Chapter 5</b> | Evaluation of polygenic risk models using multiple performance measures: a critical assessment of discordant results.<br><i>Genet Med. 2019;21(2):391-7</i>   | 89  |
| <b>Chapter 6</b> | Dealing with discordant findings from multiple performance metrics in clinical prediction studies.<br><i>Prepared for submission</i>  | 117 |
| <b>Chapter 7</b> | Letters to the editor<br>Based on:<br>Risk analysis of prostate cancer in PRACTICAL Consortium-Letter.<br><i>Cancer Epidem Biomar. 2016;25(1):222</i><br><br>External validation is only needed when prediction models are worth it.<br><i>J Clin Epidemiol. 2016;69:249-50</i> | 151 |



|                  |                        |     |
|------------------|------------------------|-----|
| <b>Chapter 8</b> | General Discussion     | 159 |
|                  | Summary   Samenvatting | 175 |
|                  | Dankwoord              | 190 |
|                  | Publications           | 192 |
|                  | About the author       | 193 |

Voor Opa



# 1

## General introduction

Most of us will encounter, at some point in life, the consequences of common diseases, be it personally or through a friend or family member. As human beings we have a tendency for wanting to know what our future beholds; hence the current interest in our DNA as predictive factor for disease is not surprising. In recent years, polygenic risk scores (PRS) were introduced in science, to summarize the effects of multiple variations in our DNA that are associated with the development of disease. Recently, some researchers have claimed that PRSs are ready for implementation in clinical practice, while others argue that the clinical usefulness has yet to be proven. This thesis focuses on the methodology that is used to evaluate the predictive performance of PRSs and other predictive tools, which is an essential step in the implementation of new health applications in practice. The introduction gives an overview of the progress in the field of risk prediction for common diseases, the evaluation of prediction models, current methodological challenges in the field of genetic risk prediction and concludes with the aims and scope of the research presented in this thesis.

### **Risk prediction for common diseases**

More than 70% of deaths globally are due to non-communicable diseases; also known as common chronic diseases (1). Among the leading causes of death are cardiovascular diseases, cancers, diabetes, chronic respiratory diseases and mental health conditions. Studying common diseases improves insight into their causes and possible preventive and therapeutic interventions. The focus of preventing common diseases is often on reducing the associated risk factors (1). Risk factors can also be used for the prediction of common diseases. Predictors can be causally related to the disease, such as risk factors typically are, but do not necessarily have to be. They are incorporated in statistical models to predict occurrence of disease, are used in differential diagnosis for patients, and are used to predict outcomes after diagnosis. This means that in healthcare, prediction models can be used to identify at-risk groups for preventive interventions, support physicians in medical decision making, and inform individuals about their risk or progression of the disease, to ultimately improve patients' health and decrease the number of deaths. For example, in young adults a prediction model could predict the 10-year risk of type 2 diabetes to stratify prevention with a supervised exercise program for the high-risk group. In this thesis I focus on the prediction of disease.

## **Personalized medicine**

More accurately assessed risks are desirable for optimizing decision making to provide the best care for each patient. Personalized-, precision- or sometimes called stratified medicine was introduced to individualize care and move away from a 'one size fits all' approach. Although the term 'personalized medicine' is relatively new, tailoring treatments and care to the individual patient is not a new approach at all, it dates back to the Greek physician Hippocrates who stated that "it is more important to know what sort of person has a disease than to know what sort of disease a person has" and later to one of the founders of the Johns Hopkins Hospital, the Canadian physician Sir William Osler, who recognized that "variability is the law of life, and as no two faces are the same, so no two bodies are alike, and no two individuals react alike and behave alike under the abnormal conditions we know as disease" and warned to "Care more particularly for the individual patient than for the special features of the disease" (2,3). Today, personalized medicine means a healthcare approach in which interventions are targeted to the individual or to subgroups rather than to the population at large by considering individual variability in genes, the environment and people's lifestyle. This is especially warranted when an intervention cannot be given to the target population at large because health care budgets are scarce or because the intervention is not beneficial for all individuals from the target population.

## **Architecture of common diseases**

Common diseases are often caused by a complex interplay between multiple genetic and nongenetic factors, such as environmental and lifestyle factors. In common diseases there is no straight link between common genetic variants and the development of disease, because for most variants the pathophysiological mechanisms have not yet been identified and the variants that have been identified are often only statistically associated to the disease. The heritability of common diseases, the proportion of phenotypic variation that is attributed to genetic variation, is for many of these diseases estimated to be moderate to high, yet only a very small amount of the genetic variants has been unraveled (4). Apart from rare mutations and copy number variations, most of the genetic contribution to common diseases that has been unraveled appears to reflect the effect of many common single nucleotide polymorphisms (SNPs) that have individually small effects. SNPs are variations occurring at a single nucleotide of

the genome (adenine, thymine, cytosine, or guanine, denoted by the letters A, T, C, or G) that are present in >1% of the general population. SNPs are biological markers and may occur in coding or non-coding regions of the DNA, which means that they may or may not play a direct role in disease.

### **Gene discovery**

Over the past decade, genome-wide association studies (GWASs) have identified many SNPs, that are robustly associated with risk of common diseases, including type 2 diabetes, cardiovascular disease, cancer, psychiatric disorders and many other diseases (3–8). Among the primary aims of the study of these SNPs is to improve understanding of the genetic architecture of common diseases, elucidate the role of relevant biological pathways, and implicate novel therapeutic targets (4). But, the study of SNPs has also fueled expectations that, one day, genetic testing can be used to predict risk of common diseases, their prognosis, and the response to treatment. The individual effects of SNPs are small; in the early days of GWAS they reflected odds ratios often close to 1.1<sup>1</sup> (3–5), and only a small fraction of the heritability could be explained (9). Initially it was hoped that when many GWAS samples were collected, a larger fraction of the heritability could be explained. Although these samples were collected, and many more SNPs were found with even smaller effects, the fraction of heritability that could be explained remained small. For a while, this suggested that a realization of the expectations of SNPs for the prediction of common diseases would never be possible. Until a few years ago, when several developments led to a resurgence of interest in using SNPs for the prediction of common diseases (5).

### **Polygenic risk scores**

For common diseases it is agreed that SNPs with small effects will have no useful predictive value on their own, therefore multiple SNPs are combined into one score, frequently referred to as polygenic risk score (PRS) or genetic risk score (GRS). PRSs quantify the combined contribution of multiple SNPs to the risk of common diseases. The scores are calculated by 1) multiplying the number of risk alleles of a SNP (0, 1 or 2 alleles) with the effect of the variant on the risk of disease, and then 2) adding these products. Although the concept of a sum score was proposed in the early 20th century (6), in the start of the 21st century

---

<sup>1</sup> Individual with risk allele 'X' is 1.1 times more likely to develop disease 'Y' than individual without risk allele 'X'.

PRSs were used as a solution to include larger numbers of genetic variants in the prediction model instead of adding each identified SNP as separate variable. Recently the construction of PRSs has become more advanced by the introduction of new statistical methods such as LDpred (7), which contributed to the renewed interest in utilizing SNPs to predict common diseases (8). Whereas earlier studies included only SNPs that passed the genome-wide significance line ( $P < 5 \cdot 10^{-8}$ ) (9), later studies allowed PRSs to also include variants that are below the traditional significance line. This has resulted in PRSs made up of millions of SNPs (10–12) and it is argued that these may be used to estimate an individual's genetic risk of disease and identify groups that are most at risk and may benefit most from preventive interventions (13). Another factor contributing to the resurgence of PRSs for common diseases is recently published articles (11, 14), including the publication in Nature Genetics by Khera et al., who claimed that the PRSs identified individuals with “risk equivalent to monogenic mutations” (11). The studies were enthusiastically received in the field of genomic medicine, but received criticism as well (5).

## **Predictors**

Despite some have argued that PRSs have shown to be promising for future clinical applications for common diseases including breast cancer (15), diabetes and cardiovascular disease (11), PRSs on their own often have low predictive ability. Because the factors that contribute to the development of common diseases are multifactorial, also other predictors are relevant to consider in the prediction of these diseases, such as lifestyle, demographics, family history, and biomarkers. PRSs are often added to prediction models containing multiple of those predictors to improve the predictive ability of the model. In this thesis I consider clinical or genetic models that provide risk predictions for a dichotomous outcome (event vs. no event), since these are most commonly used in prediction studies.

## **Clinical application of PRSs**

Due to the recent developments in genetics some researchers are now, as mentioned earlier, convinced that PRSs could be a promising personalized application for common diseases (11, 15). The ongoing interest in this application of PRSs has also been fueled by the expanding offer of genetic tests by direct-to-consumer (DTC) companies as 23andMe, Helix, DNAfit, Ancestry, and



MyHeritage. Whilst some of the consumers of these genetic tests are initially interested in their ancestry, some people use their retrieved genetic data for health (16). Another boost to the provision of PRS by DTC companies was given in 2017 by the U.S. Food and Drug Administration, by allowing DTC companies to offer genetic test for disease prediction directly to consumers (17). There are now several companies who directly offer PRSs, for example a PRS for type 2 diabetes by 23andMe and a PRS for heart disease by MyHeritage (18,19). Next, I describe some examples of possible future clinical application of PRSs that have been suggested in scientific literature as such, including applications for breast cancer and other cancers, cardiovascular disease, and psychiatric disorders.

Today, apart from population screening for all women in a certain age category (e.g. 50-75y), the risk assessment of breast cancer to determine the optimal strategy for surveillance is mainly focused on clinical risk factors and high-penetrant genetic risk factors such as BRCA1 and BRCA2. Carriers of the BRCA1 and BRCA2 genetic variants, and women with a breast cancer family history in general, have a higher risk of breast cancer compared to non-carriers or women without family history, to such an extent that earlier and more frequent screening including MRI and mammography is proven to decrease breast cancer mortality (20). Moreover, prophylactic surgery is available for women in the highest risk stratum. In the past decade many common low risk genetic variants have been identified that together may be of clinical interest to improve prediction of breast cancer and hence the management of the disease (21). Combining traditional risk factors for the prediction of breast cancer with PRSs might improve the predictive ability (11,22) and influence the management of the disease. For instance, women with a high PRS for breast cancer could be advised to start breast cancer screening at a younger age or be under surveillance more frequently (23). The risk assessment of other cancers, such as prostate- and colorectal cancer, are also being investigated to include PRSs. For example, to incorporate the score into risk stratified population screening programs (24–28), to refine risk assessment for high-risk families (29,30) and improve prostate-specific antigen (PSA) testing in screening for prostate cancer (31,32). But, the emerging evidence is not entirely positive about the harms, benefits and costs of using PRSs for screening. For example, risk-stratified colon cancer screening is unlikely to be cost-effective in comparison with uniform screening (33), and although results of a modeling study suggest that a

breast cancer PRS combined with family history could have greater benefit than screening based on family history alone, more overdiagnoses and false positive results should be expected (34).

The risk assessment of cardiovascular diseases is traditionally done with well-known risk factors, such as age, sex, hypertension, smoking and obesity (35). In recent studies researchers suggested that the use of PRSs may improve the prediction of, for example, coronary heart disease (11,14,24). It is argued that improved predictive ability could support management of cardiovascular diseases, for example, by providing preventive therapies or lifestyle advice to individuals with a high PRS. The recent studies were covered positively by the media, however, the analysis that formed the basis for their conclusions about PRSs for the prediction of coronary heart disease were unconventional (36). The added value of PRSs has yet to be proven and the results of these studies will likely not hold up when conventional approaches are used. Previously it has been stated that especially for individuals with high PRSs, adherence to a healthy lifestyle was related to a significantly decreased risk of cardiovascular disease, but at the same time it was emphasized that a healthy lifestyle is recommended for everyone (37). The utility of a PRSs thus depends also on the availability of effective interventions for each risk group. When these are present, the PRS could possibly support physicians in decision making about, for example, additional statin treatment as preventive measure for cardiovascular disease.

Furthermore, the use of PRSs in psychiatric disorders (such as major depression, schizophrenia, bipolar disorder, psychosis and Alzheimer's disease) is being investigated, for example, to improve diagnosis, predict diagnostic and treatment outcomes (38–41). However, for most applications evidence of clinical validity and utility is currently lacking.

## **Evaluation of prediction models for common diseases**

### **From gene discovery to health application**

Successful and responsible implementation of new health applications requires the necessary research. The continuum of this translational research process starts with association studies, such as GWAS, from which the candidate predictors (SNPs) are selected, followed by prediction studies. Prediction studies

focus on the use of risk prediction in health care, by assessing the predictive performance and utility of the prediction model (e.g., a PRS or a clinical prediction model) in the intended setting. Once a new or updated prediction model is worth implementing in healthcare, several other types of studies that together should prepare the implementation and use of the model, should be conducted. Examples of these studies include risk communication studies, behavioral and psychological research, implementation and cost-effectiveness studies. For example, before safe implementation of a breast cancer PRS can be guaranteed, i.e., to rule out any adverse events, questions about how it may change the patient's perception of risk need to be answered.

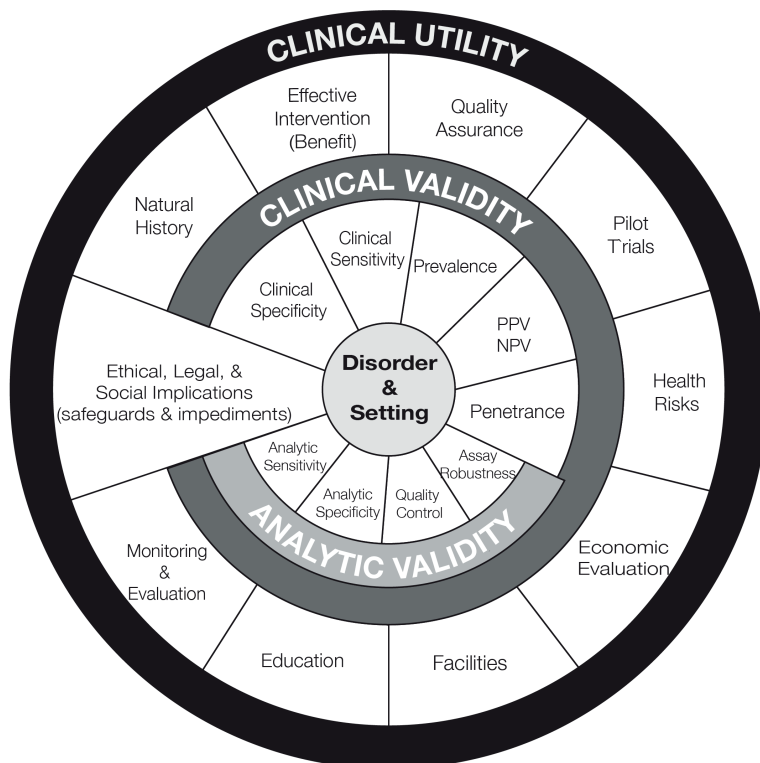
From 2000 to 2004, the CDC's office of Public Health Genomics developed a framework, the ACCE model, for collecting, evaluating, interpreting, and reporting data about genetic tests in a format that could support policy makers in decision making. Meanwhile the framework has been expanded (42), however its main components form the basis of every analytical process that is followed to evaluate scientific data on emerging (genetic) applications. The ACCE model (Figure 1) is composed of a standard set of questions that address disease and clinical setting, the analytical validity, clinical validity, clinical utility and associated ethical, legal and social issues (ELSI) (43). The clinical setting refers to the intended use of the prediction model; analytical validity to how well the model performs in the laboratory (addressed in association studies); the clinical validity to how well the model performs in the clinical setting; the clinical utility to how useful the model is; and the ELSIs to the ethical, legal and social implications of the prediction model. Clinical validity is typically addressed in prediction studies.

### **Designing prediction studies**

The main objective of prediction studies is to determine the probability of an outcome with a set of predictors in a population (44). Prediction studies address the development of a prediction model, the validation of a prediction model or both. A development study selects predictors, estimates relative weights and assesses the model performance. A validation study re-applies a prediction model in another population using the same relative weights to reassess the model performance.

When prediction models are foreseen to be implemented in healthcare it is important that the prediction study is designed with the intended use in

mind. The healthcare scenario specifies what needs to be predicted, in whom, how and for what purpose. This means that the intended use informs what the outcome, study sample, and predictors need to be in the prediction study. All these, and other key elements such as the study design, statistical model, and statistical analysis should be well defined beforehand and reported following existing guidelines (45–48), such as ‘The Genetic Risk Prediction Studies’ (GRIPS) statement, to maximize the transparency, quality, and completeness of reporting on the research methodology and findings in prediction studies. The reporting guidelines are in line with the set of standard questions accompanying the ACCE model.



**Figure 1.** ACCE Model for the evaluation of genetic testing. CDC’s Office of Public Health Genomics supported the establishment of the first publicly available analytical model for evaluating scientific data on emerging genetic tests. ACCE stands for the main criteria for evaluating genetic tests: analytical validity, clinical validity, clinical utility, and ethical, legal and social implications. At the heart of the model is the ‘disorder and setting’, which refers to the intended use of the test. (source: <https://www.cdc.gov/genomics/gtesting/acce/index.htm>).

### ***Outcome, study population and selection of predictors***

Prediction studies should focus on outcomes that are clinically relevant to the stakeholders involved (providers, patients) and should include a risk period, for example, the 10-year risk of type 2 diabetes. The study population needs to be representative of the population in which the model will be used, the target population. The study sample includes a selection of the general population, a subgroup defined by, for example, age, gender, or the presence of certain risk factors. The best design to answer prediction questions is a cohort study, preferably a longitudinal prospective study as it allows to measure the outcome and predictors over time. Case-control studies are sometimes used, but as the design is not longitudinal it does not consider the risk period, and as participants are selected based on the presence or absence of disease, absolute risks cannot be calculated, hence, this design is not preferred for prediction research. The selection of predictors refers to the selection of candidate variables for inclusion in the prediction model, and is based on their association with the development of disease and with the intended use in mind. As indicated above, they include demographics, type/severity of disease, history characteristics, comorbidity, physical functional status, subjective health status, and genetic predisposition (49).

### ***Developing a prediction model***

Prediction models express the relation between the predictors in the model and the selected outcome. The most commonly used statistical models in empirical prediction studies are logistic regression and Cox proportional hazards regression. Other methods, such as machine learning, have been investigated, but it has been argued that these do not outperform traditional regression approaches (50). Before the development of the prediction model is started, several decisions need to be made concerning the selection of candidate predictors, the quality of data, missing data, outliers, data handling decisions, how to model continuous variables, studying possible interaction between predictors, and variable reduction (51).

### **Metrics to evaluate the clinical validity and clinical utility of genetic prediction models**

An important step in the translation of prediction models to useful applications in healthcare is the evaluation of the predictive performance. The ACCE model

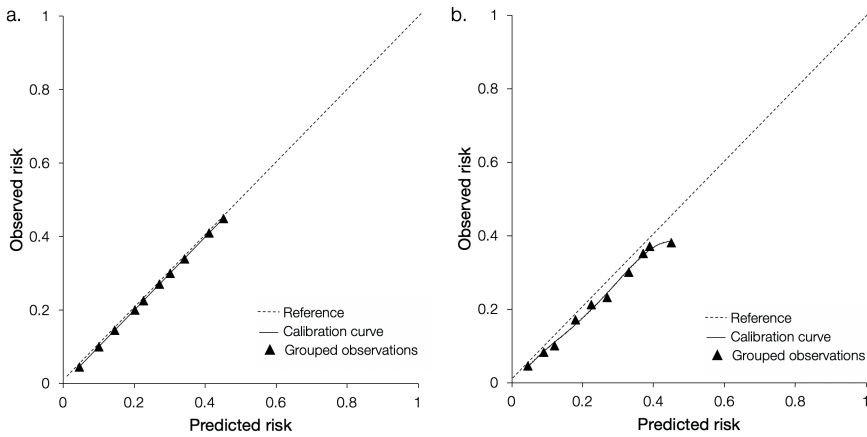
and GRIPS guideline both recommend the assessment of several metrics to evaluate the clinical validity and clinical utility of prediction models. These include the following metrics of model assessment and validation: model fit, calibration metrics, positive and negative predictive value, sensitivity, specificity, discriminative ability and reclassification metrics. Next, I discuss each of these metrics briefly.

### ***Model fit and calibration***

The goodness of fit of a genetic prediction model describes how well the model fits the observations; it indicates how likely the estimates from the prediction model would conform to the observed data. Goodness of fit is assessed by the Akaike Information Criterion (AIC) and calibration metrics. The AIC estimates the model fit, plus it takes into account the number of predictors in the model (52). The metric can be used to compare models; the model with the lower AIC has a favorable balance between fitting and overfitting the data. Calibration commonly refers to how well the predicted risks from the prediction model match the actual observed event rates (53) and is often graphically displayed in a calibration plot (Figure 2) (54). Calibration can be quantified by ‘calibration in the large’ and the Hosmer-Lemeshow test. Calibration in the large measures the difference between the average of all predicted risks and the average risk of disease in the study population. The Hosmer-Lemeshow test compares observed and expected outcomes within deciles of predicted risk. A disadvantage of the latter is its inability to detect substantial miscalibration in small samples and oversensitiveness to minor miscalibration in large samples (55).

### ***Clinical validity***

Clinical validity refers to how well the prediction model estimates risks and is indicated by the predictive ability and discriminative ability. Predictive ability refers to the variation in predicted risks and is indicated by the distribution of predicted risks and by the positive predictive value (PPV) and negative predictive value (NPV) at (possible) risk thresholds. A risk distribution refers to the frequencies of the predicted risks in the population. Higher predictive ability requires more variation in predicted risk. The PPV and NPV indicate, respectively, the risk of disease and 1-risk of disease for risk groups that are defined by a certain risk threshold (Figure 3). The PPV is the percentage of individuals with the event among all individuals who test positive. The NPV is the percentage of individuals that remain free of the event among those with a negative test.



**Figure 2.** Calibration plot

Predicted risks (x-axis) against the observed outcomes (y-axis) for groups defined by for example, deciles of predicted risks. Figure 2a shows the calibration curve of a well calibrated prediction model, i.e. the predicted and observed risks agree, yielding a calibration curve that follows a 45-degree line (slope = 1) (53). This suggests that the predicted risks are correct, for example, among patients with a predicted risk of 20% to develop breast cancer in 5 years, 2 out of 10 indeed develop breast cancer in 5 years. Figure 2b shows is poorly calibrated model. The deviations from the reference line indicate underestimations (or in other cases overestimations) of the predicted risks by the prediction model.

Discriminative ability indicates how well a prediction model can distinguish between patients and nonpatients. The discriminative ability is assessed by the area under the receiver operating characteristic curve (ROC curve; AUC) (56) and by the sensitivity and specificity for specific risk thresholds. Metrics of discrimination are best understood when the risk distribution is presented separately for patients and nonpatients (Figure 3). Sensitivity is the percentage of patients that test positive, and specificity is the percentage of nonpatients that test negative (Figure 3). Lowering the risk threshold, i.e., moving the risk threshold to the left in the figure, typically increases sensitivity and decreases specificity. Depending on the intended use of the model, the minimal or sufficient level of sensitivity and specificity is determined. There is no general level for what sensitivity and specificity is good or excellent. If they would both be 100%, the prediction model would not produce any false positive and false negative predictions. For the prediction of common diseases this is never seen, therefore, the required level of sensitivity and specificity is based on the percentage of false positive and false negative predictions that are considered

acceptable. Some applications require prediction models with a risk threshold that has high specificity with acceptable sensitivity. For example, for newborn screening on cystic fibrosis, a low percentage of false positives is desirable, to minimize unnecessary follow-up testing and negative effects of the false positive test results for infants and parents (57). Other applications (such as first tier screening tests) require high sensitivity with acceptable specificity, in order to maximize the detection of the disease and therefore the need for an as low as possible percentage of individuals who will receive false negative result. For again other applications (such as selection for invasive, irreversible procedures or the non-invasive prenatal test, NIPT) both sensitivity and specificity need to be very high. False negative and false positive results in the latter are to be avoided as the choices made based on these test results may have far reaching consequences, for example, termination of pregnancy.

The most well-known metric of discrimination for binary outcomes is the AUC (also seen as AUROC or c-statistic). The ROC curve is drawn in a ROC plot that presents sensitivity against 1-specificity (Figure 4). The curve connects the combinations of sensitivity and specificity for all possible risk thresholds. AUC is the magnitude of the area under this curve and is explained as the probability that a randomly chosen patient has a higher predicted risk than a randomly chosen nonpatient (56). Frequently, PRSs are evaluated for their ability to improve existing clinical prediction models, as PRS alone generally have lower predictive ability compared to clinical models. For example, AUCs of 0.61, 0.66, and 0.62 compared to 0.76, 0.73, and 0.71 for coronary artery disease, type 2 diabetes, and breast cancer, respectively (58–60). When a PRS is added to a clinical model, the difference in AUC, denoted as  $\Delta$ AUC, between the clinical model and the updated model with a PRS is used to assess the improvement in discriminative ability (61,62). The increment in AUC from PRSs is generally low, often below 0.02 (24,63,64).

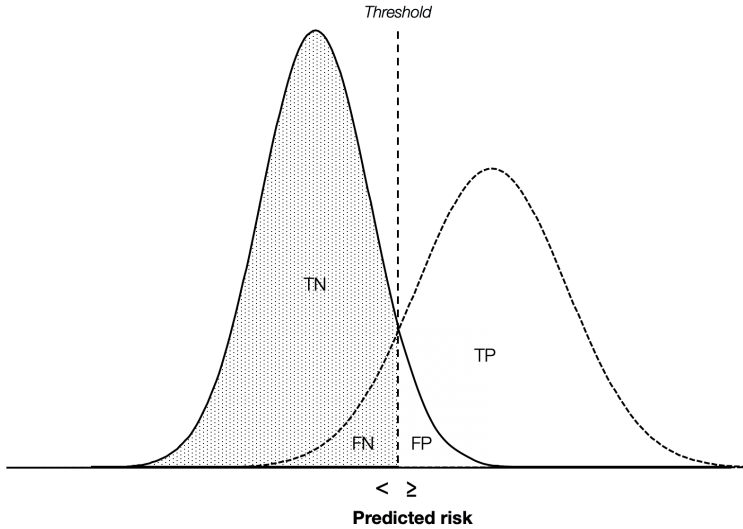
There are many other ways of expressing the predictive ability and discriminative ability, but these are less frequently used. For example, box plots can be used to show the means and distributions of risks in patients and nonpatients, and the difference in means is known as the discrimination slope. When a prediction model is extended by adding predictors, the integrated discrimination improvement (IDI) assesses the improvement in the discrimination slope. IDI is calculated by taking the difference of the risk difference between patients and nonpatients for the initial and extended models (65).



a.

| Predicted risk |               | Events                                  | Nonevents                               | Total   |
|----------------|---------------|---|---|---|
| Test           | + ≥ threshold | true positive (TP)                      | false positive (FP)                     | TP+FP <b>PPV</b> Positive predictive value = $TP/(TP+FP)$ |
|                | - < threshold | false negative (FN)                     | true negative (TN)                      | FN+TN <b>NPV</b> Negative predictive value = $TN/(FN+TN)$ |
| Total          |               | TP+FN                                   | FP+TN                                   | 100%  |
|                |               | <b>Se</b><br>Sensitivity = $TP/(TP+FN)$ | <b>Sp</b><br>Specificity = $TN/(FP+TN)$ |   |

b.

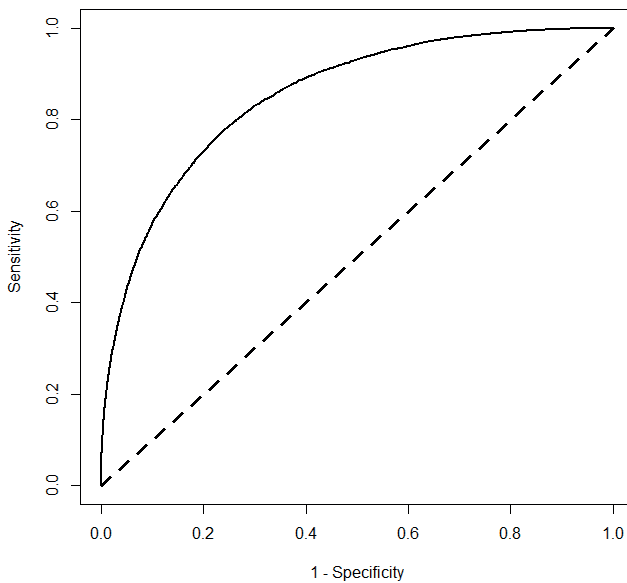


**Figure 3.** From risk distributions to a contingency table for a certain risk threshold. The threshold determines whether a positive or negative test result is reported; moving the threshold to the left means that more individuals have a predicted risk above the threshold and hence test positive. Positive predictive value (PPV), negative predictive value (NPV), sensitivity (se) and specificity (sp) are metrics to calculate the clinical validity of the test. Sensitivity and specificity indicate the test's ability to detect the presence of disease in people with the disease and its absence in those without. Positive and negative predictive values represent the probability of having the disease when the test result is positive and the probability of not having the disease when the result is negative. Figure 3a shows a 2 by 2 contingency table from which all metrics can be calculated. Figure 3b shows how the same percentages are calculated from the separate risk distributions of patients and nonpatients.

**Validation**

The model fit and performance (clinical validity) of genetic prediction models tend to be highest in the population that was used to develop the model (66).

Validation of a prediction model in the target population is therefore warranted and refers to the re-assessment of the prediction model. This can be done using other data within the same population (internal validation) or in an independent population (external validation).



**Figure 4.** Receiver operating characteristic curve. In general, specificity and sensitivity are related. Choosing a risk threshold with higher sensitivity, comes at the cost of lower specificity and vice versa. To choose a threshold for reporting a result as positive or negative, a balance between the two has to be achieved, which depends on the intended use of the test. A test performs better, when the ROC curve is placed more to the left and upper lines of the figure, i.e. has a bigger area under the curve.

Internal validation uses a subset of the same population that was used for the development of the prediction model. There are several methods for internal validation, including a split sample method and bootstrap sample method (53). Internal validation is a good first step to preventing overoptimistic interpretations about the predictive ability, but it is not sufficient. Internal validation still tends to give an optimistic predictive accuracy because the data in the development and validation study come from the same population, collected by the same methods and researchers, using the same variable definitions. Temporal validation concerns the validation of the model on individuals of a cohort at another point

in time, for instance, if the prediction model is developed on patients that are treated between 2007 and 2014, the validation of the model could be executed in patients that are treated at the same hospital, but between 2014 and 2021. Temporal validation is sometimes regarded as a validation approach that lies between internal- and external validation (67).

External validation is done to investigate the generalizability of the prediction model by evaluating its performance in a different, but similar target population. When a prediction model is constructed, its regression coefficients are estimated in such a way that the model best fits the data. External validation evaluates whether the model is robust to changes in the population and measurement of variables. Even with the same inclusion criteria for the study population and the same definition and measurement of variables, differences occur when other researchers collect the data at a different location. External validation may concern validation of the model in a different region or country (geographical validation) and fully independent external validation, which means that the performance of the model is tested in data collected by independent researchers, generally at a different site (68). Generally, the predictive ability and discriminative ability are lower when assessed in an independent population. For example, the discriminative ability (AUC) of the PREDICT model for breast cancer was 0.82 at development stage and 0.72 when validated (69). Ethnicity also plays a role in the predictive ability of the model, due to possible differences in genetic make-up between populations. PRSs developed from data of Caucasian individuals may not hold up for other ethnicities, for instance, risk estimates for individuals of European ancestry were less accurate in individuals of African ancestry (39,70). Approximately 80% of GWAS is conducted in populations of European ancestry (70), which unfortunately means that PRSs based on these studies are less accurate for other ethnic populations and more research should be focused on non-Caucasian populations (29,71).

### ***Changing risk category***

The clinical utility of a prediction model provides information about the usefulness of the prediction model in health care, usually referring to the ability of the model to improve health outcomes (72). When considering to add a prediction model to an existing health care service, a first step in the evaluation of clinical utility is the assessment of reclassification metrics and proof of positive reclassification is a prerequisite for clinical utility. The rationale for the use of reclassification statistics

is that updating a prediction model by adding new predictors is only warranted when people change between risk categories and therefore receive different health care, for instance in terms of surveillance or prophylactic medication. When updating a prediction model changes predicted risks but people will still be classified in the same risk category, the new model does not lead to different decisions about treatment. There are several different metrics of reclassification that quantify how many people change between risk categories. These metrics are calculated from a reclassification table and include total reclassification and the net reclassification improvement. Total reclassification calculates the percentage of individuals who change risk category, in any direction (73). This metric does not consider whether individuals move in the 'right' direction and its use has therefore been discouraged. The net reclassification improvement (NRI) does consider whether people move in the right direction (65). For patients and nonpatients separately, it counts the 'net' good moves (good moves – bad moves), meaning that more people move in the right than wrong direction. Good moves mean that patients move to a higher risk category and nonpatients to a lower; bad moves are vice versa. NRI is the sum of the percentage of net good moves in patients and nonpatients. Because the bases for these percentages are different when the number of patients and nonpatients not equal, the metric is generally not easy to interpret and is advised to report separately for patients and nonpatients (74).

## **Challenges in genetic risk prediction studies**

During the past decades, the field of genetic risk prediction for common diseases has developed rapidly. Upfront genotyping costs have dropped and several researchers have expressed the readiness of PRSs for implementation in the field, however, the field still faces many challenges before PRSs can successfully be implemented in practice. Several of the challenges concerning the design and evaluation of prediction studies and metrics are addressed in this thesis.

### **The rationale behind developing genetic prediction studies**

With the ongoing discovery of genetic variants that are statistically significantly associated with common diseases, it is expected that even more prediction

models will be developed, and existing models updated. For research this means that empirical studies are needed to investigate the predictive performance of the (updated) genetic prediction models. Preferably these are conducted in prospective cohort studies, with a study population that is unselected for the outcome of interest and the outcome measured over time. The study population, predictors and the outcome are determined by the intended use of the prediction model. Without taking into account the intended use for the development of prediction models, models may or may not be clinically relevant and of interest to the public. Furthermore, understanding whether the predictive performance of the model is high enough becomes hard when the purpose of using the prediction model is unknown. As conducting prospective studies can be time consuming and expensive, researchers often rely on readily available data from convenience cohorts. These cohorts were developed for studying epidemiological questions, not prediction. When these cohorts are used for prediction studies, they rely on data from the same sample as used for the discovery of genetic variants or PRSs, or publicly made available datasets such as from the UK biobank cohort. Performing the prediction analyses on these data means that the predictive performance of the models in the target population remains unknown. Reporting guidelines and frameworks mention the importance of describing the key information that is relevant for the interpretation and validation of genetic prediction models (43,75), but an explanation of why these are important and how the intended use of a prediction model determines design choices and informs interpretations of genetic prediction studies is lacking.

### **The area under the receiver operating characteristic curve**

There is a general agreement among researchers that methodological and reporting standards for prediction studies are often not met and should be improved (46,76–81). For example, the interpretation of one of the most commonly used metrics of predictive performance, the AUC, has been a challenge since its introduction in medicine (56). The AUC value is generally described as the probability that predicted risks correctly identify a random pair of a patient and nonpatient, but this explanation is perceived clinically irrelevant as a physician does not see two random people during a consultation. A more intuitive explanation of the AUC value is lacking, which could improve the understanding of the metric and possibly nullify the argued limitations. Subsequently, because the AUC is considered not intuitive, researchers often

think of AUC as an insensitive metric and criticize its usefulness because even statistically significant new risk factors or PRSs yield a minimal improvement in the AUC when added to existing prediction models, especially when the AUC of this model is already high (72,73,82,83). Previous studies have shown that both  $\Delta$ AUC and IDI are higher when the effect size of the added risk factor is higher (84–86). However, little is known about the size of IDI in the situation when  $\Delta$ AUC is small, for example, lower than 0.01, at which it is generally concluded that the discriminative ability of the model is not improved (84). Insight into the characteristics of the metrics may explain whether the argued insensitivity of the AUC is justified.

### **Assessing multiple metrics of predictive performance**

Partly because AUC is considered insensitive, other metrics gained popularity since their introduction in 2008 (65). These metrics include the NRI and the IDI that focus on reclassification and risk differences between patients and nonpatients respectively. Both metrics have been argued to be too sensitive for identifying changes in predicted risks (87–89). NRI has been shown to reflect improvement when AUC does not, but whether this means that there is indeed improvement or whether the NRI reports noise complicates interpretation. The focus in the interpretation of the improvement in performance is often, surprisingly, on the statistical significance of NRI rather than on the AUC value, especially when the latter shows approximately no improvement. This suggests that researchers may not know the difference between the AUC and NRI, and do not know how to interpret the absolute values. For IDI, it remains unknown whether researchers also emphasize the statistical significance of IDI in the absence of statistically significant improvement in AUC. Moreover, researchers often use NRI and IDI in addition to AUC in the assessment of the predictive performance of genetic and clinical prediction models, which may complicate and deteriorate interpretation even more. Although the three metrics provide complementary information and it therefore has been advised (84) to report all three alongside, whether researchers are aware of the differences between the metrics when they use all three, and how they deal with discording findings remains unknown.

## Scope of this thesis

### Aim, objectives and research questions

The overall aim driving the research described in this thesis is to contribute to the understanding of the design, evaluation and interpretation of genetic risk prediction studies for common diseases. I want to provide insight and guidance to support researchers, physicians, and policymakers who work with (genetic) risk prediction models. The objective of this thesis is to improve understanding and use of multiple metrics that are used to assess the predictive performance of genetic risk prediction models and to provide insight into key topics and considerations that are made in (genetic) prediction research. The main research questions that will be addressed in this thesis are:

1. How does the intended use of risk prediction models determine the design and interpretation of prediction studies?
2. Why is the area under the ROC curve a metric of discrimination?
3. What do different metrics of predictive performance measure?
  - a. Can the predictive ability of a model improve when discrimination does not?
  - b. How do researchers describe the use and interpret the results of multiple metrics in the assessment of improvement in predictive performance of risk prediction models?

### Outline

An overview of the key topics and considerations in the design and evaluation of genetic prediction studies is presented in Chapter 2. The purpose of this chapter was to explain how the intended use of PRSs in health care guides the design and evaluation of prediction studies (Question 1). Chapter 3 continues with the exploration of the most commonly used metric for the assessment of genetic prediction models, the ROC. We investigated how the ROC is another way of presenting the risk distributions of patients and nonpatients and how the shape of the ROC curve is informative of these underlying risk distributions (Question 2). In Chapter 4 we investigated using simulated data whether a genetic risk factor that minimally improves the AUC ( $\Delta$ AUC) may nevertheless improve the predictive ability of the model, assessed by the IDI. Additionally, we investigated the assessment of  $\Delta$ AUC and IDI empirically in prediction studies that had investigated the addition of SNP(s) to a model containing

clinical risk factors (Question 3a). In Chapter 5 we reviewed how researchers defined and calculated multiple metrics of predictive performance (AUC, NRI and IDI) and how they interpreted their results when simultaneously used in the assessment of genetic prediction models (Question 3b). Chapter 6 continues with an assessment of the generalizability of the results of Chapter 5 to non-genetic prediction models by evaluating the simultaneous use of multiple performance metrics in non-genetic prediction studies (Question 3b). Chapter 7 includes two letters to the editor, to demonstrate the importance of scientific communication and to point out how understanding of the main concepts and metrics in prediction research may lead to different conclusions of the prediction studies. In Chapter 8 we conclude with a general discussion of our findings and provide guidance for future design and evaluation of prediction studies as well as directions for future research.



## References

1. WHO. (<https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases>) (Accessed October 7, 2021)
2. Egnew TR. Suffering, meaning, and healing: Challenges of contemporary medicine. *Ann Fam Med*. 2009;7(2):170–175.
3. Cushing H. *The life of Sir William Osler, Volume 2*. 2013.
4. Visscher PM, Wray NR, Zhang Q, et al. 10 Years of GWAS Discovery: Biology, function, and translation. *Am J Hum Genet*. 2017;101(1):5–22.
5. Warren M. The approach to predictive medicine that is taking genomics research by storm. *Nature*. 2018;562(7726):181–183.
6. Visscher PM, Goddard ME. From R.A. Fisher's 1918 paper to GWAS a century later. *Genetics*. 2019;211(4):1125–1130.
7. Vilhjálmsson BJ, Yang J, Finucane HK, et al. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am J Hum Genet*. 2015;97(4):576–92.
8. De La Vega FM, Bustamante CD. Polygenic risk scores: A biased prediction? *Genome Med*. 2018;10(1):100.
9. Machiela MJ, Chen C-Y, Chen C, et al. Evaluation of polygenic risk scores for predicting breast and prostate cancer risk. *Genet Epidemiol*. 2011;35(6):506–514.
10. Natarajan P, Peloso GM, Zekavat SM, et al. Deep-coverage whole genome sequences and blood lipids among 16,324 individuals. *Nat Commun*. 2018;9(1):3391.
11. Khera A V, Chaffin M, Aragam KG, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet*. 2018;50(9):1219–1224.
12. Thomas M, Sakoda LC, Hoffmeister M, et al. Genome-wide modeling of polygenic risk score in colorectal cancer risk. *Am J Hum Genet*. 2020;107(3):432–444.
13. Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. *Nat Rev Genet*. 2018;19(9):581–590.
14. Inouye M, Abraham G, Nelson CP, et al. Genomic risk prediction of coronary artery disease in 480,000 adults. *J Am Coll Cardiol*. 2018;72(16):1883–1893.
15. Mavaddat N, Michailidou K, Dennis J, et al. Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. *Am J Hum Genet*. 2019;104(1):21–34.
16. Nelson SC, Bowen DJ, Fullerton SM. Third-party genetic interpretation tools: A mixed-methods study of consumer motivation and behavior. *Am J Hum Genet*. 2019;105(1):122–131.
17. Food and Drug Administration United States. FDA allows marketing of first direct-to-consumer tests that provide genetic risk information for certain conditions | FDA. (<https://www.fda.gov/news-events/press-announcements/fda-allows-marketing-first-direct-consumer-tests-provide-genetic-risk-information-certain-conditions>). (Accessed March 31, 2021)
18. 23andMe. (<https://www.23andme.com/en-eu/>) (Accessed October 7, 2021)
19. MyHeritage. 6 New Reports Added to MyHeritage Health including glaucoma and prostate cancer - MyHeritage Blog. (<https://blog.myheritage.com/2020/12/6-new-reports-added-to-myheritage-health-including-glaucoma-and-prostate-cancer/>). (Accessed April 20, 2021)
20. Chiarelli AM, Prummel M V., Muradali D, et al. Effectiveness of screening with annual magnetic resonance imaging and mammography: Results of the initial screen from the Ontario High Risk Breast Screening Program. *J Clin Oncol*. 2014;32(21):2224–2230.
21. Michailidou K, Lindström S, Dennis J, et al. Association analysis identifies 65 new breast cancer risk loci. *Nature*. 2017;551(7678):92–94.
22. Lee A, Mavaddat N, Wilcox AN, et al. BOADICEA: a comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors. *Genet Med*. 2019;21(8):1708–1718.
23. Li H, Feng B, Miron A, et al. Breast cancer risk prediction using a polygenic risk score in the familial setting: a prospective study from the Breast Cancer Family Registry and kConFab. *Genet Med*. 2017;19(1):30–35.

24. Mars N, Koskela JT, Ripatti P, et al. Polygenic and clinical risk scores and their impact on age at onset and prediction of cardiometabolic diseases and common cancers. *Nat Med.* 2020;26(4):549–557.
25. Pashayan N, Duffy SW, Neal DE, et al. Implications of polygenic risk-stratified screening for prostate cancer on overdiagnosis. *Genet Med.* 2015;17(10):789–795.
26. Na R, Labbate C, Yu H, et al. Single-nucleotide polymorphism-based genetic risk score and patient age at prostate cancer diagnosis. *JAMA Netw open.* 2019;2(12):e1918145.
27. Jeon J, Du M, Schoen RE, et al. Determining risk of colorectal cancer and starting age of screening based on lifestyle, environmental, and genetic factors. *Gastroenterology.* 2018;154(8):2152-2164. e19.
28. Jenkins MA, Makalic E, Dowty JG, et al. Quantifying the utility of single nucleotide polymorphisms to guide colorectal cancer screening. *Futur Oncol.* 2016;12(4):503–513.
29. Fahed AC, Wang M, Homburger JR, et al. Polygenic background modifies penetrance of monogenic variants for tier 1 genomic conditions. *Nat Commun.* 2020;11(1):3635.
30. Lecarpentier J, Silvestri V, Kuchenbaecker KB, et al. Prediction of breast and prostate cancer risks in male BRCA1 and BRCA2 mutation carriers using polygenic risk scores. *J Clin Oncol.* 2017;35(20):2240.
31. Seibert TM, Fan CC, Wang Y, et al. Polygenic hazard score to guide screening for aggressive prostate cancer: development and validation in large scale cohorts. *BMJ.* 2018;j5757.
32. Li-Sheng Chen S, Ching-Yuan Fann J, Sipeky C, et al. Risk prediction of prostate cancer with single nucleotide polymorphisms and prostate specific antigen. *J Urol.* 2019;201(3):486–495.
33. Naber SK, Kundu S, Kuntz KM, et al. Cost-effectiveness of risk-stratified colorectal cancer screening based on polygenic risk: Current status and future potential. *JNCI Cancer Spectr.* 2020;4(1).
34. van den Broek JJ, Schechter CB, van Ravesteyn NT, et al. Personalizing breast cancer screening based on polygenic risk and family history. *JNCI J Natl Cancer Inst.* 2020; 113(4):434–442.
35. Wilson PWF, D'Agostino RB, Levy D, et al. Prediction of coronary heart disease using risk factor categories. *Circulation.* 1998;97(18):1837–1847.
36. Janssens AC, Joyner MJ. Polygenic risk scores that predict common diseases using millions of single nucleotide polymorphisms: Is more, better? *Clin Chem.* 2019;65(5):609–611.
37. Khera A V., Emdin CA, Drake I, et al. Genetic risk, adherence to a healthy lifestyle, and coronary disease. *N Engl J Med.* 2016;375(24):2349–2358.
38. Desikan RS, Fan CC, Wang Y, et al. Genetic assessment of age-associated Alzheimer disease risk: Development and validation of a polygenic hazard score. *PLoS Med.* 2017;14(3).
39. Vassos E, Di Forti M, Coleman J, et al. An Examination of polygenic score risk prediction in individuals with first-episode psychosis. *Biol Psychiatry.* 2017;81(6):470–477.
40. Calafato MS, Thygesen JH, Ranlund S, et al. Use of schizophrenia and bipolar disorder polygenic risk scores to identify psychotic disorders. *Br J Psychiatry.* 2018;213(3):535–541.
41. Amare AT, Schubert KO, Hou L, et al. Association of polygenic score for major depression with response to lithium in patients with bipolar disorder. *Mol Psychiatry.* 2020; 26(6):2457–2470.
42. Burke W, Zimmern R. Moving beyond ACCE: An expanded framework for genetic test evaluation. a paper for the United Kingdom genetic testing network. This report is the result of work funded by the Department of Health for the UK Genetic Testing Network (UKGTN). 2007.
43. Haddow J, Palomaki G. ACCE: a model process for evaluating data on emerging genetic tests. In: Khoury M, Little J, Burke W editors. *Human genome epidemiology: a scientific foundation for using genetic information to improve health and prevent disease.* 2004;217–33.
44. Moons KGM, Royston P, Vergouwe Y, et al. Prognosis and prognostic research: what, why, and how? *BMJ.* 2009;338(1):b375–b375.
45. Janssens ACJW, Ioannidis JPA, Van Duijn CM, et al. Strengthening the reporting of Genetic Risk Prediction Studies: The GRIPS statement. *Genet Med.* 2011;13(5):453–456.
46. Collins GS, Reitsma JB, Altman DG, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. *Ann Intern Med.* 2015;162(1):55.
47. Moons KGM, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. *Ann Intern Med.* 2015;162(1):W1.

