Evaluating generalisability and clinical utility of risk prediction model developed from routinely collected electronic health records and longitudinal cohort using Cardiovascular disease as exemplar

A dissertation submitted to The University of Manchester for the degree of
Doctor of Philosophy
in the faculty of biology, medicine and health

2020

Yan Li

School of Health Sciences

Division of Informatics, Imaging and Data Science

List of tables

List of figures

Abstract

Risk prediction models are mathematical formulas which use disease as outcome variable and individual's characteristics or risk factors as predictors to predict a risk of the individual having the disease in future. They are used in health care system to assist clinicians to make treatment decisions for patients. Healthcare guideline such as NICE recommends prescribing statins to patients who have QRISK3 predicted risk above 10%. These models are developed from routinely collected electronic health records or longitudinal cohort. However, they are validated on population level but being used on individual level for individual patients. Generalisability and clinical utility reflect whether a model developed in one setting could be generalised to other setting and still being clinical useful. Current statistical framework for the model development and validation does not consider reporting or minimally assessing the generalisability and clinical utility of risk prediction models especially on individual level. The objective of this PhD is to assess the generalisability and clinical utility of risk prediction models in different settings especially on accurately predicting high risk patients who are missed by the model, with Cardiovascular disease risk prediction as exemplar.

There are 6 main chapters in this PhD. Chapter 2 evaluated generalisability of QRISK3 by assessing the effects of practice variability on individual risk prediction. Chapter 3 assessed the effects of data quality and variation of association between disease outcome and predictor on the risk predictions of individual patients. Chapter 4 assessed clinical utility of machine learning models and Cox models on both population and individual level. Chapter 5 implemented ClinRisk's QRISK3 algorithm into R. Chapter 6 assessed whether a new individual level measurement may improve clinical utility of risk prediction model. Chapter 7 discussed all the identified generalisability and clinical utility issues for current models and possible solutions.

This PhD found that risk prediction models may have good performance on population level but with limited generalisability and clinical utility especially on individual level. The reason is that prediction models based on different techniques or modelling decisions can yield inconsistent individual results. Risk prediction models should be used in conjunction with additional clinical tests and clinical judgement.

Declaration

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright Statement

i. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the “Copyright”) and he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

ii. Copies of this thesis, either in full or in extracts and whether hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issues under it or, where appropriate, in accordance with licensing agreement which the University has from time to time. This page must form part of any such copies made.

iii. The ownership of certain Copyright, patents, designs, trademarks and other intellectual property (the “Intellectual Property”) and any reproductions of copyright works in the thesis, for example graphs and tables (“Reproductions”), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

iv. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=24420), in any relevant Thesis restriction declarations deposited in the University Library, The University Library’s regulations (see http://www.library.manchester.ac.uk/about/regulations/) and in The University’s policy on Presentation of Theses.

Acknowledgements

I am extremely thankful and appreciated to the help from my supervisors Prof. Tjeerd van Staa and Dr. Matthew Sperrin. Not only they provided me with full professional academic support but also guidance for life. Prof. Tjeerd guided me to understand how to do research and write academic papers, and Dr. Mathew Sperrin helped me to learn advanced statistics and modelling. These are extreme useful for my whole life, which I cannot thank you enough.

Without help and support from colleagues, this PhD is impossible. Here, I wish to thank Prof. Darren Ashcroft to provide support especially on interpretation of statistical results for this PhD. Thank Dr. Alexander Pate for sharing his experience using CPRD database. Also, Thank Dr. Glen Martin for his useful life advice, reviewing literature review and Chapter 3 of this PhD and sharing his structure of PhD thesis to me.

I also wish to thank my close family. To my grandmother JiaZhu Liu who encouraged me to pursuit academic career. Though She passes away three years ago, her love and wish helped me through the most difficult time of this PhD. To my mother Shu Chen who provided me life support and guidance for last three years. To my father ShaoLong Li who also helped me during the PhD.

At last, I wish to thank all the students, staff and colleagues from University of Manchester and other UK universities. Working with you opened my eyes and helps me grow. I also wish to thank my funder China Scholarship Council (CSC), without your financial support this PhD is not possible.

Thanks to you all,

Yan Li

About the author

Degree and work experience

2016 - 2017 MSc health data science (Merit), University of Manchester

2014 - 2016 Work as statistical analyst in clinical trails

2009 - 2013 BSc Mathematics, University of Sichuan

Research interests

1. The generalisability and clinical usefulness of traditional statistical risk prediction model and how could we improve its performance on individual level. Mainly with Cardiovascular disease as exemplar.
2. The clinical implementation of machine learning (AI) models for risk prediction modelling and how could we take advantage of the strength of machine learning models and walk around from their disadvantages.
3. The drivers of antibiotics over-prescription including what are the main drivers of antibiotic prescription, how could we best decrease the antibiotic prescription and any other impact of these interventions of antibiotic prescribing.

Publications

1. Li Y, Sperrin M, Belmonte M, Pate A, Ashcroft DM, van Staa TP. Do population-level risk prediction models that use routinely collected health data reliably predict individual risks? *Sci Rep*. 2019;9(1):11222. doi:10.1038/s41598-019-47712-5
2. Li, Y., Sperrin, M., Martin, G. P., Ashcroft, D. M. & van Staa, T. P. Examining the impact of data quality and completeness of electronic health records on predictions of patients' risks of cardiovascular disease. *Int. J. Med. Inform.* 104033 (2019). doi:10.1016/j.ijmedinf.2019.104033
3. Li, Y., Sperrin, M., Ashcroft, D. M. & van Staa, T. P. Does machine-learning improve the accuracy of the predictions of clinical risks? An exemplar around cardiovascular risk prediction. *submitted*
4. Li Y, Mölter A, White A, et al. Relationship between prescribing of antibiotics and other medicines in primary care: a cross-sectional study. *Br J Gen Pract.* 2019;69(678):e42-e51. doi:10.3399/bjgp18X700457

5. Li, Y., Sperrin, M. & van Staa, T. R package "QRISK3": an unofficial research purposed implementation of ClinRisk's QRISK3 algorithm into R. *F1000Research* **8,** 2139 (2019).