48. Altman DG, McShane LM, Sauerbrei W, et al. Reporting Recommendations for Tumor Marker Prognostic Studies (REMARK): Explanation and elaboration. *PLoS Med.* 2012;9(5):e1001216.
49. Harrell, FE. *Regression Modeling strategies.* Cham: Springer International Publishing; 2015.
50. Gravesteyn BY, Nieboer D, Ercole A, et al. Machine learning algorithms performed no better than regression models for prognostication in traumatic brain injury. *J Clin Epidemiol.* 2020;122:95–107.
51. Royston P, Moons KGM, Altman DG, et al. Prognosis and prognostic research: Developing a prognostic model. *BMJ.* 2009;338(7707):1373–1377.
52. Akaike H. A new look at the statistical model identification. *IEEE Trans Automat Contr.* 1974;19(6):716–723.
53. Steyerberg EW. *Clinical Prediction Models: A Practical approach to development, validation, and updating.* New York, NY: Springer US; 2009.
54. Austin PC, Steyerberg EW. Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. *Stat Med.* 2014;33(3):517–535.
55. Hosmer DW, Hjort NL. Goodness-of-fit processes for logistic regression: simulation results. *Stat Med.* 2002;21(18):2723–38.
56. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology.* 1982;143(1):29–36.
57. Dankert-Roelse JE, Bouva MJ, Jakobs BS, et al. Newborn blood spot screening for cystic fibrosis with a four-step screening strategy in the Netherlands. *J Cyst Fibros.* 2019;18(1):54–63.
58. Elliott J, Bodinier B, Bond TA, et al. Predictive accuracy of a polygenic risk score-enhanced prediction model vs a clinical risk score for coronary artery disease. *JAMA - J Am Med Assoc.* 2020;323(7):636–645.
59. Mahajan A, Taliun D, Thurner M, et al. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat Genet.* 2018;50(11):1505–1513.
60. Garcia-Closas M, Gunsoy NB, Chatterjee N. Combined associations of genetic and environmental risk factors: Implications for prevention of breast cancer. *J Natl Cancer Inst.* 2014;106(11).
61. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology.* 1983;148(3):839–843.
62. Steyerberg EW, Pencina MJ, Lingsma HF, et al. Assessing the incremental value of diagnostic and prognostic markers: a review and illustration. *Eur J Clin Invest.* 2012;42(2):216–28.
63. Zhang X, Rice M, Tworoger SS, et al. Addition of a polygenic risk score, mammographic density, and endogenous hormones to existing breast cancer risk prediction models: A nested case-control study. *PLoS Med.* 2018;15(9):e1002644.
64. Mosley JD, Gupta DK, Tan J, et al. Predictive accuracy of a polygenic risk score compared with a clinical risk score for incident coronary heart disease. *JAMA - J Am Med Assoc.* 2020;323(7):627–635.
65. Pencina MJ, D'Agostino Sr. RB, D'Agostino Jr. RB, et al. Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. *Stat Med.* 2008;27(2):157–172.
66. Moons KG, Kengne AP, Grobbee DE, et al. Risk prediction models: II. External validation, model updating, and impact assessment. *Heart.* 2012;98(9):691–698.
67. Altman DG, Vergouwe Y, Royston P, et al. Prognosis and prognostic research: Validating a prognostic model. *BMJ.* 2009;338(7708):1432–1435.
68. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Ann Intern Med.* 1999;130(6):515–524.
69. Wishart GC, Azzato EM, Greenberg DC, et al. PREDICT: A new UK prognostic model that predicts survival following surgery for invasive breast cancer. *Breast Cancer Res.* 2010;12(1):R1.
70. Martin AR, Kanai M, Kamatani Y, et al. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet.* 2019;51(4):584–591.
71. Dikilitas O, Schaid DJ, Kosel ML, et al. Predictive utility of polygenic risk scores for coronary heart disease in three major racial and ethnic groups. *Am J Hum Genet.* 2020;106(5):707–716.
72. Grosse SD, Khoury MJ. What is the clinical utility of genetic testing? 2006;8(7).
73. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction.

- Circulation. 2007;115(7):928–935.
74. Leening MJG, Vedder MM, Witteman JCM, et al. Net reclassification improvement: computation, interpretation, and controversies: a literature review and clinician's guide. *Ann Intern Med*. 2014;160(2):122–31.
  75. Janssens ACJW, Ioannidis JPA, Bedrosian S, et al. Strengthening the reporting of Genetic Risk Prediction Studies (GRIPS): explanation and elaboration. *J Clin Epidemiol*. 2011;64(8):e1–e22.
  76. Damen JAAG, Hooft L, Schuit E, et al. Prediction models for cardiovascular disease risk in the general population: Systematic review. *BMJ*. 2016;353:i2416.
  77. Studerus E, Ramyeed A, Riecher-Rössler A. Prediction of transition to psychosis in patients with a clinical high risk for psychosis: a systematic review of methodology and reporting. *Psychol Med*. 2017;47(7):1163–1178.
  78. Bouwmeester W, Zuihoff NPA, Mallett S, et al. Reporting and methods in clinical prediction research: a systematic review. *PLoS Med*. 2012;9(5):1–12.
  79. Mallett S, Royston P, Dutton S, et al. Reporting methods in studies developing prognostic models in cancer: a review. *BMC Med*. 2010;8:20.
  80. Kleinrouweler CE, Cheong-See FM, Collins GS, et al. Prognostic models in obstetrics: available, but far from applicable. *Am J Obstet Gynecol*. 2016;214(1):79–90.e36.
  81. Wand H, Lambert SA, Tamburro C, Iacocca MA, et al. Improving reporting standards for polygenic scores in risk 1 prediction studies. *Nature*. 2021; 591(7849):211–219.
  82. Pepe MS, Janes HE. Gauging the performance of SNPs, biomarkers, and clinical factors for predicting risk of breast cancer. *J Natl Cancer Inst*. 2008;100(14):978–979.
  83. Ware JH. The limitations of risk factors as prognostic tools. *N Engl J Med*. 2006;355(25):2615–2617.
  84. Pencina MJ, D'Agostino RB, Pencina KM, et al. Interpreting incremental value of markers added to risk prediction models. *Am J Epidemiol*. 2012;176(6):473–481.
  85. Mihaescu R, Pencina MJ, Alonso A, et al. Incremental value of rare genetic variants for the prediction of multifactorial diseases. *Genome Med*. 2013;5(8):76.
  86. Kerr KF, Bansal A, Pepe MS. Further insight into the incremental value of new markers: the interpretation of performance measures and the importance of clinical context. *Am J Epidemiol*. 2012;176(6):482–487.
  87. Pepe MS, Janes H, Li CI. Net risk reclassification P values: Valid or misleading? *JNCI J Natl Cancer Inst*. 2014;106(4):dju041.
  88. Gerds TA, Hilden J. Calibration of models is not sufficient to justify NRI. *Stat Med*. 2014;33(19):3419–3420.
  89. Hilden J, Gerds TA. A note on the evaluation of novel biomarkers: do not rely on integrated discrimination improvement and net reclassification index. *Stat Med*. 2014;33(19):3405–3414.



# 2

## **How the intended use of polygenic risk scores guides the design and evaluation of prediction studies**

Forike K. Martens and A. Cecile J.W. Janssens  
Current Epidemiology Reports 6 (2019) 184-190

## Abstract

**Purpose of review** To explain how the intended use of polygenic risk scores (PRSs) in healthcare guides the design and evaluation of prediction studies.

**Recent findings** The advances in gene discovery in common complex diseases have fueled the interest in the potential of PRSs to predict risks and improve the prevention and early detection of disease. As the predictive ability of a PRS differs between populations and settings, it is important that prediction studies are designed and evaluated with the intended use of the risk scores in mind, but this is rarely done.

**Summary** The intended use indicates in whom and how the PRS will be used in healthcare and for what purpose. This intended use dictates what outcome needs to be predicted in which population using which predictors. It also tells which other variables or clinical risk models might be available to improve the prediction. The intended use also provides the necessary context to evaluate whether the predictive ability of the PRS or the risk model that includes PRS is high enough for the score to be potentially useful in healthcare. The intended use should be leading risk prediction research.

## Introduction

Over the past decade, genome wide association studies (GWASs) have identified many genetic variants (single nucleotide polymorphisms, SNPs) that are robustly associated with the risk of common diseases. Researchers combine SNPs into polygenic risk scores (PRSs) to identify individuals at increased risk of common diseases such as type 2 diabetes, cardiovascular disease and cancer (1–3). In contrast to monogenic diseases that are caused by a single mutated gene, common diseases are caused by an interplay of multiple SNPs, and non-genetic factors such as lifestyle.

After years of polygenic risk research in which PRSs mostly had a modest predictive ability and little value added to clinical risk models (4–8), the tide recently seems to have changed. Several recent studies reported that people with the highest PRSs were at much higher risk compared to the rest of the study population (1,9,10), with risks comparable to the increased risks of genetic mutations (1,10). These studies concluded that PRSs may be useful in healthcare and that it is time to consider their implementation (1,11).

PRSs quantify the combined contribution of multiple SNPs to the risk of these diseases. The scores can consist of a few up to millions of SNPs. The scores are weighted or unweighted sums of the risk alleles across all SNPs that are included. The SNPs are generally selected from GWASs based on the GWAS effect sizes, the weights, or their *P* values. Individual risk estimates are obtained from a regression model that includes the PRS with or without clinical risk factors.

Prediction studies aim to investigate the performance of tests and models for predicting diseases in a population. In the framework of translational genomics research (12), prediction research is part of the second of four phases. Following the first phase of association studies, this second phase investigates potential applications of genetic tests: what is their predictive ability and, if applicable, what is the potential utility? If the application has potential utility, the third phase is to investigate the implementation in healthcare and the fourth to monitor whether the application delivers as anticipated.

The intended use of PRSs that informs how their predictions will be used in healthcare: which medical decisions will be supported and for whom will these decisions be relevant? This intended use specifies in whom the PRS will be used, what outcome needs to be predicted, and what the purpose of



prediction is. Therewith, the intended use of PRSs has two major implications for their scientific study.

First, as the predictive ability of PRSs varies between populations and settings, the intended use dictates the design of the study. It defines what population needs to be studied, what outcome predicted, and what additional or other opportunities for prediction are available. A PRS that predicts a disease in one population does not evidently predict the disease in another, which is relevant to keep in mind when a PRS is studied in readily available datasets as these may not be relevant for the population in which the use of the PRS is foreseen. A PRS needs to be studied in a population in which the PRS is intended to be used (13).

Second, and related, the intended use provides context for the interpretation of the results and the evaluation of the predictive ability. The intended use can serve as a benchmark to judge whether the observed performance of the risk model may be high enough to expect health benefits or improved efficiency of care when the risk model is used in clinical or public health practice.

Researchers rarely specify the intended use in their studies. They may state that the PRS can be used to identify high-risk groups that can benefit from early intervention or low-risk groups that can benefit from delaying treatment or surgery without specifying which treatment or when high-risk individuals qualify (14–16). They may state that PRS can be used to support decision making, motivate risk reduction behaviors, or impact prevention strategies without clarifying the decisions, behaviors, and preventive strategies (17, 18). And they certainly do not specify how high the predictive ability at least needs to be to make the PRS worth considering in practice. When the intended use is not specified, it is difficult to appreciate the relevance of the study and the results may need to be inferred from the design of the study.

In this paper, we elaborate on how the intended use of PRSs guides the design of studies that aim to investigate its predictive ability and how it provides context for the evaluation of the results.

## Design of PRS studies

As the goal of a prediction study is to develop and evaluate PRSs for predicting disease in clinical practice (19), the purpose of testing should be clearly specified, and the outcome of interest, the study population and predictors should be carefully chosen so that the study is relevant for the intended use of the PRS.

### Purpose of testing

Prediction models in health care are used to identify at-risk groups for preventive interventions or the early detection of disease; assist and support doctors in making medical decisions about procedures, treatments, medications, and other interventions; and inform individuals about their risks or progression of the disease to allow them making plans for the future.

Identifying and specifying the purpose of testing is essential because the same test may be predictive enough for one application, but not for another. It helps distinguishing whether, for example, the predictive ability of a PRS for breast cancer is high enough for improving the efficiency of mammography screening or even high enough for recommending prophylactic mastectomy to high-risk women, as, evidently, the latter predictive ability needs to be very high to prevent that women are erroneously classified as being at high-risk.

### Outcome

The outcome of interest specifies what needs to be predicted, such as the 10-year risk of developing type 2 diabetes (20), the 7-year risk of developing Alzheimer disease (21), or the 5-year risk of breast cancer (22). Researchers often leave the risk period unreported, referring to it as the risk of prostate cancer or breast cancer without further specification (23,24). In such instances, the risk period and its relevance may be inferred from the follow-up duration of the study.

The outcome of interest will often be self-evident, but large cohort studies may have data on related diagnoses and disease (sub)types so that the selection of the outcome may involve a decision. PRSs can, for example, be developed for the prediction of the risk of dementia (25) or for Alzheimer disease (21), vascular dementia (26), and other subtypes separately. When the purpose of testing is to identify people for preventive drug treatment, the outcome of interest may be the risk of disease, but can also be treatment response,

prognosis, or side effects, depending on what is most relevant in making the decision about treatment.

### **Study population**

The target population is the population in which the PRS will be used if proven predictive. This population is a selection of the general population that is defined by one or more risk factors of disease, such as age, sex, family history, or early symptoms. Selection of the target population is in part determined by the course of a disease process over time in the absence of interventions. This natural history tells at which ages a disease may manifest and how it may progress and is the reason why the risk of Alzheimer disease is predicted in the elderly and autism and attention-deficit/hyperactivity disorder (ADHD) in the young (27,28).

The target population consists of people who are expected to develop the disease within the risk period of interest and people who will not. The latter group also includes people who will develop the disease later, those who may already have developed risk factors that increase the risk of disease, have a preclinical stage of disease, or early stages that do not yet formally meet the diagnostic criteria. Disease diagnoses are most difficult to predict in the context of this variety in symptom and disease presentation at follow-up, which is why a PRS ideally is investigated in a prospective longitudinal cohort study: a population unselected for the outcome of interest and with predictors and outcomes measured prospectively over time. Case-control studies may overestimate the predictive ability when the selection of the two groups excludes people with ambiguous or inconclusive symptoms and when recall bias distorts the assessment of nongenetic risk factors. The latter also makes the cross-sectional study design less suitable for prediction research.

The question of what would be the optimal study population is seldom asked. Researchers do not set up data collections for investigating the predictive ability of PRSs, but they also do not weigh the pros and cons of datasets that could be available to them through collaboration either. Prediction analyses are generally performed in datasets that researchers have direct access to, and the relevance of that dataset for evidencing the clinical applicability of the PRS generally remains undiscussed.

These days, researchers frequently use the UK Biobank for assessing the predictive ability of PRSs (1,2,29,30). The UK biobank is an epidemiological cohort providing data of up to 500,000 participants (31). The prospective collection of data is ideal for prediction studies, but the wide age range, from 40

to 69 at baseline (31), makes it less a relevant population for the prediction of diseases that come with age (32–34). The younger participants are too young to develop Alzheimer disease and too young to die (21,35), whereas the older participants are too old to develop type 1 diabetes or multiple sclerosis (36,37). Using the entire UK biobank population for the development of a risk model will likely include age as a very strong predictor and inflate the performance of the risk model.

The handling of age and other covariates also alludes to an essential difference between epidemiological and prediction studies. In epidemiological research, the association of a PRS with the risk of disease is studied with adjustment for covariates, while in prediction analyses these covariates become part of the risk model. When researchers write that, for example, the area under the receiver operating characteristic curve (AUC) of a type 2 diabetes PRS was 0.75, *adjusted* for age, sex, BMI, and other factors, it means that the AUC was for a risk model that included the PRS and all other factors, not for the PRS alone (38). In such instances, a comparison between risk models with and without a PRS is warranted to evaluate the value added by including PRS in the risk model (see below).

## Predictors

When the target population is known, it follows which predictors will or can be available to predict the outcome of interest. The risk model can include a PRS, alone or in combination with other predictors such as demographic data, family history, biomarkers, comorbidity, and subjective health status. If a PRS alone can predict the risk of disease as good as a clinical model or clinical plus a PRS model, then the PRS may suffice, depending on whether the type of risk factors matters. If the aim is to monitor high-risk individuals for the early detection of disease, it is less relevant whether the high risk is due to modifiable or non-modifiable genetic risk factors. If the aim is a behavioral intervention of modifiable risk factors, then it seems counterintuitive not to include these in the risk model.

When implementation of the PRS or risk model is the goal, and generalizability of the model is essential, the selection of its predictors is preferably based on replicated associations, such as from GWAS for SNPs and from meta-analyses for nongenetic predictors. Also, and evidently, predictors need to be available or obtainable, measurable, and affordable.

## Evaluation of PRS studies

The predictive ability of risk models varies with the population, the outcome, and predictors that are used. In the evaluation of prediction studies, the following questions are relevant:

### **Are the predictions accurate?**

When a risk model predicts a risk of 25%, is the risk 25%? Accuracy of risk predictions is assessed using measures of calibration that compare predicted risks with observed risks in study data (39). Calibration of a PRS is important because the score assumes that the effects of all genetic variants can be added into a (weighted) sum score, that all variants are relevant for the risk in all people, and that, when combined with clinical risk factors, the genetic factors are an independent risk factor. When the intended use is to identify high-risk individuals, calibration of the risks in the tails is particularly important. Calibration is essential as inaccurate risk predictions may lead to wrong medical decisions and cause unnecessary harm (40).

### **What is the distribution of predicted risks?**

Risk distributions show whether the predicted risks range from 0 to 100% or spread narrowly around the average risk. They also help identify a skewed distribution with a long flat high-end tail that indicate a (small) group of people with a substantially higher risk than the rest.

### **Can predicted risks identify people who will develop the disease?**

A risk model can identify people who will develop the disease when they have higher predicted risks than those who will not develop the disease, or in statistical terms, when the distribution of risks of cases and noncases show no overlap. The farther these distributions are separated, the better the discriminative ability. This degree of separation is assessed using measures of discrimination such as the AUC or *c*-statistic (41). Measures of discrimination are rank tests, they assess whether cases tend to have a higher risk than noncases, but not how much higher. How much higher the predicted risks are is learned from the risk distributions or from the difference in average risks of the two groups.

## How well does the model classify people at risk?

Risk models are often used to classify people in risk categories by one or more risk thresholds (42). Ideally, a risk model would have a threshold that classifies all cases above and all noncases below the threshold, thus having sensitivity, specificity, positive, and negative predictive values all at 100% (43,44). When risk distributions of cases and noncases overlap, selecting an optimal threshold requires weighing the benefits and costs of true and false risk classifications.

When a PRS is added to clinical risk factors or vice versa, the same questions apply: Is the new risk model well-calibrated? Did the addition of the PRS change the distribution of risk? Did it improve discrimination? And did it change the classification of risk groups? The improvement in the AUC or *c*-statistic assesses the improvement in discrimination (45), the integrated discrimination improvement (IDI) indicates the increase in the risk difference (46), and measures of reclassification assess whether updating a risk model changes the classification of risk in the right direction (46,47). Assessing reclassification is only meaningful when the risk thresholds are clinically relevant, i.e., when management of people at risk differs between the risk categories, as the amount of reclassification varies with the cutoff thresholds (48).

The intended use provides context for evaluating whether the predictive ability will be good enough for the risk model to be used in practice and, if a PRS is added to clinical risk factors, whether the improvement provides meaningful changes in predicted risks. When there is no information about the intended use, then the performance of the risk model can only be judged by its statistical significance, as the values of the metrics have no benchmark (49). Evaluation studies of existing risk models for the same outcome may provide quantitative reference points for the interpretation. For example, a recent diabetes study reported that the AUC of the PRS was the same as that of a model with only age, sex, and body mass index (3).

When the predictive ability is promising (50), validation of the model in independent populations is warranted as the predictive ability tends to be higher in the people whose data were used to develop the PRS (51). External validation should re-assess both calibration and discrimination (51). The requirement of external validation underscores the importance of selecting established predictors: the best risk model is the one with the best predictive ability at external validation.

## Examples

The intended use of PRSs not only guides the design and evaluation of prediction studies but also helps interpreting the results of published studies when the intended use is not specified. The intended use can be inferred from the selected study population, outcome, and predictors used in the study, which informs whether the study addresses a relevant healthcare scenario. Table 1 illustrates three recent studies that did not report about the intended use of the PRS, which suggests that the studies were conducted without a specific public health scenario in mind. Here are examples of inferences and questions that follow from the study design and analyses.

The study of Zhang et al. is a case-control study, which tells that we should review the inclusion criteria for the selection of cases and controls to evaluate whether the discriminative ability of the PRS might be overestimated (22). The researchers added several risk factors to an existing model, which warrants an assessment of each separately to learn which risk factor (or all) meaningfully improved the predictive ability. And finally, since the percentage of reclassification depends on the risk threshold that is chosen, it is worth reviewing what the rationale for the 2.27% threshold was.

The study of Abraham et al. predicts the risk of coronary heart disease in two population-based cohorts, which suggest that they may or could have used one for the development of the risk model and the other for its validation (52). Their study distinguished four risk categories, and it is of interest to question how these match with the lifestyle modifications and medical interventions that are the purpose of testing.

Finally, the study of Pitkänen et al. investigates the risk of type 2 diabetes in children (38). The researchers did not specify a risk period and risk thresholds, which raises additional questions about the purpose of testing. While type 2 diabetes occurs at younger ages these days, the question is whether the risk is high enough and the clinical risk factors prevalent enough in children to be of interest for prevention.

**Table 1.** Inferring the intended use of polygenic risk score from the study methods

|                                 | <b>Zhang et al., 2018 (22)</b>   | <b>Abraham et al., 2016 (52)</b>   | <b>Pitkänen et al., 2016 (38)</b>                                    |
|---------------------------------|--|--|--|
| What is the purpose of testing? | Identification of women at higher risk who would benefit most from chemoprevention | Early identification of individuals at increased risk of coronary heart disease for preventive lifestyle modifications and medical interventions | Early identification of individuals at high risk for type 2 diabetes |
| What is predicted?              | 5-Year risk of invasive breast cancer  | 10-Year risk of incident coronary heart disease  | Risk of type 2 diabetes in adulthood                                 |
| In whom?                        | Female registered nurse cases and controls, age range 34-70 years                  | Population-based cohorts, mean age 46 and 44 years   | Population-based cohort, age 3-18 years                              |
| How?                            | Adding 67 SNP PRS, mammographic density and hormone levels to existing risk models | Adding 49K SNP PRS to existing risk models   | Adding 73 SNP PRS to clinical risk factors                           |
| Risk thresholds                 | <2.27% and ≥2.27%  | <7.5%, 7.5–10%, 10–20%, and ≥20%   | None   |

PRS polygenic risk score, SNP single nucleotide polymorphism

## Conclusions

This paper discusses considerations in the design and conduct of studies that aim to investigate the predictive ability of PRSs. The intended use of a risk model dictates the design of the study and provides context for the interpretation of the scores' predictive ability. The importance of considering the intended use applies to the study of all prediction models, also those that do not include genetic variants.

To be sure, many researchers use PRSs in epidemiological research with no interest in the predictive ability or utility of the PRS. They may report that the PRS was a “powerful predictor” of schizophrenia (53) and that the PRS “predicted educational achievement” (54). “Predict” is often used to describe a statistically significant association, without claiming that the predictive ability is high enough to identify high-risk individuals. Our paper is not about these association studies.



Risk models not only predict risk, they also inform which risk factors are (most) important. Often researchers test multiple models and select the model with the highest AUC as the best model, even though its AUC is only minimally higher than other models. The addition of variables should be worth the “costs” (55), also when the data are available and free. It is not only the financial costs and collection of data that is deciding this, but also the reception by the target population. The latter is a relevant consideration with the new methods that build PRS using millions of SNPs (1,10,56), but these million SNPs often do not outperform the statistically significant SNPs by much. Risks that are calculated from millions of SNPs may be perceived as more deterministic, and therefore, implications of a PRS on behavior should be considered when developing the score.

Scientists increasingly claim that the time is right to consider inclusion of PRSs in clinical care based on the observation that a group at the end of the risk distribution has an increased risk that is comparable to that of monogenic risk (1,11). Our paper and those of others have summarized that the predictive ability cannot be judged from a relative risk alone (39,57). When making claims about utility or implementation in health care, then the predictive ability should be investigated using the appropriate metrics and an informative comparison with other models. Uniform reporting facilitates the synthesis of research findings across studies (58,59).

Whether it is time to consider the implementation of PRSs in health care does not depend on their predictive ability, but on their usability, usefulness, and meaningfulness. Does the PRS improve prediction beyond clinical risk models? What interventions can be recommended to people at high genetic risk? And what will be offered to those with a high PRS in the absence of traditional risk factors? Can PRSs help to change behavior? Evidence of clinical utility determines when the time has come. We are not there yet.

## References

1. Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet.* 2018; 50(9):1219–1224.
2. Inouye M, Abraham G, Nelson CP, Wood AM, Sweeting MJ, Dudbridge F, et al. Genomic risk prediction of coronary artery disease in 480,000 adults. *J Am Coll Cardiol.* 2018;72(16):1883–1893.
3. Mahajan A, Taliun D, Thurner M, Robertson NR, Torres JM, Rayner NW, et al. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat Genet.* 2018;50(11):1505–1513.
4. Brautbar A, Pompeii LA, Dehghan A, Ngwa JS, Nambi V, Virani SS, et al. A genetic risk score based on direct associations with coronary heart disease improves coronary heart disease risk prediction in the Atherosclerosis Risk in Communities (ARIC), but not in the Rotterdam and Framingham Offspring, Studies. *Atherosclerosis.* 2012;223(2):421–426.
5. Havulinna AS, Kettunen J, Ukkola O, Osmond C, Eriksson JG, Kesäniemi YA, et al. A blood pressure genetic risk score is a significant predictor of incident cardiovascular events in 32,669 individuals. *Hypertens (Dallas, Tex 1979).* 2013;61(5):987–994.
6. Ripatti S, Tikkanen E, Orho-Melander M, Havulinna AS, Silander K, Sharma A, et al. A multilocus genetic risk score for coronary heart disease: case-control and prospective cohort analyses. *Lancet.* 2010;376(9750):1393–1400.
7. Lyssenko V, Jonsson A, Almgren P, Pulizzi N, Isomaa B, Tuomi T, et al. Clinical risk factors, DNA variants, and the development of type 2 diabetes. *N Engl J Med.* 2008;359(21):2220–2232.
8. Park HY, Choi HJ, Hong Y-C. Utilizing Genetic Predisposition score in predicting risk of type 2 diabetes mellitus incidence: A community-based cohort study on middle-aged Koreans. *J Korean Med Sci.* 2015;(8):1101–1109.
9. Matejčić M, Saunders EJ, Dadaev T, Brook MN, Wang K, Sheng X, et al. Germline variation at 8q24 and prostate cancer risk in men of European ancestry. *Nat Commun.* 2018;9(1):4616.
10. Natarajan P, Peloso GM, Zekavat SM, Montasser M, Ganna A, Chaffin M, et al. Deep-coverage whole genome sequences and blood lipids among 16,324 individuals. *Nat Commun.* 2018;9(1):3391.
11. Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. *Nat Rev Genet.* 2018;19(9):581–590.
12. Khoury MJ, Gwinn M, Yoon PW, Dowling N, Moore CA, Bradley L. The continuum of translation research in genomic medicine: how can we accelerate the appropriate integration of human genome discoveries into health care and disease prevention? *Genet Med.* 2007;(10):665–674.
13. Haddow J, Palomaki G. ACCE: a model process for evaluating data on emerging genetic tests. In: Khoury M, Little J, Burke W editors. *Human genome epidemiology: a scientific foundation for using genetic information to improve health and prevent disease.* Oxford, UK: Oxford University Press; 2004;217–233.
14. Yang X, Leslie G, Gentry-Maharaj A, Ryan A, Intermaggio M, Lee A, et al. Evaluation of polygenic risk scores for ovarian cancer risk prediction in a prospective cohort study. *J Med Genet.* 2018;55(8):546–554.
15. Kuchenbaecker KB, McGuffog L, Barrowdale D, Lee A, Soucy P, Dennis J, et al. Evaluation of polygenic risk scores for breast and ovarian cancer risk prediction in BRCA1 and BRCA2 mutation carriers. *J Natl Cancer Inst.* 2017;109(7).
16. Lecarpentier J, Silvestri V, Kuchenbaecker KB, Barrowdale D, Dennis J, McGuffog L, et al. Prediction of breast and prostate cancer risks in male BRCA1 and BRCA2 mutation carriers using polygenic risk scores. *J Clin Oncol.* 2017;35(20):2240.
17. Evans DG, Brentnall A, Byers H, Harkness E, Stavrinou P, Howell A, et al. The impact of a panel of 18 SNPs on breast cancer risk in women attending a UK familial screening clinic: a case-control study. *J Med Genet.* 2017;54(2):111–113.
18. Cust AE, Drummond M, Kanetsky PA, Goldstein AM, Barrett JH, MacGregor S, et al. Assessing the incremental contribution of common genomic variants to melanoma risk prediction in two

- population-based studies. *J Invest Dermatol*. 2018;138(12):2617–2624.
19. Moons KGM, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? *BMJ*. 2009;338(1):b375–b375.
  20. Kim S-H, Lee E-S, Yoo J, Kim Y. Predicting risk of type 2 diabetes mellitus in Korean adults aged 40–69 by integrating clinical and genetic factors. *Prim Care Diabetes*. 2019;13(1):3–10.
  21. Chouraki V, Reitz C, Maury F, Bis JC, Bellenguez C, Yu L, et al. Evaluation of a genetic risk score to improve risk prediction for alzheimer’s disease. Hall A, editor. *J Alzheimers Dis*. 2016;53(3):921–932.
  22. Zhang X, Rice M, Tworoger SS, Rosner BA, Eliassen AH, Tamimi RM, et al. Addition of a polygenic risk score, mammographic density, and endogenous hormones to existing breast cancer risk prediction models: A nested case–control study. Zheng W, editor. *PLoS Med*. 2018;15(9):e1002644.
  23. Szulkin R, Whittington T, Eklund M, Aly M, Eeles RA, Easton D, et al. Prediction of individual genetic risk to prostate cancer using a polygenic score. *Prostate*. 2015;75(13):1467–1474.
  24. Chan CHT, Munusamy P, Loke SY, Koh GL, Yang AZY, Law HY, et al. Evaluation of three polygenic risk score models for the prediction of breast cancer risk in Singapore Chinese. *Oncotarget*. 2018;9(16):12796–804.
  25. Tang EYH, Harrison SL, Errington L, Gordon MF, Visser PJ, Novak G, et al. Current developments in dementia risk prediction modelling: An updated systematic review. Forloni G, editor. *PLoS One*. 2015;10(9):e0136181.
  26. Lee C. Best linear unbiased prediction of individual polygenic susceptibility to sporadic vascular dementia. *J Alzheimers Dis*. 2016;53(3):1115–1119.
  27. Einziger T, Levi L, Zilberman-Hayun Y, Auerbach JG, Atzaba-Poria N, Arbelle S, et al. Predicting ADHD symptoms in adolescence from early childhood temperament traits. *J Abnorm Child Psychol*. 2018;46(2):265–276.
  28. Bussu G, Jones EJM, Charman T, Johnson MH, Buitelaar JK, BASIS Team. Prediction of autism at 3 years from behavioural and developmental measures in high-risk infants: a longitudinal cross-domain classifier analysis. *J Autism Dev Disord*. 2018;48(7):2418–2433.
  29. Smith T, Gunter MJ, Tzoulaki I, Muller DC. BRIEF COMMUNICATION The added value of genetic information in colorectal cancer risk prediction models: development and evaluation in the UK Biobank prospective cohort study. *Br J Cancer*. 2018;119.
  30. Khawaja AP, Cooke Bailey JN, Wareham NJ, Scott RA, Simcoe M, Igo RP, et al. Genome-wide analyses identify 68 new loci associated with intraocular pressure and improve risk prediction for primary open-angle glaucoma. *Nat Genet*. 2018;50(6):778–82.
  31. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK Biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Med*. 2015;12(3):e1001779.
  32. Janssens AC, Joyner MJ. Polygenic risk scores that predict common diseases using millions of single nucleotide polymorphisms: Is more, better? *Clin Chem*. 2019;65(5):609–611.
  33. Greenland P, Hassan S. Precision preventive medicine—Ready for prime time? *JAMA Intern Med*. 2019;179(5):605–606.
  34. Curtis D. Clinical relevance of genome-wide polygenic score may be less than claimed. *Ann Hum Genet*. 2019; 83(4):274–277.
  35. Ganna A, Ingelsson E. 5 year mortality predictors in 498 103 UK Biobank participants: a prospective population-based study. *Lancet*. 2015;386(9993):533–540.
  36. Jacobsen LM, Larsson HE, Tamura RN, Vehik K, Clasen J, Sosenko J, et al. Predicting progression to type 1 diabetes from ages 3 to 6 in islet autoantibody positive TEDDY children. *Pediatr Diabetes*. 2019;20(3):263–270.
  37. Martinelli V, Dalla Costa G, Messina MJ, Di Maggio G, Sangalli F, Moiola L, et al. Multiple biomarkers improve the prediction of multiple sclerosis in clinically isolated syndromes. *Acta Neurol Scand*. 2017;136(5):454–461.
  38. Pitkänen N, Juonala M, Rönnemaa T, Sabin MA, Hutri-Kähönen N, Kähönen M, et al. Role of conventional childhood risk factors versus genetic risk in the development of type 2 diabetes and impaired fasting glucose in adulthood: The cardiovascular risk in Young Finns Study. *Diabetes Care*. 2016;39(8):1393–9.

39. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*. 2010;21(1):128–138.
40. Van Calster B, Vickers AJ. Calibration of risk prediction models. *Med Decis Mak*. 2015;35(2):162–169.
41. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;143(1):29–36.
42. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation*. 2007;115(7):928–935.
43. Altman DG, Bland JM. Diagnostic tests. 1: Sensitivity and specificity. *BMJ*. 1994;308(6943):1552.
44. Altman DG, Bland JM. Diagnostic tests 2: Predictive values. *BMJ*. 1994;309(6947):102.
45. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*. 1983;148(3):839–843.
46. Pencina MJ, D'Agostino Sr. RB, D'Agostino Jr. RB, et al. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med*. 2008;27(2):112–157.
47. Cook NR. Quantifying the added value of new biomarkers: how and how not. *Diagnostic Progn Res*. 2018;2(1):14.
48. Mihaescu R, van Zitteren M, van Hoek M, Sijbrands EJJ, Uitterlinden AG, Witterman JCM, et al. Improvement of risk prediction by genomic profiling: Reclassification measures versus the area under the receiver operating characteristic curve. *Am J Epidemiol*. 2010;172(3):353–361.
49. Martens FK, Tonk ECM, Janssens ACJW. Evaluation of polygenic risk models using multiple performance measures: a critical assessment of discordant results. *Genet Med*. 2019;21(2):391–397.
50. Martens FK, Kers JG, Janssens ACJW. External validation is only needed when prediction models are worth it (Letter commenting on: *J Clin Epidemiol*. 2015;68:25–34). *J Clin Epidemiol*. 2016;69.
51. Moons KG, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: II. External validation, model updating, and impact assessment. *Heart*. 2012;98(9):691–698.
52. Abraham G, Havulinna AS, Bhalala OG, Byars SG, De Livera AM, Yetukuri L, et al. Genomic prediction of coronary heart disease. *Eur Heart J*. 2016;37(43):3267–3278.
53. Vassos E, Di Forti M, Coleman J, Iyegbe C, Prata D, Euesden J, et al. An examination of polygenic score risk prediction in individuals with first-episode psychosis. *Biol Psychiatry*. 2017;81(6):470–477.
54. Selzam S, Krapohl E, von Stumm S, O'Reilly PF, Rimfeld K, Kovas Y, et al. Predicting educational achievement from DNA. *Mol Psychiatry*. 2017;22(2):267–272.
55. Akaike H. A new look at the statistical model identification. *IEEE Trans Automat Contr*. 1974;19(6):716–723.
56. Gu F, Chen T-H, Pfeiffer RM, Fargnoli MC, Calista D, Ghorzo P, et al. Combining common genetic variants and non-genetic risk factors to predict risk of cutaneous melanoma. *Hum Mol Genet*. 2018;27(23):4145–4156.
57. Wynants L, Collins GS, Van Calster B. Key steps and common pitfalls in developing and validating risk models. *BJOG*. 2017;124(3):423–432.
58. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. *Ann Intern Med*. 2015;162(1):55.
59. Janssens ACJW, Ioannidis JPA, Bedrosian S, Boffetta P, Dolan SM, Dowling N, et al. Strengthening the reporting of Genetic Risk Prediction Studies (GRIPS): explanation and elaboration. *J Clin Epidemiol*. 2011;64(8):e1–22.



# 3

## **Reflection on modern methods: Revisiting the area under the ROC Curve**

A. Cecile J.W. Janssens and Forike K. Martens

International Journal of Epidemiology 49-4 (2020) 1397–1403

## Summary

The area under the receiver operating characteristic (ROC) curve (AUC) is commonly used for assessing the discriminative ability of prediction models even though the measure is criticized for being clinically irrelevant and lacking an intuitive interpretation. Every tutorial explains how the coordinates of the ROC curve are obtained from the risk distributions of diseased and non-diseased individuals, but it has not become common sense that therewith the ROC plot is just another way of presenting these risk distributions. We show how the ROC curve is an alternative way to present risk distributions of diseased and non-diseased individuals and how the shape of the ROC curve informs about the overlap of the risk distributions. For example, ROC curves are rounded when the prediction model included variables with similar effect on disease risk and have an angle when, for example, one binary risk factor has a stronger effect; and ROC curves are stepped rather than smooth when the sample size or incidence is low, when the prediction model is based on a relatively small set of categorical predictors. This alternative perspective on the ROC plot invalidates most purported limitations of the AUC and attributes others to the underlying risk distributions. AUC is a measure of the discriminative ability of prediction models. The assessment of prediction models should be supplemented with other metrics to assess their clinical utility.

In 1971, Lee Lusted introduced the receiver operating characteristic (ROC) curve in medicine to contrast the percentage of true positive against false positive diagnoses for different decision criteria applied by a radiologist (1). A decade later, Hanley and McNeil proposed the area under this ROC curve (AUC) as a single metric of diagnostic accuracy for 'rating methods or mathematical predictions based on patient characteristics' (2). The AUC is the most commonly used metric for assessing the ability of predictive and prognostic models to discriminate between individuals who will or will not develop the disease (here referred to as diseased and non-diseased individuals).

Despite its popularity, the AUC is frequently criticized and its interpretation has been a challenge since its introduction in medicine (2). The AUC value is generally described as the probability that predicted risks correctly identify a random pair of a diseased and non-diseased individual. This probability is considered clinically irrelevant as doctors never have two random people in their office (3, 4); they are only interested in the clinically relevant thresholds of the ROC curve, not in others (5); and they often want to distinguish multiple risk categories for which they need more than one threshold (6). Also, the AUC is considered insensitive, as the addition of substantial risk factors may improve AUC only minimally when they are added to a baseline model that already has good discrimination (4, 7-9). Most of this criticism on the AUC concerns the irrelevance of the ROC curve suggesting that a more intuitive interpretation of the ROC could change the appreciation of the AUC.

Every tutorial explains how the coordinates of the ROC curve are obtained from the risk distributions of diseased and non-diseased individuals. In this paper, we show that the ROC curve is an alternative graphical presentation of these risk distributions. We explain how the ROC curve gives information about the shapes and overlap of the underlying risk distributions, and re-evaluate the interpretation and purported limitations of the AUC from this alternative perspective.

## From risk distributions to ROC curve

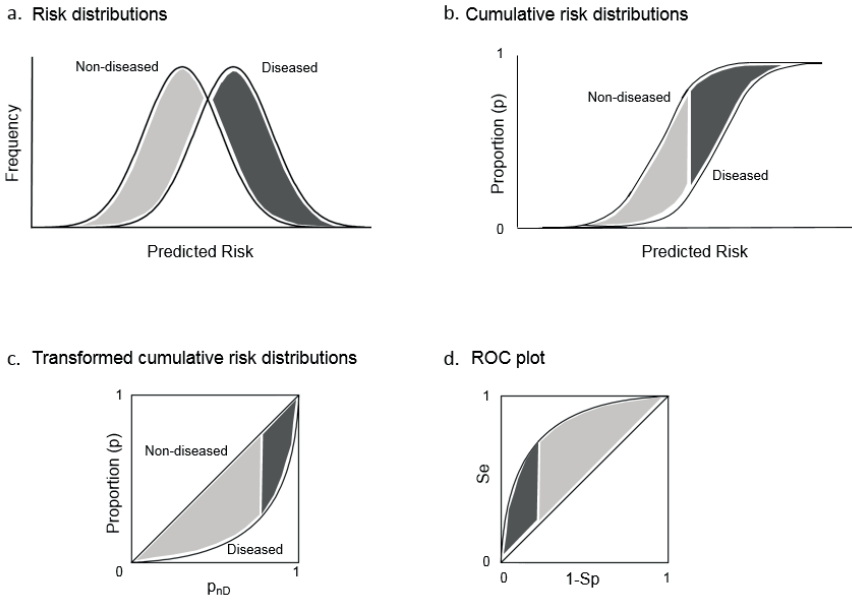
In empirical studies that investigate the development or validation of prediction models, predicted risks can be presented as separate distributions for diseased and non-diseased individuals (Figure 1a). The separation between the



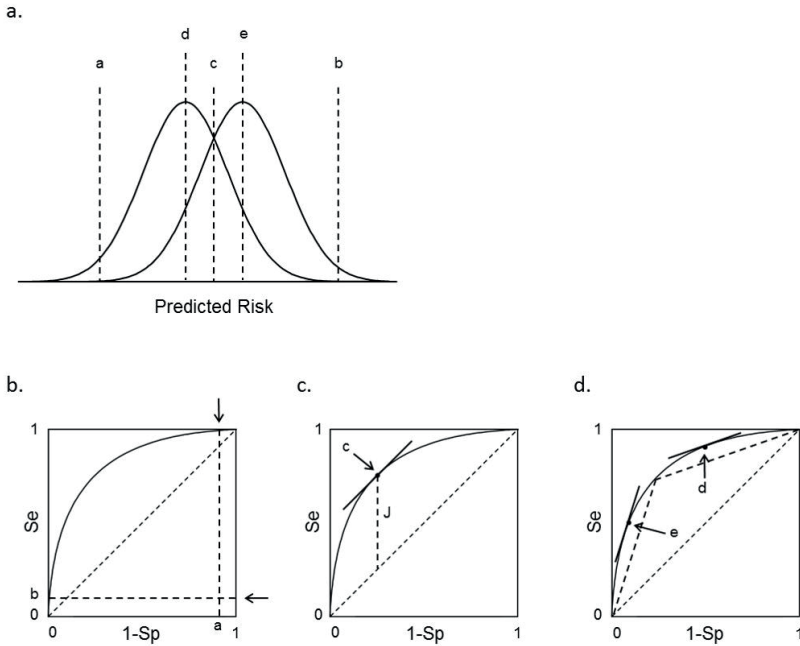
distributions, indicated by the non-overlapping areas, gives a prediction model its discriminative ability: the further the distributions are separated, the better the model can differentiate between the two populations because more diseased individuals have higher risks than the non-diseased.

These risk distributions can also be presented as cumulative distributions, where the y-axis presents the proportion of individuals who have equal or lower predicted risks at each predicted risk (Figure 1b). The separation between the distributions of diseased and non-diseased reflects the same separation as the distributions in Figure 1a. The two non-overlapping areas are now one area, 'connected' at the same predicted risk that separated them in the previous figure. At each predicted risk, if interpreted as a threshold, the proportion of diseased individuals is the sensitivity and the proportion of non-diseased individuals is 1 minus the specificity. Calculating the sensitivity and specificity for every possible risk threshold and plotting them is the best known method for constructing the ROC curve.

In a further transformation, the predicted risks on the x-axis can be replaced by the (cumulative) proportion of non-diseased individuals at each predicted risk (Figure 1c). With this proportion on the x-axis, the distribution of non-diseased individuals is now a diagonal line as its x and y-axes are the same, and the distribution of diseased individuals is the curved line. This transformation shows that the diagonal line is not just a reference line of no discrimination (2), but represents one of the two risk distributions. The difference between the curve and the diagonal line still reflects the separation between the risk distributions in Figure 1a. In a final transformation, the ROC plot is obtained by flipping both axes (Figure 1d).



**Figure 1.** From risk distributions to the receiver operating characteristic (ROC) curve. (a) Risk distributions of diseased and non-diseased individuals. Separation of the distributions creates two nonoverlapping (grey) and one overlapping (white) areas. (b) Cumulative risk distributions. The two nonoverlapping areas are now one area, connected at the same predicted risk that separated them in (a). (c) Transformed cumulative risk distributions. The x-axis presents the proportion of non-diseased individuals ( $p_{nD}$ ) at each predicted risk instead of the predicted risk. The proportion  $p$  equals  $p_D$  for diseased and  $p_{nD}$  for non-diseased individuals. (d) ROC plot. This plot is obtained by reversing both the x-axis and y-axis of (c). The same ROC plot is obtained when the x-axis in (c) has shown the proportion of diseased individuals. Sensitivity (Se) is the percentage of diseased individuals who have predicted risks higher than the threshold ( $1-p_D$ ). Specificity (Sp) is the percentage of non-diseased who have predicted risks lower than the threshold ( $p_{nD}$ ).



**Figure 2.** Inferring the risk distributions of diseased and non-diseased individuals from the receiver operating characteristic (ROC) curve. (a) Risk distributions of diseased (right) and non-diseased individuals (left) with the thresholds that can be inferred from the ROC curve. (b) Thresholds of risks that mark where the risk distributions do and do not overlap. (c) Threshold at which the risk distributions 'cross'. (d) Modus of each risk distribution. Se, sensitivity; Sp, specificity.

## From ROC curve to risk distributions

When the ROC plot is an alternative way of presenting the risk distributions of diseased and non-diseased individuals, it follows that the shapes and overlap of the distributions can be deduced from the ROC curve. This can only approximate the risk distributions; the information is not enough to draw the exact risk distributions on a probability x-axis. This would require the presentation of risk thresholds on the ROC curves or further information about population risk, the effect sizes of individual predictors and calibration.

First, the extremes of the ROC curve represent the tails of the risk distributions: the lowest possible risk threshold is in the upper right corner of the ROC plot and the highest possible threshold in the lower left corner (Figures 2a

and 2b). The ROC curve follows the border of the plot when the risk distributions do not overlap in the tail: the sensitivity remains at 1 (100%) while specificity is gradually increasing until threshold A; and the specificity is at 1 (100%) while sensitivity is still decreasing beyond threshold B. The risk distributions overlap across the entire range of predicted risks when changing the threshold in the tails changes both sensitivity and specificity.

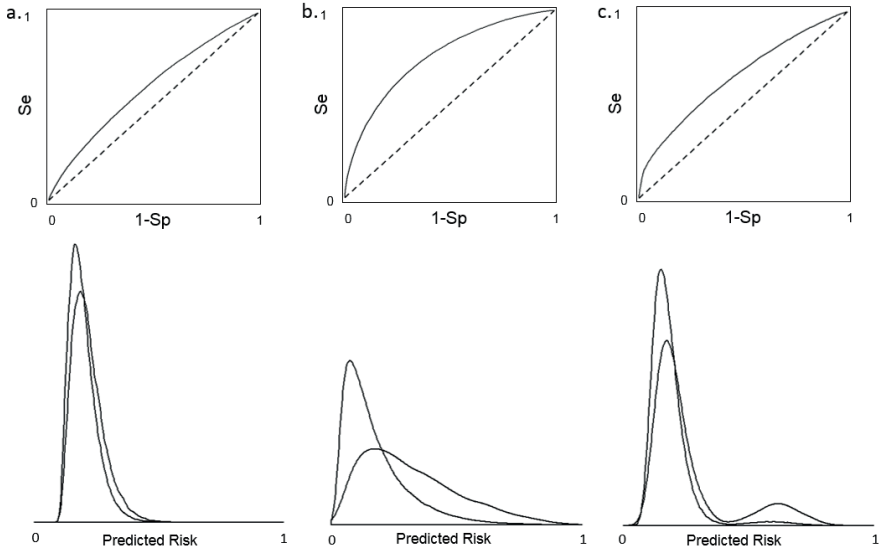
Second, the changes in sensitivity is equal to the change in 1- specificity between all two points on the diagonal line. The tangent line of the ROC curve that runs parallel to the diagonal line (Figure 2c) identifies the threshold where the risk distributions 'cross' (threshold C in Figure 2a). The change in specificity is larger than the change in sensitivity on the left of this threshold and vice versa on the right. This threshold is the one with the highest discriminative ability, where sensitivity + specificity – 1, known as Youden index, has its maximum value (Supplementary Figure 1, available as Supplementary data at *IJE* online) (10). The higher the Youden index, the more the distributions are separated, the higher the AUC.

Third, when we draw straight lines from this 'optimal' threshold to both ends of the ROC curve (Figure 2d), we see that the ROC curve moves away from the straight line and then reconvenes at each end of the ROC curve. The tangent line that runs parallel to each straight line indicates the highest point (modus) of each distribution: at the right (point d) the modus of the non-diseased, and on the left (point e) of the diseased populations. The modus and median are equal when the tangent lines touch the ROC curve where the sensitivity for diseased or the specificity for non-diseased individuals is 0.50 (50%).

Fourth, ROC curves have a 'rounded' shape when prediction models are constructed from continuous variables or binary variables that have similar effects on disease risk (Figure 2), but they may have an 'angle' (Figure 3) when, for example, one binary predictor has a stronger effect on disease risk than all other variables in the prediction model or one category of a categorical variable has a stronger effect on disease risk than the others (11). When ROC curves have an angle, the risk distributions of diseases and non-diseased individuals do not cross there where sensitivity and specificity are equal.

Finally, ROC curves differ in the smoothness of the curve. When a ROC curve is stepped rather than smooth (Figure 4), it may be that the overall sample size of the study is low, that the incidence is low or that the prediction model is based on a relatively small set of categorical predictors that generate a small number predictor combinations.

Figure 5 gives two examples of ROC curves from published empirical studies (12,13). In Figure 5a we see, starting in the lower left corner of the plot, that the ROC curve follows the border until sensitivity is approximately 40%. This pattern is not seen at the upper right corner of the plot. The skewed shape of the curve suggests that there is a categorical predictor that has a strong impact on disease risk that may put 40% of the diseased individuals at higher risk than all non-diseased. In Figure 5b, we see a ROC curve that is stepped. This study had a sample size of only 57 lesions: 28 verruca and 29 clavus lesions. As a result, each verruca and clavus lesion contribute 3% to the sensitivity and specificity. When changing the risk threshold moves one or more lesions to the other side of the threshold, the change in sensitivity or specificity is at least 3%.



**Figure 3.** Rounded and non-rounded shapes of receiver operating characteristic (ROC) curves and their underlying risk distributions. (a), (b) Rounded ROC curve when the prediction model includes continuous variables or multiple categorical variables that have a similar effect on disease risk. (c) ROC curve when (here) one binary predictor has a stronger effect on disease risk than other variables in the model. Se, sensitivity; Sp, specificity.

## Reappraisal of AUC limitations

We explained that the ROC curve is an alternative way of presenting risk distributions and cumulative risk distributions and that the diagonal line is not

merely a reference line but it is the risk distribution of non-diseased individuals (Figure 1). The separation of the risk distributions is indicated by the area between the ROC curve and the diagonal: the larger the area, the more separation between the distributions and the higher the discriminative ability. The size of the area is related to Somers'  $D$  (14), a non-parametric rank correlation that can be used to obtain the AUC as  $(D+1)/2$  (15).

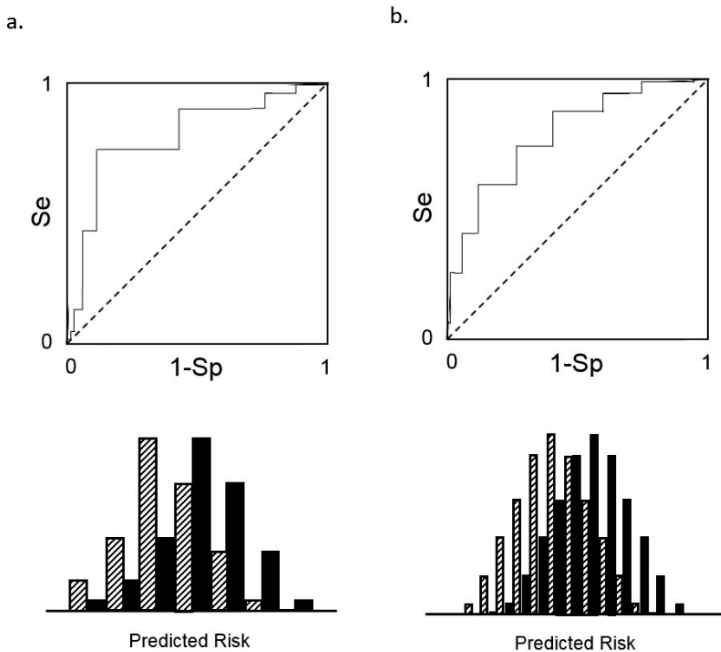
When the ROC plot is nothing more than an alternative graphical presentation of risk distributions, it follows that the ROC curve does not need to assume risk thresholds. The ROC curve can be used to determine the sensitivity and specificity of a single risk threshold, but this does not need to be its primary and only interpretation. The risk distributions of diseased and non-diseased individuals and the separation between them are relevant for prediction models, irrespective of the number of thresholds that is considered.

The AUC is commonly described as the probability that a random individual from the diseased population is more likely to have a higher predicted risk than a random individual from the non-diseased population. This explanation still holds: this probability is higher when the risk distributions are further separated. These random individuals can be considered as pairs, which is how the AUC value is calculated from Somers'  $D$  (14), but the consideration of pairs is not essential or required for the interpretation of the AUC.

AUC has been criticized for being insensitive to detect improvements in the prediction that result from adding risk factors with stronger effects (7-9, 16). As the ROC curve is nothing more than an alternative presentation of the risk distributions, it follows that this insensitivity is not a limitation of the metric: when a predictor does not change the ROC, it does not change the underlying risk distributions. Improving prediction models requires adding common predictors with strong impact on disease risk to further separate the risk distributions, which is difficult especially when prediction models have higher 'baseline' AUC and their risk distributions are already separated. When adding predictors does not improve the AUC, it means that the ROC curves of the baseline and updated models are virtually the same. Adding the predictors may have changed the predicted risks, and individuals may have moved between risk categories, but each sensitivity comes with the 'same' specificity and vice versa. That said, AUC is a metric for the big picture. The metric is unable to detect the improved prediction due to rare risk factors with strong effects. When changes in predicted risks are of interest, other metrics such as the integrated discrimination

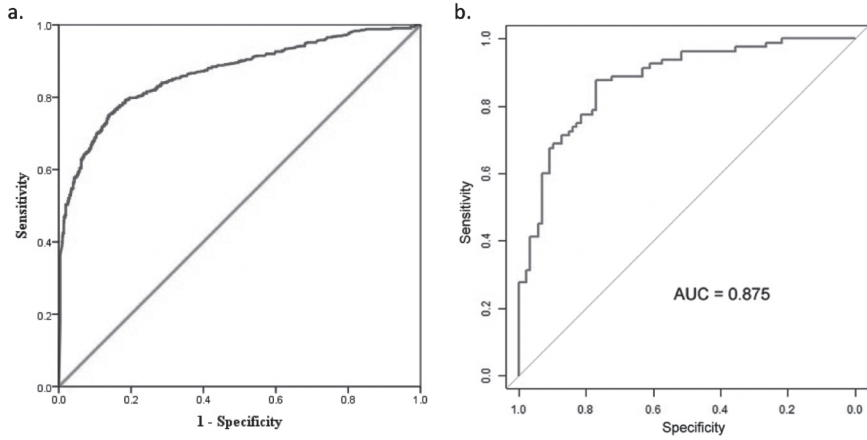
improvement (IDI) or Brier score need to be considered (17).

Finally, the criticism that the AUC lacks clinical relevance and omits the consideration of costs and harms in weighing false-positives against false-negatives (18, 19) is valid, but concerns the inappropriate use of the measure rather than its shortcomings. The AUC is a measure of the discriminative ability of a prediction model or continuous test in a certain population, quantifying the separation of the risk distributions of diseased and non-diseased individuals. It is not a measure of utility. For some clinical applications, an AUC of 0.65 will be high enough, whereas for others 0.90 might be too low. Also, the optimal threshold on the ROC curve (Youden index) may be irrelevant and suboptimal from a clinical perspective.



**Figure 4.** Examples of ‘stepped’ receiver operating characteristic (ROC) curves and their underlying risk distributions. ROC curve when overall sample size or incidence is low. Se, sensitivity; Sp, specificity.

The decision whether a prediction model is useful to guide medical decisions is not determined by its discriminative ability alone, but requires additional evaluations such as the prevalence, predictive value, the decision impact of the test results, the implications of false-positive and false-negative results, and others.



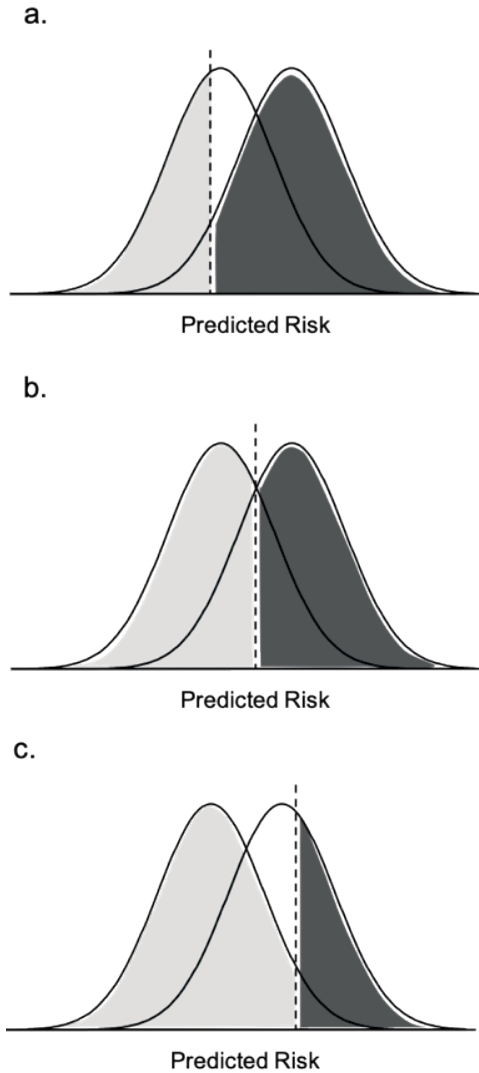
**Figure 5.** Examples of empirical ROC curves. Receiver operating characteristic curves for (a) the diagnosis of hepatitis B virus infection-related hepatocellular carcinoma using a serum marker, reprinted under Creative Commons license CC BY 3.0 from Yao et al. 2016 (12) and b) a predictive model for differentiating between two skin diseases, verruca and clavus, using electrical impedance indices, reprinted under Creative Commons license CC BY 4.0 from Hung et al. 2014 (13).



## References

1. Lusted LB. Decision-making studies in patient management. *N Engl J Med.* 1971;284(8):416-424.
2. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology.* 1982;143(1):29-36.
3. Parikh CR, Thiessen-Philbrook H. Key Concepts and limitations of statistical methods for evaluating biomarkers of kidney disease. *J Am Soc Nephrol.* 2014;25(8):1621-1629.
4. Pepe MS, Janes HE. Gauging the performance of SNPs, biomarkers, and clinical factors for predicting risk of breast cancer. *J Natl Cancer Inst.* 2008;100(14):978-979.
5. Halligan S, Altman DG, Mallett S. Disadvantages of using the area under the receiver operating characteristic curve to assess imaging tests: A discussion and proposal for an alternative approach. *Eur Radiol.* 2015;25(4):932-939.
6. Flach P. ROC Analysis. In: Sammut C, Webb G, editors. *Encyclopedia of machine learning*; Springer US; 2010;869-875.
7. Pencina MJ, D'Agostino Sr. RB, D'Agostino Jr. RB, et al. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med.* 2008;27(2):112-157.
8. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation.* 2007;115(7):928-935.
9. Ware JH. The limitations of risk factors as prognostic tools. *N Engl J Med.* 2006;355(25):2615-2617.
10. Youden WJ. Index for rating diagnostic tests. *Cancer.* 1950;3(1):32-5.
11. Kundu S, Kers JG, Janssens AC. Constructing hypothetical risk data from the area under the ROC curve: Modelling distributions of polygenic risk. *PLoS One.* 2016;11(3):e0152359.
12. Yao M, Zhao J, Lu F. Alpha-fetoprotein still is a valuable diagnostic and prognosis predicting biomarker in hepatitis B virus infection-related hepatocellular carcinoma. *Oncotarget.* 2016;7(4):3702-3708.
13. Hung CY, Sun PL, Chiang SJ, Jaw FS. In vitro differential diagnosis of clavus and verruca by a predictive model generated from electrical impedance. *PLoS One.* 2014;9(4):e93647.
14. Somers RH. A new asymmetric measure of association for ordinal variables. *Am Sociol Rev.* 1962;27(6):799-811.
15. Steyerberg E. *Clinical Prediction Models - A practical approach to development, validation, and updating.* New York, NY: Springer US; 2009.
16. Pepe MS. Limitations of the Odds Ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *Am J of Epidemiol.* 2004;159(9):882-890.
17. Austin PC, Steyerberg EW. Predictive accuracy of risk factors and markers: a simulation study of the effect of novel markers on different performance measures for logistic regression models. *Stat Med.* 2013;32(4):661-672.
18. Hand DJ. Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Mach Learn.* 2009;77(1):103-23.
19. Samawi HM, Yin J, Rochani H, Panchal V. Notes on the overlap measure as an alternative to the Youden index: How are they related? *Stat Med.* 2017;36(26):4230-4240.

## Supplementary data



**Supplementary Figure 1.** Risk distributions and Youden index. Sensitivity (light grey) and specificity (dark grey) for different risk thresholds, showing that sensitivity + specificity is optimal when the threshold is where the risk distributions 'cross' (Figure b). Youden index is sensitivity + specificity - 1.



# 4

## **Small improvement in the area under the receiver operating characteristic curve indicated small changes in predicted risks**

Forike K. Martens, Elisa C. M. Tonk, Jannigje G. Kers, and A. Cecile J.W. Janssens  
*Journal of Clinical Epidemiology* 79 (2016) 159-164

## Abstract

**Objective** Adding risk factors to a prediction model often increases the area under the receiver operating characteristic curve (AUC) only slightly, particularly when the AUC of the model was already high. We investigated whether a risk factor that minimally improves the AUC may nevertheless improve the predictive ability of the model, assessed by integrated discrimination improvement (IDI).

**Study Design and Setting** We simulated data sets with risk factors and event status for 100,000 hypothetical individuals and created prediction models with AUCs between 0.50 and 0.95. We added a single risk factor for which the effect was modeled as a certain odds ratio (OR 2, 4, 8) or AUC increment ( $\Delta$ AUC 0.01, 0.02, 0.03).

**Results** Across all AUC values of the baseline model, for a risk factor with the same OR, both  $\Delta$ AUC and IDI were lower when the AUC of the baseline model was higher. When the increment in AUC was small ( $\Delta$ AUC 0.01), the IDI was also small, except when the AUC of the baseline model was  $> 0.90$ .

**Conclusion** When the addition of a risk factor shows minimal improvement in AUC, predicted risks generally show minimal changes too. Updating risk models with strong risk factors may be informative for a subgroup of individuals, but not at the population level. The AUC may not be as insensitive as is frequently argued.

## Introduction

The area under the receiver operating characteristic (ROC) curve (AUC, or c-statistic) is the most commonly used metric to evaluate prediction models for their ability to discriminate between individuals with and without an event, and improvement in AUC ( $\Delta$ AUC) is the standard for assessing the value of adding new risk factors to prediction models (1-4). Yet,  $\Delta$ AUC has been criticized for being insensitive to detect improvements in prediction that result from adding clinically established risk factors (2, 5-9).

In recent years, researchers have widely adopted novel measures such as the net reclassification improvement (NRI) and integrated discrimination improvement (IDI) (2, 5). These measures can produce statistically significant results even when  $\Delta$ AUC is small, which may explain their popularity. Inferences about improvement in prediction are generally based on the statistical significance of the NRI and IDI, and the small absolute values for the measures are often ignored (article in preparation).

The argument that AUC is insensitive suggests that there may be improvements in the predictive performance that are not detected by AUC. AUC is a measure of discrimination. It considers the rank of the predicted risks of individuals who will develop an event and those who will not, not the absolute values of the predicted probabilities. A measure that does focus on changes in the absolute predicted risks is the IDI (2). IDI quantifies the improvement of the risk difference between individuals with and without an event that results from adding risk factors to a prediction model (2). In the case of a risk factor that adds to a model's predictive ability, the risk factor increases predicted risks for individuals who will develop the event and decreases predicted risks for those who will not, leading to a larger risk difference, and hence a positive IDI value. However, the absolute value of IDI is strongly determined by the overall event rate in the population (10), which hampers a clear and uniform interpretation of IDI across studies with different event rates.

Previous studies have shown that both  $\Delta$ AUC and IDI are higher when the effect size of the added risk factor is higher (10-12). However, little is known about the size of IDI in the situation when  $\Delta$ AUC is small, for example, lower than 0.01, at which it is generally concluded that the discriminative ability of the model is not improved (10). When AUC is insensitive, as is argued (2, 5-9), small values of  $\Delta$ AUC might still go together with IDI values that are of clinical interest.

In other words, there might be worthy improvement in predicted risks that is not apparent from the small values of  $\Delta\text{AUC}$ .

In this paper, we investigate whether and when the addition of risk factors may show minimal improvement in discrimination ( $\Delta\text{AUC}$ ) but have a major impact on the predictive ability, which we define as the difference in predicted risks between individuals who will develop the event and those who will not, assessed by IDI. Using simulated data, we assessed the improvement in AUC and IDI for prediction models that were updated by adding a single risk factor for which we varied the frequency and effect size.

## Methods

This study was conducted using simulated data, which allows us to vary the parameters that determine the predictive performance. We created datasets of risk factors and event status and constructed baseline prediction models that we updated by adding a single risk factor. Between scenarios, we varied the AUC of the baseline prediction models, the population event incidence, and the frequency and effect size of the added risk factor.

### Data simulation

To construct simulated datasets, we used a modeling procedure that has been described in detail elsewhere (11, 13). In short, the procedure creates a dataset of risk factors for a hypothetical population of 100,000 individuals. Risk factors were assigned in such a way that their frequencies matched prespecified values. By changing the number, frequency and odds ratios (ORs) of simulated risk factors, we created baseline models with an AUC ranging between 0.50 and 0.95 (see the following). We then added a single binary risk factor for which we varied the frequency and OR between scenarios. Event status was simulated on the basis of event probabilities, which were estimated using Bayes' theorem using the ORs and frequencies of the risk factors. Bayes' theorem specifies that the posterior odds of developing an event is obtained by multiplying the prior odds by the likelihood ratios of the individual's status on all risk factors. Events were then assigned based on a procedure that for each individual compared the calculated probability to a randomly drawn value between 0 and 1 from a uniform distribution. An individual was assigned to develop an event when the

Small improvement in the area under the receiver operating characteristic curve indicated small changes in predicted risks

probability was higher than the random value and to not develop an event when it was lower than the random value. This procedure ensures that the percentage of individuals who experience an event closely approximates the predicted event risk at each level of risk. This means that, for example, approximately 60% of the individuals with a predicted baseline event probability of 60% experience the event in our simulated data. Predicted risks for each individual for the baseline and updated models were obtained using logistic regression analysis, and rounded to two decimal points.

To obtain the risk data for a specific value of the AUC, we used an iterative procedure in which we added as many genetic variants until the AUC of the prediction model reached a prespecified value (14). To this end, we calculated predicted risks using Bayes' theorem, assigned disease status and obtained the AUC of the prediction model after each variant added, as described previously. The procedure was stopped when the AUC value exceeded the prespecified value, and then the risk distribution for which the AUC value was closest to the prespecified value was considered. The AUC value, the population disease risk, the ORs, and frequencies of the risk alleles that were used to construct the risk distributions, were varied between scenarios.

### **Statistical analyses**

$\Delta$ AUC was calculated as the difference between the AUC of the baseline and updated models, with AUC assessed using the c-statistic (4). IDI was calculated as improvement in the risk difference of mean predicted risks in individuals with and without events between the baseline model and the updated model (2).

To understand the relationship between AUC and IDI, we first show how the OR of the added risk factor affects both metrics. We show the relationship between  $\Delta$ AUC and the OR of the added risk factor across different AUC values of the baseline prediction model (8, 15). We show  $\Delta$ AUC when the baseline models are updated with a single risk factor, for which we considered fixed ORs of 2, 4, and 8 in separate scenarios. We also show the value of the OR that was needed to improve the prediction model by different levels of  $\Delta$ AUC, for which we considered  $\Delta$ AUC of 0.01, 0.02, and 0.03. For each value of the baseline AUC, we added a risk factor for which we increased the OR until the specified  $\Delta$ AUC ( $\pm 0.0025$ ) was reached. ORs were increased by increments of 0.1, and by 0.01 when  $\Delta$ AUC was close to the lower end of the range. For example, to simulate a  $\Delta$ AUC of 0.01, we increased the OR until the observed  $\Delta$ AUC was



between 0.0075 and 0.0125.

Next, we show the relationship between IDI and the OR of the added risk factor across different AUC values of the baseline model (8, 15). The absolute value of IDI is determined by the overall event rate (e.g., disease incidence) in the population: larger differences between subjects with and without events will be observed for the same OR when event rates are higher. In addition, when the frequency of the risk factor is higher than the event rate, we will find that also subjects without the event will have the risk factor and that the risk difference will be smaller. For these reasons, we varied the event rate and the frequency of the added risk factor between scenarios. The event rates were 5%, 10% or 20%, and the risk factor frequencies were set at half, equal or twice the value of the event rate to reflect scenarios in which the added risk factor was half, equal or more twice as frequent as the event. For example, when the event rate was 5%, the frequency of the added risk factor was 2.5%, 5%, and 10% in separate scenarios.

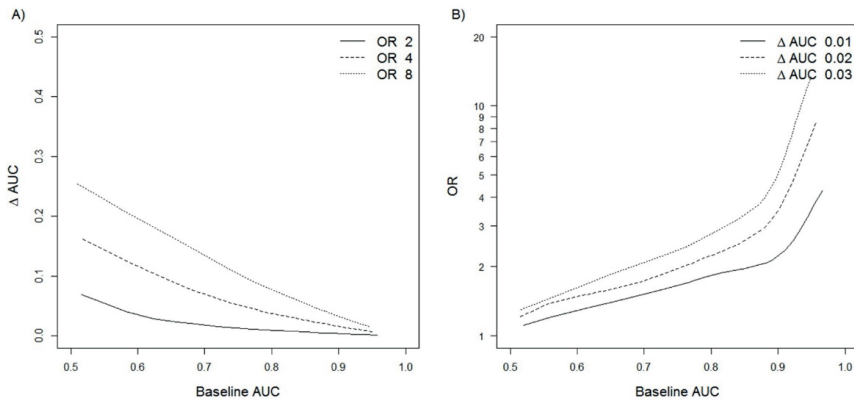
To investigate IDI for fixed increments of AUC, we constructed baseline risk models for which the AUC ranged from 0.50 to 0.95. We added a single binary risk factor to each risk model for which we varied the ORs so that  $\Delta$ AUC was 0.01, 0.02, and 0.03. In separate scenarios, we additionally varied the event rate and the frequency of the added risk factor, in a similar way as specified previously. All analyses were performed using R software version 3.1.0 (R-project.org) (16).

Finally, to investigate the assessment of  $\Delta$ AUC and IDI empirically, we conducted a literature review of genetic prediction studies in which both  $\Delta$ AUC and IDI were assessed. Using Thomson Reuters Web of Knowledge (version 5.17, Thomson Reuters, Philadelphia, USA) we retrieved all publications that cited the article by Pencina et al. that first introduced IDI (search date 28 April, 2015) (2). We limited our analysis to empirical studies that had investigated the addition of one or more single-nucleotide polymorphisms (SNPs) to a model containing clinical risk factors. The search yielded 1,962 unique publications (Appendix Fig. 1 at [www.jclinepi.com](http://www.jclinepi.com)), of which 40 reported both  $\Delta$ AUC and IDI. Seven studies were excluded because they only reported the *P*-value for the IDI without the IDI value. Thirty-three studies were included in our analysis.

Small improvement in the area under the receiver operating characteristic curve indicated small changes in predicted risks

## Results

Figure 1a shows how  $\Delta AUC$  declines with increasing baseline AUC when the effect size of the added risk factor is held constant. The OR of the added risk factor needed to be higher at higher baseline AUC to achieve a specific  $\Delta AUC$ , particularly when the AUC of the baseline model exceeded 0.90 (Figure 1b). For example, to increase the AUC by 0.03 ( $\Delta AUC$  0.03), the OR of the added risk factor needed to be 1.7 when the baseline AUC was 0.61 and 5.0 when the baseline AUC was 0.90. The exact values of  $\Delta AUC$  and the OR varied some with the event rate and the risk factor frequency (data not shown).



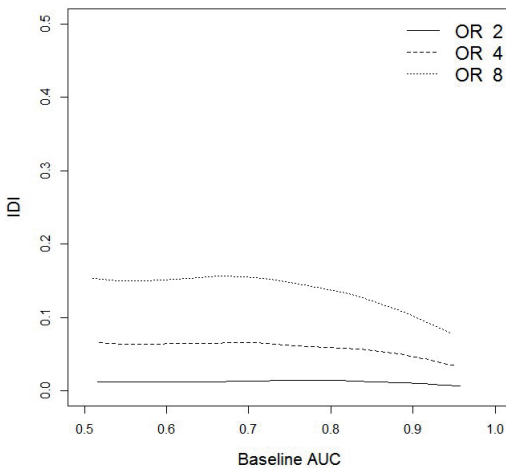
**Figure 1.** Relationship between the odds ratio (OR) of the added risk factor and the improvement in the area under the receiver operating characteristic curve ( $\Delta AUC$ ) by AUC value of the baseline prediction model. (A) Observed  $\Delta AUC$  for fixed values of OR and (B) Required OR for fixed values of  $\Delta AUC$ .  $\Delta AUC$  and odds ratios were calculated for scenarios in which the event rate was 10%, and the frequency of the added risk factor was 20%.

Figure 2 shows how IDI decreased with increasing baseline AUC. Where  $\Delta AUC$  showed a “linear” decline with increasing baseline AUC (Figure 1a) the IDI was constant for most values of baseline AUC. For higher values of ORs, this constant IDI was observed when the frequency of the risk factor was higher than the rate of the event (Appendix Figure 2 at [www.jclinepi.com](http://www.jclinepi.com)). As expected, the absolute values of IDI varied with the event rate and the frequency of the added risk factor (see Appendix Figure 2 at [www.clinepi.com](http://www.clinepi.com)).

When the effect of updating the prediction model was fixed in terms of  $\Delta AUC$ , we observed a larger IDI when the AUC of the baseline model was

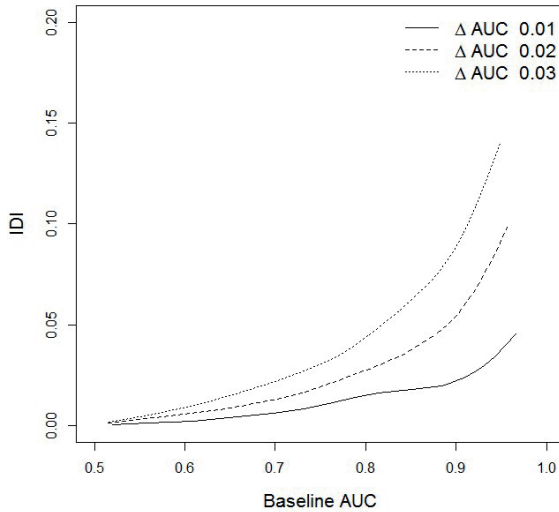
higher (Figure 3). This relation was observed across all scenarios in which the event rate and the frequency of the added risk factor were varied (Appendix Figure 3 at [www.jclinepi.com](http://www.jclinepi.com)). The higher IDI is explained by the fact that the risk factor needed to have a larger OR to yield the same  $\Delta$ AUC at higher levels of baseline AUC (Figure 1b). Yet, IDI remained low across baseline AUCs when  $\Delta$ AUC was 0.01. For example, when the AUC of the baseline model was 0.90 and the event rate was 10%, IDI was 0.02, indicating that the risk differences between individuals with and without events increased by 0.02.

Figure 4 shows a scatterplot of IDI by  $\Delta$ AUC values from 33 genetic prediction studies that reported the extension of clinical prediction model with one or more SNPs. The figure shows that IDI values tended to be higher for higher values of  $\Delta$ AUC. Yet, when  $\Delta$ AUC was lower than 0.01, IDI was also lower than 0.01, indicating a 1% absolute increase in the risk difference between individuals with and without events.

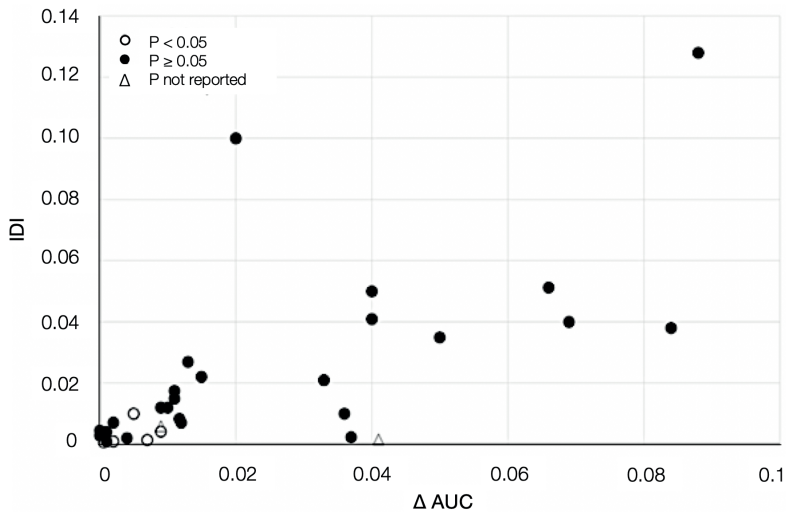


**Figure 2.** Relationship between the odds ratio (OR) of the added risk factor and integrated discrimination improvement (IDI) by the area under the receiver operating characteristic curve (AUC) of the baseline prediction model. IDI was calculated for scenarios in which the event rate was 10% and the frequency of the added risk factor was 20%.

Small improvement in the area under the receiver operating characteristic curve indicated small changes in predicted risks



**Figure 3.** Integrated discrimination improvement (IDI) for fixed increments in the area under the receiver operating characteristic curve ( $\Delta AUC$ ) by AUC value of the baseline prediction model. In all scenarios, the event rate was 10% and the frequency of the added risk factor was 20%.



**Figure 4.** Increment in the area under the receiver operating characteristic curve ( $\Delta AUC$ ) and integrated discrimination improvement (IDI) in empirical studies on genetic prediction of multifactorial diseases.

## Discussion

This study showed that adding risk factors to prediction models may improve the predictive ability when the increase in AUC is minimal, but only when the AUC of the baseline model was high ( $AUC > 0.90$ ). In the range of commonly observed AUC values, those between 0.60 and 0.80, a small increase in AUC ( $\Delta AUC < 0.01$ ) was accompanied with small improvements in predictive ability both in the simulation analyses and the review of empirical studies.

The aim of our paper was to investigate whether updating risk models may change predicted risks in the absence of an apparent improvement in AUC. The most basic metric that assesses changes in predicted risks on the group level is the comparison of the risk differences between individuals who will develop an event and those who will not, before and after updating the risk model. This change in predicted risks is indicated by the IDI. In absence of a change in predicted risks, and hence absence of improved risk differences, all other metrics that operate from the absolute predicted risks will show effectively zero improvement, such as Brier score. When updating does not improve AUC all combinations of sensitivity and specificity will remain unchanged, and their ROC curves will overlap perfectly. This also means that other measures that depend on sensitivity and specificity, such as decision curve analysis and net benefit, will also show that updating of the model will result in no improvement.

AUC has been criticized for being an insensitive metric to evaluate improvement in predictive performance (2, 5, 8, 9), as the measure was unable to detect improvements that result from adding risk factors, even those with strong effects. In line with earlier studies we found that the effect size of a risk factor needs to be higher to improve the discriminative ability further when prediction models have a higher baseline AUC (8-12). Rather than concluding that AUC is an insensitive measure, it seems more justified to infer that prediction models with higher discriminative ability at baseline need stronger risk factors to improve further.

In line with previous studies, we also observed that IDI varied with the effect size and frequency of the added risk factor as well as with the event rate (10-12). Where an earlier study suggested that IDI was relatively constant across values of baseline AUC (10), we found that IDI clearly declined (Figure 2 and Appendix Figure 2 at [www.jclinepi.com](http://www.jclinepi.com)). This might be explained by the fact that we considered higher AUC values for the baseline prediction models,

Small improvement in the area under the receiver operating characteristic curve indicated small changes in predicted risks

up to AUC 0.95 as compared to 0.90 (10).

Pepe et al. (17) demonstrated that measures for evaluating the value of adding risk factors to a prediction model have the same null hypothesis. This does not, however, mean that the measures can be used interchangeably (12). When two measures have the same null hypothesis, such as  $\Delta$ AUC and IDI, they still assess different aspects of model performance.  $\Delta$ AUC quantifies improvement in discrimination and IDI assesses improvement in predictive ability. If risk factors improve discrimination, they will also improve the predictive ability, but they inform about different aspects of model performance. In the evaluation of genetic and diagnostic tests, this distinction is commonly accepted: sensitivity and specificity are always assessed together with measures of predictive value such as penetrance or positive and negative predictive values. Statistically, to assess whether a risk factor adds information to a risk model, the likelihood ratio test is preferred (15, 17).

The most important finding from this study is that strong risk factors added to prediction models with higher baseline AUC that did not yield substantial improvement of AUC can improve predictions of risk, as assessed by the IDI. This means that adding risk factors to models that already have excellent discrimination does not further improve the discriminative ability but can improve the predictive ability: the risk difference between individuals who will develop the event and those who will not may become larger when models with high AUC are updated with strong risk factors. However, when the increment in AUC of the prediction models was small ( $\Delta$ AUC = 0.01), the changes in predicted risk were also small across all AUC values of the baseline model. The same was observed in the empirical genetic prediction studies (see Figure 4). When AUC improves by 0.01, which is commonly observed in empirical studies, the effect of added risk factor(s) was not strong enough to improve the model's discriminative and predictive ability.

The reason why adding strong risk factors can show improvement in the predictive ability and only minimal improvement in discrimination is that, at higher levels of baseline AUC, many individuals who will experience an event already have higher predicted risks than many of those who will not. Because AUC is a rank order test, which compares the average ranks of individuals with and without events, increasing the predicted risks of individuals who will experience an event and were already at higher risk or further decreasing the risks of those who will not experience an event has no or limited impact on the

ranking. In contrast, IDI specifically measures these changes in predicted risks and may observe improved prediction not indicated by  $\Delta\text{AUC}$ .

Measures of predictive performance are used to assess and identify the best prediction model for making medical decisions and informing patient in a certain health care context. The intended use of the model sets the standard to decide whether the best model is predictive enough but also which measure(s) should be used for the assessment of predictive performance and its improvement. AUC and IDI, and other measures such as NRI, are complementary, they each assess a different aspect predictive performance, and their results may not evidently lead to the same inferences about its improvement. For example, when the model is used to identify risk groups, the interest is in improving the sensitivity and specificity at various risk thresholds. Thus, the updated model should yield more favorable combinations of sensitivity and specificity, which is indicated by a change in the ROC curve and by a positive  $\Delta\text{AUC}$ . When  $\Delta\text{AUC}$  is small, then the updated model does not perform markedly better. Yet, when the model is used to inform individual patients, the interest is in improving predicted risks, which is indicated by IDI.

When IDI is used to assess the predictive performance improvement, the value may be statistically significant even when the improvement in AUC is minimal. When  $\Delta\text{AUC}$  was lower than 0.01 in the empirical studies of the literature review, IDI values were statistically significant in 7 out of 14 studies (see Appendix Table 1 at [www.jclinepi.com](http://www.jclinepi.com)). Yet, the improvement in performance should not be concluded from the statistical significance of IDI (15) but from its absolute value. The essential question is whether risk factors meaningfully improve the model's predictive ability. What degree of improvement is clinically relevant varies between scenarios. In large studies, small values of IDI may be statistically significant, but not relevant. Furthermore, note that the absolute value of IDI is affected by the event rate (10). Given the same baseline AUC and  $\Delta\text{AUC}$ , IDI will be higher for events that are more common. This dependence on the event rate hampers a clear and uniform interpretation of IDI across populations with different event rates. What level of IDI is clinically relevant depends on the specific health care scenario and by the question what is to be gained from the additional information. Note that only three studies used validation data or a validation approach (bootstrapping/cross validation) for the assessment of predictive performance (Appendix References 15, 16, 21 at [www.jclinepi.com](http://www.jclinepi.com)). Overfitting of the risk models in derivation data may lead to overestimation of the

Small improvement in the area under the receiver operating characteristic curve indicated small changes in predicted risks

improvement in predictive performance.

Our study showed that there were no major improvements in predicted risks when improvements in discrimination were minimal. Only for prediction models with exceptionally high discriminative accuracy, predictive ability may have improved when increments in AUC are small. In all other instances, small improvements in AUC indicate small changes in predicted risks. The AUC may not be as insensitive as is frequently argued.



## References

1. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;143(1):29-36.
2. Pencina MJ, D'Agostino Sr. RB, D'Agostino Jr. RB, et al. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med*. 2008;27(2):112-157.
3. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*. 2010;21(1):128-138.
4. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*. 1983;148(3):839-843.
5. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation*. 2007;115(7):928-935.
6. Pencina MJ, D'Agostino Sr. RB, Demler OV. Novel metrics for evaluating improvement in discrimination: net reclassification and integrated discrimination improvement for normal variables and nested models. *Stat Med*. 2012;31(2):101-113.
7. Pepe MS, Janes HE. Gauging the performance of SNPs, biomarkers, and clinical factors for predicting risk of breast cancer. *J Natl Cancer Inst*. 2008;100(14):978-979.
8. Pepe MS. Limitations of the Odds Ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *Am J Epidemiol*. 2004;159(9):882-890.
9. Ware JH. The limitations of risk factors as prognostic tools. *N Engl J Med*. 2006;355(25):2615-2617.
10. Pencina MJ, D'Agostino RB, Pencina KM, Janssens AC, Greenland P. Interpreting incremental value of markers added to risk prediction models. *Am J Epidemiol*. 2012;176(6):473-481.
11. Mihaescu R, Pencina MJ, Alonso A, Lunetta KL, Heckbert SR, Benjamin EJ, et al. Incremental value of rare genetic variants for the prediction of multifactorial diseases. *Genome Med*. 2013;5(8):76.
12. Kerr KF, Bansal A, Pepe MS. Further insight into the incremental value of new markers: the interpretation of performance measures and the importance of clinical context. *Am J Epidemiol*. 2012;176(6):482-487.
13. Janssens ACJW, Aulchenko YS, Elefante S, Borsboom GJJM, Steyerberg EW, van Duijn CM. Predictive testing for complex diseases using multiple genes: Fact or fiction? *Genet Med*. 2006;8(7):395-400.
14. Kundu S, Kers JG, Janssens ACJW. Constructing hypothetical risk data from the area under the ROC curve: Modelling distributions of polygenic risk. *PLoS ONE*. 2016;11(3):e0152359.
15. Kerr KF, McClelland RL, Brown ER, Lumley T. Evaluating the incremental value of new biomarkers with integrated discrimination improvement. *Am J Epidemiol*. 2011;174(3):364-374.
16. R Core Development Team. *R: A language and environment for statistical computing*. 3.1.0 ed. Vienna, Austria: R Foundation for Statistical Computing; 2015.
17. Pepe MS, Kerr KF, Longton G, Wang Z. Testing for improvement in prediction model performance. *Stat Med*. 2013;32(9):1467-1482.