6. Li, Y., Sperrin, M., Ashcroft, D. M. & van Staa, T. P. The Instability of Machine Learning and Statistical Models in Predicting Individual Patient Risks: An Approach to Improve the Clinical Utility of These Models. ***Ready to submit***

7. Li, Y., Sperrin, M., Ashcroft, D. M. & van Staa, T. P. Clinical Risk Prediction Model Using Routinely Collected Electronic Health Records or Longitudinal Cohort in Daily Practice: Are They Robust Enough for Clinical Decision Making? ***Ready to submit***

8. van Staa TP, Palin V, Li Y, et al. The effectiveness of frequent antibiotic use in reducing the risk of infection-related hospital admissions: results from two large population-based cohorts. *BMC Med*. 2020;18(1):40. doi:10.1186/s12916-020-1504-5

Chapter 1 General introduction

1.1 Introduction

Risk prediction models are mathematical formulas which use disease as outcome variable and individual's characteristics or risk factors as predictors to predict a risk of the individual having the disease in future. Clinical risk prediction models are risk prediction models used for clinical purpose [1]. Models are developed and validated on population level and derived from patients' records such as electronic health records (EHR) [2]. Models are validated on population level by the discrimination which measures the model's ability to discriminate high and low risk patients and calibration which measures the agreement between observed and predicted events. Risk prediction models could be derived from statistical models such as logistic model [3] for binary outcome and Cox proportional hazard model [4] for survival outcome or from machine learning models such as random forest [5] or neural network [6]. Clinical risk prediction models are developed to predict patients' risk for long-term chronic disease, such as Cardiovascular disease (CVD). Models are used in clinical practice to assist the clinical decision making. For example, QRISK [7] is recommended by NICE guideline [8] in CVD prevention, i.e. prescribing statins to patients who have a QRISK predicted CVD risk above 10%. Risk prediction models were also developed for other disease such as breast cancer [1], chronic obstructive pulmonary disease and diabetes [9], other healthcare outcome such as death/readmission, number of hospital visits or length of stay [9]. Besides of disease outcome and healthcare outcome, risk prediction models could predict risk of underlying disease to help decide whether further testing is needed or estimate short-term risk for surgery [1]. For medical research, prediction models could help in different ways, such as selecting patients for the study or adjusting covariates at baseline for randomised controlled trials [1].

CVD has been the top cause of death in the US, UK, Europe and China in decades and it receives considerable attention from government, healthcare institutions and researchers [10]. In 2017, there are 80 thousand deaths in the US caused

by CVD, and one in three of them is due to CVD [11]. In Europe, CVD causes 3.9 million deaths every year [12]. CVD has become the most common cause of deaths in UK since 2001[10]. In 2014, China has 29.6 million CVD caused deaths in rural areas and 26.2 million in urban areas [13]. Studies have suggested that evaluating long-term CVD risk of patients is required, and identifying higher risk patients is the first step [14] [15] [16] [17]. Several CVD risk prediction models are developed to identify high risk CVD patients such as Framingham from US [18], QRISK from UK [7] and ESC score from Europe [14].

However, these models are only statistical validated on population level with discrimination and calibration but being used on individual level for individuals and often being applied to a setting which could be much different from the development setting. Literature shows that three statistical validated models including QRISK2, Framingham and Assign score have much disagreement to predict high risk patients [19]. A systematic literature review of all the existing risk prediction models concludes that the clinical usefulness of the most existing predicting models are unclear because of "shortage of methodology, inadequate presentation, not enough external validation and unclear impact of model" [17]. The reviewer recommends rather than developing new risk prediction models, research should focus on further validate and improve existing risk models considering new clinical setting and environment of population [17].

Generalisability and clinical utility are very important aspects of model validation for risk prediction model. A model developed from one setting was supposed to be generalisable to other setting and the estimated individual risk (probability) should be robust enough for clinical decision making. Statistically, generalisability of risk prediction model is refer to external validation which validates model's performance by discrimination and calibration in a different setting [1]. While clinically, models are used for individual patients rather than populations, generalisability in this PhD refers to the robustness of model developed in one setting and being used in other setting considering both population and individual level. For

example, whether a model developed from a setting assumed homogenous of sites predicts consistent risk to the same patients in a new setting where there is huge site-heterogeneity. Whether models with similar population level model performance could predict accurate and consistent risk for the same patients. Statistically, clinical usefulness of risk prediction model is refer to decision curve analysis with net benefit [20] on population level. Net benefit measures the number of patients could be correctly identified as events by the model without adding any false positive given different thresholds, a statistical defined clinical useful model would have a high positive value of net benefit [20]. While clinically, models are used by their predicted individual risk with a selected threshold to help decide patients' treatment, clinical utility of risk prediction model in this PhD refers to whether an accurate and consistent individual risk could be predicted by current well performed models (similar high calibration and discrimination on population level), and whether these risks are robust enough to assist clinical decision making for individual patients. Table 1.1 summarises key terms in this PhD.

Overall**,** current statistical framework for the model development and validation [21] (Table 1.2) does not consider to report or minimally assessing the generalisability and clinical utility of risk models especially on individual level and it is unclear whether a model developed in one setting could be generalised to other setting and still being clinical useful. The overall objective of this PhD is to assess the generalisability and clinical utility of risk prediction models in different settings especially on accurately predicting high risk patients who are missed by the model, with CVD risk prediction as exemplar.

Table 1.1: Summary of key terms in this PhD

Key terms	Definition in statistics	Definition in this PhD	How are they related
Assessing model on population level and individual level	Statistically, model was mainly assessed by population level measurement including discrimination (the ability of model to discriminate high/low risk patients) and calibration (the agreement between predicted and observed risk).	This PhD assessed model on both population level and individual level. The consistency of individual risk prediction among models with similar population level model performance was used to assess robustness of model on individual level.	Models with good calibration and discrimination (i.e. statistical validated model) needs to be further assessed on individual level, as they were ultimately used in clinical practice to help clinicians to make treatment decisions for individual patients.
Generalisability	Statistically, generalisability of risk prediction model is refer to external validation which validates model's performance by discrimination and calibration in a different setting [1]	Generalisability in this PhD refers to the robustness of model developed in one setting and being used in other setting considering both population and individual level.	A model developed from one setting was supposed to be generalisable to other setting and the estimated individual risk (probability) should be robust enough for clinical decision making.
Clinical utility	Statistically, clinical usefulness of risk prediction model is refer to decision curve analysis with net benefit [20] on population level.	Clinical utility of risk prediction model in this PhD refers to whether an accurate and consistent individual risk could be predicted by well performed population level models, and whether these risks are robust enough to assist clinical decision making for individual patients.	Clinically, models are used by their predicted individual risk with a selected threshold to help patients' treatment decision, therefore clinical usefulness of models on population level does not guarantee their clinical utility on individual level.

Table 1.2 The TRIPOD guideline [21] to report developing and validating risk prediction model for diagnosis or prognosis

Table 1. Checklist of Items to Include When Reporting a Study Developing or Validating a Multivariable Prediction Model for Diagnosis or Prognosis*

Section/Topic	Item	Development or Validation?	Checklist Item	Page
Title and abstract				
Title	1	D;V	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted	
Abstract	2	D;V	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions	
Introduction				
Background and objectives	3a	D;V	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models	
	3b	D;V	Specify the objectives, including whether the study describes the development or validation of the model or both	
Methods				
Source of data	4a	D;V	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable	
	4b	D;V	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up	
Participants	5a	D;V	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres	
	5b	D;V	Describe eligibility criteria for participants	
	5c	D;V	Give details of treatments received, if relevant	
Outcome	6a	D;V	Clearly define the outcome that is predicted by the prediction model, including how and when assessed	
	6b	D;V	Report any actions to blind assessment of the outcome to be predicted	
Predictors	7a	D;V	Clearly define all predictors used in developing the multivariable prediction model, including how and when they were measured	
	7b	D;V	Report any actions to blind assessment of predictors for the outcome and other predictors	
Sample size	8	D;V	Explain how the study size was arrived at	
Missing data	9	D;V	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method	
Statistical analysis methods	10a	D	Describe how predictors were handled in the analyses	
	10b	D	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation	
	10c	V	For validation, describe how the predictions were calculated	
	10d	D;V	Specify all measures used to assess model performance and, if relevant, to compare multiple models	
	10e	V	Describe any model updating (e.g., recalibration) arising from the validation, if done	
Risk groups	11	D;V	Provide details on how risk groups were created, if done	
Development vs. validation	12	V	For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors	
Results				
Participants	13a	D;V	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful	
	13b	D;V	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome	
	13c	V	For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors and outcome)	
Model development	14a	D	Specify the number of participants and outcome events in each analysis	
	14b	D	If done, report the unadjusted association between each candidate predictor and outcome	
Model specification	15a	D	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point)	
	15b	D	Explain how to use the prediction model	
Model performance	16	D;V	Report performance measures (with CIs) for the prediction model	
Model updating	17	V	If done, report the results from any model updating (i.e., model specification, model performance)	
Discussion				
Limitations	18	D;V	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data)	
Interpretation	19a	V	For validation, discuss the results with reference to performance in the development data, and any other validation data	
	19b	D;V	Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence	
Implications	20	D;V	Discuss the potential clinical use of the model and implications for future research	
Other information				
Supplementary information	21	D;V	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets	
Funding	22	D;V	Give the source of funding and the role of the funders for the present study	

* Items relevant only to the development of a prediction model are denoted by *D*, items relating solely to a validation of a prediction model are denoted by *V*, and items relating to both are denoted *D;V*. We recommend using the TRIPOD Checklist in conjunction with the TRIPOD Explanation and Elaboration document.

The first specific research question is risk prediction models such as QRISK3 are developed using EHR collected from different sites (practices) while this variability

of sites (practice variability) is not considered in the model, whether practice variability has effects on individual risk prediction. The motivation is successful risk prediction models like QRISK3 is not only used in a setting which is plausible related to [1] their development settings (e.g. GP practices in UK) but also could be used in very different settings (e.g. China) which may have larger practice variability comparing to UK. It is important to assess whether models like QRISK3 could be generalised to such a heterogeneous setting before clinical implementation. The second research question (based on the conclusion of the first one) is whether data quality and variation of association between predictors and outcome (two aspects of practice variability) are related to the effects of practice variability on individual risk prediction. The motivation is if they are related to the effects of practice variability then generalisability and clinical utility of model could be improved by improving data quality or incorporating variation of association between predictors and outcome, otherwise new predictors are needed. The third research question is whether novel approach machine learning models (where literatures claim that they may start revolutionary in healthcare system [22]) outperform traditional statistical model and provide more robust individual risk prediction. The motivation is if machine learning models are indeed superior in risk prediction task then they should be further assessed in future study rather than the current risk prediction model, otherwise generalisability and clinical utility of both machine learning and traditional statistical models should be further studied. The fourth research question is whether a new approach which considers a different individual level measurement may improve the clinical utility of both machine learning and traditional models given findings from previous research that all these models have much uncertainty on individual level risk. The motivation is if the current model development framework on population level has its underlying challenge on individual level (e.g. probability itself is defined on population level), new individual level measurement may be considered to strengthen the clinical utility of models on individual level.

Table 1.3: Summary of research output in this PhD

Chapter	Paper	Status	Objective	Doi
2	Do population-level risk prediction models that use routinely collected health data reliably predict individual risks?	Published	Assess the effects of practice variability on individual risk prediction using QRISK3 as an exemplar.	https://doi.org/10.1038/s41598-019-47712-5
3	Examining the impact of data quality and completeness of electronic health records on predictions of patients' risks of cardiovascular disease	Published	Assess the effects of data quality and variation of association between disease outcome and predictor (two aspects of practice variability) on the risk predictions of individual patients.	https://doi.org/10.1016/j.ijmedinf.2019.104033
4	Does machine learning improve the accuracy of clinical risk predictions? An exemplar examining risk of cardiovascular disease	Submitted	Assess model performance and robustness of individual risk prediction of machine learning models and Cox models on both population and individual level with CVD risk prediction as exemplar	TBD
5	R package "QRISK3": an unofficial research purposed implementation of ClinRisk's QRISK3 algorithm into R	Open peer-review	Implement ClinRisk's QRISK3 algorithm into R	https://doi.org/10.12688/f1000research.21679.2
6	The instability of machine learning and statistical models in predicting individual patient risks: an approach to improve the clinical utility of these models	Ready to Submit	Assess whether a new individual level measurement may improve clinical utility of risk prediction model	TBD
7	Clinical risk prediction model using routinely collected electronic health records or longitudinal cohort in daily practice: Are they robust enough for clinical decision making?	Ready to Submit	Analysis paper to discuss whether models developed from current guideline provide robust individual risk prediction for clinical decision making	TBD