Small improvement in the area under the receiver operating characteristic curve indicated small changes in predicted risks

## Supplementary data

**Appendix Table 1.** Details of the empirical studies from the literature review.

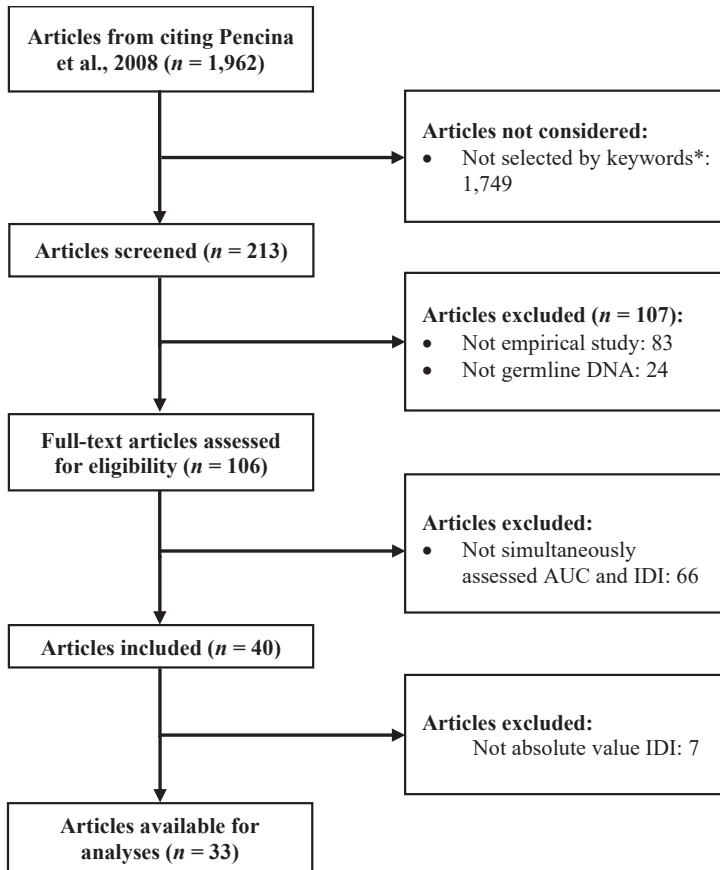
| Study              | Sample Size | Event Rate (%) | Outcome               | Clinical Risk Factors   | Number Of SNPs | Baseline AUC | $\Delta$ AUC | IDI (P value or 95% CI)   |
|--------------------|-------------|----------------|-----------------------|---|----------------|--------------|--------------|---------------------------|
| Eriksson (1)       | 9,704       | 46             | Fracture              | Age, height, weight, femoral neck BMD   | 63             | 0.63         | 0            | 0.003 (0.007)             |
| Fava (2)           | 27,003      | 13             | CVD and CVD mortality | Age, diabetes, smoking, hypertension, BMI, antilipemic medication   | 29             | 0.743        | 0            | 0.00449 (0.02)            |
| Havulinna (3)      | 17,954      | 7              | CVD                   | Diabetes, smoking, SBP, total cholesterol, HDL cholesterol, lipid-lowering medication   | 32             | 0.8445       | 0.0006       | 0.00062 (0.14)            |
| Gränsbo (4)        | 24,777      | 11             | CVD                   | Age, sex, diabetes, smoking, hypertension, BMI, lipid-lowering medication   | 1              | 0.750        | 0.001        | 0.001 (<0.001)            |
| Ripatti (5)        | 13,292      | 5              | CHD                   | Sex, diabetes, smoking, BMI, SBP, DBP, HDL and LDL cholesterol, antihypertensive medication   | 13             | 0.871        | 0.001        | 0.004 (0.0006)            |
| Thanassoulis (6)   | 3,014       | 6              | Hard CHD              | Age, sex, diabetes, smoking, SBP (and treatment), total cholesterol, HDL cholesterol, parental history of CVD   | 13             | 0.820        | 0.002        | 0.001 (-0.001 to 0.003)   |
| Mühlenbruch (7)    | 2,500       | 23             | Diabetes              | Age, smoking, hypertension, height, waist circumference, HDL cholesterol, triglycerides, alcohol intake, physical activity, intake of red meat, intake of wholegrain bread, coffee consumption, glucose, A1C, $\gamma$ -glutamyltransferase, alanine aminotransferase | 42             | 0.899        | 0.002        | 0.0071 (0.0069 to 0.0072) |
| Hivert (8)         | 2,843       | 56             | Diabetes              | Age, sex, waist circumference, ethnic background, medication  | 34             | 0.628        | 0.003        | -0.001 (0.38)             |
| Brautbar 2009 (9)  | 9,998       | 13             | CHD                   | Age, sex, diabetes, smoking, SBP, total cholesterol, HDL cholesterol, antihypertensive medication   | 1              | 0.782        | 0.004        | 0.002 (<0.015)            |
| Muehlschlegel (10) | 845         | 11             | All-cause mortality   | Age, sex, CPB duration, institution, preoperative measures, pulmonary disease, coronary stenosis  | 1              | 0.777        | 0.005        | 0.010 (0.053)             |

| Study              | Sample Size | Event Rate (%) | Outcome                            | Clinical Risk Factors  | Number Of SNPs | Baseline AUC | $\Delta$ AUC | IDI (P value or 95% CI) |
|--------------------|-------------|----------------|------------------------------------|--|----------------|--------------|--------------|-------------------------|
| Wauters (11)       | 2,009       | 13             | Recurrent MI                       | Age, sex, diabetes, smoking, SBP, total cholesterol, history of angina, prior MI, heart rate, initial serum creatinine level, elevated initial cardiac enzymes, ST segment depression, Killip class, PCI, CABG, thrombolytics, participating country | 1              | 0.637        | 0.007        | 0.0015 (0.15)           |
| Juhola (12)        | 2,119       | 34             | Adult hypertension                 | Age, sex, parental hypertension, childhood overweight/obesity status, childhood SBP, parental occupational status  | 29             | 0.733        | 0.009        | 0.012 (<0.0001)         |
| Bacci (13)         | 737         | 24             | CVD                                | Age, sex, diabetes, smoking, hypertension, BMI   | 1              | 0.704        | 0.009        | 0.0042 (0.16)           |
| Brautbar 2012 (14) | 8,542       | 13             | CHD                                | Age, sex, diabetes, smoking, SBP, total cholesterol, HDL cholesterol, antihypertensive medication  | 13             | 0.742        | 0.009        | 0.006                   |
| Lin (15)           | 5,360       | 7              | Diabetes                           | Age, sex, WHR, family history of diabetes, triacylglycerol/HDL cholesterol ratio, physical activity  | 15             | 0.86         | 0.01         | 0.012 (0.0003)          |
| Butoescu (16)      | 316         | 54*            | Prostate cancer                    | PSA, biopsy, prostate volume, DRE, and TRUS  | 9              | 0.770        | 0.011        | 0.015 (0.035)           |
| Yu (17)            | 320         | 37             | Biochemical recurrence             | PSA, stage, Gleason score, surgical margin   | 1              | 0.783        | 0.011        | 0.0175 (0.041)          |
| Zheng (18)         | 6,121       | 50*            | Breast cancer                      | Age at first live birth, age at menarche, WHR, breast cancer family history, previous benign breast disease diagnosis  | 8              | 0.6295       | 0.0117       | 0.0084 (<0.0001)        |
| Lind (19)          | 1,016       | NR             | Endothelium-dependent vasodilation | Sex, smoking, BMI, SBP, HDL and LDL cholesterol, triglycerides, antihypertensive medication, statins, CRP  | 3              | 0.640        | 0.012        | 0.0072 (0.010)          |
| Gui (20)           | 2,292       | 50*            | CHD                                | Age, sex, diabetes, smoking, hypertension, BMI, total cholesterol, triglycerides, alcohol consumption, family history of CHD   | 10             | 0.811        | 0.013        | 0.027 (<0.0001)         |
| Lobato (21)        | 1,948       | 14             | Mortality                          | Diabetes, CPB duration, EuroSCORE, grafts, beta blocker, statin  | 1              | 0.725        | 0.015        | 0.022 (0.01)            |
| Sandholt (22)      | 3,925       | 17             | Obesity                            | Age, sex, smoking, diet, physical activity, education, employment, anti-obesity medication   | 20             | 0.67         | 0.02         | 0.1 (<0.0001)           |
| Belsky (23)        | 8,286       | NR             | Obesity                            | Age, sex, center, socioeconomic status   | 32             | 0.550        | 0.024        | 0.006 (<0.0001)         |

Small improvement in the area under the receiver operating characteristic curve indicated small changes in predicted risks

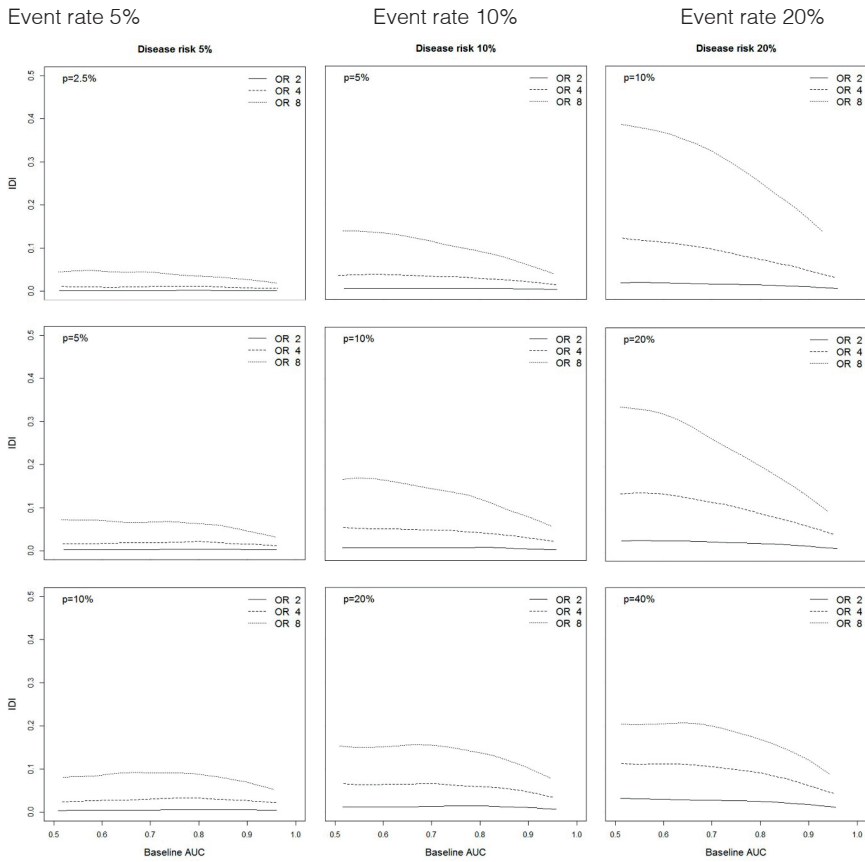
| Study                 | Sample Size | Event Rate (%) | Outcome                               | Clinical Risk Factors  | Number Of SNPs | Baseline AUC | $\Delta$ AUC | IDI (P value or 95% CI)   |
|-----------------------|-------------|----------------|---------------------------------------|--|----------------|--------------|--------------|---------------------------|
| Ribeiro (24)          | 1,006       | 45*            | Prostate cancer                       | Age, PSA   | 29             | 0.6476       | 0.033        | 0.021 (<0.0001)           |
| Rouskas (25)          | 979         | 52*            | Obesity                               | Age, sex   | 24             | 0.685        | 0.037        | 0.00242 (<0.001)          |
| Ruan (26)             | 2,846       | 49*            | Nasopharyngeal carcinoma              | Smoking, consumption of salted fish and preserved vegetables, family history of NPC  | 7              | 0.70         | 0.04         | 0.05 (<0.001)             |
| Tikkanen (27)         | 2,443       | 55             | Dyslipidemia                          | Age, sex, birth year, triglycerides  | 52             | 0.67         | 0.04         | 0.041 (0.024 to 0.059)    |
| Hüsing (28)           | 13,836      | 43*            | Breast cancer                         | Age at menarche, age at first full term pregnancy, age at menopause, smoking, BMI in interaction with menopausal status at baseline, alcohol consumption count of full term pregnancies, ever use of hormone replacement therapy | 18             | 0.564        | 0.041        | 0.0016                    |
| Fang (29)             | 1,957       | 48*            | Melanoma                              | Age, sex, pigmentation   | 11             | 0.64         | 0.05         | 0.0350 (<0.0001)          |
| del Rio-Espinoza (30) | 683         | NR             | Tissue plasminogen activator response | NIHSS score, atrial fibrillation, time from onset to treatment, DBP  | 2              | 0.654        | 0.066        | 0.0513 (0.0341 to 0.0684) |
| Bolton (31)           | 508         | 26             | CHD                                   | Age, sex, diabetes and/or glucose intolerance, smoking, SBP, total cholesterol/HDL cholesterol   | 27             | 0.671        | 0.069        | 0.04 (0.02 to 0.06)       |
| Wacholder (32)        | 11,588      | 48*            | Breast cancer                         | Age, entry year, and cohort  | 10             | 0.534        | 0.084        | 0.038 (<0.001)            |
| Chang (33)            | 268         | 22             | Edema                                 | Age, sex   | 28             | 0.702        | 0.088        | 0.128 (<0.001)            |

Legend: The table shows sample size, event rate, model characteristics and performance measures for the 33 studies from the literature review (see methods). Each study investigated the extension of a baseline prediction model containing clinical risk factors by one or more single-nucleotide polymorphisms (SNPs) and reported the increment in the area under the receiver operating characteristic curve ( $\Delta$ AUC) and integrated discrimination improvement (IDI).  $\Delta$  AUC = increment in the area under the receiver operating characteristic curve; AUC = area under the receiver operating characteristic curve; CI = confidence interval; IDI = integrated discrimination improvement; SNP = single-nucleotide polymorphism; \* = case control study; BMD = bone mineral density; BMI = body mass index; CABG = coronary artery bypass graft; CHD = coronary heart disease; CPB = cardiopulmonary bypass; CRP = C-reactive protein; CVD = cardiovascular disease; DBP = diastolic blood pressure; DRE = digital rectal examination; HDL = high-density lipoprotein; LDL = low-density lipoprotein; MI = myocardial infarction; NIHSS = National Institutes of Health Stroke Scale; NPC = nasopharyngeal carcinoma; NR = not reported; PCI = percutaneous coronary intervention; PSA = prostate-specific antigen; SBP = systolic blood pressure; ST = TRUS = transrectal ultrasound; WHR = waist-hip ratio

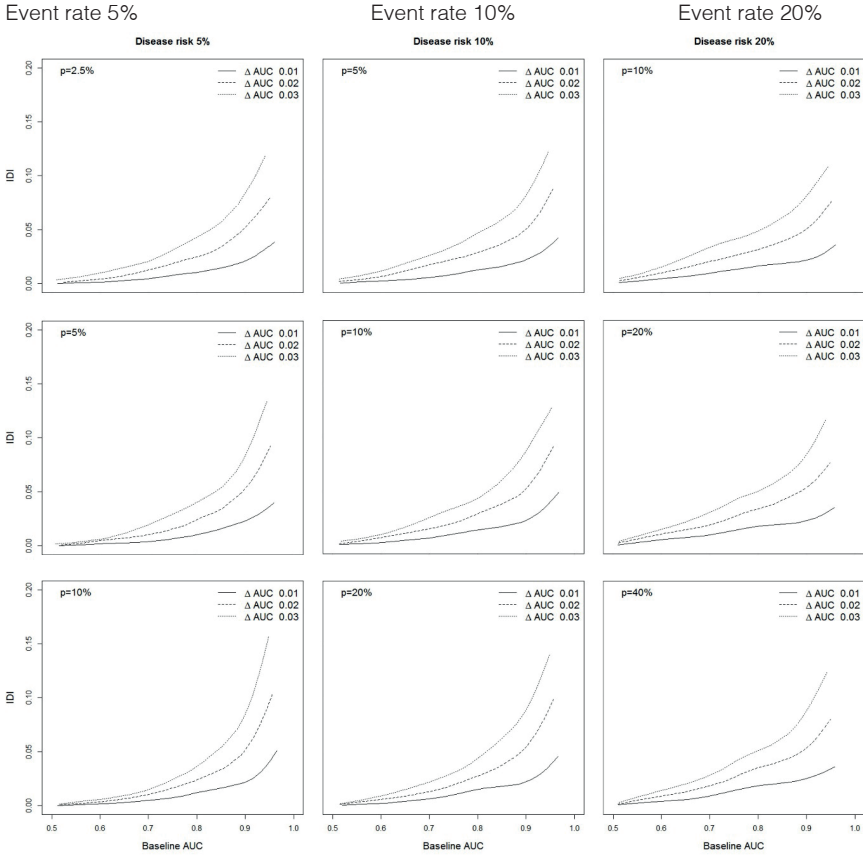


**Appendix Figure 1.** Summary of literature search and selection. \* Keywords that were used: genetic, genomic, polygenic, polymorphisms, or DNA. AUC = area under the receiver operating characteristic curve; IDI = integrated discrimination improvement; SNP = single nucleotide polymorphism

Small improvement in the area under the receiver operating characteristic curve indicated small changes in predicted risks



**Appendix Figure 2.** Relationship between the odds ratio (OR) of the added risk factor and integrated discrimination improvement (IDI) by the area under the receiver operating characteristic curve (AUC) of the baseline prediction model. IDI was calculated for scenarios in which the population disease risk (event rate) and the frequency ( $p$ ) of the added risk factor were varied.



**Appendix Figure 3.** Integrated discrimination improvement (IDI) for different increments in the area under the receiver operating characteristic curve ( $\Delta AUC$ ) by the area under the receiver operating characteristic curve of the baseline prediction model. IDI was calculated for scenarios in which the population disease risk (event rate) and the frequency ( $p$ ) of the added risk factor were varied.

Small improvement in the area under the receiver operating characteristic curve indicated small changes in predicted risks

## References (Appendix)

1. Eriksson J, Evans DS, Nielson CM, Shen J, Srikanth P, Hochberg M, et al. Limited clinical utility of a genetic risk score for the prediction of fracture risk in elderly subjects. *J Bone Miner Res.* 2015;30:184-194.
2. Fava C, Ohlsson T, Sjögren M, Tagetti A, Almgren P, Engström G, et al. Cardiovascular consequences of a polygenetic component of blood pressure in an urban-based longitudinal study. *J Hypertens.* 2014;32:1424-1428.
3. Havulinna AS, Kettunen J, Ukkola O, Osmond C, Eriksson JG, Kesaniemi YA, et al. A blood pressure genetic risk score is a significant predictor of incident cardiovascular events in 32 669 individuals. *Hypertension.* 2013;61:987-994.
4. Gränsbo K, Almgren P, Sjögren M, Smith JG, Engström G, Hedblad B, et al. Chromosome 9p21 genetic variation explains 13% of cardiovascular disease incidence but does not improve risk prediction. *J Intern Med.* 2013;274:233-240.
5. Ripatti S, Tikkanen E, Orho-Melander M, Havulinna AS, Silander K, Sharma A, et al. A multilocus genetic risk score for coronary heart disease: case-control and prospective cohort analyses. *Lancet.* 2010;376:1393-1400.
6. Thanassoulis G, Peloso GM, Pencina MJ, Hoffmann U, Fox CS, Cupples LA, et al. A genetic risk score is associated with incident cardiovascular disease and coronary artery calcium: the Framingham Heart Study. *Circ Cardiovasc Genet.* 2012;5:113-121.
7. Muhlenbruch K, Jeppesen C, Joost HG, Boeing H, Schulze MB. The value of genetic information for diabetes risk prediction - differences according to sex, age, family history and obesity. *PLoS ONE.* 2013;8:e64307.
8. Hivert MF, Jablonski KA, Perreault L, Saxena R, McAteer JB, Franks PW, et al. Updated genetic score based on 34 confirmed type 2 diabetes Loci is associated with diabetes incidence and regression to normoglycemia in the diabetes prevention program. *Diabetes.* 2011;60:1340-1348.
9. Brautbar A, Ballantyne CM, Lawson K, Nambi V, Chambless L, Folsom AR, et al. Impact of adding a single allele in the 9p21 locus to traditional risk factors on reclassification of coronary heart disease risk and implications for lipid-modifying therapy in the atherosclerosis risk in communities study. *Circ Cardiovasc Genet.* 2009;2:279-285.
10. Muehlschlegel JD, Liu KY, Perry TE, Fox AA, Collard CD, Shernan SK, et al. Chromosome 9p21 variant predicts mortality after coronary artery bypass graft surgery. *Circulation.* 2010;122.
11. Wauters E, Carruthers KF, Buysschaert I, Dunbar DR, Peuteman G, Belmans A, et al. Influence of 23 coronary artery disease variants on recurrent myocardial infarction or cardiac death: the GRACE Genetics Study. *Eur Heart J.* 2013;34:993-1001.
12. Juhola J, Oikonen M, Magnussen CG, Mikkilä V, Siitonen N, Jokinen E, et al. Childhood physical, environmental, and genetic predictors of adult hypertension: The cardiovascular risk in Young Finns Study. *Circulation.* 2012;126:402-409.
13. Bacci S, Rizza S, Prudente S, Spoto B, Powers C, Facciorusso A, et al. The ENPP1 Q121 variant predicts major cardiovascular events in high-risk individuals: evidence for interaction with obesity in diabetic patients. *Diabetes.* 2011;60:1000-1007.
14. Brautbar A, Pompeii LA, Dehghan A, Ngwa JS, Nambi V, Virani SS, et al. A genetic risk score based on direct associations with coronary heart disease improves coronary heart disease risk prediction in the Atherosclerosis Risk in Communities (ARIC), but not in the Rotterdam and Framingham Offspring. *Studies. Atherosclerosis.* 2012;223:421-426.
15. Lin X, Song K, Lim N, Yuan X, Johnson T, Abderrahmani A, et al. Risk prediction of prevalent diabetes in a Swiss population using a weighted genetic score--the CoLaus Study. *Diabetologia.* 2009;52:600-608.
16. Butoescu V, Ambroise J, Stainier A, Dekairielle A-F, Gala J-L, Tombal B. Does genotyping of risk-associated single nucleotide polymorphisms improve patient selection for prostate biopsy when combined with a prostate cancer risk calculator? *Prostate.* 2014;74:365-371.
17. Yu C-C, Lin VC, Huang C-Y, Liu C-C, Wang J-S, Wu TT, et al. Prognostic significance of Cyclin



- D1 Polymorphisms on prostate-specific antigen recurrence after radical prostatectomy. *Ann Surg Oncol.* 2013;20:492-499.
18. Zheng W, Wen W, Gao YT, Shyr Y, Zheng Y, Long J, et al. Genetic and clinical predictors for breast cancer risk assessment and stratification among Chinese women. *J Natl Cancer Inst.* 2010;102:972-981.
  19. Lind L, Syvänen AC, Axelsson T, Lundmark P, Hägg S, Larsson A. Variation in genes in the endothelin pathway and endothelium-dependent and endothelium-independent vasodilation in an elderly population. *Acta Physiol.* 2013;208:88-94.
  20. Gui L, Wu F, Han X, Dai X, Qiu G, Li J, et al. A multilocus genetic risk score predicts coronary heart disease risk in a Chinese Han population. *Atherosclerosis.* 2014;237:480-485.
  21. Lobato RL, White WD, Mathew JP, Newman MF, Smith PK, McCants CB, et al. Thrombomodulin gene variants are associated with increased mortality after coronary artery bypass surgery in replicated analyses. *Circulation.* 2011;124.
  22. Sandholt CH, Sparso T, Grarup N, Albrechtsen A, Almind K, Hansen L, et al. Combined analyses of 20 common obesity susceptibility variants. *Diabetes.* 2010;59:1667-1673.
  23. Belsky DW, Moffitt TE, Sugden K, Williams B, Houts R, McCarthy J, et al. Development and evaluation of a genetic risk score for obesity. *Biodemography Soc Biol.* 2013;59:85-100.
  24. Ribeiro RJ, Monteiro CP, Azevedo AS, Cunha VF, Ramanakumar AV, Fraga AM, et al. Performance of an adipokine pathway-based multilocus genetic risk score for prostate cancer risk prediction. *PLoS ONE.* 2012;7:e39236.
  25. Rouskas K, Kouvatzi A, Paletas K, Papazoglou D, Tsapas A, Lobbens S, et al. Common variants in FTO, MC4R, TMEM18, PRL, AIF1, and PCSK1 show evidence of association with adult obesity in the Greek population. *Obesity (Silver Spring).* 2012;20:389-395.
  26. Ruan HL, Qin HD, Shugart YY, Bei JX, Luo FT, Zeng YX, et al. Developing genetic epidemiological models to predict risk for nasopharyngeal carcinoma in high-risk population of China. *PLoS ONE.* 2013;8:e56128.
  27. Tikkanen E, Tuovinen T, Widen E, Lehtimäki T, Viikari J, Kahonen M, et al. Association of known loci with lipid levels among children and prediction of dyslipidemia in adults. *Circ Cardiovasc Genet.* 2011;4:673-680.
  28. Husing A, Canzian F, Beckmann L, Garcia-Closas M, Diver WR, Thun MJ, et al. Prediction of breast cancer risk by genetic risk factors, overall and by hormone receptor status. *J Med Genet.* 2012;49:601-608.
  29. Fang S, Han J, Zhang M, Wang LE, Wei Q, Amos CI, et al. Joint effect of multiple common SNPs predicts melanoma susceptibility. *PLoS ONE.* 2013;8:e85642.
  30. del Rio-Espinola A, Fernandez-Cadenas I, Giral D, Quiroga A, Gutierrez-Agullo M, Quintana M, et al. A predictive clinical-genetic model of tissue plasminogen activator response in acute ischemic stroke. *Ann Neurol.* 2012;72:716-729.
  31. Bolton JL, Stewart MC, Wilson JF, Anderson N, Price JF. Improvement in prediction of coronary heart disease risk over conventional risk factors using SNPs identified in genome-wide association studies. *PLoS ONE.* 2013;8:e57310.
  32. Wacholder S, Hartge P, Prentice R, Garcia-Closas M, Feigelson HS, Diver WR, et al. Performance of common genetic variants in breast-cancer risk models. *N Engl J Med.* 2010;362:986-993.
  33. Chang T-J, Liu P-H, Liang Y-C, Chang Y-C, Jiang Y-D, Li H-Y, et al. Genetic predisposition and nongenetic risk factors of thiazolidinedione-related edema in patients with type 2 diabetes. *Pharmacogenet Genomics.* 2011;21:829-836.





# 5

## **Evaluation of polygenic risk models using multiple performance measures: a critical assessment of discordant results**

Forike K. Martens, Elisa C.M. Tonk, and A. Cecile J.W. Janssens  
Genetics in Medicine 21 (2019) 391–397

## Abstract

**Purpose** The area under the receiver operating characteristic curve (AUC) is commonly used for evaluating the improvement of polygenic risk models and increasingly assessed together with the net reclassification improvement (NRI) and integrated discrimination improvement (IDI). We evaluated how researchers described and interpreted AUC, NRI, and IDI when simultaneously assessed.

**Methods** We reviewed how researchers described definitions of AUC, NRI and IDI and how they computed each metric. Next, we reviewed how the increment in AUC, NRI and IDI were interpreted; and how the overall conclusion about the improvement of the risk model was reached.

**Results** AUC, NRI and IDI were correctly defined in 63%, 70%, and 0% of the articles. All statistically significant values and almost half of the non-significant were interpreted as indicative of improvement, irrespective of the values of the metrics. Also, small, nonsignificant changes in the AUC were interpreted as indication of improvement when NRI and IDI were statistically significant.

**Conclusion** Researchers have insufficient knowledge about how to interpret the various metrics for the assessment of the predictive performance of polygenic risk models and rely on the statistical significance for their interpretation. A better understanding is needed to achieve more meaningful interpretation of polygenic prediction studies.

## Introduction

The area under the receiver operating characteristic (ROC) curve (AUC or c-statistic) (1) is the most commonly used measure for the evaluation of prediction models. AUC quantifies the ability to discriminate between individuals who will or will not manifest the outcome of interest (referred to as events and nonevents in this article). When a model is updated with new risk factors, such as biomarkers, genetic factors or imaging results, the improvement in the discriminative ability is assessed by the increment in AUC ( $\Delta$ AUC) (Box 1) (2-4).

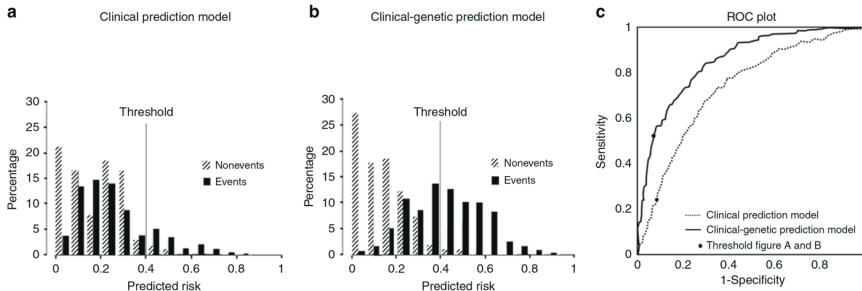
In recent years, alternative measures for the evaluation of prediction models have been proposed, including reclassification measures such as the net reclassification improvement (NRI) and integrated discrimination improvement (IDI) (2,5,6). NRI quantifies the extent to which the addition of risk factors leads to improved classification of risks, and IDI assesses the improvement of the risk difference between events and nonevents (2). NRI and IDI are increasingly used in addition to AUC, but the rationale and value of adding these metrics remain often unclear. NRI and IDI are frequently described as measures of discrimination (7,8) and IDI is often labeled as measure of reclassification (9,10). When the purpose and meaning of the metrics are unclear, it is challenging to interpret the findings, especially when these are discordant.

Discordant findings are often attributed to shortcomings of the metrics. AUC is argued to be insensitive as it often fails to detect improvements in prediction that result from adding clinically relevant risk factors (2,5,11-14). Others argue that NRI and IDI are too sensitive for identifying changes in predicted risks, which may lead to false positive conclusions about the improvement of prediction models (15-17). We earlier showed that findings might also be discordant because the metrics assess different aspects of the improvement in predictive performance:  $\Delta$ AUC assesses the gain in discriminative ability, NRI assesses changes in risk classification, and IDI assesses changes in the risk differences (18). For example, adding genetic factors might increase the risk differences without improving discriminative ability when the AUC of the baseline prediction model is already high (18).

The aim of this study was to evaluate how researchers describe and interpret the simultaneous use of multiple metrics in the assessment of improvement in predictive performance of polygenic risk models. Following the recommendations given by the Statement on the reporting of genetic risk

prediction studies (GRIPS) (19), we reviewed how researchers described what the metrics are assessing; how the metrics were obtained, how their results were interpreted, and how the overall conclusion was reached.

**Box 1.** Evaluating the predictive performance of polygenic models using AUC, NRI, and IDI: a tutorial



Genetic factors are added to clinical prediction models to improve the prediction of disease. If these genetic factors improve the model, these improvements are reflected in the distributions of predicted risks. Figure A shows the distributions of predicted risks using a clinical prediction model for participants in a hypothetical study. The participants who did not develop the disease during the duration of the study (referred to as nonevents) tended to have lower predicted risks than those who did develop the disease (events): the distribution of predicted risks for nonevents is skewed toward lower risk as compared with the distribution of predicted risks for events. When genetic factors are added to the clinical prediction model, we see that the distribution for nonevents “moves” even more toward lower risk, and the distribution for events moves toward higher risk (Figure B). There are several ways how these changes in the distributions of predicted risks can be quantified. The most commonly known is the area under the receiver operating characteristic curve (AUC) (1), but the net reclassification improvement (NRI) and integrated discrimination improvement (IDI) became popular once introduced (2). We will explain the measures in reverse order.

IDI: increase in risk difference

Instead of presenting distributions of predicted risks for events and nonevents, we can calculate the average predicted risks in both groups for each prediction model. When the risk distributions of events and nonevents entirely overlap, the difference between the averages is zero. When the risk distributions “move” further apart—in our example, because genetic factors were added—the difference between the two averages becomes larger. The increase in the risk differences between the clinical and the clinical-genetic prediction model is the IDI (2).

NRI: reclassification into correct risk category

Prediction models are often used to classify people in risk categories by setting one or more risk thresholds. In our example, we have a single threshold that divides the population

into a low- and high-risk group. The proportion of events that have predicted risks above the threshold is the sensitivity and the proportion of nonevents with predicted risks below the threshold is the specificity. The sensitivity and specificity are the proportions of correct classifications. A perfect prediction model would classify all events above the threshold and all nonevents below, and have sensitivity and specificity of 100%. When predicted risks change because genetic factors are added to the clinical model, we want the sensitivity and/or specificity to increase. The increase in sensitivity plus the increase in specificity is the NRI. In general, and if more thresholds are considered, NRI is the sum of the proportion of events that are reclassified to higher risk categories and the proportion of nonevents reclassified to lower categories (2).

AUC: classification across all risk thresholds

NRI assesses the improvement in discrimination for specific risk thresholds and varies with the number of thresholds and their values (22). When a clinical prediction model has no known risk thresholds, we can assess the improvement by calculating and comparing sensitivity and specificity across all possible risk thresholds. The lines that connect the sensitivity–specificity of all thresholds of a prediction model is the receiver operating characteristic (ROC) curve and the area underneath is the AUC (Figure C) (1). The figures show that the clinical–genetic prediction model has more favorable combinations of sensitivity and specificity than the clinical model: each sensitivity comes with a higher specificity (or each specificity with a higher sensitivity). The combinations are more favorable, because there is less overlap between the risk distributions of events and nonevents using the clinical–genetic model as compared with the clinical model. This leads to a larger area under the ROC curve and thus a higher AUC. The improvement in discriminative ability between the models is the increment in AUC ( $\Delta$ AUC) (4).

## Materials and methods

### Literature search

We performed a literature search to find empirical studies that evaluated the improvement in predictive performance of risk models by assessing  $\Delta$ AUC, NRI, and IDI. Using Thomson Reuters Web of Knowledge (version 5.17) we retrieved all publications that cited the article by Pencina et al. in which the NRI and IDI were introduced (search date 28 December 2016) (2). To limit the number of articles, we focused on studies that investigated the improved predictive performance of adding genetic variants (single nucleotide polymorphisms, or SNPs) to clinical risk models. For this purpose, we selected publications using the keywords *genetic*, *genomic*, *polygenic*, *polymorphisms*, or *DNA*. We excluded studies on non-germline DNA, such as circulating cell-free DNA or tumor DNA. Full-text articles and supplementary materials were obtained for data extraction.



## Data extraction

For each study, we recorded sample size, event rate, clinical risk factors in the clinical prediction models as well as the number of SNPs that was added. The event rate is the proportion of individuals with the outcome of interest in the study population, which was the incidence, prevalence or the size of case population, depending on the design of the study. We extracted AUC values of the baseline and updated models, as well as the values of NRI and IDI along with *P* values and confidence intervals. We recorded whether NRI was used with or without categories: categorical NRI is a metric that is based on the proportions of people that move between risk categories, and continuous NRI is based on the proportions of people that have higher or lower risks after updating the risk model. When multiple prediction models were investigated in one article, we selected the model that was described in the abstract, the model that had the highest number of risk factors in the baseline model, or the model that had the highest number of SNPs added.

We extracted, verbatim, descriptions of the definitions and calculations of AUC, NRI, and IDI from the methods section of the articles. From the results and discussion sections, we extracted descriptions of the numerical results of the metrics, the interpretation of each measure, and the general conclusions. All descriptions were imported into Microsoft Excel (Microsoft Corporation, Redmond, WA, USA).

## Analysis

We evaluated the point estimates and statistical significance of NRI and IDI in relation to  $\Delta$ AUC. Statistical significance was based on the confidence intervals or the reported *P* values using the threshold of statistical significance mentioned in the articles, which was  $P < 0.05$  in all of them.

Using the excerpts of the methods section, we reviewed how the measure and calculation of AUC, NRI and IDI were described, and evaluated whether these followed common definitions and approaches. For the latter, we required that the definition of AUC should at least have mentioned that it is a measure of discrimination or the concordance between predicted and observed survival, that NRI is a measure of reclassification, and that IDI assesses the improvement in risk differences or discrimination slopes (Box S1). Descriptions of the calculations needed to give insight in the computation. For AUC the description needed to refer to the c-statistic or nonparametric trapezoidal rule. For

NRI the description needed to include that it was the sum of the net percentage of correct reclassification in events and nonevents, with reclassification referring to changes between risk categories for categorical NRI and changes in risk for continuous NRI. The description of IDI needed to refer to the difference of the mean increments and mean decrements in estimated probabilities between models or the difference in discrimination slopes of the baseline and updated model (Box S1).

Using the excerpts of the results section, we assessed how the values of AUC, NRI and IDI were described. We documented whether the results were described by their effect sizes, *P* values or confidence intervals, or both, and whether and how the results were interpreted in terms of model improvement. We documented whether authors reported the presence or absence of improvement, and considered “minimal improvement” when they described the improvement or increase in the estimates as being small or minimal.

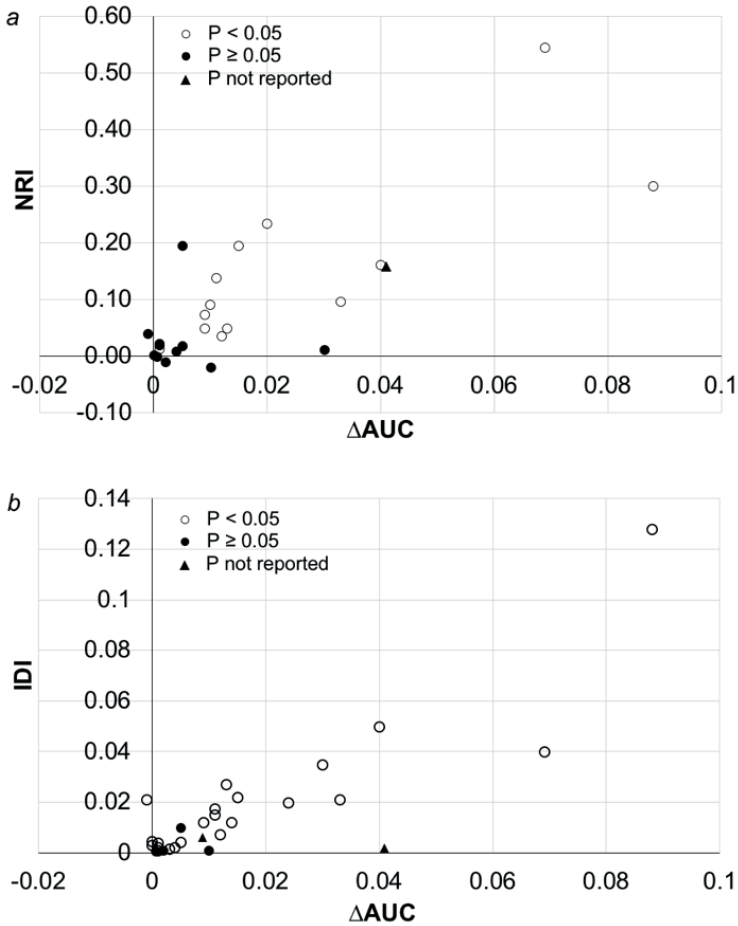
Finally, using excerpts from the discussion, we evaluated how the overall improvement of the model was interpreted. In addition to the presence or absence of improvement, we distinguished “minimal improvement” when the reported improvement was considered minimal or marginal, and “inconclusive” when the authors concluded that improvement was demonstrated from some metric(s) but not others. Two researchers independently evaluated the descriptions and disagreements were discussed to reach consensus.

## Results

Of the 2509 articles that had cited the article of Pencina et al., 250 articles reported polygenic risk studies of which 32 met the inclusion criteria (Figure S1). Most excluded articles did not report empirical analyses (such as reviews and commentaries,  $n = 94$ ) or did not report on all three measures ( $n = 83$ ). The majority of the 32 included articles evaluated cardiovascular ( $n = 15$ ) and cancer prediction models ( $n = 8$ ; Table S1).

Definitions of AUC and NRI and IDI were given in 84, 81, and 72% of the articles, of which 63, 70, and 0% were correct (Table 1). IDI was frequently described as a metric of reclassification (30%) and discrimination (22%), and five articles described NRI and IDI together, for example, as measures of “model performance” or “utility”. Half of the articles (56%) described how AUC was

obtained, of which all mentioned the c-statistic, but only three (9%) explained the calculation of NRI and three others (9%) explained IDI. The three descriptions for the calculation of IDI were correct, but none of the articles described NRI as the *sum* of two *net* percentages.



**Figure 1.** a Net reclassification improvement (NRI) and b integrated discrimination improvement (IDI) by increments in the area under the receiver operating characteristic curve ( $\Delta AUC$ ). Excluded are studies that used continuous NRI or that did not report the value of the NRI (a) and articles that did not report the value of IDI (b)

AUC values of the baseline clinical risk models ranged from 0.56 to 0.87 (Table S2), and  $\Delta$ AUC ranged from -0.001 to 0.09 (median 0.01, interquartile range [IQR] 0.002-0.02; Table 2). Most (94%)  $\Delta$ AUC values were 0.04 or lower. Of the 24 articles that computed the categorical NRI, the values ranged from -0.02 to 0.54 (median 0.044, IQR 0.012-0.142;) and the 7 articles that computed the continuous NRI reported values ranging from 0.07 to 1.24 (median 0.233; IQR 0.137-0.356; Table 2). Of the 24 articles that reported absolute IDI, values ranged from 0.00062 (a 0.062% absolute increase in risk difference between events and nonevents) to 0.128 (median 0.011; IQR 0.002-0.021). NRI and IDI values were, as expected, higher for higher values of  $\Delta$ AUC (Figure 1).

$\Delta$ AUC was statistically significant in 13 articles, NRI in 21, and IDI in 26 (Table 2). When  $\Delta$ AUC was higher than 0.01 ( $n = 15$  studies), IDI and NRI were both statistically significant in all but 1 of 14 studies (Table 2). Of the 17 studies in which  $\Delta$ AUC was equal or lower than 0.01, NRI and IDI values were still statistically significant in 7 out of 16 of them.

When the value of a metric was statistically significant, the metric was interpreted as indicating improvement of the model in all articles, with several reporting that the improvement was minimal (Table 3). When a metric was not statistically significant, almost half were still described as indicative of model improvement, now with most acknowledging that the improvement was minimal. All  $\Delta$ AUC values that were not statistically significant and interpreted as no indication of improvement were lower than 0.005, whereas those that were considered to indicate (minimal) improvement were all equal or higher than 0.005. All statistically significant  $\Delta$ AUC values were interpreted as indicating improvement of the model, irrespective of their absolute values.

In 17 of the 27 articles that reported all three values in the results section (Table 2), the authors interpreted that all three metrics showed improvement of the model. Among these were 7 studies in which all three metrics were statistically significant and 7 studies in which NRI and IDI were statistically significant but  $\Delta$ AUC was not. In 6 of the 27 articles, the authors interpreted that the  $\Delta$ AUC showed no improvement of the model but that the NRI and IDI did. In all of these,  $\Delta$ AUC was equal or lower than 0.003, and NRI was not statistically significant in 2 of them. Only 1 of the 27 articles interpreted that none of the metrics indicated an improvement of the prediction model; in this study, the absolute values of  $\Delta$ AUC, NRI and IDI were all lower than 0.001 and not statistically significant.

All but five articles concluded that, overall, the prediction model had

improved from the addition of genetic factors (Table 2). Half of them mentioned that the improvement was minimal. All articles in which the individual metrics were evaluated as indicative of improvement, also had a overall positive evaluation, except one in which all three metrics were interpreted as showing minimal improvement leading to an overall conclusion of no improvement. Of the six articles that reported improvement indicated by NRI and IDI but not by  $\Delta$ AUC, five concluded that the model had improved albeit minimally, and one refrained from making an overall conclusion.

**Table 1.** Definition and calculation method of AUC, NRI and IDI as described in included articles

| <b>Metric</b>  | <b>Definition</b>  | <b>% (Articles)</b> | <b>Calculation method</b>  | <b>% (Articles)</b> |
|--|--|---------------------|--|---------------------|
| AUC  | <i>Not reported</i>  | 16 (5)              | <i>Not described</i>   | 44 (14)             |
|  | <i>Reported</i>  | 84 (27)             | <i>Described</i>   | 56 (18)             |
|  | Discrimination   |                     |  |                     |
|  | Probability of concordance between predicted and observed survival   | 56 (15)             | C-statistic/index  | 100 (18)            |
|  | Prediction Performance   | 7 (2)               |  |                     |
| NRI  | Accuracy, classification, clinical value, incremental value, predictive value, correlation models with outcome | 7 (2)               |  |                     |
|  | <i>Not reported</i>  | 19 (6)              | <i>Not described</i>   | 91 (29)             |
|  | <i>Reported</i>  | 81 (26)             | <i>Described</i>   | 9 (3)               |
|  | Reclassification   | 70 (18)             | Comparison of proportions of correct reclassifications to either higher or lower risk  | 100 (3)             |
|  | Classification   | 7 (2)               |  |                     |
| IDI  | Discrimination, improvement, model fit, model performance, prediction, utility                                 | 23 (6)              |  |                     |
|  | <i>Not reported</i>  | 28 (9)              | <i>Not described</i>   | 91 (29)             |
|  | <i>Reported</i>  | 72 (23)             | <i>Described</i>   | 9 (3)               |
|  | Reclassification   | 30 (7)              | Difference of mean increments and decrements in estimated probabilities between models | 67 (2)              |
|  | Discrimination   | 22 (5)              | Differences in discrimination slopes between models                                    |                     |
| Improvement in average sensitivity without sacrificing average specificity | Model performance  | 13 (3)              |  |                     |
|  | Classification   | 9 (2)               |  |                     |
|  | Model fit, improvement, prediction, utility  | 9 (2)               |  |                     |
|  |  | 9 (2)               |  |                     |
|  |  | 17 (4)              |  |                     |

Abbreviations: AUC = area under the receiver operating characteristic curve; IDI = integrated discrimination improvement; NRI = net reclassification improvement

**Table 2.** Point estimates, interpretations of model improvement based on  $\Delta$ AUC, NRI and IDI values, and overall conclusions about improvement of predictive performance

| First Author                 | Point Estimates        |                             |                               | Model Improvement |       |       |       | Overall |
|------------------------------|------------------------|-----------------------------|-------------------------------|-------------------|-------|-------|-------|---------|
|                              | $\Delta$ AUC           | NRI (P value or 95% CI)     | IDI (P value or 95% CI)       | $\Delta$ AUC      | NRI   | IDI   |       |         |
| Park <sup>S1</sup>           | -0.001 (0.99)          | 0.040 (0.32)                | 0.021 (0.02)                  | No                | No    | Yes   | No    |         |
| Eriksson <sup>S2</sup>       | 0 (0.246)              | 0.11 (0.005) <sup>a</sup>   | 0.003 (0.007)                 | No                | [Yes] | [Yes] | [Yes] |         |
| Fava <sup>S3</sup>           | 0 (>0.05)              | 0.002 (0.39)                | 0.00449 (0.02)                | No                | [Yes] | [Yes] | [Yes] |         |
| Kathiresan <sup>9</sup>      | 0 (NR)                 | NR (0.01)                   | NR (0.02)                     | No                | Yes   | Yes   | [No]  |         |
| Havulinna <sup>S4</sup>      | 0.0006 (0.16)          | -0.0008 (0.92)              | 0.00062 (0.14)                | No                | No    | No    | No    |         |
| Gränsbo <sup>10</sup>        | 0.001 (NR)             | 0.012 (0.043)               | 0.001 (<0.001)                | [Yes]             | [Yes] | [Yes] | No    |         |
| Ripatti <sup>S5</sup>        | 0.001 (0.19)           | 0.022 (0.182)               | 0.004 (0.0006)                | No                | [Yes] | Yes   | [Yes] |         |
| Lim <sup>S6</sup>            | 0.001 (0.1057)         | 0.019 (0.0495)              | 0.002 (0.0131)                | No                | Yes   | Yes   | [Yes] |         |
| Thanassoulis <sup>S7</sup>   | 0.002 (NR)             | -0.01 (-0.052 to 0.033)     | 0.001 (-0.001 to 0.003)       | [Yes]             | NR    | [Yes] | [Yes] |         |
| Fava <sup>S8</sup>           | 0.003 (>0.05)          | 0.0659 (0.013) <sup>a</sup> | 0.001452 (0.003)              | No                | Yes   | Yes   | [Yes] |         |
| Brautbar <sup>S9</sup>       | 0.004 (0.001 to 0.007) | 0.008 (0.31)                | 0.002 (<0.015)                | Yes               | [Yes] | Yes   | [Yes] |         |
| Muehlschlegel <sup>S10</sup> | 0.005 (NS)             | 0.195 (0.072)               | 0.010 (0.053)                 | NR                | Yes   | Yes   | Yes   |         |
| Park <sup>S11</sup>          | 0.005 (0.050)          | 0.0173 (0.352)              | 0.0041 (0.007)                | [Yes]             | No    | NR    | [Yes] |         |
| Juhola <sup>S12</sup>        | 0.009 (0.015)          | 0.048 (0.0002)              | 0.012 (<0.0001)               | Yes               | Yes   | Yes   | Yes   |         |
| Brautbar <sup>S13</sup>      | 0.009 (0.006 to 0.014) | 0.073 (0.019 to 0.12)       | 0.006 (NR)                    | Yes               | Yes   | Yes   | Yes   |         |
| Lyssenko <sup>S14</sup>      | 0.010 (0.00001)        | 0.09 (<0.001)               | NR (<0.001)                   | [Yes]             | Yes   | Yes   | [Yes] |         |
| Krärup <sup>S15</sup>        | 0.01 (0.002)           | -0.02(NS)                   | 0.001 (NS)                    | Yes               | No    | No    | No    |         |
| Butoescu <sup>S16</sup>      | 0.011 (NR)             | 0.403 (<0.001) <sup>a</sup> | 0.015 (0.035)                 | [Yes]             | Yes   | Yes   | [Yes] |         |
| Yu <sup>S17</sup>            | 0.011 (>0.050)         | 0.137 (0.015)               | 0.0175 (0.041)                | [Yes]             | Yes   | Yes   | [Yes] |         |
| Lind <sup>S18</sup>          | 0.012 (0.09)           | 0.035 (0.047)               | 0.0072 (0.010)                | Yes               | Yes   | Yes   | Yes   |         |
| Gui <sup>S19</sup>           | 0.013 (0.17)           | 0.04850 (<0.001)            | 0.027 (<0.001)                | [Yes]             | Yes   | Yes   | [Yes] |         |
| Pitkanen <sup>S20</sup>      | 0.014 (0.007)          | 0.163 (0.001) <sup>a</sup>  | 0.012 (1.8x10 <sup>-5</sup> ) | Yes               | Yes   | Yes   | Yes   |         |
| Lobato <sup>S21</sup>        | 0.015 (NR)             | 0.194 (0.005)               | 0.022 (0.01)                  | [Yes]             | Yes   | Yes   | Yes   |         |
| Morote <sup>8</sup>          | 0.02 (0.092)           | 0.233 (0.003) <sup>a</sup>  | NR (<0.001)                   | [Yes]             | Yes   | Yes   | Yes   |         |

| First Author           | Point Estimates  |                             |                         | Model improvement |     |       |       | Overall |
|------------------------|------------------|-----------------------------|-------------------------|-------------------|-----|-------|-------|---------|
|                        | $\Delta$ AUC     | NRI (P value or 95% CI)     | IDI (P value or 95% CI) | $\Delta$ AUC      | NRI | IDI   |       |         |
| Kertai <sup>S22</sup>  | 0.024 (0.001)    | 0.308 (0.0003) <sup>a</sup> | 0.02 (0.000024)         | Yes               | Yes | Yes   | Yes   |         |
| Fang <sup>S23</sup>    | 0.03 (0.0000601) | 0.0109 (0.6076)             | 0.0350 (<0.0001)        | [Yes]             | No  | [Yes] | Yes   |         |
| Ribeiro <sup>S24</sup> | 0.033 (0.0002)   | 0.095 (<0.0001)             | 0.021 (<0.0001)         | Yes               | Yes | Yes   | Yes   |         |
| Borque <sup>S25</sup>  | 0.034 (0.025)    | 1.242 (<0.001) <sup>a</sup> | NR (<0.001)             | Yes               | Yes | Yes   | Yes   |         |
| Ruan <sup>7</sup>      | 0.04 (<0.001)    | 0.16 (<0.001)               | 0.05 (<0.001)           | Yes               | Yes | Yes   | [Yes] |         |
| Huesing <sup>S26</sup> | 0.041 (NR)       | 0.158 (NR)                  | 0.0016 (NR)             | Yes               | Yes | [Yes] | [Yes] |         |
| Bolton <sup>S27</sup>  | 0.069 (0.0001)   | 0.544 (<0.001)              | 0.04 (0.02 to 0.06)     | Yes               | NR  | NR    | Yes   |         |
| Chang <sup>S28</sup>   | 0.088 (0.002)    | 0.300 (0.005)               | 0.128 (<0.001)          | Yes               | Yes | NR    | Yes   |         |

The point estimates, *P* values and interpretations of model improvement are as reported in the results section and the overall conclusion as reported in the discussion section of the articles. Square brackets indicate that the authors had expressed hesitancy, e.g., that they considered the improvement of the model to be minimal. References S1-S28 can be found in the Supplementary data. <sup>a</sup> Continuous NRI (see Table S2). Abbreviations:  $\Delta$ AUC = increment in the area under the receiver operating characteristic curve; CI = confidence interval; IDI = integrated discrimination improvement; NR = not reported; NRI = net reclassification improvement; NS = not statistically significant.