1.2 Contents of this PhD

This PhD consists of 6 papers (Table 1.3). Chapter 2 (the 1st paper) assessed the effects of practice variability on individual risk prediction using QRISK3 as an exemplar. QRISK3 is developed from an integrated EHR cohort of UK population, while the variability of practices including coding variation, missing value and underlying heterogeneity are not considered in the model. The effects of practice variability on risk prediction must be assessed before considering generalise QRISK to a different population say Chinese population, as other population like Chinese population might have larger practice variability than UK population due to large differences of healthcare facilities from different provinces. Chapter 3 (the 2nd paper) assessed the effects of data quality and variation of association between disease outcome and predictor (two aspects of practice variability) on the risk predictions of individual patients. This aims to investigate on what aspects of practice variability contribute to the models' uncertainty on individual risk prediction. Chapter 4 (the 3rd paper) assessed model performance and robustness of individual risk prediction of machine learning models and Cox models on both population and individual level with CVD risk prediction as exemplar, as there is a hype around machine learning models that they may start revolutionary in health care [22]. This paper aims to evaluate whether innovative methods like machine learning models are truly superior than the traditional models as claimed [23], and whether machine learning models have more certainty on individual risk prediction. Chapter 5 (the 4th paper) implemented ClinRisk's QRISK3 algorithm into R, as there is a gap that QRISK3 model was written in a low level programming language C while model development needs high level programming language R. Providing a popular applied clinical risk prediction model (QRISK3) in an easy accessible way would help research community to better understand and improve generalisability and clinical utility of risk prediction model. Chapter 6 (The 5th paper) assessed whether a new individual level measurement may improve clinical utility of risk prediction model. This implies how the current individual risk prediction could be best used for individual patients with an additional

new individual measurement and provide a direction to develop new type of model based on this new individual level measurement in future. Chapter 7 (The 6th paper) discussed all the identified key challenges and possible solutions for models developed from current model guideline.

1.3 Author Contributions

Specific author contributions are mentioned in each individual chapter. In general, for all projects:

Yan Li as the PhD candidate: Designed the study; designed and conducted all statistical analysis; produced all tables and figures; wrote the main manuscript text.

Matthew Sperrin: Supervised the study; improved study design; improved statistical method; improved interpretation of statistical results; reviewed statistical results; reviewed and edited the main manuscript text.

Darren M Ashcroft: Improved the major interpretation of statistical results and discussion; reviewed and edited paper;

Tjeerd Pieter van Staa: Designed and supervised the study; Quality control of all aspects of the paper; wrote the main manuscript text;

1.4 References

1. Steyerberg EW. *Clinical Prediction Models*. New York, NY: Springer New York; 2009. doi:10.1007/978-0-387-77244-8

2. Clinical Practice Research Datalink - CPRD. https://www.cprd.com/intro.asp. Accessed August 20, 2017.

3. Sperandei S. Understanding logistic regression analysis. *Biochem medica*. 2014;24(1):12-18. doi:10.11613/BM.2014.003

4. Cox DR. Regression Models and Life-Tables Authors (s): D . R . Cox Source : Journal of the Royal Statistical Society . Series B (Methodological), Vol . 34 , No . 2 Published by : Wiley for the Royal Statistical Society Stable URL : http://www.jstor.org/stable. *J R Stat Soc Ser B*. 1972;34(2):187-220. doi:10.2307/2985181

5. Breiman L. Random Forests. *Mach Learn*. 2001;45(1):5-32. doi:10.1023/A:1010933404324

6. Hassoun MH, H. M. *Fundamentals of Artificial Neural Networks*. MIT Press; 1995. https://dl.acm.org/citation.cfm?id=526717. Accessed September 7, 2019.

7. Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. *BMJ*. 2017;357(3):j2099. doi:https://doi.org/10.1136/bmj.j2099

8. *National Clinical Guideline Centre Lipid Modification Cardiovascular Risk Assessment and the Modification of Blood Lipids for the Primary and Secondary Prevention of Cardiovascular Disease Clinical Guideline Methods, Evidence and Recommendations Lipid Modification Contents.*; 2014. https://www.nice.org.uk/guidance/cg181. Accessed August 28, 2019.

9. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JPA, Goldstein B. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review Correspondence to. doi:10.1093/jamia/ocw042

10. Phe. Action plan for cardiovascular prevention: 2017 to 2018. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/648190/cardiovascular_disease_prevention_action_plan_2017_to_2018.pdf. Accessed December 6, 2017.

11. Heart Disease and Stroke Statistics 2017 At-a-Glance. https://healthmetrics.heart.org/wp-content/uploads/2017/06/Heart-Disease-and-Stroke-Statistics-2017-ucm_491265.pdf. Accessed December 6, 2017.

12. CVD Statistics. http://www.ehnheart.org/cvd-statistics.html. Accessed December 6, 2017.

13. Chen W-W, Gao R-L, Liu L-S, et al. China cardiovascular diseases report 2015: a summary. *J Geriatr Cardiol*. 2017;14(1):1-10. doi:10.11909/j.issn.1671-5411.2017.01.012

14. Piepoli MF, Hoes AW, Agewall S, et al. 2016 European Guidelines on cardiovascular disease prevention in clinical practice. *Eur Heart J*. 2016;37(29):2315-2381. doi:10.1093/eurheartj/ehw106

15. Cardiovascular disease prevention overview - NICE Pathways. https://pathways.nice.org.uk/pathways/cardiovascular-disease-prevention. Accessed December 6, 2017.

16. Guidelines P, Risk C. Prevention of Cardiovascular Disease Prevention of Cardiovascular Disease. *World Heal Organ*. 2007:1-30. doi:10.1093/innovait/inr119

17. Damen JAAG, Hooft L, Schuit E, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. *BMJ*. 2016;353:i2416. doi:10.1136/BMJ.I2416

18. Bitton A, Gaziano TA. The Framingham Heart Study's impact on global risk assessment. *Prog Cardiovasc Dis*. 2010;53(1):68-78. doi:10.1016/j.pcad.2010.04.001

19. van Staa T-P, Gulliford M, Ng ES-W, Goldacre B, Smeeth L. Prediction of cardiovascular risk using Framingham, ASSIGN and QRISK2: how well do they predict individual rather than population risk? *PLoS One*. 2014;9(10):e106455. doi:10.1371/journal.pone.0106455

20. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. doi:10.1177/0272989X06295361

21. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD Statement. *Eur Urol*. 2015;67(6):1142-1151. doi:10.1016/j.eururo.2014.11.025

22. Hinton G. Deep Learning—A Technology With the Potential to Transform Health Care. *JAMA*. 2018;320(11):1101. doi:10.1001/jama.2018.11100

23. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? Liu B, ed. *PLoS One*. 2017;12(4):e0174944. doi:10.1371/journal.pone.0174944

Blank page

Chapter 2 Do population-level risk prediction models that use routinely collected health data reliably predict individual risks?

Yan Li[1], Matthew Sperrin[1], Miguel Belmonte[1], Alexander Pate[1], Darren M Ashcroft[2,3], Tjeerd Pieter van Staa[1,4,5]

[1]Health e-Research Centre, Farr Institute, School of Health Sciences, Faculty of Biology, Medicine and Health, the University of Manchester, Manchester Academic Health Sciences Centre (MAHSC), Oxford Road, Manchester, M13 9PL, UK
[2]Centre for Pharmacoepidemiology and Drug Safety, School of Health Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Oxford Road, Manchester, M13 9PL, UK
[3]NIHR Greater Manchester Patient Safety Translational Research Centre, School of Health Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Oxford Road, Manchester, M13 9PL, UK
[4]Utrecht Institute for Pharmaceutical Sciences, Utrecht University, Utrecht, Netherlands
[5]Alan Turing Institute, Headquartered at the British Library, London, UK

Corresponding author: Tjeerd van Staa, tjeerd.vanstaa@manchester.ac.uk

Journal title: Nature: Scientific Reports
Doi: https://doi.org/10.1038/s41598-019-47712-5

Word count: 3109
Abstract: 198
Number of tables: 2
Number of figures: 3

2.1 Abstract

The objective of this study was to assess the reliability of individual risk predictions based on routinely collected data considering the heterogeneity between clinical sites in data and populations. Cardiovascular disease (CVD) risk prediction with QRISK3 was used as exemplar. The study included 3.6 million patients in 392 sites from the Clinical Practice Research Datalink. Cox models with QRISK3 predictors and a frailty (random effect) term for each site were used to incorporate unmeasured site variability. There was considerable variation in data recording between general practices (missingness of body mass index ranged from 18.7 to 60.1%). Incidence rates varied considerably between practices (from 0.4 to 1.3 CVD events per 100 patient-years). Individual CVD risk predictions with the random effect model were inconsistent with the QRISK3 predictions. For patients with QRISK3 predicted risk of 10%, the 95% range of predicted risks were between 7.2 and 13.7% with the random effects model. Random variability only explained a small part of this. The random effects model was equivalent to QRISK3 for discrimination and calibration. Risk prediction models based on routinely collected health data perform well for populations but with great uncertainty for individuals. Clinicians and patients need to understand this uncertainty.

Key words

EHR; QRISK; practice variability; frailty model; CVD risk prediction

2.2 Introduction:

Cardiovascular disease (CVD) was the primary cause of death in USA, Europe and China in 2017[1]. Multiple studies have suggested that the identification of patients with high CVD risk is important in its prevention[2,3,4,5]. Risk prediction models are often used to predict CVD risk for individual patients[5]. Examples are the Framingham risk score (FRS) and QRISK which provide risks of developing CVD in the next 10 years. Information is used on risk factors such as age, gender, body mass index (BMI), ethnicity, smoking history and disease histories[6,7]. FRS models have good performance in the USA population, but the risk predictions may be problematic when applied to cohorts that are hugely different from the cohort used for model development[8]. In the UK, treatment guidelines for the primary prevention of CVD recommend the use of QRISK2 (second version) to identify patients with high CVD risk[9].

QRISK is based on routinely collected data from general practices in the UK[7]. Conventional approaches were used to measure discrimination and calibration in the overall population[7]. However, there can be substantial variation between general practices in the style of coding clinical information (coding style) and completeness of data recording[10]. Different coding dictionaries are also currently being used in UK primary care as the EHR systems either use Read version 2 or CTV3 codes[11]. The patient case-mix (referring to a variation in risk factors for disease) may also vary between practices. This variability in the underlying data sources is currently not routinely considered in the development of risk prediction models, but it could potentially lead to heterogeneity in the prediction model's performance[12]. The objective of this study was to assess the level of generalisability of risk prediction models that are based on routinely collected data from EHRs, and to measure the effects of practice heterogeneity on the individual predictions of risk. The QRISK3 prediction model (for the 10 year risk of CVD) was used as an exemplar.

2.3 Methods

2.3.1 Data source

This study used data from the Clinical Practice Research Datalink (CPRD) which

is a database with anonymised EHRs from 674 GP practices in the UK. The database includes 4.4 million (6.9% of the UK population) patients and is broadly representative of the UK general population in terms of age, gender and ethnicity[13]. CPRD includes patient records of demographics, symptoms, tests, diagnoses, therapies, health-related behaviours and referrals to secondary care. Data from over half of the practices have been linked using unique patient identifiers to other datasets from secondary care, disease-specific cohorts and mortality records[13]. This study was restricted to 392 general practices that have been linked to Hospital Episode Statistics (HES), Office for National Statistics (ONS) and Townsend scores[7]. Over 1,700 publications have used CPRD data[14]. Previously, CPRD data has been used to externally validate QRISK2[15].

2.3.2 QRISK prediction models

QRISK is a statistical model which is being used to predict a patient's risk over 10 years of developing CVD (including coronary heart disease, stroke or transient ischaemic attack). The second version (QRISK2) was derived in 2008 using data from 355 practices in the QResearch database[16], and validated using data from 364 practices from the THIN database[17]. QRISK3 is the latest version published in 2017, which includes more clinical variables, such as migraine and chronic kidney disease, than QRISK2[7]. The QRISK3 predicted risks were calculated using the open access algorithm[18]. Calculations were successfully verified to be the same as predictions by the online calculator. This was done for simulated different patient groups in which each risk factor was changed sequentially covering the changes of all QRISK3 risk factors.