**Table 3.** Inferences about model improvement in the results section of the article in relation to the statistical significance of the metrics

|                               | Model improvement   |                                    |                    |
|-------------------------------|---------------------|------------------------------------|--------------------|
|                               | Yes<br>% (articles) | Yes, but minimally %<br>(articles) | No<br>% (articles) |
| Statistically significant     |                     |                                    |                    |
| ΔAUC                          | 85 (11)             | 15 (2)                             | 0 (0)              |
| NRI                           | 90 (18)             | 10 (2)                             | 0 (0)              |
| IDI                           | 83 (19)             | 17 (4)                             | 0 (0)              |
| Not statistically significant |                     |                                    |                    |
| ΔAUC                          | 8 (1)               | 33 (4)                             | 59 (7)             |
| NRI                           | 11 (1)              | 33 (3)                             | 56 (5)             |
| IDI                           | 25 (1)              | 25 (1)                             | 50 (2)             |

Statistical significance was based on reported *P* values and confidence intervals and the criterion of statistical significance in the articles, which was  $P < 0.05$  in all of them. Articles that did not report *P* values or confidence intervals for ΔAUC ( $n = 6$ ), NRI ( $n = 1$ ) and IDI ( $n = 2$ ), or did not interpret ΔAUC ( $n = 1$ ), NRI ( $n = 2$ ) and IDI ( $n = 3$ ) are excluded from this table. ΔAUC = increment in the area under the receiver operating characteristic curve; IDI = integrated discrimination improvement; NRI = net reclassification improvement

## Discussion

AUC, NRI, and IDI are three metrics that are increasingly used together in the assessment of polygenic risk models. Our analysis showed that authors provided minimal information about the purpose and assessment of the three metrics and that they mostly relied on statistical significance when interpreting the results. None of the articles distinguished, in their conclusions, between the different aspects of model performance that the metrics address.

Three observations can be made from this study. First, one-third of the articles did not specify what was measured by IDI and one-fifth did not do so for AUC and NRI. When authors did describe the metrics, only two-thirds were correct about what is measured by AUC and NRI, namely discrimination and reclassification, but were mostly wrong about IDI, which they described as a metric of discrimination, reclassification, or more general as a measure of model performance. These findings suggest that researchers may not know what each of the metrics assesses, and that the measures assess different aspects of predictive performance.

Second, only roughly half of the articles reported how AUC ( $n = 18$ ) was obtained and only 9% ( $n = 3$ ) reported how NRI and IDI were calculated.

When researchers did provide details, they gave the correct description for the calculation of AUC and IDI, but not of NRI. The three studies that mentioned the calculation of NRI did not describe that NRI is obtained by the sum of the two net proportions. Mentioning the *sum* of the two *net* percentages is important to make clear that NRI is not merely the percentage of reclassified people in a population. These findings confirm that researchers may not know what is measured by NRI and IDI. Whether researchers understand AUC cannot be concluded from this review; evidently, reporting that they obtained the c-statistic may not imply that they understand how the c-statistic is calculated.

And third, inferences about each metric, and hence the overall conclusion about improvement of predictive performance, were largely based on their statistical significance while absolute values of the metrics were small. When the values of the metrics would have been rounded to two decimals, the estimates would be 0.00 for 11 AUC, 2 NRI, and 12 IDI values. Of these, 3 AUC, 1 NRI, and 9 IDI values were interpreted as showing improvement of the model. Small values of AUC, IDI, and NRI may be statistically significant in large studies, but not clinically relevant. Relying on the statistical significance may lead to false claims about the improvement of prediction. Therefore, the interpretation should focus on the absolute values of the metrics rather than the statistical significance of their estimates (20,21). What degree of improvement is clinically relevant varies between scenarios and by the answer to the question what is to be gained from the additional information.

The interpretation of polygenic risk studies is straightforward when all measures show the same large and statistically significant improvement in predictive performance. When values are small and inferences are discordant, the question is whether the discordance is due to limitations in the assessment of the metrics or reflecting differential impact on the various aspects of predictive performance. For example, AUC is often criticized for being an insensitive metric to evaluate improvement in predictive performance (2,5,11-14), but improving discrimination requires a substantial change in the rank order of predicted risks that should not be expected when minor risk factors are added to the risk model. In such instances, IDI, which assesses the mean of predicted risks between events and nonevents before and after updating of the risk model, might still be able to show improvement in risk differentiation. Another example is that changes in risk classification as indicated by NRI may not imply that discrimination is improved as well. NRI has been shown to be too sensitive for

identifying minor changes in predicted risks (15-17) and it may be statistically significant, while AUC remains virtually unchanged (22,23).

All but four studies concluded that the addition of genes to clinical risk models improved the predictive performance of clinical risk models. In most studies, the values of  $\Delta$ AUC, NRI, and IDI were small and none of them were externally validated. The latter is relevant for the few studies in which the improvement in predictive performance would be of interest if it were replicated in independent data. Judging if clinical risk models improve by the addition of genes is challenging when researchers have limited understanding of the metrics used for evaluation of the models. Our study suggests that this limited understanding leads to false positive conclusions about the value of adding genes to clinical risk models.

Interpretation of polygenic risk studies is straightforward when there is no or substantial improvement in predictive performance, but it is challenging in between. Discordant results from multiple metrics may indicate that there is no improvement but that some metrics are sensitive enough to detect very small effects. Yet, it may also mean that there is improvement in prediction but not on all aspects of predictive performance. A better understanding is needed to achieve more meaningful interpretations of prediction studies. Overinterpretation of small improvements in predictive ability will unlikely improve the management of people at risk in public health practice.

## References

1. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;143(1):29-36.
2. Pencina MJ, D'Agostino Sr. RB, D'Agostino Jr. RB, et al. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med*. 2008;27(2):157-172.
3. Steyerberg EW, Pencina MJ, Lingsma HF, et al. Assessing the incremental value of diagnostic and prognostic markers: a review and illustration. *Eur J Clin Invest*. 2012;42(2):216-228.
4. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*. 1983;148(3):839-843.
5. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation*. 2007;115(7):928-935.
6. Pencina MJ, D'Agostino RB, Sr., Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Stat Med*. 2011;30(1):11-21.
7. Ruan HL, Qin HD, Shugart YY, et al. Developing genetic epidemiological models to predict risk for nasopharyngeal carcinoma in high-risk population of China. *PLoS ONE*. 2013;8(2):e56128.
8. Morote J, del Amo J, Borque A, et al. Improved prediction of biochemical recurrence after radical prostatectomy by genetic polymorphisms. *J Urol*. 2010;184(2):506-511.
9. Kathiresan S, Melander O, Anefski D, et al. Polymorphisms associated with cholesterol and risk of cardiovascular events. *N Engl J Med*. 2008;358(12):1240-1249.
10. Gränsbo K, Almgren P, Sjögren M, et al. Chromosome 9p21 genetic variation explains 13% of cardiovascular disease incidence but does not improve risk prediction. *J Intern Med*. 2013;274(3):233-240.
11. Pencina MJ, D'Agostino RB, Sr., Demler OV. Novel metrics for evaluating improvement in discrimination: net reclassification and integrated discrimination improvement for normal variables and nested models. *Stat Med*. 2012;31(2):101-113.
12. Pepe MS, Janes HE. Gauging the performance of SNPs, biomarkers, and clinical factors for predicting risk of breast cancer. *J Natl Cancer Inst*. 2008;100(14):978-979.
13. Pepe MS. Limitations of the Odds Ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *Am J Epidemiol*. 2004;159(9):882-890.
14. Ware JH. The limitations of risk factors as prognostic tools. *N Engl J Med*. 2006;355(25):2615-2617.
15. Pepe MS, Janes H, Li CI. Net risk reclassification P values: valid or misleading? *J Natl Cancer Inst*. 2014;106(4):dju041.
16. Gerds TA, Hilden J. Calibration of models is not sufficient to justify NRI. *Stat Med*. 2014;33(19):3419-3420.
17. Hilden J, Gerds TA. A note on the evaluation of novel biomarkers: do not rely on integrated discrimination improvement and net reclassification index. *Stat Med*. 2014;33(19):3405-3414.
18. Martens FK, Tonk EC, Kers JG, et al. Small improvement in the area under the receiver operating characteristic curve indicated small changes in predicted risks. *J Clin Epidemiol*. 2016; 79:159-164.
19. Janssens AC, Ioannidis JP, Bedrosian S, et al. Strengthening the reporting of Genetic Risk Prediction Studies (GRIPS): explanation and elaboration. *J Clin Epidemiol*. 2011;64(8):e1-e22.
20. Pepe MS, Kerr KF, Longton G, et al. Testing for improvement in prediction model performance. *Stat Med*. 2013;32(9):1467-1482.
21. Vickers AJ, Cronin AM, Begg CB. One statistical test is sufficient for assessing new predictive markers. *BMC Med Res Methodol*. 2011;11:13.
22. Mihaescu R, van Zitteren M, van Hoek M, et al. Improvement of risk prediction by genomic profiling: reclassification measures versus the area under the receiver operating characteristic curve. *Am J of Epidemiol*. 2010;172(3):353-361.
23. Janssens AC, Khoury MJ. Assessment of improved prediction beyond traditional risk factors: when

- does a difference make a difference? *Circ Cardiovasc Genet.* 2010;3(1):3-5.
24. Harrell FE, Jr., Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med.* 1996;15(4):361-387.

## Supplementary data

### Box S1. Definitions and calculation methods of AUC, NRI and IDI

| Metric | Definition   | Calculation method   |
|--------|--|--|
| AUC    | Discrimination   | C-statistic, or trapezoidal rule   |
| NRI    | Reclassification   | Categorical: sum of net percentages of correctly reclassified persons with and without an event;<br>Continuous: sum of net percentages of persons with and without an event correctly assigned a higher (event) or lower (no event) predicted risk |
| IDI    | Improvement in discrimination slopes or risk differences | Difference between discrimination slopes of baseline and updated models<br>Difference of mean predicted risks of persons with and without an event between models  |

Definitions and calculations are based on references (1,2,3). Abbreviations: AUC = area under the receiver operating characteristic curve; IDI = integrated discrimination improvement; NRI = net reclassification improvement

### References

1. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;143(1):29-36.
2. Pencina MJ, D'Agostino RB, Sr., D'Agostino RB, Jr., et al. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Statistics in medicine*. 2008;27(2):157-172; discussion 207-112.
3. Harrell FE Jr., Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med*. 1996;15:361-87.

**Table S1.** Characteristics of the prediction models in the literature review

| First Author                 | Sample Size | Event rate (%) | Outcome  | Clinical Risk Factors  | Number of Added SNPs |
|------------------------------|-------------|----------------|--|--|----------------------|
| Park <sup>S1</sup>           | 2,188       | 4              | Major adverse cardiac and cerebrovascular events | Age, sex, myocardial infarction, chronic kidney disease, stroke history, left ventricular ejection fraction <40%   | 2                    |
| Eriksson <sup>S2</sup>       | 9,704       | 46             | Fracture   | Age, height, weight, femoral neck BMD  | 63                   |
| Fava <sup>S3</sup>           | 27,003      | 13             | CVD and CVD mortality                            | Age, diabetes, smoking, hypertension, BMI, antilipemic medication  | 29                   |
| Kathiresan <sup>9</sup>      | 4,232       | 6              | CVD  | Age, sex, diabetes, smoking, BMI, SBP, DBP, HDL cholesterol, LDL cholesterol, triglycerides, family history of MI, lipid-lowering medication, antihypertensive medication, CRP | 9                    |
| Havulinna <sup>S4</sup>      | 17,954      | 7              | CVD  | Diabetes, smoking, SBP, total cholesterol, HDL cholesterol, lipid-lowering medication  | 32                   |
| Gränsbo <sup>10</sup>        | 24,777      | 11             | CVD  | Age, sex, diabetes, smoking, hypertension, BMI, lipid-lowering medication  | 1                    |
| Ripatti <sup>S5</sup>        | 13,292      | 5              | CHD  | Sex, diabetes, smoking, BMI, SBP, DBP, HDL and LDL cholesterol, antihypertensive medication  | 13                   |
| Lim <sup>S6</sup>            | 5,632       | NR             | Hypertension                                     | Age, sex, smoking, SBP, DBP, parental hypertension   | 4                    |
| Thanassoulis <sup>S7</sup>   | 3,014       | 6              | CVD  | Age, sex, smoking, SBP (and treatment), total cholesterol, HDL cholesterol, parental history of CVD  | 13                   |
| Fava <sup>S8</sup>           | 6,092       | 60             | Stroke   | Age, sex, smoking, hypertension, diabetes  | 29                   |
| Brautbar <sup>S9</sup>       | 9,998       | 13             | CHD  | Age, sex, smoking, SBP, total cholesterol, HDL cholesterol, antihypertensive medication, diabetes  | 1                    |
| Muehlschlegel <sup>S10</sup> | 845         | 11             | Mortality  | Age, sex, CPB duration, institution, preoperative measures, pulmonary disease, coronary stenosis   | 1                    |
| Park <sup>S11</sup>          | 6,910       | 18             | Diabetes   | Age, family history of diabetes, BMI, history of hypertension, exercise, triglycerides, fasting plasma glucose, HDL-C, HbA1c   | 36                   |
| Juhola <sup>S12</sup>        | 1,939       | 34             | Adult hypertension                               | Age, sex, parental hypertension, childhood overweight/obesity status, childhood SBP, parental occupational status  | 29                   |
| Brautbar <sup>S13</sup>      | 8,542       | 13             | CHD  | Age, sex, diabetes, smoking, SBP, total cholesterol, HDL cholesterol, antihypertensive medication  | 13                   |
| Lyssenko <sup>S14</sup>      | 16,061      | 13             | Diabetes   | Age, sex, family history of diabetes, BMI, blood pressure, triglycerides, fasting plasma glucose   | 11                   |
| Kranup <sup>S15</sup>        | 5,791       | 2              | MI   | Age, sex, smoking, blood pressure, total cholesterol   | 45                   |

| First Author            | Sample Size | Event rate (%)  | Outcome                                 | Clinical Risk Factors   | Number of Added SNPs |
|-------------------------|-------------|-----------------|---|---|----------------------|
| Butoescu <sup>S16</sup> | 316         | 54 <sup>a</sup> | Prostate cancer                         | PSA, prostate volume, DRE, TRUS   | 9                    |
| Yu <sup>S17</sup>       | 320         | 37              | Prostate cancer recurrence <sup>b</sup> | PSA, stage, Gleason score, surgical margin  | 1                    |
| Lind <sup>S18</sup>     | 1,016       | NR              | Endothelium-dependent vasodilation      | Sex, smoking, BMI, SBP, HDL and LDL cholesterol, triglycerides, antihypertensive medication, statins, CRP   | 3                    |
| Gui <sup>S19</sup>      | 2,292       | 50 <sup>a</sup> | CHD                                     | Age, sex, diabetes, smoking, hypertension, BMI, total cholesterol, triglycerides, alcohol consumption, family history of CHD  | 10                   |
| Pitkanen <sup>S20</sup> | 2,298       | 25              | Impaired fasting glucose                | Age, sex, smoking, BMI, mother's BMI, parental diabetes, SBP, fasting insulin   | 73                   |
| Lobato <sup>S21</sup>   | 1,948       | 14              | Mortality                               | Diabetes, CPB duration, EuroSCORE, grafts, beta blocker, statin   | 1                    |
| Morote <sup>8</sup>     | 703         | 35              | Prostate cancer recurrence <sup>b</sup> | Preoperative PSA, stage, Gleason score, surgical margin, lymph node involvement   | 3                    |
| Kertai <sup>S22</sup>   | 957         | 27              | Postoperative QTc prolongation          | Age, sex, self-reported ethnicity, comorbidities, medication use at hospital admission, type of cardiac surgical procedure, duration of aortic cross-clamp  | 2                    |
| Fang <sup>S23</sup>     | 1,957       | 48 <sup>a</sup> | Melanoma                                | Age, sex, pigmentation  | 11                   |
| Ribeiro <sup>S24</sup>  | 1,006       | 45 <sup>a</sup> | Prostate cancer                         | Age, PSA  | 29                   |
| Borque <sup>S25</sup>   | 670         |                 | Prostate cancer recurrence <sup>b</sup> | PSA, stage, Gleason score   | 4                    |
| Ruan <sup>7</sup>       | 2,846       | 49 <sup>a</sup> | Nasopharyngeal carcinoma                | Smoking, consumption of salted fish and preserved vegetables, family history of NPC   | 7                    |
| Huesing <sup>S26</sup>  | 13,636      | 43 <sup>a</sup> | Breast cancer                           | BMI, age at menarche, age at first full term pregnancy, age at menopause, smoking, menopausal status, alcohol consumption count of full term pregnancies, ever use of hormone replacement therapy | 18                   |
| Bolton <sup>S27</sup>   | 508         | 26              | CHD                                     | Age, sex, diabetes and/or glucose intolerance, smoking, SBP, total cholesterol/HDL cholesterol  | 27                   |
| Chang <sup>S28</sup>    | 268         | 22              | Edema                                   | Age, sex  | 28                   |

<sup>a</sup> Case control study, <sup>b</sup> assessed as biochemical recurrence. Abbreviations: BMD = bone mineral density; BMI = body mass index; CHD = coronary heart disease; CPB = cardiopulmonary bypass; CRP = C-reactive protein; CVD = cardiovascular disease; DBP = diastolic blood pressure; DRE = digital rectal examination; HbA1c = glycated hemoglobin A1c; HDL = high-density lipoprotein; HDL-C = high-density lipoprotein cholesterol; LDL = low-density lipoprotein; MI = myocardial infarction; NPC = nasopharyngeal carcinoma; PSA = prostate-specific antigen; NR = not reported; SBP = systolic blood pressure; SNP = single-nucleotide polymorphism; TRUS = transrectal ultrasound.



**Table S2.** AUC, NRI and IDI of the prediction models in the literature review

| First author                 | Baseline AUC | $\Delta$ AUC (P value or 95% CI) | NRI (P value or 95% CI) | Type of NRI              | Cut-offs used for NRI    | IDI (P value or 95% CI)       |
|------------------------------|--------------|----------------------------------|-------------------------|--------------------------|--------------------------|-------------------------------|
| Park <sup>S1</sup>           | 0.786        | -0.001 (0.99)                    | 0.040 (0.32)            | Categorical              | 2.20%, 16.44% and 16.44% | 0.021 (0.02)                  |
| Eriksson <sup>S2</sup>       | 0.63         | 0 (0.246)                        | 0.11 (0.005)            | Continuous               | NA                       | 0.003 (0.007)                 |
| Fava <sup>S3</sup>           | 0.743        | 0 (>0.05)                        | 0.002 (0.39)            | Categorical              | 6% and 20%               | 0.00449 (0.02)                |
| Kathiresan <sup>9</sup>      | 0.80         | 0 (NR)                           | NR (0.01)               | Categorical              | 10% and 20%              | NR (0.02)                     |
| Havulinna <sup>S4</sup>      | 0.8445       | 0.0006 (0.16)                    | -0.0008 (0.92)          | Categorical              | 5%, 10% and 20%          | 0.00062 (0.14)                |
| Grånsbo <sup>10</sup>        | 0.750        | 0.001 (NR)                       | 0.012 (0.043)           | Categorical              | 5%, 10% and 20%          | 0.001 (<0.001)                |
| Ripatti <sup>S5</sup>        | 0.871        | 0.001 (0.19)                     | 0.022 (0.182)           | Categorical              | 5%, 10% and 20%          | 0.004 (0.0006)                |
| Lim <sup>S6</sup>            | 0.810        | 0.001 (0.1057)                   | 0.019 (0.0495)          | Categorical              | 4%, 8%, 12% and 16%      | 0.002 (0.0131)                |
| Thanassoulis <sup>S7</sup>   | 0.820        | 0.002 (NR)                       | -0.01 (-0.052 to 0.033) | Categorical              | 6% and 20%               | 0.001 (-0.001 to 0.003)       |
| Fava <sup>S8</sup>           | 0.669        | 0.003 (>0.05)                    | 0.0659 (0.013)          | Continuous               | NA                       | 0.001452 (0.003)              |
| Brautbar <sup>S9</sup>       | 0.782        | 0.004 (0.001 to 0.007)           | 0.008 (0.31)            | Categorical              | 5%, 10% and 20%          | 0.002 (<0.015)                |
| Muehlschlegel <sup>S10</sup> | 0.777        | 0.005 (NS)                       | 0.195 (0.072)           | Categorical <sup>a</sup> | NR                       | 0.010 (0.053)                 |
| Park <sup>S11</sup>          | 0.735        | 0.005 (0.050)                    | 0.0173 (0.352)          | Categorical              | 10%, 20% and 30%         | 0.0041 (0.007)                |
| Juhola <sup>S12</sup>        | 0.733        | 0.009 (0.015)                    | 0.048 (0.0002)          | Categorical              | 5%, 10% and 20%          | 0.012 (<0.0001)               |
| Brautbar <sup>S13</sup>      | 0.742        | 0.009 (0.006 to 0.014)           | 0.073 (0.019 to 0.12)   | Categorical              | 5%, 10% and 20%          | 0.006 (NR)                    |
| Lyssenko <sup>S14</sup>      | 0.743        | 0.010 (0.0001)                   | 0.09 (<0.001)           | Categorical              | 10% and 20%              | NR (<0.001)                   |
| Kraruup <sup>S15</sup>       | 0.68         | 0.01 (0.002)                     | -0.02(NS)               | Categorical              | 5% and 10%               | 0.001 (NS)                    |
| Butoescu <sup>S16</sup>      | 0.770        | 0.011 (NR)                       | 0.403 (<0.001)          | Continuous               | NA                       | 0.015 (0.035)                 |
| Yu <sup>S17</sup>            | 0.783        | 0.011 (>0.050)                   | 0.137 (0.015)           | Categorical <sup>a</sup> | NR                       | 0.0175 (0.041)                |
| Lind <sup>S18</sup>          | 0.640        | 0.012 (0.09)                     | 0.035 (0.047)           | Categorical              | 30% and 50%              | 0.0072 (0.010)                |
| Gui <sup>S19</sup>           | 0.811        | 0.013 (0.17)                     | 0.04850 (<0.001)        | Categorical              | 20%                      | 0.027 (<0.001)                |
| Pitkanen <sup>S20</sup>      | 0.678        | 0.014 (0.007)                    | 0.163 (0.001)           | Continuous               | NA                       | 0.012 (1.8x10 <sup>-5</sup> ) |
| Lobato <sup>S21</sup>        | 0.725        | 0.015 (NR)                       | 0.194 (0.005)           | Categorical <sup>a</sup> | NR                       | 0.022 (0.01)                  |
| Morote <sup>8</sup>          | 0.76         | 0.02 (0.092)                     | 0.233 (0.003)           | Continuous <sup>a</sup>  | NR                       | NR (<0.001)                   |
| Kertai <sup>S22</sup>        | 0.749        | 0.024 (0.001)                    | 0.308 (0.0003)          | Continuous               | NA                       | 0.02 (0.000024)               |
| Fang <sup>S23</sup>          | 0.64         | 0.03 (0.0000601)                 | 0.0109 (0.6076)         | Categorical              | 20% and 50%              | 0.0350 (<0.0001)              |
| Ribeiro <sup>S24</sup>       | 0.6476       | 0.033 (0.0002)                   | 0.095 (<0.0001)         | Categorical              | 15% and 45%              | 0.021 (<0.0001)               |
| Borque <sup>S25</sup>        | 0.728        | 0.034 (0.025)                    | 1.242 (<0.001)          | Continuous               | NA                       | NR (<0.001)                   |

| First author           | Baseline AUC | $\Delta$ AUC (P value or 95% CI) | NRI (P value or 95% CI) | Type of NRI | Cut-offs used for NRI | IDI (P value or 95% CI) |
|------------------------|--------------|----------------------------------|-------------------------|-------------|-----------------------|-------------------------|
| Ruan <sup>7</sup>      | 0.70         | 0.04 (<0.001)                    | 0.16 (<0.001)           | Categorical | 20% and 30%           | 0.05 (<0.001)           |
| Huesing <sup>S26</sup> | 0.564        | 0.041 (NR)                       | 0.158 (NR)              | Categorical | 1%, 1.66% and 3.5%    | 0.0016 (NR)             |
| Bolton <sup>S27</sup>  | 0.671        | 0.069 (0.0001)                   | 0.544 (<0.001)          | Categorical | 20%                   | 0.04 (0.02 to 0.06)     |
| Chang <sup>S28</sup>   | 0.702        | 0.088 (0.002)                    | 0.300 (0.005)           | NR          | NR                    | 0.128 (<0.001)          |

<sup>a</sup> This study did not report whether the categorical or continuous NRI was used. The type of NRI was inferred from the text. Abbreviations:  $\Delta$  AUC = increment in the area under the receiver operating characteristic curve; AUC = area under the receiver operating characteristic curve; CI = confidence interval; IDI = integrated discrimination improvement; NA = not applicable; NR = not reported; NRI = net reclassification improvement; NS = not statistically significant



**Figure S1.** Summary of literature search and selection. <sup>a</sup> Keywords: genetic, genomic, polygenic, polymorphisms, DNA. Abbreviations: AUC = area under the receiver operating characteristic curve; IDI = integrated discrimination improvement; NRI = net reclassification improvement; SNP = single nucleotide polymorphism

## Reference list of the 32 included articles in this study

- S1. Park MW, Her SH, Kim CJ, et al. Evaluation of the incremental prognostic value of the combination of CYP2C19 poor metabolizer status and ABCB1 3435 TT polymorphism over conventional risk factors for cardiovascular events after drug-eluting stent implantation in East Asians. *Genet Med*. 2016;18(8):833-841.
- S2. Eriksson J, Evans DS, Nielson CM, et al. Limited clinical utility of a genetic risk score for the prediction of fracture risk in elderly subjects. *J Bone Miner Res*. 2015;30(1):184-194.
- S3. Fava C, Ohlsson T, Sjögren M, et al. Cardiovascular consequences of a polygenetic component of blood pressure in an urban-based longitudinal study. *J Hypertens*. 2014;32(7):1424-1428.
9. Kathiresan S, Melander O, Anevski D, et al. Polymorphisms associated with cholesterol and risk of cardiovascular events. *N Engl J Med*. 2008;358(12):1240-1249.
- S4. Havulinna AS, Kettunen J, Ukkola O, et al. A blood pressure genetic risk score is a significant predictor of incident cardiovascular events in 32 669 individuals. *Hypertension*. 2013;61(5):987-994.
10. Gränsbo K, Almgren P, Sjögren M, et al. Chromosome 9p21 genetic variation explains 13% of cardiovascular disease incidence but does not improve risk prediction. *J Intern Med*. 2013;274(3):233-240.
- S5. Ripatti S, Tikkanen E, Orho-Melander M, et al. A multilocus genetic risk score for coronary heart disease: case-control and prospective cohort analyses. *Lancet*. 2010;376(9750):1393-1400.
- S6. Lim NK, Lee JY, Lee JY, et al. The role of genetic risk score in predicting the risk of hypertension in the Korean population: Korean genome and epidemiology study. *PLoS One*. 2015;10(6):e0131603.
- S7. Thanassoulis G, Peloso GM, Pencina MJ, et al. A genetic risk score is associated with incident cardiovascular disease and coronary artery calcium: the Framingham Heart Study. *Circ Cardiovasc Genet*. 2012;5(1):113-121.
- S8. Fava C, Sjogren M, Olsson S, et al. A genetic risk score for hypertension associates with the risk of ischemic stroke in a Swedish case-control study. *Eur J Hum Genet*. 2015;23(7):969-974.
- S9. Brautbar A, Ballantyne CM, Lawson K, et al. Impact of adding a single allele in the 9p21 locus to traditional risk factors on reclassification of coronary heart disease risk and implications for lipid-modifying therapy in the atherosclerosis risk in communities study. *Circ Cardiovasc Genet*. 2009;2(3):279-285.
- S10. Muehlschlegel JD, Liu KY, Perry TE, et al. Chromosome 9p21 variant predicts mortality after coronary artery bypass graft surgery. *Circulation*. 2010;122(11\_suppl\_1):S60-S65.
- S11. Park HY, Choi HJ, Hong YC. Utilizing genetic predisposition score in predicting risk of type 2 diabetes mellitus incidence: A community-based cohort study on middle-aged Koreans. *J Korean Med Sci*. 2015;30(8):1101-1109.
- S12. Juhola J, Oikonen M, Magnussen CG, et al. Childhood physical, environmental, and genetic predictors of adult hypertension: The cardiovascular risk in Young Finns Study. *Circulation*. 2012;126(4):402-409.
- S13. Brautbar A, Pompeii LA, Dehghan A, et al. A genetic risk score based on direct associations with coronary heart disease improves coronary heart disease risk prediction in the Atherosclerosis Risk in Communities (ARIC), but not in the Rotterdam and Framingham Offspring, Studies. *Atherosclerosis*. 2012;223(2):421-426.
- S14. Lyssenko V, Jonsson A, Almgren P, et al. Clinical risk factors, DNA variants, and the development of type 2 diabetes. *N Engl J Med*. 2008;359(21):2220-2232.
- S15. Krarup NT, Borglykke A, Allin KH, et al. A genetic risk score of 45 coronary artery disease risk variants associates with increased risk of myocardial infarction in 6041 Danish individuals. *Atherosclerosis*. 2015;240(2):305-310.
- S16. Butoescu V, Ambroise J, Stainier A, et al. Does genotyping of risk-associated single nucleotide polymorphisms improve patient selection for prostate biopsy when combined with a prostate cancer risk calculator? *Prostate*. 2014;74(4):365-371.
- S17. Yu C-C, Lin VC, Huang C-Y, et al. Prognostic significance of Cyclin D1 polymorphisms on prostate-

- specific antigen recurrence after radical prostatectomy. *Ann Surg Oncol*. 2013;20(S3):492-499.
- S18. Lind L, Syvänen AC, Axelsson T, et al. Variation in genes in the endothelin pathway and endothelium-dependent and endothelium-independent vasodilation in an elderly population. *Acta Physiol*. 2013;208(1):88-94.
- S19. Gui L, Wu F, Han X, et al. A multilocus genetic risk score predicts coronary heart disease risk in a Chinese Han population. *Atherosclerosis*. 2014;237(2):480-485.
- S20. Pitkanen N, Juonala M, Ronnema T, et al. Role of conventional childhood risk factors versus genetic risk in the development of type 2 diabetes and impaired fasting glucose in adulthood: The cardiovascular risk in Young Finns Study. *Diabetes care*. 2016;39(8):1393-1399.
- S21. Lobato RL, White WD, Mathew JP, et al. Thrombomodulin gene variants are associated with increased mortality after coronary artery bypass surgery in replicated analyses. *Circulation*. 2011;124(11 Suppl):S143-148.
8. Morote J, del Amo J, Borque A, et al. Improved prediction of biochemical recurrence after radical prostatectomy by genetic polymorphisms. *J Urol*. 2010;184(2):506-511.
- S22. Kertai MD, Ji Y, Li Y-J, et al. Interleukin-1 $\beta$  gene variants are associated with QTc interval prolongation following cardiac surgery: A prospective observational study. *Can J Anaesth*. 2016;63(4):397-410.
- S23. Fang S, Han J, Zhang M, et al. Joint effect of multiple common SNPs predicts melanoma susceptibility. *PLoS ONE*. 2013;8(12):e85642.
- S24. Ribeiro RJ, Monteiro CP, Azevedo AS, et al. Performance of an adipokine pathway-based multilocus genetic risk score for prostate cancer risk prediction. *PLoS ONE*. 2012;7(6):e39236.
- S25. Borque Á, del Amo J, Esteban LM, et al. Genetic predisposition to early recurrence in clinically localized prostate cancer. *BJU Int*. 2013;111(4):549-558.
7. Ruan HL, Qin HD, Shugart YY, et al. Developing genetic epidemiological models to predict risk for nasopharyngeal carcinoma in high-risk population of China. *PLoS One*. 2013;8(2):e56128.
- S26. Husing A, Canzian F, Beckmann L, et al. Prediction of breast cancer risk by genetic risk factors, overall and by hormone receptor status. *J Med Genet*. 2012;49(9):601-608.
- S27. Bolton JL, Stewart MC, Wilson JF, et al. Improvement in prediction of coronary heart disease risk over conventional risk factors using SNPs identified in genome-wide association studies. *PLoS One*. 2013;8(2):e57310.
- S28. Chang T-J, Liu P-H, Liang Y-C, et al. Genetic predisposition and nongenetic risk factors of thiazolidinedione-related edema in patients with type 2 diabetes. *Pharmacogenet Genomics*. 2011;21(12):829-836.





# 6

## **Simultaneous use of AUC, NRI and IDI for the evaluation of clinical prediction models: reporting and interpretation practices**

Forike K. Martens, Elisa C. M. Tonk, and A. Cecile J.W. Janssens  
Prepared for submission



## Abstract

For evaluating the improvement in risk predictions, the net reclassification improvement (NRI) and integrated discrimination improvement (IDI) are increasingly used in addition to the area under the receiver operating characteristic (ROC) curve (AUC or c-statistic). We evaluated how researchers defined, calculated and interpreted these when simultaneously used in the assessment of the improvement in predictive performance of clinical prediction models. Fifty-six articles met our inclusion criteria. Researchers defined the AUC as measure of discriminative ability in over 69% of the articles, the NRI in 17% and the IDI 22%. Values of the metrics were interpreted as indicative of improvement when they were statistically significant, irrespective of their values. Hence, also the overall conclusions were based on the statistical significance of the metrics. When the interpretations were discordant ( $n = 9$ ) the conclusion appears based on the statistical significance of the NRI or IDI values in most of them (7 out of 9). Better understanding of the meaning and relevance of the metrics can facilitate more meaningful interpretation of prediction studies.

## Introduction

The area under the receiver operating characteristic (ROC) curve (AUC or c-statistic) (1) is the most commonly used metric for the evaluation of prediction models for their ability to discriminate between individuals who will or will not manifest an outcome of interest (referred to as events and nonevents in this article). The increment in AUC ( $\Delta$ AUC) is the standard for assessing the improvement in discrimination after adding new risk factors, such as biomarkers, genetic factors or imaging results to existing models (2–4).

In the past decade, researchers have widely adopted new metrics for evaluating the improved predictive performance of updated prediction models, including the net reclassification improvement (NRI) and integrated discrimination improvement (IDI) (2,5,6). While these are all metrics of discrimination (7), they are computationally different. AUC gives the probability that predicted risks correctly identify a random pair of an event and nonevent (rank order of events and nonevents), NRI quantifies the improvement in the classification of risks and IDI assesses the increase in the risk difference between events and nonevents (2,7). Often IDI is described as a measure to assess improvement in integrated sensitivity without compromising integrated specificity (2), but this does not immediately provide insight in what is measured, and may hamper an easy interpretation. It also appears that researchers frequently use NRI and IDI in addition to AUC without explaining the differences between the metrics and why each metric is assessed (8). For example, NRI is often referred to as metric of the discriminative ability of a model without elaborating on what is specifically quantified by the measure, namely reclassification (9,10). Moreover, researchers often define, for example, IDI as a measure of reclassification instead of improvement in the risk differences between events and nonevents (11,12). Interpreting findings is challenging when it is not clear among researchers what each metric adds to the evaluation of prediction models, especially when the findings contradict.

Contradictory findings are frequently attributed to limitations of the metrics. The AUC is criticized for being insensitive to detect improvements in prediction that result from adding clinically relevant risk factors (2,5,13–16), and the NRI and IDI may pick up subtle changes in predicted risks suggesting improvement in prediction while the rank order of events and nonevents has not changed (17–19). In previous work, we showed that results can contradict

as each metric assesses a different aspect of the improvement in predictive performance (20).  $\Delta$ AUC quantifies changes in rank order, NRI assesses changes in risk classification, and IDI changes in the risk differences. Because of this different emphasis, the addition of novel risk factors might increase the risk differences between events and nonevents measured by the IDI, but not change the rank order of events and nonevents, for example when the AUC of the baseline model is high (20).

We recently evaluated how researchers describe and interpret results of the  $\Delta$ AUC, NRI and IDI when the three metrics are simultaneously used to assess the improvement in predictive performance (further referred to as “improvement”) of polygenic prediction models (8). We found that AUC and NRI were defined as discrimination and reclassification in two thirds of the articles (63% and 70%), but that none of the definitions for IDI referred to improvement in risk differences. However, we have not evaluated whether researchers elaborated on what the AUC measures and how the c-statistic is calculated. Furthermore, we observed that the evaluation of the metrics generally followed their statistical significance irrespective of their values and small non-statistically significant  $\Delta$ AUC values were interpreted as indicative of improvement when NRI and IDI were significant. It is unknown whether the results of our previous study are specific to polygenic risk studies, because the field of polygenic prediction is new and researchers may have relatively little experience with these metrics. In this paper, we evaluate the simultaneous use of the three metrics in the assessment of improvement in prediction of disease and elaborate on researchers’ understanding of the AUC, focusing on recently published clinical prediction studies.

## Methods

### Literature search

We collected empirical studies that stated in the methods that the improvement in predictive performance of clinical prediction models was evaluated by assessing  $\Delta$ AUC, NRI, and IDI and that reported the results of all metrics in the article. Using Thomson Reuters Web of Knowledge (version 5.23) we retrieved all publications from 2016 that cited the article by Pencina et al. that introduced the NRI and IDI (search date 28 December 2016) (2). Articles were excluded when they performed a simulation or methodological study or discussed a

genetic prediction model. For articles in which multiple prediction models were discussed we chose one baseline model with its updated model. For example, when different risk factors were added to the same baseline model, the same risk factors were added to different baseline models, the same baseline and updated models were used but in different study populations or when the same models were used for different outcomes. Our selection of models was done in the following order: the outcome of the model was the main focus of the paper (i.e., the main conclusions were drawn for this model), the baseline model with the highest number of risk factors included, the highest number of risk factors added to the baseline model, or the model for which the largest sample was used.

**Box 1.** Definitions and calculation methods of AUC, NRI and IDI

| <b>Metric</b>   | <b>Definition</b>  | <b>Calculation method</b>  |
|---|--|--|
| Area Under the Receiver Operating Characteristic Curve (AUC)* | Discrimination / the probability that predicted risks correctly identify a random pair of an event and nonevent  | C-statistic / c-index / trapezoidal rule / the proportion of all possible pairs (an event and nonevent) in which the event had a higher predicted risk than the nonevent   |
| Net reclassification improvement (NRI)                        | Reclassification / improvement in risk classification  | Categorical: sum of net percentages of correctly reclassified persons with and without an event;<br><br>Continuous: sum of net percentages of persons with and without an event correctly assigned a higher (event) or lower (no event) predicted risk |
| Integrated discrimination improvement (IDI)                   | Improvement in discrimination slopes** / improvement in risk differences improvement in integrated sensitivity without compromising integrated specificity | Difference between discrimination slopes of baseline and updated models / difference between mean predicted risks of persons with and without an event between models  |

Definitions and calculations are based on references (1,2,37).

\*  $\Delta$ AUC is the AUC of the updated model minus the AUC of the baseline model; \*\* discrimination slope is the mean predicted risk of events minus the mean predicted risk of nonevents

## Data extraction

For all selected models, we recorded study characteristics, including sample size, event rate, clinical risk factors in the baseline prediction models and the risk factor(s) added. Depending on the study design that was used, the event rate was the incidence, prevalence or the proportion of cases in the population. Furthermore, we extracted AUC values of the baseline and updated models,  $\Delta$ AUC, NRI and IDI with corresponding *P* values or confidence intervals, and the version of NRI that was used: categorical or continuous (Box 1).

Definitions of AUC, NRI and IDI and their calculation methods were extracted verbatim from the methods section of the included articles. The numerical results, the interpretation of AUC, NRI and IDI and the overall conclusions were extracted from the results and discussion sections. We imported all extracted texts into a spreadsheet for a content analysis.

**Table 1.** Examples of concordant and discordant interpretations of the performance metrics

| Study            | From the publication         |                           |                              | Our assessment  |       |       |                             |
|------------------|------------------------------|---------------------------|------------------------------|-----------------|-------|-------|-----------------------------|
|                  | Metrics                      |                           |                              | Model improved? |       |       | Interpretations discordant? |
|                  | $\Delta$ AUC                 | NRI                       | IDI                          | $\Delta$ AUC    | NRI   | IDI   |                             |
| Dhana (32)       | 0.001 (NR)                   | 0.05<br>(-0.01 to 0.12)   | 0.001<br>(-0.001 to 0.001)   | No              | No    | No    | No                          |
| Kim (33)         | 0.0047<br>(0.0001 to 0.0128) | 0.104<br>(0.031 to 0.247) | 0.0041<br>(0.0001 to 0.0120) | Yes             | Yes   | Yes   | No                          |
| Gravensen (34)   | 0.006 (0.032)                | 0.01<br>(0.718)           | 0.006<br>(0.029)             | Yes             | No    | Yes   | Yes                         |
| Wotherspoon (35) | 0.008 (0.17)                 | 0.306<br>(0.003)          | 0.009<br>(0.11)              | No              | Yes   | Yes   | Yes                         |
| Vandenput (36)   | 0.01 (NS)                    | 0.178<br>(S)              | 0.004<br>(S)                 | No              | [Yes] | [Yes] | Yes                         |
| Nagahara (10)    | 0.06 (0.07)                  | 0.60<br>(0.0049)          | 0.054<br>(0.0072)            | No              | Yes   | Yes   | Yes                         |

Values are point estimates with *P* values or 95% confidence intervals between brackets. Labeling of researchers' interpretations of the metrics is described in the Methods. Square brackets indicate that the researchers considered the observed improvement of the model to be minimal.

Abbreviations:  $\Delta$ AUC = increment in the area under the receiver operating characteristic curve; IDI = integrated discrimination improvement; NR = not reported; NRI = net reclassification improvement; NS = not statistically significant; S = statistically significant (per researchers' reporting).

## Content analysis

We evaluated whether common definitions and approaches were used (Box 1). We used excerpts of the results section or from the discussions section when no description was given, to assess how the values of  $\Delta$ AUC, NRI and IDI were described and interpreted. We documented effect sizes, *P* values or confidence intervals, as well as whether and how the values were interpreted in terms of improvement in predictive performance. We documented “not reported” when no interpretation was described, and “yes” or “no” when the researchers wrote that the value of the metric was or was not indicative of improvement. When they wrote that a metric indicated slight or marginal improvement, we documented “minimal improvement.” Interpretations were considered discordant when some of the metrics were described as indicative of improvement and others were not (see Table 1 for examples). Lastly, using excerpts from the discussion, we evaluated how the overall improvement of the predictive performance of the model was concluded. Descriptions of improvement were categorized as “improvement”, “minimal improvement”, “no improvement” or “inconclusive”. We marked conclusions as “inconclusive” when researchers could not come to an overall conclusion because of discordant observations. Two reviewers (F.K.M and E.C.M.T) independently evaluated the descriptions and disagreements were resolved by discussion.