2.3.3 Study population

The study population in this study was similar to that used for the development cohort for QRISK3[7]. Patients were included if they were aged between 25 and 84 years, had no CVD history or prescribing of statins prior to the index date. The follow-up of patients in CPRD cohort started one year after start of data collection, patient's registration date, date of reaching age 25 years, or January 1 1998 (whatever came last) and it ended at the end of data collection, a patient leaving the practice, date patient's death or the CVD outcome (whatever came first). Patients were censored by the earliest date among the first statin prescription, transfer or the end of

follow-up[19]. The index date (as the start date for evaluating CVD and the baseline date for assessing a patient's history) was chosen randomly from the period of follow-up. The random index date[19] was preferred, because it gets a better spread of calendar time and age, and captures the time-relevant practice variability (e.g., change of recording and second trend of CVD incidence rate). This study considered the same risk factors as in QRISK3[7].

2.3.4 Statistical analysis

The QRISK3 predicted risks were estimated for each patient and were also averaged within each practice. Averaged predicted risks were compared to the observed risks at year 10 which were based on Kaplan Meier life tables. The observed risks were extrapolated for the 13.5% of practices with less than 10 years of follow-up. It was assumed that the life tables of these practices followed the pattern of the overall population life table. We calculated each year's CVD relative risk (RR) by dividing the current year's CVD proportion by the next year's CVD proportion. The extrapolation was verified using practices with 10 years follow-up. Specifically, we randomly remove records to make these practices have less than 10 years follow-up and then compared the extrapolated risk to the observed risk. We found no evidence[20] that the extrapolated risks were statistically significant to the actual observed risks.

A Cox model with a frailty (random effect) term for each practice was fitted to assess the effects of practice heterogeneity[21]. Patient survival time (time until censoring or CVD) was the outcome (dependent variable) and the linear predictor from the QRISK3 model was included as an offset. Each patient's linear predictor was calculated using the patient's risk factors and corresponding QRISK3 coefficients. Each practice's random effects on individual risk prediction and the standard deviation of all practices' random effects were extracted from the frailty model. Patient QRISK3 predictions and their corresponding practice random effects were combined to calculate a random effects model predicted risk. These were compared with the QRISK3 predicted risks. The distribution of the differences between the QRISK3 and the random effects model's predicted risks were plotted.

Limited practice size or duration of follow-up could contribute to the unknown variability between risks predicted by QRISK3 and the random effects model. In order to measure this random error, we simulated data under a null hypothesis of no practice level variability and estimated the distribution of the practice level random

effects, and compared this with the distribution of the practice level random effects observed in the CPRD data (i.e. a permutation test). Specifically, simulations were conducted using 2,000 datasets of the same size and follow-up as the CPRD data. The CVD outcomes were simulated by assigning a random probability from a uniform distribution (0, 1) to each patient. The random effects model was then fitted to these simulated data in order to quantify the random variability. The comparison between effects of unknown random variability and effects of practice level variability on individual patients was plotted using one million patients (50% male and 50% female) who had a QRISK3 predicted risk of 10%.

We used classical model performance measurements to compare QRISK3 with the random effects model. The data from each practice were randomly divided into two (70% and 30%) stratified by gender. The first part was used to develop the random effects model and the second part to test and calculate model performance measurements including the C-statistic[22], brier score[23,24] and net benefit[25]. These measurements were calculated using QRISK3 predictions, predictions of random effects model, patient follow-up time and patient status at the time of censoring. Empirical confidence intervals were calculated using 1,000 bootstrap samples.

Missing values for ethnicity, BMI, Townsend score, systolic blood pressure (SBP), standard deviation of SBP, cholesterol, High-Density Lipoprotein (HDL) and smoking status (only these have missing values) were imputed using Markov chain Monte Carlo (MCMC) method with monotone style[26]. The QRISK3 and random effects risks were then averaged based on ten imputations. We calculated random effects of CPRD practices and random effects separately for females and males consistent with QRISK3 development. The random effects of practices were calculated independently by both SAS and R with almost identical results. The random effects model used procedures from SAS 9.4 and “coxme” package for the R 3.4.2. The analyses of the datasets, missing value imputation, extrapolation validation and life tables were produced by SAS. R was used to model the data. The protocol for this work was approved by the independent scientific advisory committee for Clinical Practice Research Datalink research (protocol No 17_125RMn2). We confirm that all methods were performed in accordance with the relevant guidelines and regulations.

2.4 Results

Table 2.1 shows the patient characteristics and level of data recording across the 392 general practices. The mean age of patients varied between practices (5% percentile was 40.0 years and 95% percentile was 49.8 years). Presence of CVD risk factors also varied between practices. The 5-95% range between practices was 1.9 to 16.4 for recorded history of severe mental illness. The level of data completeness also varied substantially between practices. Ethnicity was not recorded for 19.6% of patients in the 5th percentile of practices compared to 93.9% in the 95% percentiles. Life table analysis are shown in eTable2.11.1 in the Supplement.

Table 2.1: Characteristics of the general practices included in the study and the distribution of data recording

		Distribution of characteristics across practices: Percentiles				
	Mean (SD)	**5th**	**25th**	**50th**	**75th**	**95th**
General characteristics of practices						
Total number of CVD events over 10 years in each practice	266.4 (176.5)	17.0	129.5	251.5	376.5	581.0
Average age of patients in each practice	44.9 (3.0)	40.0	42.9	45.0	46.7	49.8
% female patients	51.2 (2.1)	47.5	50.1	51.2	52.4	54.5
Total number of patients in each practice	9262.3 (5072.9)	2305.0	5292.5	8792.5	12180.0	17616.0
CVD risk factors						
% patients with alcohol abuse	1.4 (1.2)	0.5	0.8	1.1	1.6	3.0
% patients with anxiety	13.8 (5.3)	6.5	10.0	13.1	16.9	23.4
% patients with HIV	0.1 (0.1)	0.0	0.0	0.1	0.1	0.3
% patients with left ventricular hypertrophy	0.2 (0.1)	0.1	0.1	0.2	0.3	0.5
% patients with atrial fibrillation	0.7 (0.3)	0.3	0.5	0.7	0.9	1.3
% patients on atypical antipsychotic medication	0.4 (0.2)	0.2	0.3	0.4	0.6	0.9
% patients with Chronic kidney disease (stage 3, 4 or 5)	1.0 (0.9)	0.3	0.6	0.9	1.3	2.1
% patients on regular steroid tablets	0.1 (0.1)	0.0	0.0	0.1	0.1	0.2
% patients with erectile dysfunction	1.5 (0.6)	0.7	1.1	1.5	1.8	2.4
% patients with angina or heart attack in a 1st degree relative < 60	3.6 (3.0)	0.7	1.8	2.9	4.4	8.7
% patients on blood pressure treatment	6.8 (1.9)	3.8	5.6	6.7	8.2	9.9
% patients with migraines	6.4 (2.1)	3.2	4.8	6.4	7.8	9.6
% patients with rheumatoid arthritis	0.6 (0.2)	0.3	0.5	0.6	0.7	1.0
% patients with severe mental illness (this includes schizophrenia, bipolar disorder and moderate/severe depression)	7.8 (4.5)	1.9	4.2	7.2	10.8	16.4
% patients with Systemic Lupus Erythematosus	0.1 (0.0)	0.0	0.0	0.1	0.1	0.1
SBP						
Average SBP within practice	126.8 (2.8)	122.3	125.1	126.8	128.8	131.0
% patients with missing SBP	25.5 (7.3)	13.9	20.7	25.3	30.0	38.5
Average SBP standard deviation within practice	9.9 (0.7)	8.9	9.5	9.9	10.3	11.0
% patients with missing SBP standard deviation	52.7 (7.7)	39.0	48.3	53.1	57.3	64.7
BMI						
Average BMI when recorded	26.4 (0.7)	25.0	25.9	26.4	26.9	27.5
% patients with missing BMI	39.2 (11.8)	18.7	31.2	39.1	46.6	60.1

Table 2.1 (continued)

		Distribution of characteristics across practices: Percentiles				
	Mean (SD)	**5th**	**25th**	**50th**	**75th**	**95th**
Cholesterol/HDL ratio						
Average Cholesterol/HDL ratio	4.0 (0.2)	3.6	3.8	4.0	4.1	4.4
% patients with missing Cholesterol/HDL ratio	64.4 (10.0)	48.2	57.6	63.9	70.4	81.6
Smoking						
% patients who never smoked	47.8 (7.6)	36.0	43.3	47.9	52.7	59.4
% ex-smokers	22.3 (5.2)	13.8	19.0	22.5	25.5	30.9
% current-smokers	29.8 (7.0)	19.9	25.1	29.2	33.8	42.7
% patients with missing smoking status	24.2 (8.6)	10.3	18.6	23.8	29.5	39.4
Diabetes						
% patients with type 1 diabetes	0.2 (0.1)	0.1	0.2	0.2	0.3	0.4
% patients with type 2 diabetes	1.3 (0.4)	0.6	1.0	1.3	1.6	2.0
Ethnicity						
% other Asian patients	1.9 (3.2)	0.0	0.3	0.9	1.9	7.6
% Bangladeshi patients	0.4 (1.3)	0.0	0.0	0.2	0.4	1.4
% Black patients	3.5 (5.9)	0.1	0.5	1.3	3.4	15.3
% Chinese patients	0.7 (0.7)	0.0	0.2	0.5	1.0	2.0
% Indian patients	2.7 (5.3)	0.0	0.3	1.1	2.9	10.6
% patients with other ethnicity	2.9 (3.0)	0.3	0.9	2.0	3.6	9.1
% Pakistani patients	1.2 (3.6)	0.0	0.1	0.3	0.9	4.7
% White patients	86.7 (15.5)	48.2	83.4	92.3	96.8	98.8
% patients with missing ethnicity	58.5 (23.7)	19.6	38.5	62.5	77.5	93.9
Townsend score (Socioeconomic Status)						
% patients with Townsend score 1 (the least deprived)	20.3 (19.2)	0.1	4.1	14.7	31.1	59.7
% patients with Townsend score 2 (less deprived)	21.3 (16.4)	0.6	8.8	18.6	30.3	51.8
% patients with Townsend score 3 (deprived)	21.2 (13.1)	2.4	12.1	18.5	29.4	44.8
% patients with Townsend score 4 (more deprived)	21.1 (15.5)	0.3	8.6	19.9	29.5	52.9
% patients with Townsend score 5 (the most deprived)	16.1 (21.8)	0.0	0.4	7.6	22.3	66.3
% patients with Townsend score missing	0.1 (0.6)	0.0	0.0	0.1	0.1	0.3

Figure 2.1 shows the variation of CVD incidence rate among practices by plotting CVD incidence rate per 100 person years against the total follow-up time. A large amount of variation of CVD incidence rate were found between practices.