## Results

In 2016, 309 publications cited the 2008 article that introduced the NRI and IDI (Appendix Figure 1). Of these, 182 were excluded because they did not use all three metrics, 47 because they did not present empirical data, and 24 because they did not present non-genetic clinical risk models that were updated with clinical risk factors. Fifty-six articles were included. Outcomes of the included prediction models were cardiovascular related diseases (*n* = 20), mortality (*n* = 20), diabetes related (*n* = 3) and other disease related outcomes (*n* = 13; Appendix Table 1).

In most of the 56 included articles, researchers reported a definition of AUC (*n* = 47; 84%), NRI (*n* = 49; 88%) and IDI (*n* = 45; 80%; Table 2). In all others, they merely stated that the measures were calculated. AUC was described as discrimination in 69% of the 47 articles that gave its definition and

as the probability that the predicted risks correctly identify a random pair of an event and nonevent in 2%. NRI was described as a measure of reclassification in 24% of the 49 articles, and IDI as improvement in discrimination slopes or risk differences in 18% of the 45 articles. In the 66% of the articles that reported definitions of all three metrics, the same definition was used for all three, namely a metric of discrimination ( $n = 5$ ) or more general descriptions such as risk estimation, model performance, predictive ability or improvement ( $n=4$ ).

In 57% of the articles, researchers indicated how AUC was calculated and all of these mentioned the c-statistic of which only one added how this was done (Table 2). Only in 16% and 21% of the articles, researchers explained how NRI and IDI were calculated, of which 56% were correct for NRI, and 67% for IDI. When the formula of NRI was not correctly described, the 'errors' were often in the details, for example, omitting to mention that NRI is the *sum of net* percentages of individuals *with and without an event* or the *proportion* of participants reclassified.

When estimates of the  $\Delta$ AUC, NRI, or IDI were statistically significant, they were always interpreted as indicating that the model had improved (Table 3). Only in four articles, researchers added that the improvement in AUC, while statistically significant, was minimal and in one article they considered the improvement in NRI minimal. When values were not statistically significant, six out of 13  $\Delta$ AUC values, one out of six NRI values and five out of six IDI values were still interpreted as being indicative of improvement. For these, three of the  $\Delta$ AUC values were low (0.01, 0.01, 0.02) and three were higher (0.03, 0.04, 0.05), the NRI was 0.0027, and the IDI values were all 0.03 or lower, meaning a less than 3% absolute increase in the risk differences between events and nonevents. All 5 non-statistically significant IDI values were accompanied by a statistically significant NRI.

In 45 (80%) out of 56 articles, researchers had interpreted whether all three values of the metrics were indicative of improvement; in others, they only interpreted some of the values ( $n = 8$ ) or none of them ( $n = 3$ ) (Appendix Table 2). In 35 (78%) of the 45 articles, researchers reported that  $\Delta$ AUC, NRI and IDI values all showed evidence for improvement of the predictive performance. Only in one article, in which reported values for  $\Delta$ AUC and IDI were virtually zero and NRI 0.05, researchers reported that none of the metrics indicated improvement. In nine (20%) of the 45 articles, the interpretations of metric improvement as described by the researchers were discordant. In seven of

these, the researchers wrote that NRI and IDI suggested that the model had improved, but  $\Delta$ AUC did not.

Finally, researchers concluded in 48 (86%) of the 56 articles that the predictive performance of the clinical model had improved from the additional risk factor(s), three of which commented that the improvement was minimal (Appendix Table 2). Others concluded that the model did not improve ( $n = 4$ ), were inconclusive ( $n = 1$ ) or refrained from making an overall conclusion ( $n = 3$ ). As expected, most ( $n = 32$ ) of the 35 articles in which the three metrics were considered to be improved concluded overall improvement of the model. Also, when the interpretations of the metrics were discordant ( $n = 9$ ), in all but one of the articles, researchers concluded that the prediction model improved from the additional risk factor(s).



**Table 2.** Definition and calculation method of AUC, NRI and IDI as reported in the 56 articles

| <b>Metric</b>   | <b>Definition</b>   | <b>Percentage of articles (number)</b> | <b>Calculation method</b>  | <b>Percentage of articles (number)</b> |
|---|---|--|--|--|
| AUC   | <i>Not reported</i>   | 16 (9)                                 | <i>Not reported</i>  | 43 (24)                                |
|   | <i>Reported</i>   | 84 (47)                                | <i>Reported</i>  | 57 (32)                                |
|   | * The probability that predicted risks correctly identify a random pair of an event and nonevent                              | 2 (1)                                  | * C-statistic / c-index  | 100 (32)                               |
|   | Discrimination  | 69 (32)                                |  |  |
|   | Diagnostic accuracy / prognostic accuracy   | 4 (2)                                  |  |  |
|   | Prognostic / predictive utility   | 4 (2)                                  |  |  |
| NRI   | Performance/model performance/prediction performance  | 6 (3)                                  |  |  |
|   | Improvement, predictive power, predictive value, prognostic ability, predictive ability, risk classification, risk estimation | 15 (7)                                 |  |  |
|   | <i>Not reported</i>   | 12 (7)                                 | <i>Not reported</i>  | 84 (47)                                |
|   | <i>Reported</i>   | 88 (49)                                | <i>Reported</i>  | 16 (9)                                 |
|   | * Reclassification / improvement in risk classification   | 49 (24)                                | * Sum of net percentages of correctly reclassified persons with and without an event / sum of the percentage of events assigned a higher probability and nonevents a lower probability   | 56 (5)                                 |
|   | Discrimination  | 17 (8)                                 | Difference of the difference between participants moving up and down for events and nonevents / proportion of participants reclassified / twice the difference in the probabilities of upward reclassification for events minus that for nonevents / sum of upward proportions of events and the downward proportions of nonevents | 44 (4)                                 |
|   | Model performance   | 6 (3)                                  |  |  |
|   | Predictive ability  | 4 (2)                                  |  |  |
|   | Prognostic ability  | 4 (2)                                  |  |  |
|   | (Risk) classification, utility on risk classification, risk stratification  | 8 (4)                                  |  |  |
| Accuracy, changes in estimated prediction probabilities, clinical utility, model performance, prediction, risk estimation | 12 (6)  |  |  |  |

| <b>Metric</b> | <b>Definition</b>  | <b>Percentage of articles (number)</b> | <b>Calculation method</b>   | <b>Percentage of articles (number)</b> |
|---------------|--|--|---|--|
| IDI           | <i>Not reported</i>  | 20 (11)                                | <i>Not reported</i>   | 79 (44)                                |
|               | <i>Reported</i>  | 80 (45)                                | <i>Reported</i>   | 21 (12)                                |
|               | * Improvement in discrimination slopes / improvement in risk differences / improvement in integrated sensitivity without compromising integrated specificity / increase average sensitivity without reducing specificity | 18 (8)                                 | * Differences in discrimination slopes between models / the difference between the mean of the estimated prediction probabilities for persons with and without an event   | 67 (8)                                 |
|               | Discrimination   | 22 (10)                                | The sum of the average increase in sensitivity and specificity / difference between the integrated sensitivity gain and the integrated specificity loss between models / comparing the integrals of sensitivity and specificity between models / difference between the integrated difference in sensitivity and one minus specificity between models | 33 (4)                                 |
|               | Reclassification / (risk) classification / classification ability, risk stratification, utility on risk classification   | 27 (12)                                |   |  |
|               | Model performance  | 7 (3)                                  |   |  |
|               | Sensitivity  | 4 (2)                                  |   |  |
|               | Predictive ability   | 7 (3)                                  |   |  |
|               | Accuracy, clinical utility, differentiation along a continuum of predicted risks, improvement, predictive utility, prognostic value, risk estimation   | 15 (7)                                 |   |  |

AUC = area under the receiver operating characteristic curve; IDI = integrated discrimination improvement; NRI = net reclassification improvement.

\* Definitions considered in line with those in Box 1.

**Table 3.** Improvement in predictive performance based on reported interpretations of  $\Delta$ AUC, NRI and IDI values by the statistical significance of the point estimates

|                               | Model improved? |                    |    |
|-------------------------------|-----------------|--------------------|----|
|                               | Yes             | Yes, but minimally | No |
| Statistically significant     |                 |                    |    |
| $\Delta$ AUC                  | 26              | 4                  | 0  |
| NRI                           | 44              | 1                  | 0  |
| IDI                           | 40              | 0                  | 0  |
| Not statistically significant |                 |                    |    |
| $\Delta$ AUC                  | 5               | 1                  | 7  |
| NRI                           | 1               | 0                  | 5  |
| IDI                           | 5               | 0                  | 1  |

Values are number of articles. Interpretations of  $\Delta$ AUC, NRI or IDI not counted in this table when the articles did not interpret the metrics ( $\Delta$ AUC,  $n = 4$ ; NRI,  $n = 5$ ; IDI,  $n = 9$ ) or did not report  $P$  values or confidence intervals ( $\Delta$ AUC,  $n = 11$ ; IDI,  $n = 1$ ). Abbreviations:  $\Delta$ AUC = increment in the area under the receiver operating characteristic curve; IDI = integrated discrimination improvement; NRI = net reclassification improvement

## Discussion

In the evaluation of the predictive performance of prediction models, the AUC is frequently complemented with NRI and IDI. When the results of the metrics are contradictory about the improvement in prediction, the interpretation of the findings is challenging. In this study we observed that articles often lack information about the meaning and calculation of AUC, NRI and IDI and what the added value is of using all three metrics. Researchers heavily relied on the statistical significance of the metrics to interpret their findings and reach their conclusions. When interpretations of the values of the metrics were discordant, researchers often concluded that the predictive performance of the prediction model was improved by the addition of the risk factor(s). In none of the articles, the researchers critically reflected on the different aspects of performance that are assessed by the three metrics.

Before interpreting the observations of our study, a limitation of the study needs a mention. We inferred researchers' knowledge about the metrics based on what they reported, but researchers may have a better understanding of the metrics that they didn't display in their articles (21). This may change the number of more extensive definitions and calculation methods, but does not change how their interpretation of the improved predictive performance mainly

followed the statistical significance.

Several observations in this study suggest that researchers have limited understanding of what aspects of the performance are measured by each of the metrics and how their values should be interpreted. First, in approximately 16%, 12% and 20% of the articles, researchers did not provide any definition of the AUC, NRI or IDI. When AUC was defined, only one article (2%) gave a more extensive description, where others defined AUC as discrimination (69%). Also, NRI was only defined as reclassification in half of the articles that gave a definition (49%) and IDI as improvement in risk differences (or related definition) in 18%. Moreover, in 5 out of 10 articles where IDI was defined as discrimination also NRI and AUC were, which suggests that researchers may not be aware or not care that IDI quantifies the improvement in risk differences and NRI the improvement in risk classification. When researchers provided definitions for all three metrics (66%), they often (24%) did not distinguish between the three as they described them with the same term, such as metrics of “discrimination”, “risk estimation”, “predictive ability”, “improvement” or “model performance”. The variety of definitions for NRI and IDI suggests that researchers may have insufficient understanding of the aspects of predictive performance that are assessed by each metric.

Second, researchers rarely described how NRI and IDI were calculated. While the calculation method of AUC was described in almost two thirds, the methods of NRI and IDI were only described in one fifth of the articles. It should be noted that the descriptions for the calculation method of AUC was generally no more than a mention of the c-statistic; whether researchers understand what exactly is calculated by the c-statistic cannot be concluded from our study. Since, the description of the calculation method of IDI was often taken verbatim from the article that introduced the metric (2), it cannot be concluded either whether researchers understand how the IDI is calculated.

Third, the statistical significance of the individual metrics, not the values of the metrics was the basis for inferences about the improvement in prediction and hence the overall conclusions were based on the statistical significance of the metrics, even when the values were low. As small values may be statistically significant in larger studies but of limited utility in clinical or public health practice, emphasis should be on the values rather than their statistical significance when making conclusions about the improvement in predictive performance of prediction models (22,23).

In comparison with our previous published article about the simultaneous use of AUC, NRI and IDI in polygenic prediction studies (8), researchers of the clinical prediction studies in the present article described more often how they calculated NRI and IDI. Definitions of IDI were more extensively described here, compared to the previous article (18% to 0%), while only one article provided a definition of AUC beyond discrimination, even though more articles defined AUC (71% to 56%). The increase in IDI definitions may be explained by the fact that the prediction studies in this article were more recently published and hence might have gained more insight in the IDI. However, this does explain the increase in AUC definitions, because this has been the standard for long. Additionally, there seems no difference in the understanding of the AUC between researchers of polygenic prediction studies and clinical prediction studies. Furthermore, in our other study (8) researchers also followed the statistical significance of the values in their interpretations of the metrics, however added more often that the values of the metrics were indicative of a *minimal* improvement, and more often considered in the conclusion that the overall improvement was only minimal. These reservations in the interpretations may be due to the lower  $\Delta$ AUC values in the polygenic prediction studies (median 0.01; IQR 0.002-0.02) (24) compared to the  $\Delta$ AUC values in the present study (median 0.02; IQR 0.01-0.04).

The fact that some metrics indicate improvement of the model and others do not is generally considered a problem of the metrics (17–19), whereas it may also reflect that the addition of variables improves certain aspects of predictive performance but not others. For instance, AUC has been criticized for being insensitive and not intuitive (2,5,13–16), but improving the rank order of events and nonevents requires a risk factor that can substantially change the rank order when baseline AUC is higher. As a result, adding a strong risk factor may not easily change that ranking, showing minimal improvement in AUC, but it may widen the risk differences between events and nonevents, as indicated by a positive IDI (20). Also, when a risk factor does not increase AUC, we may see a positive NRI when risk thresholds are in the center of the risk distribution where many individuals can move across thresholds with minimal changes in predicted risks (25). That is why NRI is sensitive in identifying minor changes in predicted risks (17–19) and may be statistically significant, while AUC remains virtually unchanged (25,26).

The difficulty that researchers may have with the interpretation of the metrics is understandable as the metrics are not intrinsically intuitive. The

interpretation of the NRI is difficult because it is the sum of two fractions with different denominators (the number of events and nonevents) and the value cannot be interpreted as a percentage. Because there is no clear meaning of the number itself, it has been recommended to report the NRI for events and nonevents separately (27). Also, when IDI is explained as the differences in discrimination slopes, it may not be obvious that it is a metric of improvement in the risk differences between events and nonevents. Similarly, when AUC is explained as the probability that predicted risks correctly identify a random pair of an event and nonevent, it may not be apparent that AUC informs about the shape and overlap of risk distributions of events and non-events (28).

Pepe et al. (22) demonstrated that metrics for evaluating the value of adding risk factors to a prediction model have the same null hypothesis, however, this does not mean that these metrics can be used interchangeably (29), because they assess different aspects of model performance. Which metric would be of interest is determined by the research question. When the question is whether a prediction model can stratify a population in certain risk groups, the primary interest is in how well the prediction model can classify events above a threshold and nonevents below. Because the magnitude of the categorical NRI depends on the number of thresholds, it is recommended to only use the NRI with established clinically meaningful risk thresholds and report the NRI for events and nonevents separately to facilitate interpretation (27,30). When the interest is in whether individual risks improve, the IDI should be used; and when the question is whether overall the ability of the model to discriminate events and nonevents improved in the updated model, the  $\Delta$ AUC is the preferred measure.

Determining whether the improvement in predictive performance is high enough, depends on what the model will be used for (31). Relying on the statistical significance when improvements are minimal leads to false positive conclusions about the added value of the risk factor, because very small effects that are statistically significant in large studies may have no clinical value. Insight in the different aspects of the predictive performance and the meaning and applicability of the metrics can facilitate the right use of the metrics and enhance the interpretation of prediction studies.

## References

1. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982 Apr;143(1):29–36.
2. Pencina MJ, D'Agostino Sr. RB, D'Agostino Jr. RB, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med*. 2008;27(2):157–172.
3. Steyerberg EW, Pencina MJ, Lingsma HF, Kattan MW, Vickers AJ, Van Calster B. Assessing the incremental value of diagnostic and prognostic markers: a review and illustration. *Eur J Clin Invest*. 2012;42(2):216–228.
4. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*. 1983;148(3):839–843.
5. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation*. 2007;115(7):928–935.
6. Pencina MJ, D'Agostino RB, Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Stat Med*. 2011;30(1):11–21.
7. Pencina MJ, D'Agostino RB, Pencina KM, Janssens ACJW, Greenland P. Interpreting incremental value of markers added to risk prediction models. *Am J Epidemiol*. 2012;176(6):473–481.
8. Martens FK, Tonk ECM, Janssens ACJW. Evaluation of polygenic risk models using multiple performance measures: A critical assessment of discordant results. *Genet Med*. 2019;21(2):391–397.
9. Mise K, Hoshino J, Ueno T, Hazue R, Hasegawa J, Sekine A, et al. Prognostic value of tubulointerstitial lesions, Urinary N-Acetyl- $\beta$ -d-Glucosaminidase, and Urinary  $\beta$ 2-Microglobulin in patients with type 2 diabetes and biopsy-proven diabetic nephropathy. *Clin J Am Soc Nephrol*. 2016;11(4):593–601.
10. Nagahara Y, Motoyama S, Sarai M, Ito H, Kawai H, Takakuwa Y, et al. Eicosapentaenoic acid to arachidonic acid (EPA/AA) ratio as an associated factor of high risk plaque on coronary computed tomography in patients without coronary artery disease. *Atherosclerosis*. 2016;250:30–37.
11. Park M-W, Her SH, Kim CJ, SunCho J, Park G-M, Kim T-S, et al. Evaluation of the incremental prognostic value of the combination of CYP2C19 poor metabolizer status and ABCB1 3435 TT polymorphism over conventional risk factors for cardiovascular events after drug-eluting stent implantation in East Asians. *Genet Med*. 2016;18(8):833–841.
12. Lee JM, Kim CH, Koo B-K, Hwang D, Park J, Zhang J, et al. Integrated myocardial perfusion imaging diagnostics improve detection of functionally significant coronary artery stenosis by  $^{13}\text{N}$ -ammonia positron emission tomography. *Circ Cardiovasc Imaging*. 2016;9(9):e004768.
13. Pencina MJ, D'Agostino RB, Demler O V. Novel metrics for evaluating improvement in discrimination: net reclassification and integrated discrimination improvement for normal variables and nested models. *Stat Med*. 2012;31(2):101–113.
14. Pepe MS, Janes HE. Gauging the performance of SNPs, biomarkers, and clinical factors for predicting risk of breast cancer. *J Natl Cancer Inst*. 2008;100(14):978–979.
15. Pepe MS. Limitations of the Odds Ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *Am J Epidemiol*. 2004;159(9):882–890.
16. Ware JH. The limitations of risk factors as prognostic tools. *N Engl J Med*. 2006;355(25):2615–2617.
17. Pepe MS, Janes H, Li CI. Net risk reclassification P values: Valid or misleading? *JNCI J Natl Cancer Inst*. 2014;106(4):dju041.
18. Gerds TA, Hilden J. Calibration of models is not sufficient to justify NRI. *Stat Med*. 2014;33(19):3419–20.
19. Hilden J, Gerds TA. A note on the evaluation of novel biomarkers: do not rely on integrated discrimination improvement and net reclassification index. *Stat Med*. 2014;33(19):3405–3414.
20. Martens FK, Tonk ECM, Kers JG, Janssens ACJW. Small improvement in the area under the receiver operating characteristic curve indicated small changes in predicted risks. *J Clin Epidemiol*. 2016;79:159–164.

21. Iglesias AI, Mihaescu R, Ioannidis JPA, Khoury MJ, Little J, van Duijn CM, et al. Scientific reporting is suboptimal for aspects that characterize genetic risk prediction studies: a review of published articles based on the Genetic Risk Prediction Studies statement. *J Clin Epidemiol*. 2014;67(5):487–499.
22. Pepe MS, Kerr KF, Longton G, Wang Z. Testing for improvement in prediction model performance. *Stat Med*. 2013;32(9):1467–1482.
23. Vickers AJ, Cronin AM, Begg CB. One statistical test is sufficient for assessing new predictive markers. *BMC Med Res Methodol*. 2011;11(1):13.
24. Martens FK, Tonk ECM, Janssens ACJW. Evaluation of polygenic risk models using multiple performance measures: a critical assessment of discordant results. *Genet Med*. 2019;21(2):391–397.
25. Mihaescu R, van Zitteren M, van Hoek M, Sijbrands EJG, Uitterlinden AG, Witteman JCM, et al. Improvement of risk prediction by genomic profiling: Reclassification measures versus the area under the receiver operating characteristic curve. *Am J Epidemiol*. 2010;172(3):353–361.
26. Janssens ACJW, Khoury MJ. Assessment of improved prediction beyond traditional risk factors: When does a difference make a difference? *Circ Cardiovasc Genet*. 2010;3(1):3–5.
27. Leening MJG, Vedder MM, Witteman JCM, Pencina MJ, Steyerberg EW. Net reclassification improvement: computation, interpretation, and controversies: a literature review and clinician's guide. *Ann Intern Med*. 2014;160(2):122–131.
28. Janssens ACJW, Martens F.K. Reflection on modern methods: Revisiting the area under the ROC Curve. *Int J Epidemiol*. 2020;49(4):1397:1403.
29. Kerr KF, Bansal A, Pepe MS. Further insight into the incremental value of new markers: the interpretation of performance measures and the importance of clinical context. *Am J Epidemiol*. 2012;176(6):482–487.
30. Kerr KF, Wang Z, Janes H, McClelland RL, Psaty BM, Pepe MS. Net reclassification indices for evaluating risk prediction instruments: a critical review. *Epidemiology*. 2014;25(1):114–121.
31. Martens FK, Janssens ACJW. How the intended use of polygenic risk scores guides the design and evaluation of prediction studies. *Curr Epidemiol Rep*. 2019;6(2):184–90.
32. Dhana K, Kavousi M, Ikram MA, Tiemeier HW, Hofman A, Franco OH. Body shape index in comparison with other anthropometric measures in prediction of total and cause-specific mortality. *J Epidemiol Community Health*. 2016;70(1):90–96.
33. Kim S-H, Lee E-S, Yoo J, Kim Y. Predicting risk of type 2 diabetes mellitus in Korean adults aged 40–69 by integrating clinical and genetic factors. *Prim Care Diabetes*. 2019;13(1):3–10.
34. Graversen P, Abildstrøm SZ, Jespersen L, Borglykke A, Prescott E. Cardiovascular risk prediction: Can Systematic Coronary Risk Evaluation (SCORE) be improved by adding simple risk markers? Results from the Copenhagen City Heart Study. *Eur J Prev Cardiol*. 2016;23(14):1546–1556.
35. Wotherspoon AC, Young IS, McCance DR, Patterson CC, Maresh MJA, Pearson DWM, et al. Serum Fatty Acid Binding Protein 4 (FABP4) Predicts pre-eclampsia in women with type 1 diabetes. *Diabetes Care*. 2016;39(10):1827–1829.
36. Vandenput L, Mellström D, Kindmark A, Johansson H, Lorentzon M, Leung J, et al. High serum SHBG predicts incident vertebral fractures in elderly men. *J Bone Miner Res*. 2016;31(3):683–689.
37. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med*. 1996;15(4):361–387.



## Supplementary data

**Appendix Table 1.** Characteristics of the prediction models in the literature review

| Study          | Sample size | Event rate (%) | Outcome                             | Clinical risk factors   | Added risk factor  |
|----------------|-------------|----------------|-------------------------------------|---|--|
| An (A1)        | 2902        | 10             | Allograft failure                   | Donor age, recipient sex, induction therapy, donor type, recurrent glomerulonephritis, biopsy-proven acute rejection  | Pre-transplant cardio vascular risk score, vascular disease    |
| Dhana (29)     | 3740        | 53             | Mortality                           | Age, diabetes, smoking, SBP, hypertension medication, HDL cholesterol, total cholesterol  | A body shape index   |
| Chuang (A2)    | 3146        | 14             | CVD                                 | Age, sex, smoking, SBP, BMI, HDL cholesterol, total cholesterol, triglycerides, DBP, fasting glucose, history of heart disease, history of stroke   | End-diastolic velocity   |
| Bonaca (A3)    | 21162       | 5              | Major adverse cardiovascular events | Age, diabetes, smoking, hypertension, eGFR, race, myocardial infarction type, second prior myocardial infarction, multivessel disease, hypercholesterolemia, congestive heart failure, chronic obstructive pulmonary disease, history of stroke or transient ischemic attack, angina, coronary artery bypass graft, time from P2Y12 withdrawal, history of percutaneous coronary intervention with stenting, region | Peripheral artery disease                                      |
| Parikh (A4)    | 72982       | 6              | CHD                                 | Age, diabetes, smoking, SBP, antihypertension medication, high cholesterol  | Age at first birth, still births, miscarriages, breastfeeding  |
| Kim (30)       | 2529        | 10             | Ischemic Events                     | Age, diabetes, SBP, eGFR, history of cardiovascular disease, albumin, hemoglobin, phosphate, sodium   | Trimethylamine N-oxide   |
| Graverson (31) | 8476        | 8              | CVD mortality                       | Age, sex, smoking, SBP, total cholesterol   | Forced expiratory volume                                       |
| Akazawa (A5)   | 332         | 60             | CAD                                 | Age, sex, diabetes, smoking, hypertension, BMI, HbA1c, hyperlipidemia, plaque thickness   | Common carotid artery, common carotid artery *plaque thickness |
| O'Neal (A6)    | 6394        | 3              | Heart failure                       | Age, diabetes, SBP, BMI, heart rate, CHD, valve disease   | Delayed intrinsicoid deflection                                |

| Study            | Sample size | Event rate (%) | Outcome                             | Clinical risk factors  | Added risk factor   |
|------------------|-------------|----------------|-------------------------------------|--|---|
| Ioakeimidis (A7) | 298         | 7              | Major adverse cardiovascular events | Age, smoking, hypertension, HDL cholesterol, total cholesterol   | Dynamic penile peak systolic velocity   |
| Wotherspoon (32) | 710         | 17             | Pre-eclampsia                       | Age, diabetes, smoking, BMI, blood pressure, HbA1c, gestational age, parity, history of pre-eclampsia, renal status, diabetes and pre-eclampsia intervention trial treatment group   | Fatty Acid-Binding Protein 4  |
| Vandenput (33)   | 2847        | 7              | Fracture                            | Country-specific calculated estimate of the 10-year risk of a major osteoporotic fracture with bone mineral density  | Sex hormone binding globulin  |
| Hayek (A8)       | 1497        | 21             | Peripheral arterial disease         | Age, sex, diabetes, smoking, hypertension, BMI, eGFR, race, hyperlipidemia, history of heart failure, statin use, angiotensin pathway antagonist use, obstructive CAD  | Cluster of differentiation 34+, cluster of differentiation 34+/vascular endothelial growth factor receptor-2+ cells |
| Brownrigg (A9)   | 49027       | 6              | Cardiovascular events               | Age, sex, diabetes, smoking, SBP, BMI, HbA1c, HDL cholesterol, total cholesterol, LDL cholesterol, eGFR, DBP, race, statin use, angiotensin-converting enzyme inhibitor or angiotensin II receptor blocker, blood pressure medication, antiplatelet  | Microvascular disease variables   |
| Chen (A10)       | 213         | 23             | Acute kidney injury progression     | Age, sex, diabetes, hypertension, eGFR, N-terminal pro-B-type natriuretic peptide, serum albumin, hemoglobin, diuretic dosage before acute kidney injury, spironolactone use before acute kidney injury, renin-angiotensin system inhibitors use before acute kidney injury, change of serum creatinine from baseline at the time of acute kidney injury diagnosis | Urinary angiotensinogen   |
| Marcus (A11)     | 356         | NR             | 1-year mortality                    | Age, Karnofsky performance status score, absence of extracranial metastases, brain metastasis  | Cumulative intracranial tumor volume  |
| Wang (A12)       | 1142        | 50             | Type 2 diabetes                     | History of hypertension, BMI, education, HDL cholesterol, triglycerides, exercise  | Alanine aminotransferase  |

| Study         | Sample size | Event rate (%) | Outcome   | Clinical risk factors  | Added risk factor  |
|---------------|-------------|----------------|---|--|--|
| Oikawa (A13)  | 259         | 29             | Aortic valve calcification                      | Age, diabetes  | Abdominal visceral adipose tissue  |
| Ahn (A14)     | 25859       | 8              | Diabetes  | Fasting glucose, HbA1c   | Age, smoking, family history of diabetes and hypertension, waist circumference, alcohol intake |
| Obokata (A15) | 423         | 11             | All-cause mortality                             | A mortality risk score, N-terminal pro-B-type natriuretic peptide  | Galectin-3   |
| Gori (A16)    | 8402        | 23             | Fatal and nonfatal heart failure, CHD or stroke | Age, sex, smoking, SBP, hypertension medication, BMI, total/HDL cholesterol, triglycerides, race, education, center, pack-years of cigarettes, waist-to-hip ratio, lipid-lowering medication, aspirin use, duration of disease, electrocardiogram abnormalities, retinopathy, nephropathy, peripheral arterial disease | Hs-TnT, NTproBNP   |
| Long (A17)    | 1181        | 27             | Hepatic steatosis                               | Age, sex, diabetes, hypertension, BMI, triglycerides   | Aspartate aminotransferase/alanine aminotransferase  |
| Yadav (A18)   | 2276        | 17             | Metabolic syndrome                              | HDL cholesterol, waist circumference, blood pressure, triglycerides, fasting glucose   | Aspartate aminotransferase/alanine aminotransferase  |
| Yadav (A19)   | 2784        | 3              | Diabetes  | Age, sex, smoking, SBP, HDL cholesterol, total cholesterol, fasting glucose, exercise, family history, alcohol intake, insulin resistance  | Fatty liver index  |
| Patel (A20)   | 1411        | 16             | Mortality                                       | Age, sex, diabetes, smoking, hypertension, BMI, GFR, HDL cholesterol, total cholesterol, statin use, acute myocardial infarction, left ventricular function, Gensini score   | High or low C-reactive protein, high or low cysteine/glutathione                               |
| Jani (A21)    | 425         | 23             | Mortality                                       | Age, sex, diabetes, SBP, GFR, serum sodium, blood urea nitrogen, ejection fraction, ischaemic aetiology, B-type natriuretic peptide or N-Terminal pro-BNP  | Depressive symptoms  |

| Study             | Sample size | Event rate (%) | Outcome  | Clinical risk factors   | Added risk factor   |
|-------------------|-------------|----------------|--|---|---|
| Kozminski (A22)   | 2837        | 9              | Biochemical recurrence                         | Age, preoperative PSA, surgical margin status   | Primary Gleason grade $\geq 4$ or stage $\geq pT3$  |
| Mise (A23)        | 149         | 63             | Renal progression                              | Age, sex, BMI, SBP, eGFR, diabetic retinopathy, urinary protein excretion   | Interstitial fibrosis and tubular atrophy score   |
| Alshehry (A24)    | 3779        | 18             | Cardiovascular events                          | Age, sex, diabetes, SBP, antihypertensive medication, BMI, eGFR, HDL cholesterol, HbA1c, exercise, C-reactive protein, history of macrovascular disease, history of heart failure, antiplatelet medication  | Alkylphosphatidylcholine, cholesteryl ester, alkylphosphatidylethanolamine, phosphatidylcholine and lysophosphatidylcholine |
| Iribarren (A25)   | 1135        | 14             | Coronary heart disease                         | Age, sex, smoking, SBP, HDL cholesterol, total cholesterol, blood pressure medication   | High-sensitivity cardiac troponin I   |
| Li (A26)          | 870         | 7              | All-cause mortality                            | Age, history of diabetes, SBP, history of hypertension, BMI, eGFR, pulse, anterior myocardial infarction, total ischaemic time, glucose, N-terminal pro-brain natriuretic peptide                           | Presence of older thrombus  |
| Proietti (A27)    | 3551        | NR             | Bleeding                                       | Age $\geq 75$ years, anemia, severe renal disease, history of hemorrhage, history of hypertension   | Therapeutic range $< 65\%$  |
| Stone (A28)       | 263         | 2              | Mortality or hospitalization for heart failure | Age, sex, diabetes, smoking, hypertension, hyperlipidemia, left anterior descending versus non-left anterior descending infarct vessel, symptom-to-first device time, thrombolysis in myocardial infarction | Infarct size  |
| Rydingsward (A29) | 6895        | 10             | 90-day post-discharge mortality                | Sex, acute organ failure score  | Functional status   |
| Keyzer (A30)      | 697         | 12             | Mortality                                      | Age, sex, eGFR  | Serum calciprotein particle maturation time   |

| Study         | Sample size | Event rate (%) | Outcome                          | Clinical risk factors  | Added risk factor  |
|---------------|-------------|----------------|----------------------------------|--|--|
| Liu (A31)     | 1737        | 10             | Mortality                        | Age, sex, diabetes, smoking, hypertension, BMI, dyslipidemia, alcohol intake, lipid-lowering medication  | Cholesterol efflux capacity  |
| Peetz (A32)   | 3,565       | 23             | 720-day post-discharge mortality | Sex, acute organ failure score, derived injury severity score  | Functional status  |
| May (A33)     | 50908       | 2              | All-cause mortality              | Age, sex, glucose, calcium, bicarbonate, potassium, sodium, mean platelet volume, red cell distribution width, mean corpuscular hemoglobin concentration, mean corpuscular volume, platelet, white blood cell count, hematocrit  | Anion gap, albumin, alkaline phosphatase, alanine aminotransferase, aspartate aminotransferase, bilirubin                |
| Behnes (A34)  | 401         | 14             | All-cause mortality              | Age, sex, diabetes, CAD, left ventricular function, New York heart association functional class, creatinine, angiotensin-converting enzyme inhibitor or angiotensin II receptor blocker medication, beta-blocker medication, sodium, hemoglobin, N-terminal pro-B-type natriuretic peptide | Galectin-3   |
| Garde (A35)   | 2051        | 30             | Hospitalization                  | Respiratory rate, median blood oxygen saturation value   | Age, sex, heart rate, time peripheral capillary oxygen saturation <98%, time peripheral capillary oxygen saturation <94% |
| Deo (A36)     | 3587        | 6              | Cardiovascular mortality         | Age, sex, diabetes, smoking, SBP, BMI, eGFR, HDL cholesterol, total cholesterol, DBP, race, history of cardiovascular disease, proteinuria   | Five electrocardiogram metrics   |
| Haybar (A37)  | 183         | NR             | CAD                              | Age, sex, diabetes, smoking, hypertension, hyperlipidemia  | Neutrophil/lymphocyte, fibrinogen  |
| Hoshida (A38) | 4278        | 2              | Stroke                           | Age, sex, diabetes, smoking, office SBP, antihypertensive medication, BMI, HDL cholesterol, total cholesterol, history of CVD, statin use  | Morning home SBP   |

| Study          | Sample size | Event rate (%) | Outcome  | Clinical risk factors   | Added risk factor   |
|----------------|-------------|----------------|--|---|---|
| Barbour (A39)  | 901         | 18             | End-stage renal disease or 50% reduction in eGFR | eGFR, mean arterial blood pressure, proteinuria   | Mesangial and endocapillary hypercellularity, segmental sclerosis and interstitial fibrosis/tubular atrophy |
| Eggers (A40)   | 1016        | 16             | Mortality  | Sex, diabetes, smoking, hypertension, BMI, eGFR, HDL cholesterol, C-reactive protein, LDL cholesterol, CVD  | Growth/differentiation factor 15  |
| Neyra (A41)    | 846         | 30             | Hospital mortality                               | Sequential Organ Failure Assessment   | Cumulative fluid balance  |
| Nagahara (34)  | 139         | 19             | High risk plaque                                 | Age, diabetes, smoking, coronary artery calcium score, plaque thickness   | Eicosapentaenoic acid/arachidonic acid  |
| Ferreira (A42) | 28771       | 15             | Cardiovascular mortality                         | Sex, diabetes, smoking, SBP, hypertension, heart rate   | Body surface area adjusted-Cockcroft-Gault  |
| Mentias (A43)  | 737         | 9              | Mortality  | Society of thoracic surgeons score, left ventricular end-systolic dimension, right ventricular systolic pressure, mitral effective regurgitant orifice                              | % Predicted metabolic equivalents   |
| Li (A44)       | 512         | 7              | Bleeding   | Sex, diabetes, SBP, heart rate, hematocrit, creatinine heart failure, history of vascular disease   | Platelet reactivity   |
| Elias (A45)    | 810         | 17             | Significant colorectal diseases                  | Age, C-reactive protein, abdominal pain, rectal blood loss, rectal mucus, weight loss, change in bowel habit, abdominal bloating, constipation, abnormal digital rectal examination | Point-of-care faecal immunochemical test for haemoglobin and calprotectin point-of-care test                |
| Ko (A46)       | 27          | NR             | Fractional flow reserve                          | Computed tomography angiography   | Rest transmural attenuation gradient S20  |

| Study             | Sample size | Event rate (%) | Outcome   | Clinical risk factors  | Added risk factor                                   |
|-------------------|-------------|----------------|---|--|---|
| Kalim (A47)       | 366         | 33             | Mortality   | History of diabetes, SBP, BMI, DBP, residual renal function, initial vascular access type, and history of congestive heart failure, average serum albumin, transferrin saturation, phosphorus, hemoglobin, ferritin, parathyroid hormone level, Kt/V, blood urea nitrogen, normalized protein catabolic rate, history of CAD | Carbamylated albumin                                |
| Tribouilloy (A48) | 289         | 43             | Aortic valve replacement or all-cause death       | Age, sex, hypertension, Charlson comorbidity index, CAD, atrial fibrillation, left ventricular ejection fraction   | Aortic valve area/height                            |
| Li (A49)          | 596         | 12             | CVD   | Age, SBP, heart rate, creatinine, congestive heart failure, cardiac arrest at admission, ST-segment deviation, elevated cardiac enzyme or biomarker levels   | Platelet reactivity                                 |
| Lee (A50)         | 130         | NR             | Functionally significant coronary artery stenosis | Relative perfusion defect, coronary flow reserve   | Stress myocardial blood flow, relative flow reserve |

Abbreviations: BMI = body mass index; CAD = coronary artery disease; CHD = coronary heart disease; CVD = cardiovascular disease; DBP = diastolic blood pressure; eGFR = estimated glomerular filtration rate; GFR = glomerular filtration rate; HbA1c = glycated hemoglobin; HDL = high-density lipoprotein; LDL = low-density lipoprotein; NR = not reported; SBP = systolic blood pressure.

**Appendix Table 2.** AUC, NRI and IDI of the prediction models in the literature review

| Study            | Estimates    |                           |                           | Model improved?        |              |                           |       |                 |       |              |
|------------------|--------------|---------------------------|---------------------------|------------------------|--------------|---------------------------|-------|-----------------|-------|--------------|
|                  | Baseline AUC | Δ AUC                     | IDI                       | NRI                    | Type of NRI  | NRI cut-offs              | Δ AUC | NRI             | IDI   | Overall      |
| An (A1)          | 0.789        | 0.001 (0.576)             | 0.0046 (0.005)            | 0.128 (0.004)          | NR           | NA                        | No    | Yes             | Yes   | Yes          |
| Dhana (29)       | 0.783        | 0.001 (NR)                | 0.001 (-0.001 to 0.001)   | 0.05 (-0.01 to 0.12)   | Continuous   | NA                        | No    | No              | No    | [No]         |
| Chuang (A2)      | 0.7848       | 0.0026 (0.34)             | 0.0056 (0.016)            | 0.0027 (0.446)         | NR           | NA                        | No    | Yes             | Yes   | Yes          |
| Bonaca (A3)      | 0.667        | 0.003 (0.211)             | 0.002 (0.144)             | 0.016 (0.733)          | Continuous   | NA                        | NR    | NR              | NR    | NR           |
| Parikh (A4)      | 0.726        | 0.0033 (0.0022 to 0.0051) | 0.0013 (<0.0001)          | 0.005 (0.37)           | Categorical  | <5%, 5-10% and ≥10%       | [Yes] | No <sup>d</sup> | NR    | Inconclusive |
| Kim (30)         | NR           | 0.0047 (0.0001 to 0.0128) | 0.0041 (0.0001 to 0.0120) | 0.104 (0.031 to 0.247) | NR           | NA                        | Yes   | Yes             | Yes   | Yes          |
| Gravensen (31)   | 0.837        | 0.006 (0.032)             | 0.006 (0.029)             | 0.01 (0.718)           | Categorical* | NR                        | Yes   | No              | Yes   | No           |
| Akazawa (A5)     | 0.827        | 0.006 (NR)                | 0.042 (<0.001)            | 0.411 (<0.001)         | Continuous   | NA                        | Yes   | Yes             | Yes   | Yes          |
| O'Neal (A6)      | 0.773        | 0.007 (NR)                | 0.0034 (0.0010 to 0.0059) | 0.069 (0.015 to 0.13)  | Categorical  | <2.5%, 2.5% to 5%, >5%    | Yes   | Yes             | Yes   | Yes          |
| Ioakeimidis (A7) | 0.767        | 0.007 (0.44)              | 0.047 (0.038)             | 0.081 (0.27)           | Categorical  | <3.5%, 3.5-8.5% and >8.5% | [Yes] | No              | Yes   | Yes          |
| Wotherspoon (32) | 0.793        | 0.008 (0.17)              | 0.009 (0.11)              | 0.306 (0.003)          | Continuous   | NA                        | No    | Yes             | Yes   | Yes          |
| Vandempuut (33)  | 0.61         | 0.01 (NS)                 | 0.004 (S)                 | 0.178 (S)              | Continuous   | NA                        | No    | [Yes]           | [Yes] | Yes          |
| Hayek (A8)       | 0.717        | 0.01 (-0.001 to 0.020)    | 0.027 (0.017 to 0.036)    | 0.39 (0.234 to 0.546)  | Continuous   | NA                        | No    | Yes             | Yes   | Yes          |
| Brownrigg (A9)   | 0.679        | 0.01 (NR)                 | 0.003 (<0.0001)           | 0.036 (<0.0001)        | Categorical  | <7.5%, ≥ 7.5%             | Yes   | Yes             | Yes   | No           |
| Chen (A10)       | 0.77         | 0.01 (NR)                 | 0.14 (<0.001)             | 0.76 (<0.001)          | Continuous   | NA                        | Yes   | Yes             | Yes   | Yes          |
| Marcus (A11)     | NR           | 0.01 (0.04)               | 0.007 (0.001 to 0.018)    | 0.304 (0.198 to 0.407) | Continuous   | NA                        | Yes   | Yes             | Yes   | Yes          |
| Wang (A12)       | 0.74         | 0.01 (0.02)               | 0.02 (<0.001)             | 0.53 (<0.001)          | Continuous   | NA                        | [Yes] | Yes             | Yes   | Yes          |
| Oikawa (A13)     | NR           | 0.011 (NS)                | 0.0116 (0.07)             | 0.5093 (<0.001)        | NR           | NA                        | Yes   | Yes             | Yes   | NR           |



| Study             | Estimates    |                        |                        |                        | Model improved? |                                    |                  |     | Overall          |                  |
|-------------------|--------------|------------------------|------------------------|------------------------|-----------------|------------------------------------|------------------|-----|------------------|------------------|
|                   | Baseline AUC | $\Delta$ AUC           | IDI                    | NRI                    | Type of NRI     | NRI cut-offs                       | $\Delta$ AUC     | NRI |                  | IDI              |
| Ahn (A14)         | 0.838        | 0.011 (0.016)          | 0.006 (0.176)          | 0.046 (0.264)          | Categorical     | <5%, 5-10%, 10-15% and $\geq 15\%$ | [Yes]            | No  | NR               | Yes              |
| Obokata (A15)     | 0.815        | 0.013 (NR)             | 0.053 (<0.001)         | 0.345 (0.030)          | Continuous      | NA                                 | Yes              | Yes | Yes              | Yes              |
| Gori (A16)        | 0.688        | 0.015 (0.004)          | 0.03 (<0.001)          | 0.16 (<0.001)          | Continuous      | NA                                 | [Yes]            | Yes | Yes              | Yes              |
| Long (A17)        | 0.83         | 0.015 (<0.001)         | 0.03 (<0.001)          | 0.688 (<0.0001)        | Continuous      | NA                                 | Yes              | NR  | NR               | Yes              |
| Yadav (A18)       | 0.715        | 0.017 (0.0043)         | 0.0094 (<0.0001)       | 0.23 (<0.0001)         | Continuous      | NA                                 | Yes              | Yes | Yes              | Yes              |
| Yadav (A19)       | 0.818        | 0.017 (0.0289)         | 0.015 (0.0121)         | 0.417 (0.0002)         | Continuous      | NA                                 | Yes              | Yes | Yes              | Yes              |
| Patel (A20)       | 0.717        | 0.018 (0.002 to 0.035) | 0.018 (0.004 to 0.043) | 0.124 (0.026 to 0.191) | Continuous      | NA                                 | Yes              | Yes | Yes              | Yes              |
| Jani (A21)        | 0.781        | 0.019 (0.06)           | 0.031 (0.001)          | 0.3504 (0.002)         | Continuous      | NA                                 | No               | Yes | Yes              | Yes              |
| Kozminski (A22)   | 0.70         | 0.02 (NR)              | 0.036 (NR)             | 0.65 (0.52 to 0.78)    | NR              | NA                                 | NR               | NR  | NR               | Yes <sup>d</sup> |
| Mise (A23)        | 0.82         | 0.02 (0.00 to 0.05)    | 0.03 (<0.03 to 0.08)   | 0.54 (0.03 to 1.05)    | Continuous      | NA                                 | Yes              | Yes | Yes              | Yes              |
| Alshetry (A24)    | 0.68         | 0.02 (<0.0001)         | 0.024 (0.023 to 0.024) | 0.227 (0.219 to 0.235) | Continuous      | NA                                 | Yes              | NR  | Yes              | Yes              |
| Iribarren (A25)   | 0.68         | 0.02 (0.16)            | 0.0267 (<0.0001)       | 0.18 (0.08 to 0.30)    | Categorical     | <10%, 10-20% and >20%              | Yes              | Yes | NR               | Yes              |
| Li (A26)          | 0.72         | 0.02 (NR)              | 0.02227 (0.04)         | 0.42 (<0.01)           | Continuous      | NA                                 | Yes              | Yes | Yes              | Yes              |
| Proietti (A27)    | 0.590        | 0.021 (0.052)          | 0.0020 (0.0014)        | 0.250 (0.0054)         | Continuous      | NA                                 | [Yes]            | Yes | Yes              | Yes              |
| Stone (A28)       | 0.83         | 0.03 (0.19)            | 0.015 (0.07)           | 0.472 (0.004)          | Continuous      | NA                                 | Yes <sup>d</sup> | Yes | Yes <sup>d</sup> | Yes              |
| Rydingsward (A29) | 0.70         | 0.03 (<0.001)          | 0.024 (>0.001)         | 0.039 (0.016)          | Categorical     | 50%                                | Yes              | Yes | Yes              | NR               |
| Keyzer (A30)      | 0.72         | 0.030 (0.004 to 0.056) | 0.016 (0.03)           | 0.14 (0.002)           | Categorical     | <5%, 5-10%, >10%                   | Yes              | Yes | Yes              | Yes              |
| Liu (A31)         | 0.76         | 0.03 (0.0001)          | 0.016 (<0.001)         | 0.166 (0.001)          | Categorical*    | <5%, 5-10%, 10-15% and >15%        | Yes              | Yes | Yes              | Yes              |
| Peetz (A32)       | 0.74         | 0.03 (S)               | 0.029 (<0.001)         | 0.063 (<0.001)         | NR              | NA                                 | Yes              | Yes | Yes              | Yes              |
| May (A33)         | 0.872        | 0.031 (S)              | 0.05 (<0.0001)         | 0.155 (<0.0001)        | Categorical     | NR                                 | Yes              | Yes | Yes              | Yes              |
| Behnes (A34)      | 0.8146       | 0.0367 (0.0826)        | 0.11157 (0.3004)       | 0.5643 (0.0001)        | NR              | NA                                 | Yes              | Yes | Yes              | [Yes]            |
| Garde (A35)       | 0.73         | 0.04 (NR)              | 0.0466 (0)             | 0.0389 (0.02799)       | Categorical     | 0.25                               | NR               | NR  | NR               | Yes              |

| Study             | Estimates    |                     |                        | Model improved?           |             |                     |       |     |     |         |
|-------------------|--------------|---------------------|------------------------|---------------------------|-------------|---------------------|-------|-----|-----|---------|
|                   | Baseline AUC | Δ AUC               | IDI                    | NRI                       | Type of NRI | NRI cut-offs        | Δ AUC | NRI | IDI | Overall |
| Deo (A36)         | 0.77         | 0.04 (NR)           | 0.027 (0.001)          | 0.121 (0.081 to 0.160)    | Categorical | <5%, 5-25% and >25% | Yes   | Yes | NR  | Yes     |
| Haybar (A37)      | 0.76         | 0.04 (0.004)        | 0.05 (<0.001)          | 0.130 (<0.001)            | Categorical | NR                  | Yes   | Yes | Yes | Yes     |
| Hoshide (A38)     | 0.756        | 0.046 (0.077)       | 0.015 (0.005)          | 0.3606 (0.1317 to 0.5896) | Continuous  | NA                  | Yes   | Yes | Yes | Yes     |
| Barbour (A39)     | 0.75         | 0.05 (0.03 to 0.08) | 0.06 (0.04 to 0.11)    | 0.28 (0.16 to 0.43)       | Continuous  | NA                  | Yes   | Yes | Yes | No      |
| Eggers (A40)      | 0.66         | 0.05 (0.003)        | 0.038 (<0.001)         | 0.334 (<0.001)            | Continuous  | NA                  | Yes   | Yes | Yes | Yes     |
| Neyra (A41)       | 0.66         | 0.06 (<0.001)       | 0.053 (0.036 to 0.070) | 0.46 (0.32 to 0.61)       | Continuous  | NA                  | Yes   | Yes | Yes | Yes     |
| Nagahara (34)     | 0.74         | 0.06 (0.07)         | 0.054 (0.0072)         | 0.60 (0.0049)             | Continuous* | NA                  | No    | Yes | Yes | Yes     |
| Ferreira (A42)    | 0.617        | 0.071 (<0.001)      | 0.049 (<0.001)         | 0.0249 (<0.001)           | Continuous  | NA                  | Yes   | Yes | Yes | [Yes]   |
| Mentias (A43)     | 0.69         | 0.09 (<0.01)        | 0.08 (<0.001)          | 0.6 (<0.001)              | Continuous  | NA                  | Yes   | Yes | Yes | Yes     |
| Li (A44)          | 0.732        | 0.095 (0.011)       | 0.022 (0.002)          | 0.258 (<0.001)            | NR          | NA                  | Yes   | Yes | Yes | Yes     |
| Elias (A45)       | 0.741        | 0.096 (<0.001)      | 0.14 (<0.001)          | 0.38 (<0.001)             | Categorical | 5.0, 50.0%          | Yes   | Yes | NR  | Yes     |
| Ko (A46)          | 0.66         | 0.1 (NR)            | 0.33 (<0.0001)         | 1.24 (<0.0001)            | Continuous  | NA                  | NR    | Yes | Yes | [Yes]   |
| Kalim (A47)       | 0.76         | 0.11 (0.03)         | 0.22 (<0.001)          | 0.6 (0.002)               | Continuous  | NA                  | Yes   | Yes | Yes | Yes     |
| Tribouilloy (A48) | 0.56         | 0.11 (0.002)        | 0.10 (0.001)           | 0.33 (<0.00001)           | Continuous  | NA                  | Yes   | Yes | Yes | Yes     |
| Li (A49)          | 0.749        | 0.122 (0.002)       | 0.018 (0.002)          | 0.263 (<0.001)            | Continuous  | NA                  | Yes   | Yes | Yes | Yes     |
| Lee (A50)         | 0.74         | 0.172 (<0.001)      | 0.352 (<0.001)         | 1.232 (<0.001)            | Continuous  | NA                  | Yes   | Yes | Yes | Yes     |

\* This study did not report whether the categorical or continuous NRI was used. The type of NRI was inferred from the text.

<sup>a</sup> Retrieved from the discussion section of the article.

Values are point estimates with *P* values or 95% confidence intervals between brackets. Labeling of researchers' interpretations of the metrics is described in the Methods. Square brackets indicate that the researchers commented that the observed improvement of the model was minimal. Abbreviations: ΔAUC = increment in the area under the receiver operating characteristic curve; AUC = area under the receiver operating characteristic curve; CI = confidence interval; IDI = integrated discrimination improvement; NA = not applicable; NR = not reported; NRI = net reclassification improvement; NS = not statistically significant; S = statistically significant (per researchers' reporting).

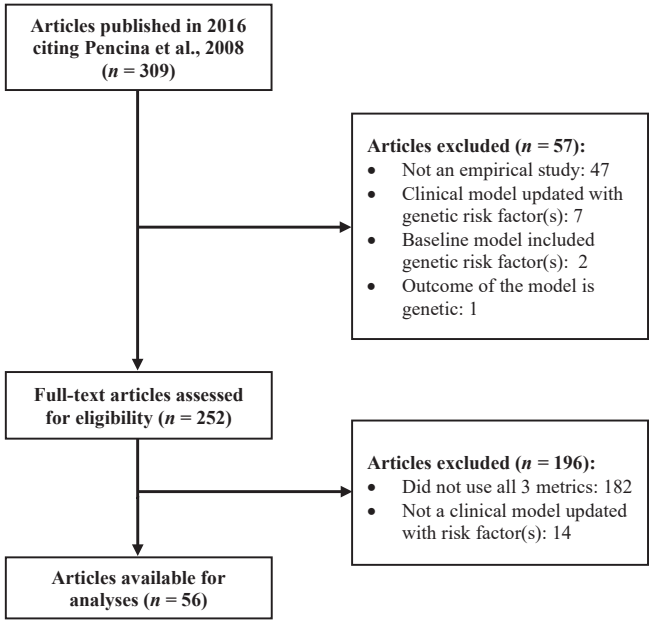
## References

- A1. An JN, Ahn SV, Lee JP, Bae E, Kang E, Kim H-L, et al. Pre-transplant cardiovascular risk factors affect kidney allograft survival: A multi-center study in Korea. Marson L, editor. *PLoS One*. 2016;11(8):e0160607.
- A2. Chuang S-Y, Bai C-H, Cheng H-M, Chen J-R, Yeh W-T, Hsu P-F, et al. Common carotid artery end-diastolic velocity is independently associated with future cardiovascular events. *Eur J Prev Cardiol*. 2016;23(2):116–124.
- A3. Bonaca MP, Bhatt DL, Storey RF, Steg PG, Cohen M, Kuder J, et al. Ticagrelor for Prevention of ischemic events after myocardial infarction in patients with peripheral artery disease. *J Am Coll Cardiol*. 2016;67(23):2719–2728.
- A4. Parikh NI, Jeppson RP, Berger JS, Eaton CB, Kroenke CH, LeBlanc ES, et al. Reproductive risk factors and coronary heart disease in the Women's Health Initiative Observational Study. *Circulation*. 2016;133(22):2149–2158.
- A5. Akazawa S, Tojikubo M, Nakano Y, Nakamura S, Tamai H, Yonemoto K, et al. Usefulness of carotid plaque (sum and maximum of plaque thickness) in combination with intima-media thickness for the detection of coronary artery disease in asymptomatic patients with diabetes. *J Diabetes Investig*. 2016;7(3):396–403.
- A6. O'Neal WT, Qureshi WT, Nazarian S, Kawel-Boehm N, Bluemke DA, Lima JAC, et al. Electrocardiographic time to intrinsicoid deflection and heart failure: The multi-ethnic study of atherosclerosis. *Clin Cardiol*. 2016;39(9):531–536.
- A7. Ioakeimidis N, Vlachopoulos C, Rokkas K, Kratiras Z, Angelis A, Samentzas A, et al. Dynamic penile peak systolic velocity predicts major adverse cardiovascular events in hypertensive patients with erectile dysfunction. *J Hypertens*. 2016;34(5):860–868.
- A8. Hayek SS, MacNamara J, Tahhan AS, Awad M, Yadalam A, Ko Y-A, et al. Circulating progenitor cells identify peripheral arterial disease in patients with coronary artery disease. *Circ Res*. 2016;119(4):564–571.
- A9. Brownrigg JRW, Hughes CO, Burleigh D, Karthikesalingam A, Patterson BO, Holt PJ, et al. Microvascular disease and risk of cardiovascular events among individuals with type 2 diabetes: a population-level cohort study. *lancet Diabetes Endocrinol*. 2016;4(7):588–597.
- A10. Chen C, Yang X, Lei Y, Zha Y, Liu H, Ma C, et al. Urinary biomarkers at the time of AKI diagnosis as predictors of progression of AKI among patients with acute cardiorenal syndrome. *Clin J Am Soc Nephrol*. 2016;11(9):1536–1544.
- A11. Marcus LP, Marshall D, Hirshman BR, McCutcheon BA, Gonda DD, Koiso T, et al. Cumulative Intracranial Tumor Volume (CITV) enhances the prognostic value of the lung-specific Graded Prognostic Assessment (GPA) Model. *Neurosurgery*. 2016;79(2):246–252.
- A12. Wang Y-L, Koh W-P, Yuan J-M, Pan A. Association between liver enzymes and incident type 2 diabetes in Singapore Chinese men and women. *BMJ open diabetes Res care*. 2016;4(1):e000296.
- A13. Oikawa M, Owada T, Yamauchi H, Misaka T, Machii H, Yamaki T, et al. Predominance of abdominal visceral adipose tissue reflects the presence of aortic valve calcification. *Biomed Res Int*. 2016;2016:2174657.
- A14. Ahn CH, Yoon JW, Hahn S, Moon MK, Park KS, Cho YM. Evaluation of non-laboratory and laboratory prediction models for current and future diabetes mellitus: A Cross-sectional and retrospective cohort study. Chen C-L, editor. *PLoS One*. 2016;11(5):e0156155.
- A15. Obokata M, Sunaga H, Ishida H, Ito K, Ogawa T, Ando Y, et al. Independent and incremental prognostic value of novel cardiac biomarkers in chronic hemodialysis patients. *Am Heart J*. 2016;179:29–41.
- A16. Gori M, Gupta DK, Claggett B, Selvin E, Folsom AR, Matsushita K, et al. Natriuretic peptide and high-sensitivity troponin for cardiovascular risk prediction in diabetes: The Atherosclerosis Risk in Communities (ARIC) Study. *Diabetes Care*. 2016;39(5):677–685.
- A17. Long MT, Pedley A, Colantonio LD, Massaro JM, Hoffmann U, Muntner P, et al. Development and validation of the Framingham Steatosis Index to identify persons with hepatic steatosis. *Clin*

- Gastroenterol Hepatol. 2016;14(8):1172-1180.e2.
- A18. Yadav D, Choi E, Ahn SV, Baik SK, Cho YZ, Koh SB, et al. Incremental predictive value of serum AST-to-ALT ratio for incident metabolic syndrome: The ARIRANG Study. Lu S-N, editor. *PLoS One*. 2016;11(8):e0161304.
- A19. Yadav D, Choi E, Ahn SV, Koh SB, Sung K-C, Kim J-Y, et al. Fatty liver index as a simple predictor of incident diabetes from the KoGES-ARIRANG study. *Medicine (Baltimore)*. 2016; 95(31):e4447.
- A20. Patel RS, Ghasemzadeh N, Eapen DJ, Sher S, Arshad S, Ko Y, et al. Novel biomarker of oxidative stress is associated with risk of death in patients with coronary artery disease. *Circulation*. 2016;133(4):361-369.
- A21. Jani BD, Mair FS, Roger VL, Weston SA, Jiang R, Chamberlain AM. Comorbid depression and heart failure: A community cohort study. Hosoda T, editor. *PLoS One*. 2016;11(6):e0158570.
- A22. Kozminski MA, Tomlins S, Cole A, Singhal U, Lu L, Skolarus TA, et al. Standardizing the definition of adverse pathology for lower risk men undergoing radical prostatectomy. *Urol Oncol*. 2016;34(9):415.e1-6.
- A23. Mise K, Hoshino J, Ueno T, Hazue R, Hasegawa J, Sekine A, et al. Prognostic value of tubulointerstitial lesions, Urinary N-Acetyl- $\beta$ -D-Glucosaminidase, and Urinary  $\beta$ 2-Microglobulin in patients with type 2 diabetes and biopsy-proven diabetic nephropathy. *Clin J Am Soc Nephrol*. 2016;11(4):593-601.
- A24. Alshehry ZH, Mundra PA, Barlow CK, Mellett NA, Wong G, McConville MJ, et al. Plasma lipidomic profiles improve on traditional risk factors for the prediction of cardiovascular events in type 2 diabetes mellitus. *Circulation*. 2016;134(21):1637-1650.
- A25. Iribarren C, Chandra M, Rana JS, Hlatky MA, Fortmann SP, Quertermous T, et al. High-sensitivity cardiac troponin I and incident coronary heart disease among asymptomatic older adults. *Heart*. 2016;102(15):1177-1182.
- A26. Li X, Kramer MC, Damman P, van der Wal AC, Grundeken MJ, van Straalen JP, et al. Older coronary thrombus is an independent predictor of 1-year mortality in acute myocardial infarction. *Eur J Clin Invest*. 2016;46(6):501-510.
- A27. Proietti M, Senoo K, Lane DA, Lip GYH. Major Bleeding in patients with non-valvular atrial fibrillation: Impact of time in therapeutic range on contemporary bleeding risk scores. *Sci Rep*. 2016;6(1):24376.
- A28. Stone GW, Selker HP, Thiele H, Patel MR, Udelson JE, Ohman EM, et al. Relationship between infarct size and outcomes following primary PCI: Patient-level analysis from 10 randomized trials. *J Am Coll Cardiol*. 2016; 67(14):1674-1683.
29. Dhana K, Kavousi M, Ikram MA, Tiemeier HW, Hofman A, Franco OH. Body shape index in comparison with other anthropometric measures in prediction of total and cause-specific mortality. *J Epidemiol Community Health*. 2016;70(1):90-96.
- A29. Rydingsward JE, Horkan CM, Mogensen KM, Quraishi SA, Amrein K, Christopher KB. Functional status in ICU survivors and out of hospital outcomes. *Crit Care Med*. 2016;44(5):869-879.
30. Kim S-H, Lee E-S, Yoo J, Kim Y. Predicting risk of type 2 diabetes mellitus in Korean adults aged 40-69 by integrating clinical and genetic factors. *Prim Care Diabetes*. 2019;13(1):3-10.
- A30. Keyzer CA, de Borst MH, van den Berg E, Jahnen-Dechent W, Arampatzis S, Farese S, et al. Calcification propensity and survival among renal transplant recipients. *J Am Soc Nephrol*. 2016;27(1):239-248.
31. Graversen P, Abildstrøm SZ, Jespersen L, Borglykke A, Prescott E. Cardiovascular risk prediction: Can Systematic Coronary Risk Evaluation (SCORE) be improved by adding simple risk markers? Results from the Copenhagen City Heart Study. *Eur J Prev Cardiol*. 2016;23(14):1546-1556.
- A31. Liu C, Zhang Y, Ding D, Li X, Yang Y, Li Q, et al. Cholesterol efflux capacity is an independent predictor of all-cause and cardiovascular mortality in patients with coronary artery disease: A prospective cohort study. *Atherosclerosis*. 2016;249:116-124.
32. Wotherspoon AC, Young IS, McCance DR, Patterson CC, Maresh MJA, Pearson DWM, et al. Serum Fatty Acid Binding Protein 4 (FABP4) predicts pre-eclampsia in women with type 1 diabetes. *Diabetes Care*. 2016;39(10):1827-1829.
- A32. Peetz AB, Brat GA, Rydingsward J, Askari R, Olufajo OA, Elias KM, et al. Functional status, age, and long-term survival after trauma. *Surgery*. 2016;160(3):762-770.

33. Vandenput L, Mellström D, Kindmark A, Johansson H, Lorentzon M, Leung J, et al. High serum SHBG predicts incident vertebral fractures in elderly men. *J Bone Miner Res.* 2016;31(3):683–689.
- A33. May HT, Anderson JL, Muhlestein JB, Lappé DL, Ronnow BS, Horne BD. Improvement in the predictive ability of the intermountain mortality risk score by adding routinely collected laboratory tests such as albumin, bilirubin, and white cell differential count. *Clin Chem Lab Med.* 2016;54(10):1619–1628.
34. Nagahara Y, Motoyama S, Sarai M, Ito H, Kawai H, Takakuwa Y, et al. Eicosapentaenoic acid to arachidonic acid (EPA/AA) ratio as an associated factor of high risk plaque on coronary computed tomography in patients without coronary artery disease. *Atherosclerosis.* 2016;250:30–37.
- A34. Behnes M, Bertsch T, Weiss C, Ahmad-Nejad P, Akin I, Fastner C, et al. Triple head-to-head comparison of fibrotic biomarkers galectin-3, osteopontin and gremlin-1 for long-term prognosis in suspected and proven acute heart failure patients. *Int J Cardiol.* 2016;203:398–406.
- A35. Garde A, Zhou G, Raihana S, Dunsmuir D, Karlen W, Dekhordi P, et al. Respiratory rate and pulse oximetry derived information as predictors of hospital admission in young children in Bangladesh: a prospective observational study. *BMJ Open.* 2016;6(8):e011094.
- A36. Deo R, Shou H, Soliman EZ, Yang W, Arkin JM, Zhang X, et al. Electrocardiographic measures and prediction of cardiovascular and noncardiovascular death in CKD. *J Am Soc Nephrol.* 2016;27(2):559–569.
- A37. Haybar H, Ahmadzadeh A, Assareh A, Afshari N, Bozorgmanesh M, Vakili M. Intermediate-risk chronic stable angina: Neutrophil-lymphocyte ratio and fibrinogen levels improved predicting angiographically-detected coronary artery disease. *Iran Red Crescent Med J.* 2016;18(9):e18570.
- A38. Hoshida S, Yano Y, Haimoto H, Yamagiwa K, Uchiba K, Nagasaka S, et al. Morning and evening home blood pressure and risks of incident stroke and coronary artery disease in the Japanese general practice population: The Japan Morning Surge-Home Blood Pressure Study. *Hypertens.* 2016;68(1):54–61.
- A39. Barbour SJ, Espino-Hernandez G, Reich HN, Coppo R, Roberts ISD, Feehally J, et al. The MEST score provides earlier risk prediction in IgA nephropathy. *Kidney Int.* 2016;89(1):167–175.
- A40. Eggers KM, Kempf T, Larsson A, Lindahl B, Venge P, Wallentin L, et al. Evaluation of temporal changes in cardiovascular biomarker concentrations improves risk prediction in an elderly population from the community. *Clin Chem.* 2016;62(3):485–493.
- A41. Neyra JA, Li X, Canepa-Escaró F, Adams-Huet B, Toto RD, Yee J, et al. Cumulative fluid balance and mortality in septic patients with or without acute kidney injury and chronic kidney disease. *Crit Care Med.* 2016;44(10):1891–1900.
- A42. Ferreira JP, Girerd N, Pellicori P, Duarte K, Girerd S, Pfeffer MA, et al. Renal function estimation and Cockcroft-Gault formulas for predicting cardiovascular mortality in population-based, cardiovascular risk, heart failure and post-myocardial infarction cohorts: The Heart “OMics” in AGEing (HOMAGE) and the high-risk myocardial infarction database initiatives. *BMC Med.* 2016;14(1):181.
- A43. Mentias A, Naji P, Gillinov AM, Rodriguez LL, Reed G, Mihaljevic T, et al. Strain echocardiography and functional capacity in asymptomatic primary mitral regurgitation with preserved ejection fraction. *J Am Coll Cardiol.* 2016;68(18):1974–1986.
- A44. Li S, Liu H, Liu J. Predictive performance of adding platelet reactivity on top of CRUSADE score for 1-year bleeding risk in patients with acute coronary syndrome. *J Thromb Thrombolysis.* 2016;42(3):360–368.
- A45. Elias SG, Kok L, de Wit NJ, Witterman BJM, Goedhard JG, Romberg-Camps MJL, et al. Is there an added value of faecal calprotectin and haemoglobin in the diagnostic work-up for primary care patients suspected of significant colorectal disease? A cross-sectional diagnostic study. *BMC Med.* 2016;14(1):141.
- A46. Ko BS, Seneviratne S, Cameron JD, Gutman S, Crossett M, Munnur K, et al. Rest and stress transluminal attenuation gradient and contrast opacification difference for detection of hemodynamically significant stenoses in patients with suspected coronary artery disease. *Int J Cardiovasc Imaging.* 2016;32(7):1131–1141.
- A47. Kalim S, Trottier CA, Wenger JB, Wibecan J, Ahmed R, Ankers E, et al. Longitudinal changes in protein carbamylation and mortality risk after initiation of hemodialysis. *Clin J Am Soc Nephrol.* 2016;11(10):1809–1816.

- A48. Tribouilloy C, Bohbot Y, Maréchaux S, Debry N, Delpierre Q, Peltier M, et al. Outcome implication of aortic valve area normalized to body size in asymptomatic aortic stenosis. *Circ Cardiovasc Imaging* . 2016;9(11):e005121.
- A49. Li S, Liu H, Liu J, Wang H. Improved predictive value of GRACE risk score combined with platelet reactivity for 1-year cardiovascular risk in patients with acute coronary syndrome who underwent coronary stent implantation. *Platelets*. 2016;27(7):650–657.
- A50. Lee JM, Kim CH, Koo B-K, Hwang D, Park J, Zhang J, et al. Integrated myocardial perfusion imaging diagnostics improve detection of functionally significant coronary artery stenosis by <sup>13</sup>N-ammonia positron emission tomography. *Circ Cardiovasc Imaging*. 2016;9(9):e004768.



**Appendix Figure 1.** Literature search and selection.

Abbreviations: AUC = area under the receiver operating characteristic curve; IDI = integrated discrimination improvement; NRI = net reclassification improvement







# 7

## Letters to the editor

Based on:

External validation is only needed when prediction models are worth it

Forike K. Martens, Jannigje G. Kers, and A. Cecile J.W. Janssens

Journal of Clinical Epidemiology 69 (2016) 249-250

Risk Analysis of Prostate Cancer in PRACTICAL Consortium—Letter

Forike K. Martens, Jannigje G. Kers, and A. Cecile J.W. Janssens

Cancer Epidemiol Biomarkers Prev 25 (2016) 222-223

A research article that is published in a scientific journal has usually been peer reviewed prior to publication. Peer review is the process of subjecting research to the scrutiny of experts in the same field (1,2). A post-publication way of peer review is the letter to the editor of the scientific journal that published the article. Besides contributing to scientific discourse, they could also be of benefit to other readers as it may provide additional insights and evidence that could help understand the article (3). In this Chapter we discuss two letters to the editor on two different topics, namely the external validation of prediction models and the assessment of calibration and discriminative ability.

## **External validation of prediction models**

As described in Chapter 1, external validation is required when prediction models are planned to be used in healthcare. External validation determines the replicability and generalizability of the prediction model to new and different patients (4). This refers to the validation of the prediction model in a completely new population or setting, which is similar to the original population. Although temporal and geographical validation are regarded as an approach in between internal- and external validation (5), sometimes they are considered a type of external validation (6,7) and are included as such in the publication of Siontis et al. that is subject of one of the two letters to the editor. Temporal validation means that the prediction model is assessed in newer collected data within the same care center, for example, among more recently included participants in the study. Geographical validation means that the prediction model is assessed in a same population but in a different place than where the prediction model was developed, for example, in another region or a different care center. Siontis et al. wrote a review about the external evaluation practices of newly developed prediction models, in which they concluded that many prediction models lacked external validation (8). Moreover, Siontis et al. describe that of the large number of prediction models that are being developed, only a few are used in clinical practice. The authors evaluated how often the validations were performed by authors that did not publish the derivation model, and subsequently, how well the prediction models performed in these validation studies.

The authors executed a literature search to find articles published until 2010 in which a new prediction model was presented. Articles that published an

external validation of the selected eligible derivation studies were retrieved by searching articles by either an overlapping author group or completely different authors, that cited the derivation studies. External validation studies were selected when the authors claimed to have validated the derivation model (same model, disease and outcome) in different populations. For each derivation- and validation study, the listed authors, several study- and model characteristics, and performance metrics were recorded.

In their review of 88 derivations studies describing 127 newly developed risk prediction models, Siontis et al. (8) found that only 32 models (25%) were externally validated. Siontis et al. conclude that 'the majority of the newly proposed risk prediction models never undergo an external independent validation' (8). From their results, the authors argue that external validation 'should be done by default for all risk prediction models' (8) and that in the absence of external independent validation misleading high expectations are offered. Based on their data, however, we conclude that the percentage of external validation may be as high as 83% as many prediction models were already externally validated, as explained below, and many others were not worth it.

First, the authors used a rather narrow and uncommon definition of external validation, namely that the prediction models had to be independently validated in a subsequent study; the common definition does not require that the external validation is published separately (6). Siontis et al. provides 'details of the derivation studies of newly introduced risk prediction models without any further validation studies' in the supplementary eTable 1 and 'details of the derivation studies of newly introduced risk prediction models that were further validated' in eTable 2. Of the 62 studies that were not externally validated according to the authors, seven had reported validation in entirely independent data sets in the same article. Fourteen other studies had included independent temporal and geographical validation efforts in the same article (Table 1).

Second, many of the remaining 41 studies (62 minus 21) that were not externally validated may not have been worth validating. Twenty-eight studies were conducted in less than 500 people, of which 19 studies in less than 200 people (Table 1). The prediction models estimated in these populations first need to be re-estimated in larger data sets to obtain more robust coefficients for the variables before external validation is warranted. In addition, the authors of nine other articles warned that their results should be interpreted with caution because of study limitations, such as a retrospective study design,

nonrepresentative population, selection bias, missing relevant predictors, and issues around variable assessment. These limitations also require re-estimation of the prediction models before external validation. Hence, only four studies remain that were not externally validated but potentially worth it. Assuming that all externally validated studies in eTable 2 conducted in >500 individuals (n = 19; Table 2) were worth validating, we calculate that the percentage of externally validated studies is 83% (19 of 23).

Our reanalysis shows that the lack of external validation of the studies reviewed by Siontis et al. seems entirely justified. External validation is crucial before prediction models can be implemented in health care, but these efforts should only be done for studies that are worth it.

**Table 1.** Sample sizes and validation of studies included in eTable 1 of Siontis et al.

| <b>Sample size</b> | <b>Number of studies</b> | <b>Number of studies without duplicates <sup>(1)</sup></b> | <b>Independent validation</b> | <b>Temporal and geographical <sup>(2)</sup> validation</b> | <b>Number of studies without duplicates and validated studies</b> |
|--------------------|--------------------------|--|-------------------------------|--|---|
| 0-100              | 13                       | 13   | 0                             | 2  | 11  |
| 101-200            | 10                       | 10   | 0                             | 2  | 8   |
| 201-300            | 6                        | 6  | 0                             | 1  | 5   |
| 301-400            | 6                        | 6  | 2                             | 1  | 3   |
| 401-500            | 3                        | 2  | 1                             | 0  | 1   |
| >500               | 28                       | 25   | 4                             | 8  | 13  |
| <i>Total</i>       | <i>66</i>                | <i>62</i>  | <i>7</i>                      | <i>14</i>  | <i>41</i>   |

(1) Duplicate with study in eTable 2; (2) One geographically validated study among those with sample size >500

**Table 2.** Sample sizes of studies included in eTable 2 of Siontis et al.

| <b>Sample sizes</b> | <b>Number of studies</b> |
|---------------------|--------------------------|
| 0-100               | 2                        |
| 101-200             | 1                        |
| 201-300             | 0                        |
| 301-400             | 2                        |
| 401-500             | 2                        |
| >500                | 19                       |
| <i>Total</i>        | <i>26</i>                |

## Assessment of calibration and discriminative ability

Chapter 1 and 2 of this thesis describe that the evaluation of prediction models should include the assessment of calibration and discrimination (9). Calibration refers to how well the predicted risks from the prediction model match the actual observed risks and discrimination how well a prediction model can distinguish between patients and nonpatients. The second letter to the editor is a comment on an article in which the authors investigated the predictive ability of a polygenic risk score (PRS) and concluded that the risk of the top 1% of the study population was more than 30-fold compared to the bottom 1%. The evaluation of calibration and the discriminative ability was, however, not reported on. The article by Amin Al Olama et al. was driven by the fact that the risks associated with genetic variants that have been discovered in genome wide association studies (GWASs) are argued to be useful for targeted prevention (10). They reason that, because the associated risks are modest, large studies are needed to provide more precise estimation of these risks. This is what they aim to contribute to in their study, by genotyping 25 prostate cancer susceptibility single nucleotide polymorphisms (SNPs) in studies from the international prostate cancer consortium (PRACTICAL) (11).

Amin Al Olama et al. combined data of 25 studies from PRACTICAL and GWASs, and genotyped 25 SNPs when these were not yet available. A total of 40,414 samples (20,288 cases and 20,126 controls) were included in the analyses. A PRS was derived based on the assumption of a log-additive model, by summing the genotypes weighted by the per-allele log odds ratios (ORs) for each of the SNPs, as estimated by logistic regression (10). The risk of prostate cancer was estimated for percentiles of the PRS distribution, categorized into: <1%, 1–10%, 10–25%, 25–75% (“median risk”), 75–90%, 90–99%, and >99%.

Amin Al Olama and colleagues investigated the predictive ability of the PRS and observed that the risk of men in the top 1% of the distribution was 30.6 fold compared with men in the bottom 1% and 4.2 fold compared with the median risk (10). The authors conclude that ‘genetic risk profiling using SNPs could be useful in defining men at high risk for the disease for targeted prevention and screening programs’. Yet, such conclusion warrants a formal assessment of calibration and discriminative ability.

First, assessment of calibration is essential because the reported risks were not based on empirical observations but calculated from a risk model that

was built assuming multiplicative effects between the SNPs. The authors verified whether allelic effects within each SNP could be considered as multiplicative, but not whether multiplicativity between SNPs could be assumed. Multiplicative models are known to under- and overestimate risks at the extremes of the risk distribution, especially when they include a large number of SNPs (12,13). While the authors mentioned that “the predicted ORs for the top 1% and the bottom 1% of the population, based on a log-linear model, did not differ from that observed”, this needs to be evidenced by a formal calibration analysis of the entire risk distribution and of the extremes if these are of special interest.

Second, the discriminative ability of the model should be assessed by examining how well the predicted risks distinguish between men who did or did not develop prostate cancer, quantified by the area under the receiver operating characteristic curve (AUC), to compare its performance with other models. Using the SNP data reported in their Table 2 and applying a validated simulation algorithm (14), we estimated that the AUC of the polygenic risk score would be 0.64. If confirmed by their data, this AUC would be lower than other models, including the prediction model from the Prostate Cancer Prevention Trial, which AUC was 0.66 for any prostate cancer and 0.71 for clinical significant prostate cancer (15).

Finally, the predictive performance is generally highest in the population in which the prediction model is developed, because the coefficients of the model are fitted to the data. The researchers have enough data to split their sample in two and perform both the development and validation analyses in one study. Independent validation of both calibration and discrimination will likely lead to a more modest perspective of the predictive performance of the polygenic risk score.

## References

1. What is peer review? (<https://www.elsevier.com/reviewers/what-is-peer-review>). (Accessed May 25, 2021)
2. Mark Ware Consulting. Publishing Research Consortium Peer review survey 2015. 2016.
3. Johnson C, Green B. How to write a letter to the editor: an author's guide. *J Chiropr Med.* 2006;5(4):144–147.
4. Ramspek CL, Jager KJ, Dekker FW, et al. External validation of prognostic models: what, why, how, when and where? *Clin Kidney J.* 2021;14(1):49–58.
5. Altman DG, Vergouwe Y, Royston P, et al. Prognosis and prognostic research: Validating a prognostic model. *BMJ.* 2009;338(7708):1432–1435.
6. Moons KG, Kengne AP, Grobbee DE, et al. Risk prediction models: II. External validation, model updating, and impact assessment. *Heart.* 2012;98(9):691–698.
7. Collins GS, de Groot JA, Dutton S, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC Med Res Methodol.* 2014;14:40.
8. Siontis GCM, Tzoulaki I, Castaldi PJ, et al. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *J Clin Epidemiol.* 2015;68(1):25–34.
9. Martens FK, Janssens ACJW. How the intended use of polygenic risk scores guides the design and evaluation of prediction studies. *Curr Epidemiol Rep.* 2019;6(2):184–90.
10. Al Olama AA, Benlloch S, Antoniou AC, et al. Risk analysis of prostate cancer in practical, a multinational consortium, using 25 known prostate cancer susceptibility loci. *Cancer Epidemiol Biomarkers Prev.* 2015;24(7):1121–1129.
11. PRACTICAL. (<http://practical.icr.ac.uk/blog/>). (Accessed May 29, 2021)
12. Moonesinghe R, Khoury MJ, Liu T, et al. Discriminative accuracy of genomic profiling comparing multiplicative and additive risk models. *Eur J Hum Genet.* 2011;19(2):180–5.
13. Song M, Kraft P, Joshi AD, et al. Testing calibration of risk models at extremes of disease risk. *Biostatistics.* 2015;16(1):143–154.
14. Kundu S, Mihaescu R, Meijer CMC, et al. Estimating the predictive ability of genetic risk models in simulated data based on published results from genome-wide association studies. *Front Genet.* 2014;5:179.
15. Louie KS, Seigneurin A, Cathcart P, et al. Do prostate cancer risk models improve the predictive accuracy of PSA screening? A meta-analysis. *Ann Oncol.* 2015;26(5):848–864.





# 8

## General discussion

After almost a century of scientific breakthroughs in genetics, the development of polygenic risk scores (PRSs) has fueled the interest in the use of genetic information for personalized medicine and the management of common diseases in healthcare practice. Moreover, genotyping is becoming cheaper, making genotype information obtainable for billions of individuals. In both scientific and public debate about PRSs, promises have been discussed, but much less has been debated about the evidence that is needed to make claims about the value of PRSs for the prediction of common diseases. The goal of this thesis was to improve understanding of the design, evaluation and interpretation of genetic risk prediction studies for common diseases. In this chapter the main findings of this thesis are addressed and results discussed in a broader perspective, followed by implications for methodology and practice. Finally, conclusions are drawn and recommendations proposed.

## **Main findings**

### **How does the intended use of risk prediction models determine the design and interpretation of prediction studies?**

Prediction models need to be usable and useful. This means that the models should be designed with the healthcare scenario in which the application of the prediction model is foreseen, in mind (Chapter 2). The intended use specifies what needs to be predicted, in whom, how and for what purpose in practice (Table 1). The outcome that is predicted, the target population and the selected predictors determine the predictive ability of the model, while the purpose of testing provides the context for deciding if the predictive ability is high enough to be useful in health care. The outcome of interest, for instance, the risk of developing type 2 diabetes should also include a relevant risk period as the predictive performance may vary with the duration of the follow-up. The target population defines in which population the prediction model should be studied, for example, when the 10-year risk of type 2 diabetes for young adults is of interest, then the study population could consist of individuals between, say, ages 18 and 25. When the target population is decided, it follows which predictors are available to predict the outcome of interest and which predictors might be less feasible or not affordable in the intended setting (1). For example, a prediction model for type 2 diabetes that is to be used by primary care physicians should

not include imaging variables, such as abdominal magnetic resonance imaging (2), as this requires that individuals are first referred to an imaging center before the model can be used.

The purpose of testing, for example, identifying young adults at high risk for type 2 diabetes to offer special exercise programs (Table 1) or improving the efficiency of breast cancer surveillance or prostate-specific antigen (PSA) testing in screening for prostate cancer (3,4), is crucial for deciding whether the predictive performance of the prediction model is high enough. For its intended use, a prediction model requires a certain minimum sensitivity and specificity, it needs to perform at least as good or better than existing stratification strategies (5,6). In large scale screening programs the overall benefits of screening must outweigh the harms, for which a minimum sensitivity and specificity are needed. In current practice, breast cancer surveillance with mammography screening is informed by a women's breast cancer risk, divided into different risk categories (e.g., in the Netherlands into categories <20%, 20-30%, 30-60%, and >60% lifetime risk (7)). The breast cancer risk is determined based on the presence of traditional risk factors (especially age), family history, and, if pertinent, high-penetrant genetic pathogenic variants. It is argued that a PRS consisting of many low-risk genetic variants could lead to improved risk stratification of the existing strategy (8). Concluding, for both design and interpretation of prediction studies the intended use of the prediction model is key.

**Table 1.** What is predicted, in whom, how, for what purpose?

| <b>Health care scenario</b> | <b>Implications for research</b>  | <b>Example</b>   |
|-----------------------------|-----------------------------------|--|
| What is predicted,          | Selection of outcome              | 10-year risk of type 2 diabetes  |
| in whom,                    | Selection of population           | Young adults   |
| how,                        | Selection of predictors and model | Age, sex, and 37 genetic susceptibility variants, in logistic regression model |
| for what purpose?           | Specification of aim              | Stratify prevention with supervised exercise program for the high-risk group   |

## Evaluation of the predictive performance

The evaluation of prediction models often includes the assessment of discrimination with the area under the receiver operating characteristic (ROC) curve (AUC), reclassification with the net reclassification improvement (NRI), and

predictive ability with integrated discrimination improvement (IDI) (9–11). Partly in response to the criticism on the AUC, Pencina et al. introduced the NRI and IDI (16). Since the introduction of the two metrics in 2008, both gained popularity. The AUC, NRI and IDI are commonly used in prediction studies, but they are also criticized (12–18). AUC is criticized for lacking an intuitive interpretation and being insensitive to new risk factors; NRI and IDI for being overly sensitive to the addition of new risk factors. For example, the NRI may easily be non-zero due to the reclassification of many individuals with a predicted risk close to the risk threshold.

### **Why is the area under the ROC curve a metric of discrimination?**

The interpretation of the AUC has been a challenge ever since its introduction in medicine (10). Generally, the AUC is described as the probability that predicted risks correctly identify a random pair of a patient and nonpatient, but this explanation seems clinically irrelevant (12) and does not clarify *why* the AUC, as the area under the ROC curve, is a metric of discrimination. The area under the ROC curve is visualized in the ROC plot. We showed that the ROC curve is a transformation of the distribution of predicted risk for patients (Chapter 3) and the diagonal line in the plot, of the distribution of the nonpatients. The latter is not simply a reference line. The space between the diagonal line and the curve reflects the separation between the risk distributions of patients and nonpatients and therewith the discriminative ability of the prediction model.

### **Can the predictive ability of a model improve when discrimination does not?**

Prediction models are updated, with a PRS or other risk factors, to improve clinical care or prevention. To achieve this, new risk factors need, at least, to improve the discriminative ability of the prediction model. The AUC has been criticized that it is unable to show a change in discrimination even when strong risk factors are added to the model (16-19). Is it possible that the predictive ability improves when discrimination does not?

Using simulated data, we found that discrimination, assessed by the AUC, and predictive ability, assessed by IDI, both increased when a strong risk factor was added to a model that had an AUC up to approximately 0.80-0.90 (Chapter 4). Thus, in this case, updating prediction models with new risk factors that do not improve the discriminative ability of a model, do not improve the risk

difference between patients and nonpatients. When the AUC of the initial model is already high, say above 0.90, we observed that in these instances the risk differences between patients and nonpatients became wider, even when the changes in predicted risks did not result noticeably in an increase of the AUC (Chapter 4). Practically, as the baseline AUCs of prediction models for common diseases do not often rise above 0.90, if the AUC does not improve from the added risk factor this indicates that there is no significant improvement in the predictive ability.

When AUC improves only minimally, IDI and NRI may be statistically significant (Chapter 5-6). That is why others have argued that the metrics are too sensitive for identifying changes in predicted risks (20–22). NRI may easily be non-zero and statistically significant due to minor changes in predicted risks resulting in the reclassification of many individuals with a predicted risk close to the risk threshold (23) and when sample sizes are large. Also, NRI may be positive when calibration of the models is poor (22,24,25). Reclassification without improvement of the discriminative ability also implies that the model did not make fewer but different errors than the initial model (23). In healthcare, a positive NRI in absence of improvement in AUC means that, for example, individuals may receive a different recommendation for breast cancer screening but that at the population level no reduction in morbidity and mortality will be observed.

### **How do researchers describe the use and interpret the results of multiple metrics in the assessment of improvement in predictive performance of risk prediction models?**

The  $\Delta$ AUC, NRI and IDI all have the same null hypothesis that the new PRS or risk factor causes no incremental predictive information (24), but they measure different aspects of predictive performance (27). The three metrics are often used simultaneously, but it is unclear whether researchers know how to interpret eventual discordant findings. Our study showed that most researchers give a correct definition of the AUC and NRI, but that IDI often is wrongly described as a metric of discrimination or reclassification. About half of the authors described how AUC was calculated, but only few reported the formulas for the NRI and IDI (Chapter 5-6). Authors of clinical prediction studies more often described how they calculated the NRI and IDI and these were also more often correct as compared to authors of polygenic prediction studies (Chapter 6). Based on

these observations, we concluded that some researchers may not know what each of the metrics assesses.

We found that the inferences from each of the metrics were largely based on the statistical significance, also when their absolute values were small (Chapter 5 and 6). Using statistical significance as a basis for conclusions about the improvement in predictive performance is problematic, because in large studies small values can be statistically significant easily. These small values do not indicate clinically relevant improvement in the predictive performance of the model.

### **Estimating predictive performance in simulated data**

Simulation studies can be useful to explore the characteristics of performance metrics (Chapter 4). Furthermore, when epidemiological information such as effect sizes and frequencies of predictors and the event rate are known, simulation studies can be used to calculate the AUC and other metrics. Having an estimate of the discriminative ability of a prediction model allows to interpret the performance of the model and compare it to similar prediction models in the absence of real data. In a letter to the editor, we applied a simulation algorithm for a prediction study in which the researchers did not report the AUC of their model for the prediction of prostate cancer (28). We found it to be lower than AUCs of similar already existing models, and concluded that a more modest conclusion about the usefulness of the PRS for defining men at high risk for prostate cancer would have been in place (Chapter 7).

## **Implications for research**

The 21st century started off with some bold predictions about how genomic medicine could possibly revolutionize the personalization of medicine (29). Advances in DNA sequencing, dropped costs, the Human Genome Project, and discoveries of many common genetic variants in GWAS, have fueled the interest in genetic risk prediction for common diseases. And even today, several leading organizations in the development of personalized medicine have PRSs on their agenda for the coming years (30–32). For many applications of PRSs the evidence is still to be gathered and the usefulness yet to be proven. Some researchers, however, have expressed that PRSs may be ready for implementation in clinical

care, for example, for breast cancer and cardiovascular disease (33,34). Here I describe several implications for the design of prediction studies and guidance for the assessment of evidence presented in (polygenic) prediction articles.

### **Designing a prediction study**

The intended use of prediction models in healthcare has implications for the design of the prediction study. As described in Chapter 2, the intended use should be the starting point of every prediction study design, which is visually shown in the ACCE model where ‘disorder & setting’, referring to the intended use and setting, are at the center of the ACCE model (Chapter 1). Data gathered in prediction studies should be relevant for the intended use of the prediction model. This means that when available data are used, the population, outcome and available predictors should be evaluated in order to know how the data match the intended setting. Today, many genetic prediction studies use data from the UK biobank (35), but this dataset may overestimate the predictive performance. The population of the UK biobank consist of individuals within a wide range of age, from 40 to 69 at baseline (36), which inflates the AUC of prediction models of age-related diseases that include age as a predictor. The chances for older individuals to develop a common disease within the short follow-up time of the cohort (6-7 years (37)) are higher for older than for younger individuals. Also from the intended use perspective, the wide age range does not make sense, because prediction models are generally used in individuals at a specific age. For example, a cardiovascular risk profiling program in health care might invite specific cohorts, for instance men aged 50 and women aged 60 years (38) while a PRS for coronary heart disease developed in the UK biobank (33,39) implies a target population between 40 and 69 years old. In this case, the wide age range of the UK biobank is not representative of the target population, and hence the performance of the model still undetermined.

When the design of the prediction study does not reflect the intended use, it should be anticipated that the future predictive performance in the target population may deviate from the study. A recent study applied a poststratification method to match individuals from the UK biobank cohort to the target population and concluded that the lack of cohort representativeness in the UK biobank may lead to false effect estimates (40). The design of prediction studies should be guided by the intended use. When existing cohort data are used, researchers should consider to only use a selection of the study population that reflects the



target population.

## **Evaluating a prediction study**

### ***External validation***

Most articles describe the development of a new prediction model instead of validating existing models externally (41–43). Only when prediction models show sufficient (improvement in) predictive performance and can potentially improve health outcomes or the efficiency of care, need external validation before they can be considered further (Chapter 7). The performance of the model should be reassessed in an independent, clinically relevant population (Chapter 1, 2 and 7) to investigate the generalizability of the prediction model. Validation of the prediction models is needed, because predictive performance is usually higher in the population that was used to fit the prediction model.

### ***Calibration***

To ensure that predicted risks agree with the observed event rates, prediction models need to be well calibrated. This is graphically displayed in a calibration plot and quantified by calibration in the large and the calibration slope (9,44). The Hosmer-Lemeshow test is also often used as a calibration test, but, because the metric is unable to detect substantial miscalibration in small samples and is over-sensitive to minor miscalibration in large samples, its use is discouraged (45,46). Under- or overestimation of risk may lead to under- and overtreatment (45). Poor calibration may affect the values of all metrics, but NRI and IDI in particular (22,25). Reporting calibration metrics is hence important (Chapter 7).

### ***The area under the ROC curve***

The AUC is a suitable and relevant metric for the evaluation of the discriminative ability of prediction models for common diseases. The alternative explanation of the ROC plot as an alternative way of presenting risk distributions (Chapter 3) invalidates most purported limitations of the AUC (Chapter 3). Criticism remains to whether the ROC plot provides information beyond the value of the AUC (47). We argue that the curve may show an 'angle' which tells that the model includes a binary predictor with a stronger effect on disease risk than all other variables (48) and the curve may be stepped rather than smooth which tells that the sample size is too low, the incidence is low, or that the prediction model is based on a relatively small set of categorical predictors that generate a small

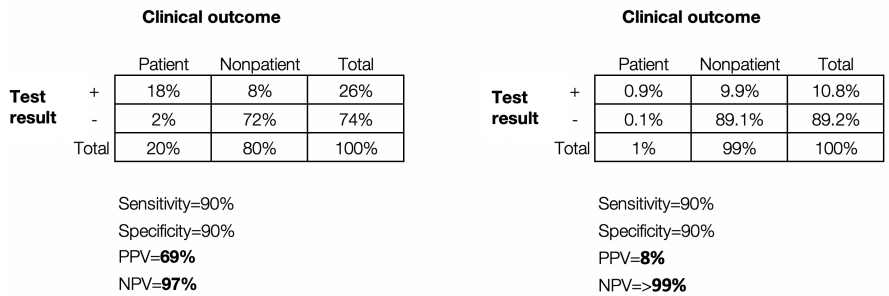
number of predictor combinations. Knowing that the curve is skewed to the right or left of the ROC plot informs whether the effect of the risk factor concerns mostly individuals at high- or low risk, respectively. This provides insight into the underlying risk distributions. Of course, information about predictors and the required sample size should be discussed in the prediction article. Additionally, the ROC plot can assist interpretation of the results by providing visual information.

For rare diseases and low incidence, the AUC should be interpreted with caution as a high AUC may be accompanied with low predictive values (49,50). The predictive values are influenced by disease incidence; even excellent models with very high sensitivity and specificity across relevant risk thresholds may have poor positive predictive values (PPV is the risk of disease and NPV 1-risk of disease for risk groups defined by a certain risk threshold, see Figure 3 Chapter 1) when used in populations where the incidence of the disease is low. For instance, when a test with sensitivity and specificity of 90% for a certain risk threshold is used in a population in which the disease occurs in 20%, PPV will be 69% and NPV will be 97%. Yet, when the same test is used in a population in which the incidence of the disease is 1%, PPV will be 8% while NPV will be higher than 99% (Figure 1). In other words, when the test is used in a population in which the disease is less frequent, more individuals test falsely positive. False positive test results may have negative psychosocial consequences for individuals who receive such a false positive result and may lead to unnecessary treatment with its associated costs and risks.

### ***Interpretation of metrics of reclassification***

AUC, NRI and IDI provide complementary information about the improvement in predictive performance of prediction models (Chapter 5 and 6). They should not be interpreted individually without regard of the results of the others. The NRI and IDI are easily statistically significant in large studies, which means that focusing on the statistical significance of the NRI and IDI without evaluating their values could lead to the conclusion that the model improved, while the values indicated minimal or no improvement. The evaluation should focus on the values of the metrics, not on the statistical significance. We argue that the intended use might determine which metric can be the decisive factor for the conclusion about the improvement of the prediction model, for instance, the NRI when the interest is improving classification.

The NRI has two versions: a categorical and continuous one. Because the categorical NRI evaluates the net changes between risk categories, the selected risk thresholds should be well established and motivated in the prediction article as the NRI varies with the chosen cutoff values (23). Justifying thresholds is often omitted in empirical studies (51,52). When well established thresholds are not available, the cutoff should at least be chosen such that it potentially results in a meaningful change of medical decisions. The use of the NRI is discouraged as the metric is often positive and statistically significant from added risk factors with weak effects (25,53). The categorical NRI is the sum of two fractions with different denominators, which is impossible to interpret as it is a meaningless number, therefore, it is urged to report the reclassification of events and nonevents separately (51).



**Figure 1.** Influence of incidence on predictive value of a test or model in a population in which the disease occurs in 20% (left) versus 1% (right). Positive predictive values (PPV) and negative predictive values (NPV) represent the probability of having the disease when the test result is positive and the probability of not having the disease when the result is negative. Sensitivity and specificity indicate the test’s ability to detect the presence of disease in people with the disease and its absence in those without.

### Reporting practices

The literature on polygenic risk prediction research is growing rapidly, but suffers from a great variability in terminology, lack of information provided in the articles and metrics reported. The intended use of the prediction models is rarely elaborated on (Chapter 2) and definitions and calculation methods of metrics insufficiently reported (Chapter 5 and 6). To provide the needed evidence for the prediction models and to allow comparison between models, it is very important that guidelines such as GRIPS and the GRIPS update, Polygenic Risk

Score Reporting Standards (54,55) are followed and analysis described with care. Better reporting hopefully contributes to improving the quality of prediction studies.

### **What promising risk models have in common**

Promising risk models that include a PRS show substantial improvement in discrimination compared to current models, investigated in a population that reflects the target population. What promising models tend to have in common is the presence of several SNPs with strong effect on the development of disease, and the availability of preventive measures and treatments for different risk groups. Eventually, the ability to improve current models and the availability of interventions determines whether PRSs could be a fruitful application for personalized medicine. And, of course, a thorough evaluation of the PRS following the ACCE model is needed to provide evidence of their utility, including how PRSs are accepted by clinicians, patients and citizens, the ability to resolve ethical aspects of genetic tests, social effects, accessibility, and more practical aspect such as integration with electronic health records (56) as these aspects will actually determine whether PRSs will be a success in practice.

## **Concluding remarks and recommendations**

From the results of the studies presented in this thesis, I have the following conclusions and recommendations:

- The intended use of a prediction model has a pivotal role in the design and evaluation of prediction studies and should be clearly described in scientific prediction articles, including specification of what needs to be predicted, in whom, how and for what purpose.
- The ROC curve is an alternative way of presenting risk distributions and the diagonal line is not only a reference line, but it is the risk distribution of the nonpatients. The separation between the risk distributions represents the discriminative ability of the model.
- The AUC is not insensitive; when a risk factor increases the AUC minimally also a minimal improvement in predictive ability should be expected. Only when the AUC of the initial model is high, say above 0.90, the predictive ability may still improve while the improvement in discrimination does not.

- The evaluation of prediction models, including definitions and calculation methods of required metrics, should be clearly described in prediction articles to improve their quality.
- The interpretation of the metrics of predictive performance should be based on their absolute values and not on the statistical significance.
- PRSs can be taken into consideration for follow-up studies such as cost-effectiveness, and implementation studies when improvement in the discriminative ability of a model and calibration is proven and promising.

It is hoped that this thesis advances knowledge about prediction studies and helps to promote better evaluation and understanding of prediction models in the attempt to improve the prediction of common diseases and translate prediction models into valuable applications in healthcare.

## References

1. Moons KGM, Royston P, Vergouwe Y, et al. Prognosis and prognostic research: what, why, and how? *BMJ*. 2009;338(feb23 1):b375–b375.
2. Wang M, Luo Y, Cai H, et al. Prediction of type 2 diabetes mellitus using noninvasive MRI quantitation of visceral abdominal adiposity tissue volume. *Quant Imaging Med Surg*. 2019;9(6):1076–1086.
3. Chiarelli AM, Prummel M V., Muradali D, et al. Effectiveness of screening with annual magnetic resonance imaging and mammography: Results of the initial screen from the Ontario High Risk Breast Screening Program. *J Clin Oncol*. 2014;32(21):2224–2230.
4. Seibert TM, Fan CC, Wang Y, et al. Polygenic hazard score to guide screening for aggressive prostate cancer: development and validation in large scale cohorts. *BMJ*. 2018;j5757.
5. Andermann A, Blancquaert I, Beauchamp S, et al. Revisiting Wilson and Jungner in the genomic age: A review of screening criteria over the past 40 years. *Bull World Health Organ*. 2008;86(4):317–319.
6. Calonge N, Green NS, Rinaldo P, et al. Committee report: Method for evaluating conditions nominated for population-based screening of newborns and children. *Genet Med*. 2010;12(3):153–159.
7. Borstkanker - Algemeen - Richtlijn - Richtlijndatabase. (<https://richtlijndatabase.nl/richtlijn/borstkanker/algemeen.html>). (Accessed July 20, 2021)
8. Lee A, Mavaddat N, Wilcox AN, et al. BOADICEA: a comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors. *Genet Med*. 2019;21(8):1708–1718.
9. Austin PC, Steyerberg EW. Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. *Stat Med*. 2014;33(3):517–535.
10. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;143(1):29–36.
11. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*. 1983;148(3):839–843.
12. Parikh CR, Thiessen-Philbrook H. Key concepts and limitations of statistical methods for evaluating biomarkers of kidney disease. *J Am Soc Nephrol*. 2014;25(8):1621–1629.
13. Pepe MS, Janes HE. Gauging the performance of SNPs, biomarkers, and clinical factors for predicting risk of breast cancer. *J Natl Cancer Inst*. 2008;100(14):978–979.
14. Halligan S, Altman DG, Mallett S. Disadvantages of using the area under the receiver operating characteristic curve to assess imaging tests: A discussion and proposal for an alternative approach. *Eur Radiol*. 2015;25(4):932–939.
15. Flach PA. ROC Analysis. In: Sammut C, Webb G, editors. *Encyclopedia of Machine Learning*: Springer US. 2016;869–875.
16. Pencina MJ, D'Agostino Sr. RB, D'Agostino Jr. RB, et al. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med*. 2008;27(2):112–157.
17. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation*. 2007;115(7):928–935.
18. Ware JH. The limitations of risk factors as prognostic tools. *N Engl J Med*. 2006;355(25):2615–2617.
19. Pepe MS. Limitations of the Odds Ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *Am J Epidemiol*. 2004;159(9):882–890.
20. Pepe MS, Janes H, Li CI. Net risk reclassification P values: Valid or misleading? *JNCI J Natl Cancer Inst*. 2014;106(4):dju041.
21. Gerds TA, Hilden J. Calibration of models is not sufficient to justify NRI. *Stat Med*. 2014;33(19):3419–3420.
22. Hilden J, Gerds TA. A note on the evaluation of novel biomarkers: Do not rely on integrated discrimination improvement and net reclassification index. *Stat Med*. 2014;33(19):3405–3414.
23. Mihaescu R, van Zitteren M, van Hoek M, et al. Improvement of risk prediction by genomic profiling:

- Reclassification measures versus the area under the receiver operating characteristic curve. *Am J Epidemiol.* 2010;172(3):353–361.
24. Leening MJG, Steyerberg EW, Van Calster B, et al. Net reclassification improvement and integrated discrimination improvement require calibrated models: Relevance from a marker and model perspective. *Stat Med.* 2014;33(19):3415–3418.
  25. Pepe MS, Fan J, Feng Z, et al. The net reclassification index (NRI): A misleading measure of prediction improvement even with independent test data sets. *Stat Biosci.* 2015;7(2):282–295.
  26. Pepe MS, Kerr KF, Longton G, et al. Testing for improvement in prediction model performance. *Stat Med.* 2013;32(9):1467–1482.
  27. Kerr KF, Bansal A, Pepe MS. Further insight into the incremental value of new markers: the interpretation of performance measures and the importance of clinical context. *Am J Epidemiol.* 2012;176(6):482–487.
  28. Al Olama AA, Benlloch S, Antoniou AC, et al. Risk analysis of prostate cancer in practical, a multinational consortium, using 25 known prostate cancer susceptibility loci. *Cancer Epidemiol Biomarkers Prev.* 2015;24(7):1121–1129.
  29. Collins F. Has the revolution arrived? *Nature.* 2010;464(7289):674–675.
  30. Williams GA, Liede S, Fahy N, et al. Regulating the unknown POLICY BRIEF 38 guide to regulating genomics for health policy-makers HEALTH SYSTEMS AND POLICY ANALYSIS. 2020.
  31. Government H. *Genome UK: the future of healthcare.* 2020.
  32. Green ED, Gunter C, Biesecker LG, et al. Strategic vision for improving human health at the forefront of genomics. *Nature.* 2020;586(7831):683–692.
  33. Khera A V, Chaffin M, Aragam KG, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet.* 2018; 50(9):1219–1224.
  34. Mavaddat N, Michailidou K, Dennis J, et al. Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. *Am J Hum Genet.* 2019;104(1):21–34.
  35. UK Biobank (<https://www.ukbiobank.ac.uk/>) (Accessed October 7, 2021)
  36. Sudlow C, Gallacher J, Allen N, et al. UK Biobank: An Open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Med.* 2015;12(3):e1001779.
  37. Fry A, Littlejohns TJ, Sudlow C, et al. Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. *Am J Epidemiol.* 2017;186(9):1026–1034.
  38. Cardiovasculair risicomanagement | NHG-Richtlijnen. (<https://richtlijnen.nhg.org/standaarden/cardiovasculair-risicomanagement#volledige-tekst>). (Accessed June 14, 2021)
  39. Inouye M, Abraham G, Nelson CP, et al. Genomic risk prediction of coronary artery disease in 480,000 adults. *J Am Coll Cardiol.* 2018;72(16):1883–1893.
  40. Stamatakis E, Owen KB, Shepherd L, et al. Is cohort representativeness passé? Poststratified associations of lifestyle risk factors with mortality in the UK Biobank. *Epidemiology.* 2021;179–188.
  41. Collins GS, de Groot JA, Dutton S, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC Med Res Methodol.* 2014;14:40.
  42. Steyerberg EW, Moons KGM, van der Windt DA, et al. Prognosis Research Strategy (PROGRESS) 3: Prognostic model research. *PLoS Med.* 2013;10(2).
  43. Siontis GCM, Tzoulaki I, Castaldi PJ, et al. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *J Clin Epidemiol.* 2015;68(1):25–34.
  44. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J.* 2014;35(29):1925–31.
  45. Calster B Van, McLernon DJ, Smeden M van, et al. Calibration: the Achilles heel of predictive analytics. *BMC Med* 2019 171. 2019;17(1):1–7.
  46. Hosmer DW, Hjort NL. Goodness-of-fit processes for logistic regression: simulation results. *Stat Med.* 2002;21(18):2723–38.
  47. Verbakel JY, Steyerberg EW, Uno H, et al. ROC curves for clinical prediction models part 1.

- ROC plots showed no added value above the AUC when evaluating the performance of clinical prediction models. *J Clin Epidemiol*. 2020;126:207–216.
48. Kundu S, Kers JG, Janssens ACJW. Constructing hypothetical risk data from the area under the ROC Curve: Modelling distributions of polygenic risk. *PLoS One*. 2016;11(3):e0152359.
  49. Fawcett T. An introduction to ROC analysis. *Pattern Recognit Lett*. 2006;27(8):861–874.
  50. Davis J, Goadrich M. The relationship between precision-recall and ROC curves. In: *ACM International Conference Proceeding Series*. New York, New York, USA: ACM Press; 2006:233–240.
  51. Leening MJG, Vedder MM, Wittman JCM, et al. Net reclassification improvement: computation, interpretation, and controversies: a literature review and clinician's guide. *Ann Intern Med*. 2014;160(2):122–31.
  52. Tzoulaki I, Liberopoulos G, Ioannidis JPA. Use of reclassification for assessment of improved prediction: An empirical evaluation. *Int J Epidemiol*. 2011;40(4):1094–1105.
  53. Pencina MJ, D'Agostino RB, Pencina KM, et al. Interpreting incremental value of markers added to risk prediction models. *Am J Epidemiol*. 2012;176(6):473–481.
  54. Janssens ACJW, Ioannidis JPA, Bedrosian S, et al. Strengthening the reporting of Genetic Risk Prediction Studies (GRIPS): explanation and elaboration. *J Clin Epidemiol*. 2011;64(8):e1–e22.
  55. Wand H, Lambert SA, Tamburro C, et al. Improving reporting standards for polygenic scores in risk 1 prediction studies. *Nature*. 2021;591(7849):211–219.
  56. Sharma V, Ali I, Van Der Veer S, et al. Adoption of clinical risk prediction tools is limited by a lack of integration with electronic health records. *BMJ Heal Care Informatics*. 2021;28(1):100253.





## Summary | Samenvatting

For decades, researchers have been putting effort in advancing the prediction of common diseases to improve the identification of at-risk groups for preventive interventions, support physicians in medical decision making, and inform individuals about their risk or progression of disease. Ultimately, this would lead to health gain for many individuals. Current prediction models for common diseases typically include clinical, demographical, environmental and lifestyle predictors, but due to the multifactorial etiology of these diseases and the discoveries of many common single nucleotide polymorphisms (SNP; a variation occurring at a single nucleotide of the genome) over the past decades, there has been a great interest in adding polygenic risk scores (PRSs) to the clinical models. PRSs quantify the combined contribution of multiple SNPs to the risk of common diseases. Since prediction models are developed with the aim of applying them in healthcare, and hence medical decisions are based on the risk estimates, adequate risk predictions are of great importance. Therefore, prediction studies are needed to evaluate the predictive performance of prediction models and provide the necessary evidence for claims about the clinical validity and utility. This thesis describes methodological studies on (genetic) risk prediction of common diseases and aims to improve understanding and use of traditional and newer metrics of model performance and to provide insight into key concepts and considerations in prediction research.

**Chapter 1** comprises a general introduction of the progress in the field of risk prediction for common diseases and offers an overview of the evaluation of prediction models. It describes current methodological challenges and the three research questions that have driven this project: 1) How does the intended use of risk prediction models determine the design and interpretation of prediction studies?, 2) Why is the area under the receiver operating characteristic (ROC) curve (AUC) a metric of discrimination?, and 3) What do different metrics of predictive performance measure?

The first research question pertains to the intended use of risk prediction models. **Chapter 2** describes how the intended use is defined and what the main considerations are in prediction research that are of importance in the design and evaluation of prediction studies. The intended use indicates in which healthcare setting the prediction model is foreseen. This should include a description of in whom and how the model will be used and for what purpose it will be implemented. The described healthcare setting has two major implications for scientific study. First, because the predictive ability of prediction

models vary between populations and settings, the intended use dictates the design of the prediction study; it defines the outcome that needs to be predicted, the population that needs to be studied and with what predictors. Second, the intended use of the model also provides the necessary context to decide whether the predictive ability is high enough for the model to be potentially useful in healthcare, because the same model may be predictive enough for one application, but not for another. This is why the intended use should guide the design of risk prediction research.

Our second objective is to explain how the area under the ROC curve (AUC) is a metric of discrimination. To this day, the AUC is the most commonly used metric for the evaluation of the discriminative ability of risk models, but is the most criticized as well. It has been argued that the AUC is clinically irrelevant and lacks an intuitive interpretation. Therefore, in **Chapter 3** we explain the relevance of the AUC as metric of discrimination by describing 1) how the ROC curve can be seen as an alternative way of presenting the risk distributions of patients and nonpatients. The separation between the distributions simply determines the discriminative ability of the model. And 2) how the shape of the ROC curve is informative of these underlying risk distributions. For example, ROC curves are rounded when the prediction model included variables with similar effect on disease risk; ROC curves have an angle when, for example, one binary risk factor has a stronger effect; and ROC curves are stepped rather than smooth when the sample size or incidence is low, or when the prediction model is based on a relatively small set of categorical predictors. We show that this perspective on the ROC plot invalidates most purported limitations of the AUC and attributes other argued limitations to the underlying risk distributions. As the AUC is a metric of the discriminative ability of prediction models, the model assessment should be supplemented with other metrics to evaluate the clinical utility before the decision can be made to implement a risk model in practice. Clinical utility depends on effectiveness of interventions, so the evaluation should include metrics of health gain.

Our third research question concerns what different metrics of predictive performance measure. The AUC is also criticized because the metric would be insensitive and unable to detect moderate improvements in discriminative ability of prediction models. Adding SNPs to a clinical prediction model often increases the AUC only slightly. The first sub question is whether the AUC hides improvement from additional risk factors. In **Chapter 4** we investigated with a

simulation study whether risk factors that minimally improve the AUC, may still improve the risk difference between people who will develop the disease and those who will not. We found that risk factors with stronger effects on disease risk resulted in larger increments in AUC ( $\Delta$ AUC) and risk differences, as shown by the integrated discrimination improvement (IDI). Across baseline AUC, for a risk factor with the same odds ratio, both the  $\Delta$ AUC and IDI were smaller when the AUC of the baseline model was higher. When the  $\Delta$ AUC was smaller than 0.01, the improvement in the risk differences was also small, except when the AUC of the baseline model was  $>0.90$ . Similarly, in 33 empirical genetic prediction studies we observed that  $\Delta$ AUC below 0.01 also yielded minimal improvements of the risk difference. In the range of AUC values typically observed in studies on polygenic risk prediction, small improvements in discrimination can only lead to also small improvements in the risk difference between people who develop the disease and those who will not. We argue that the AUC is not as insensitive as thought.

The AUC is increasingly assessed together with the net reclassification improvement (NRI) and IDI in the evaluation of the improvement of polygenic risk prediction models. The NRI assesses the improvement in classification in the updated model compared to the initial model. The second sub question concerns the knowledge and use of the multiple metrics of predictive performance in prediction studies. The aim of **Chapter 5** is to evaluate how researchers defined, calculated and interpreted  $\Delta$ AUC, NRI, and IDI in polygenic prediction studies where these three metrics are simultaneously assessed. We performed a literature search and included 32 articles that met the inclusion criteria (an empirical study that evaluated the improvement in predictive performance from SNPs added to clinical risk models by assessing  $\Delta$ AUC, NRI, and IDI). In the review of the articles we found that most authors correctly defined the AUC, NRI, but none defined IDI correctly and in half of the articles it was correctly described how the AUC was obtained, but only few authors described the calculation methods of NRI and IDI. The interpretation of the values of the metrics, almost all followed the statistical significance; when a metric was statistically significant the values were interpreted as indicative of improvement, irrespective of the absolute values of the metrics. Also, small, nonsignificant changes in the AUC were interpreted as indication of improvement when NRI and IDI were statistically significant.

**Chapter 6** describes the evaluation of how researchers of non-genetic

clinical prediction studies defined, calculated and interpreted the simultaneous assessment of the  $\Delta$ AUC, NRI, and IDI. In most of the fifty-six included articles (an empirical study that evaluated the improvement in predictive performance from added non-genetic factors to clinical risk models by assessing  $\Delta$ AUC, NRI, and IDI), researchers provided a correct definition of the AUC, about half the articles for NRI and few authors correctly defined the IDI. In half of the articles researchers correctly indicated how AUC was obtained. In fewer articles, researchers described the calculation methods of NRI and IDI, and when a description was provided more than half were correct. Similar to the study presented in **Chapter 5**, the values of the metrics were interpreted as indicative of improvement when they were statistically significant, irrespective of the values' magnitudes. The studies of **Chapter 5 and 6** both show that there is scope for improvement among researchers whom interpret the various metrics for the assessment of the predictive performance of prediction models, as they often rely solely on the statistical significance for their interpretation. Hence, a better understanding of the metrics is needed to achieve more meaningful interpretation of prediction studies.

In **Chapter 7** we discuss two letters to the editor on two different topics, namely the external validation of prediction models and the assessment of calibration and the discriminative ability. The first example is a letter to the editor in response to an article that found that only 25% of the risk models were externally validated. The authors used a rather uncommon definition of external validation. Based on a reanalysis of their data we conclude that the percentage of external validation may be as high as 83% as many models were already externally validated and many others were not worth it. We point out that external validation is only needed when prediction models are worth it. This is determined by, for example, the expected improvement of the current model in use or current practice, the desirability of the model to the public, the intended use and the estimated health gain. The second example is a letter to the editor in response to an article, in which the predictive ability of a PRS for the prediction of prostate cancer was investigated. The assessment of calibration and discrimination were both not reported in the article. When we used the data presented in the article and applied a validated simulation method we found that the AUC of the PRS would be lower than other known models.

## **Concluding remarks**

The intended use of prediction models has a pivotal role in the design and interpretation of prediction studies. As the predictive ability of prediction models varies between populations and settings, the prediction study should be conducted with the targeted healthcare setting in mind, and claims about the readiness of PRSs for implementation in clinical care should be supported with evidence of well calibrated models and improved discriminative ability of the model compared to currently used prediction models. For the assessment of discrimination we have shown that the AUC is the separation between the risk distributions of patients and nonpatients. For the evaluation of all metrics applies that the interpretation should not only rely on the statistical significance, but also on their values in context of the intended use. The field of prediction research could be improved by using the intended use as guidance and by explaining prediction metrics more intuitively so that more researchers could have a greater understanding of them. Whether it is time to consider the implementation of PRSs in health care does not depend solely on the predictive performance of prediction models, but proof of sufficient predictive performance is essential before executing further studies on the usability, usefulness, and meaningfulness of PRS in healthcare.

Al tientallen jaren spannen onderzoekers zich in om vooruitgang te boeken in het voorspellen van veelvoorkomende ziekten om het identificeren van risicogroepen voor preventieve interventies te verbeteren, om artsen te ondersteunen in hun medische besluitvorming en om individuen voor te lichten over hun risico op of het verloop van een ziekte. Uiteindelijk zou dit kunnen leiden tot gezondheidswinst voor velen. Huidige predictiemodellen voor veelvoorkomende ziekten omvatten meestal klinische, demografische, omgevings- en leefstijl voorspellers, maar door de multifactoriële aard van deze ziekten en het ontdekken van talloze veelvoorkomende genetische varianten ("single nucleotide polymorphisms", SNPs, een variatie van een enkele nucleotide in het genoom) gedurende de afgelopen decennia, is er grote interesse om polygene risico scores (PRSs) aan de klinische predictiemodellen toe te voegen. PRSs kwantificeren de gecombineerde bijdrage van meerdere van deze SNPs in het risico op veelvoorkomende ziekten. Aangezien predictiemodellen worden ontwikkeld voor toepassing in de gezondheidszorg, en men medische beslissingen baseert op de gemaakte risicoschattingen, zijn kloppende risicovoorspellingen van groot belang. Daarom zijn er predictiestudies nodig die het voorspellend vermogen van de modellen kunnen evalueren en ook het noodzakelijke bewijs kunnen leveren voor beweringen over de klinische validiteit en het klinisch nut. Dit proefschrift beschrijft methodologische onderzoeken over (genetische) risicovoorspelling van veelvoorkomende ziekten en stelt zich ten doel het begrip en gebruik van traditionele en nieuwere maten van modelprestatie te verbeteren en inzicht te verschaffen in de kernbegrippen en overwegingen in predictieonderzoek.

**Hoofdstuk 1** bevat een algemene inleiding met betrekking tot de voortgang op het gebied van risicopredictie bij veelvoorkomende ziekten en het geeft een overzicht over de evaluatie van predictiemodellen die aan de risicovoorspellingen ten grondslag liggen. Het beschrijft de methodologische uitdagingen van dit moment en de drie onderzoeksvragen die de basis vormen van dit proefschrift: 1) Hoe bepaalt het beoogde gebruik van de risicopredictiemodellen het ontwerp en de interpretatie van predictiestudies? 2) Waarom is het gebied onder de ROC-curve ("receiver operating characteristic curve"), de AUC ("area under the curve"), een maat voor discriminatie? en 3) Wat meten verschillende maten van voorspellend vermogen?

De eerste onderzoeksvraag richt zich op het beoogde gebruik van de risicopredictiemodellen. **Hoofdstuk 2** beschrijft hoe het beoogde



gebruik wordt gedefinieerd en welke de voornaamste overwegingen zijn in het predictieonderzoek die van belang zijn bij ontwerp en evaluatie van predictiestudies. Het beoogde gebruik laat zien op welk deel van de gezondheidszorg het predictiemodel is gericht. Dit zou een beschrijving moeten bevatten bij welke mensen en op welke manier het model zal worden gebruikt en met welk doel het zal worden geïmplementeerd. De beschreven gezondheidszorg setting heeft twee belangrijke implicaties voor wetenschappelijk onderzoek. Ten eerste, omdat het voorspellend vermogen van predictiemodellen varieert tussen verschillende populaties en settings, dicteert het beoogde gebruik het ontwerp van de predictiestudie; het bepaalt de uitkomst die moet worden voorspeld, de populatie die moet worden onderzocht en welke voorspellers worden gebruikt. Ten tweede, het beoogde gebruik verschaft ook de benodigde context om te beslissen of het voorspellende vermogen van het model groot genoeg is om potentieel nuttig te zijn in de gezondheidszorg, aangezien een bepaald model voorspellend genoeg zou kunnen zijn voor de ene toepassing, maar niet genoeg voor een andere. Dit is waarom het beoogde gebruik, het ontwerp van risicopredictieonderzoek, zou moeten leiden.

Onze tweede doelstelling is om uit te leggen hoe de oppervlakte onder de ROC-curve (AUC) een maat voor discriminatie is. Tot de dag van vandaag is de AUC de meest algemeen gebruikte maat voor de evaluatie van het discriminerend vermogen van predictiemodellen, maar tegelijk de maat waar de meeste kritiek op is. Er wordt beweerd dat de AUC klinisch irrelevant is en geen intuïtieve interpretatie kent. Daarom leggen we in **Hoofdstuk 3** uit wat het belang is van de AUC als maat van discriminatie door te beschrijven 1) hoe de ROC-curve kan worden gezien als een alternatieve manier om de risicoverdeling van patiënten en niet-patiënten weer te geven. De scheiding tussen de verdelingen bepaalt simpelweg het onderscheidend vermogen van het model. En 2) hoe de vorm van de curve een beeld geeft van de onderliggende risicoverdelingen. Zo worden, bijvoorbeeld, ROC-curves rond als het predictiemodel variabelen in zich had met een overeenkomstig effect op risico op ziekte; ROC-curves hebben een hoek als, bijvoorbeeld, één binaire risicofactor een sterker effect heeft; en ROC-curves zijn eerder getrapt dan gelijkmatig als de grootte van de steekproef of de incidentie klein is, of als het predictiemodel gebaseerd is op een relatief klein aantal categorische voorspellers. We laten zien, dat dit perspectief op de ROC plot de meeste aangevoerde beperkingen van de AUC minder valide maakt, en wijst andere beargumenteerde beperkingen toe

aan de onderliggende risicoverdelingen. Aangezien de AUC een maat is van het discriminerend vermogen van de predictiemodellen, zou de beoordeling van het model moeten worden aangevuld met andere maten om de klinische bruikbaarheid te evalueren, voordat het besluit kan worden genomen om het risicopredictiemodel te implementeren in de praktijk. Klinische bruikbaarheid hangt af van de effectiviteit van interventies, dus de evaluatie zou maten voor gezondheidswinst moeten bevatten.

Onze derde onderzoeksvraag betreft wat precies de verschillende maten van voorspellend vermogen meten. Er is ook kritiek op de AUC, omdat de maat ongevoelig zou zijn en niet in staat bescheiden verbeteringen in het discriminerend vermogen van predictiemodellen te onderscheiden. SNPs toevoegen aan een klinisch predictiemodel doet de AUC vaak maar minimaal toenemen. De eerste sub onderzoeksvraag is of de AUC, verbetering door extra risicofactoren verbergt. In **Hoofdstuk 4** onderzochten we met een simulatiestudie of risicofactoren die de AUC maar minimaal verbeteren, toch het verschil in risico tussen mensen die de ziekte zullen krijgen en die het niet krijgen zou kunnen verbeteren. We ontdekten, dat risico factoren met sterkere effecten op het risico op ziekte resulteerden in een grotere toename in de AUC ( $\Delta$ AUC) en risicoverschillen, zoals wordt getoond met de geïntegreerde discriminatie verbetering (IDI). Over alle baseline AUCs, voor een risico factor met dezelfde odds ratio, waren zowel de  $\Delta$ AUC en IDI kleiner als de AUC van het baseline model hoger was. Wanneer de  $\Delta$ AUC kleiner was dan 0,01, dan was de verbetering in de risicoverschillen tussen patiënten en niet-patiënten ook klein, behalve als de AUC van het baselinemodel groter was dan 0,90. Op dezelfde manier zagen we dat in 33 empirisch genetische predictiestudies een  $\Delta$ AUC onder de 0,01 ook slechts minimale verbeteringen van het risicoverschil opleverden. Binnen het bereik van AUC-waarden die typisch gezien worden in onderzoeken over polygene risicopredictie, kunnen kleine verbeteringen in de discriminatie slechts leiden tot kleine verbeteringen in het risicoverschil tussen mensen die de ziekte krijgen en zij die niet ziek worden. Wij beweren, dat de AUC niet zo ongevoelig is als gedacht wordt.

De AUC wordt steeds meer beoordeeld samen met de NRI (“net reclassification improvement” = netto reclassificatieverbetering) en IDI bij de evaluatie van de verbetering van de polygene risicopredictiemodellen. De NRI beoordeelt de verbetering in de classificatie in het vernieuwde model vergeleken met het oorspronkelijke model. De tweede sub onderzoeksvraag betreft de

kennis en het gebruik van meerdere maten van voorspellend vermogen in predictieonderzoeken. Het doel van **Hoofdstuk 5** is om te evalueren hoe onderzoekers  $\Delta$ AUC, NRI en IDI in polygene predictieonderzoeken waarin deze drie maten tegelijkertijd worden beoordeeld, definieerden, berekenden en interpreteerden. We deden een literatuuronderzoek en includeerden 32 artikelen die aan de criteria voldeden (een empirische studie die de verbetering in het voorspellend vermogen evalueerde van de SNP(s) toegevoegd aan de klinische risicopredictiemodellen middels  $\Delta$ AUC, NRI en IDI). In het beschouwen van de artikelen ontdekten we, dat de meeste auteurs de AUC en NRI correct definieerden, maar geen van de auteurs de IDI correct definieerde. In de helft van de artikelen werd correct beschreven hoe de AUC werd verkregen, maar slechts enkele auteurs beschreven hoe zij NRI en IDI berekenden. De interpretaties over de waarde van de maten volgden bijna allemaal de statistische significantie; wanneer een maat statistisch significant was, werden de waarden geïnterpreteerd als zijnde een indicatie van verbetering, los van de absolute waarden van de maten. Kleine, niet significante veranderingen in de AUC werden ook geïnterpreteerd als een aanwijzing van verbetering als de NRI en IDI statistisch significant waren.

**Hoofdstuk 6** beschrijft de evaluatie van hoe onderzoekers van niet-genetische klinische predictieonderzoeken de gelijktijdige beoordeling van de  $\Delta$ AUC, NRI en IDI definieerden, berekenden en interpreteerden. In de meeste van de 56 geïncludeerde artikelen (een empirische studie die de verbetering in het voorspellend vermogen evalueerde van toegevoegde niet-genetische factoren aan klinische risico modellen middels  $\Delta$ AUC, NRI en IDI), rapporteerden de onderzoekers een correcte definitie van de AUC, voor ongeveer de helft van de artikelen was dat zo voor de NRI en weinig auteurs definieerden de IDI correct. In de helft van de artikelen gaven de auteurs correct aan hoe de AUC werd verkregen. In nog minder artikelen beschreven de auteurs de rekenmethodes voor de NRI en IDI, en wanneer een beschrijving werd gegeven, dan was meer dan de helft incorrect. Vergelijkbaar met het onderzoek, dat in Hoofdstuk 5 werd gepresenteerd, werden de waarden van de maten geïnterpreteerd als wijzend op verbetering wanneer ze statistisch significant waren, los van de absolute waarden van de maten. De onderzoeken van **Hoofdstuk 5 en 6** laten beide zien, dat er ruimte is voor verbetering bij onderzoekers die de verschillende maten voor de beoordeling van het voorspellend vermogen van predictiemodellen interpreteren, omdat zij hun interpretaties vaak enkel baseren op de statistische

significantie. Er is dus een beter begrip nodig van de maten om tot een meer betekenisvolle interpretatie van predictiestudies te komen.

In **Hoofdstuk 7** bespreken we twee brieven aan de “redactie” over twee verschillende onderwerpen, namelijk de externe validatie van predictiemodellen en de beoordeling van kalibratie en discriminerend vermogen. Het eerste voorbeeld is een brief aan de redactie in antwoord op een artikel, waarin gevonden werd, dat slechts 25% van de risicopredictiemodellen extern gevalideerd werden. De auteurs gebruikten een tamelijk ongebruikelijke definitie van externe validatie. Op basis van een heranalyse van hun data concluderen wij, dat het percentage van externe validatie tot wel 83% hoog kan zijn, omdat veel modellen al extern gevalideerd waren en veel andere het niet waard waren. We wijzen erop dat externe validatie alleen nodig is wanneer de predictiemodellen het waard zijn. Dat laatste wordt, bijvoorbeeld, bepaald door de verwachte verbetering van het huidige model of de huidige praktijk, de wenselijkheid van het model voor het publiek, het beoogde gebruik en de verwachte gezondheidswinst. Het tweede voorbeeld is een brief aan de redactie in antwoord op een artikel, waarin het voorspellende vermogen van een PRS voor de voorspelling van prostaatanker werd onderzocht. Zowel over de beoordeling van de kalibratie als over de discriminatie werd niets gerapporteerd in het artikel. Met behulp van een gevalideerde simulatiemethode en de data die in het artikel gepresenteerd werden, ontdekten we dat de AUC van de PRS lager zou zijn dan in andere bestaande modellen.

## Slotopmerkingen

Het beoogde gebruik van predictiemodellen zou een centrale rol in het ontwerp en de interpretatie van predictiestudies moeten hebben. Aangezien het voorspellende vermogen van predictiemodellen varieert tussen populaties en settings, zou het predictieonderzoek moeten worden uitgevoerd met de gezondheidszorg setting waarop het gericht is in het achterhoofd en beweringen over de geschiktheid van PRSs om in de klinische zorg te worden geïmplementeerd zouden moeten worden ondersteund met bewijs van goed gekalibreerde modellen en een verbeterd discriminerend vermogen van het model vergeleken met modellen die momenteel in gebruik zijn. Voor de beoordeling van discriminatie hebben we laten zien, dat de AUC de scheiding

is tussen de risicoverdelingen van patiënten en niet-patiënten. Voor de evaluatie van alle maten geldt, dat de interpretatie zich niet alleen zou moeten baseren op statistische significantie, maar op hun waarde in relatie tot het beoogde gebruik. Het veld van predictieonderzoek zou kunnen worden verbeterd door het beoogde gebruik als leidraad te gebruiken en door de predictiematen meer intuïtief te verklaren, zodat meer onderzoekers er een beter begrip van kunnen krijgen. Of het tijd is om de implementatie van PRS in de gezondheidszorg te overwegen hangt niet alleen af van het voorspellende vermogen van de predictiemodellen, maar bewijs van voldoende voorspellend vermogen is een belangrijke stap voordat kan worden voortgegaan naar verdere studies over bruikbaarheid, nut en zinvolheid van de PRS in de gezondheidszorg.





**Dankwoord**

**Publications**

**About the author**



## Dankwoord

Een lange, intensieve en leerzame periode is nu afgesloten. Ik ben dankbaar voor iedereen die hieraan bijgedragen heeft!

Veel dank ben ik verschuldigd aan Cecile Janssens en Martina Cornel. Beide mijn promotoren, beide eigen en geheel op hun eigen manier betrokken. Cecile vanuit Atlanta, Martina vanuit Amsterdam. Cecile, je scherpzame, enthousiasme en precisie waren erg leerzaam. Je bent zeer gastvrij geweest in Atlanta, wat heeft gezorgd voor een onvergetelijke tijd. Dankjewel voor je begeleiding gedurende deze jaren! Martina, jouw snelheid, lichtheid en oplossingsgerichtheid zijn bewonderingswaardig en ik ben blij dat ik dat uiteindelijk toch nog wat intensiever heb mogen meemaken. Dank voor alle steun ook. Dan mijn copromotor, Ilse, dankjewel voor de gezelligheid in Atlanta en nuttige gesprekken daarna.

Beste promotiecommissie, ik heb het als een eer ervaren dat jullie mijn proefschrift hebben beoordeeld. Heel veel dank voor jullie tijd. Beste Hans Meij, we hebben elkaar leren kennen op de poli, gedurende een roerige en interessante tijd. Ontzettend leuk dat je hebt willen opponeren. Hartelijk dank daarvoor!

Ik wil graag al mijn oud-collega's Community Genetics bedanken. In het bijzonder Lidewij, dankjewel dat je voorzitter van de promotiecommissie wilde zijn. Dat heeft de verbinding met de afdeling mooi rond gemaakt. Ik wil je ook bedanken voor het aanbod van koffietjes drinken, dat heeft me zeker geholpen in het doorzetten. Jouw enorme gedrevenheid inspireert mij. Ook wil ik Carla, Tessel en Anke bedanken. Jullie waren fijne collega's en hebben mij op de juiste manier aangemoedigd. Dank daarvoor.

Lieve Ivy en Karuna, zonder jullie eindeloze aanmoediging was dit niet gelukt. Jullie zijn echt enorm lief geweest! Ik hoop dat we onze etentjes een traditie kunnen maken en dat we binnenkort de fles ook voor jullie kunnen opentrekken!

Lieve oud-collega's van de poli, ik wil jullie bedanken voor de gezelligheid en leerzame tijd van hoe het nu echt op de werkvloer werkt. Dat was een fijne afwisseling naast het schrijven van mijn proefschrift.

Lieve Esther, lieve Helga, mijn paranimfen, ik ben dankbaar dat jullie naast mij hebben gestaan tijdens mijn promotie. Jullie zijn beide op jullie eigen manier inspirerend en heerlijk om tijd mee door te brengen. Dank jullie wel voor jullie vriendschap.

Lieve andere vrienden, uit Nederland, Zwitserland, United States, Ghana, dank voor jullie eindeloze interesse in mij en mijn proefschrift. Ik ben blij dat jullie in mijn leven zijn. Dank voor jullie vriendschap.

Lieve familie van der Wiel, bedankt voor jullie interesse en steun. Ik ben blij dat ik jullie ken!

Lieve familie Martens en lieve familie Aalhuizen, in Nederland, Verenigde Staten en Canada. Ik ben zo trots dat jullie mijn familie zijn! Dank voor jullie steun, ieder op z'n eigen manier.

Lieve pap en mam, zoveel dank voor al jullie steun en liefde. Zonder jullie was ik niet wie ik ben.

Liefste Sander, hoe vaak jij al niet hebt gehoord "het is bijna klaar" heb ik niet kunnen bijhouden. Maar, het is nu echt klaar! Ik ben je dankbaar voor al je steun, je liefde, je wijsheid, je kracht. Wat een prachtig mens ben je, ik ben eindeloos blij dat jij er bent. Je kent me nog niet 'zonder proefschrift', maar nu heb ik eindelijk een proefschrift in de kast staan en hoef ik het niet elke dag bij me te dragen. Ik kijk uit naar de avonturen die we samen zullen delen, leven en beleven!

## Publications

**Martens FK**, Kers JG, Janssens ACJW. External validation is only needed when prediction models are worth it (Letter). *Journal of clinical epidemiology*. 2016;69:249-250.

**Martens FK**, Kers JG, Janssens ACJW. Risk Analysis of prostate Cancer in PRACTICAL Consortium (Letter). *Cancer epidemiology, biomarkers & prevention*. 2016;25(1):222.

**Martens FK**, Tonk EC, Kers JG, Janssens ACJW. Small improvement in the area under the receiver operating characteristic curve indicated small changes in predicted risks. *Journal of clinical epidemiology*. 2016;79:159-164.

**Martens FK**, Tonk ECM, Janssens ACJW. Evaluation of polygenic risk models using multiple performance measures: a critical assessment of discordant results. *Genetics in medicine*. 2019;21(2):391-397.

**Martens FK**, Janssens ACJW. How the intended use of polygenic risk scores guides the design and evaluation of prediction studies. *Current Epidemiology Reports*. 2019;6(2):184-190.

Janssens ACJW, **Martens FK**. Reflection on modern methods: Revisiting the area under the ROC Curve. *International Journal of Epidemiology*. 2020;49(4):1397-1403.

**Martens FK**, Tonk ECM, Janssens ACJW. Simultaneous use of AUC, NRI and IDI for the evaluation of clinical prediction models: reporting and interpretation practices.  
*Prepared for submission*

### OTHER PUBLICATIONS

Janssens ACJW, **Martens FK**. Prediction research: an introduction. *Published online*. 2018

**Martens FK**, Huntjens DW, Rigter T, Bartels M, Bet PM, Cornel MC. DPD testing before treatment with fluoropyrimidines in the Amsterdam UMCs: an evaluation of current pharmacogenetic practice. *Frontiers in Pharmacology*. 2020;10:1609.

## About the author

Forike Kirsten Martens was born in Nueneen, the Netherlands in 1987. After she graduated from the Waldorf high school in Zeist, she studied English language and Culture at Leiden University and violin at the Royal Conservatory in The Hague. In the first year of her studies, she found out that her passion for science and interest in 'health' and 'disease' could not be ignored. So, after traveling the USA for a while, in 2007 she started her bachelor in Health and Life Sciences at the VU University in Amsterdam. During her time here, she conducted research internships on the effectiveness of collaborative care in the treatment of depression at Trimbos Institute Utrecht and on the functioning of ethical review boards in social sciences at the Royal Tropical Institute Amsterdam. After her bachelor's she went on to pursue her master's degree in Management, Policy Analysis and Entrepreneurship in Health & Life Sciences, specializing in International Public Health, from which she graduated in 2013. During this time, she conducted an internship on the effectiveness of a walk-in center for minor ailments at TNO Innovation for Life and wrote her thesis on 'contact interventions' at the Athena Institute. After traveling the USA some more, in April 2014 Forike started as a junior researcher at the Amsterdam UMC, Section of Community Genetics & Public Health Genomics and Amsterdam Public Health Research Institute, and was a Visiting Scholar at Emory University, Rollins School of Public Health, Epidemiology, Atlanta, USA (Oct 2014 – March 2016). From April 2016 on she proceeded her research in her own time while having several other jobs, including training hospital personnel, supporting the administrative office of the outpatients' clinic 'Clinical Genetics', and conducting a scientific study on the uptake- and stakeholder experiences of DPD testing before treatment with fluoropyrimidines at Amsterdam UMC and the Amsterdam Public Health Research Institute. Currently, Forike is project manager for the regional care network antibiotic resistance at UMC Utrecht and currently lives alternately in Tirana, Albania and Amsterdam.