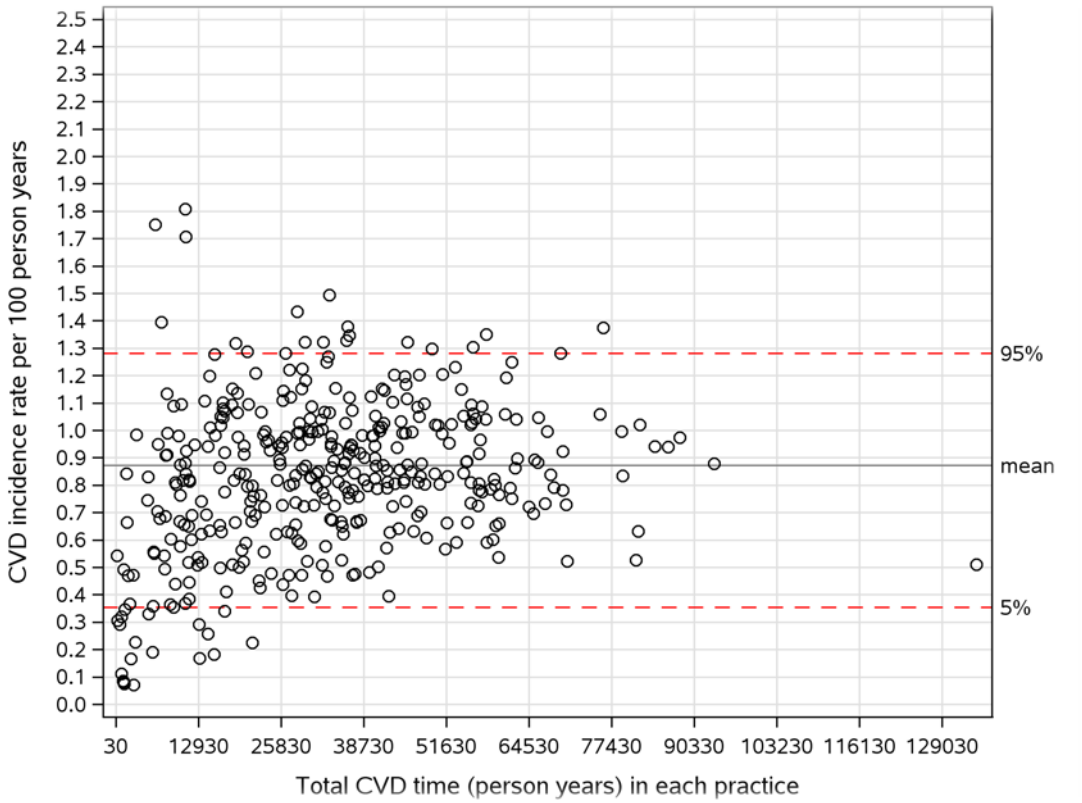


Figure 2.1: Variation of CVD incidence rate (per 100 person years) across practice

Figure 2.2 shows that the random effects model has less variation of differences between observed and predicted risk on practice level than QRISK3.

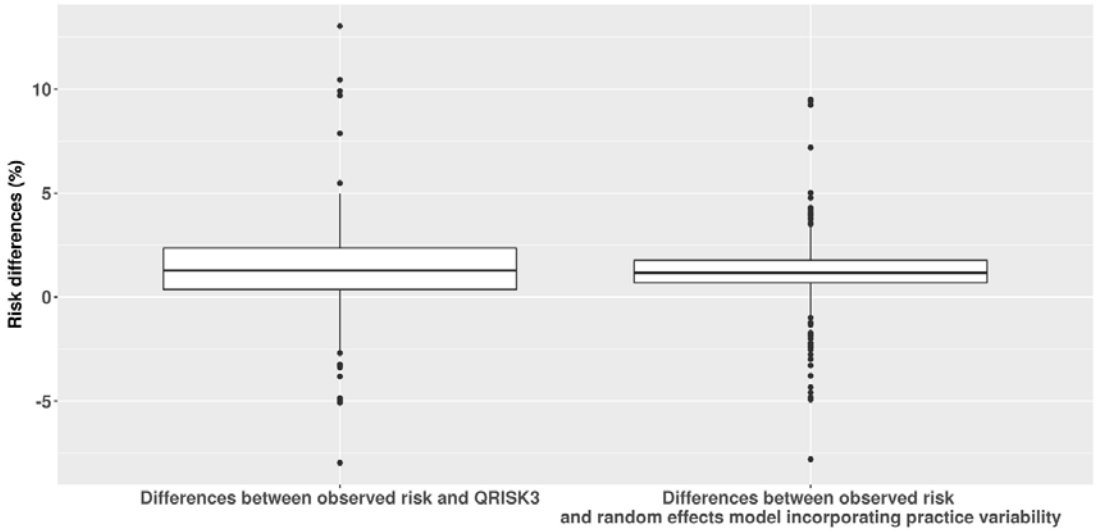


Figure 2.2: Comparison of differences between observed and QRISK3 (random effects) mode

Random effects model's Brier score (0.067 (95% CI: 0. 0667, 0. 0682)) was close to QRISK3's brier score (0.067 (95% CI: 0. 0666, 0. 0680)). The difference of Brier score between random effects model and QRISK3 was 0.002 (95% CI: 0.00008, 0.0023). Random effects model's C-statistic (0.852 (95% CI: 0.850, 0.854)) was also close to QRISK3's C-statistic (0.850 (95% CI: 0.848, 0.852)). The difference of C-statistic between the two models was 0.0017 (95% CI: 0.0015, 0.0020). The net benefit analysis [25] shows that both of models could predict three true CVD events without adding a false negative CVD events in every 100 patients with a given threshold of 10% (visualised in eFigure 2.11.2 in the Supplement). Standard deviation of random effects of CPRD practice between females (0.174) and males (0.177) were close to each other.

Table 2.2 shows the inconsistencies between the risks predicted for the same group of individual patients by QRISK3 and the random effects model (visualised in eFigure 2.11.1 in the Supplement). Patients with a predicted QRISK3 risk between 9.5% ~ 10.5% were found to have a much larger range of risks in the random effects model (between about 7.6% ~ 13.3%). Table 2.2 also shows the level of reclassification to below or above the treatment risk threshold of 10% when using the random effect model instead of the QRISK3 predicted risk. It was found that 19.7% patients with QRISK3 predicted risk between 8.5-9.5% had a risk above the treatment threshold when using the random effects model. For patients with QRISK3 predicted score between 10.5-11.5%, 24.4% of patients were reclassified to below the treatment threshold when using the different model.

Table 2.2: Inconsistencies between individual CVD risks as predicted by QRISK3 or by random effects model that incorporated practice variability

	Predicted risk according to random effects model incorporating practice variability							
QRISK3 predicted CVD risk (over 10 years)	**Percentile**					**% below /above treatment threshold of 10 year CVD risk (10%)**		**Total number of patients**
	2.5th~97.5th	**5th**	**25th**	**75th**	**95th**	**≤ 10**	**> 10**	
<6.5	0.1~6.0	0.1	0.4	2.6	5.4	100.0	0.0	2561602
6.5~7.5	5.3~9.4	5.5	6.3	7.6	8.9	99.0	1.0	96981
7.5~8.5	6.0~10.7	6.3	7.2	8.7	10.2	94.0	6.0	82768
8.5~9.5	6.8~12.0	7.1	8.2	9.7	11.4	80.3	19.7	72098
9.5~10.5	7.6~13.3	7.9	9.1	10.8	12.6	54.0	46.0	64477
10.5~11.5	8.4~14.6	8.8	10.0	11.9	13.9	24.4	75.6	56550
11.5~12.5	9.2~15.8	9.6	11.0	13.0	15.1	9.1	90.9	50278
12.5~13.5	10.0~17.1	10.4	11.9	14.0	16.3	2.4	97.6	45126
≥13.5	12.7~55.4	13.5	17.8	34.7	50.2	0.1	99.9	600938

Figure 2.3 plots the distribution of risks predicted with the random effect model for those with a QRISK3 predicted risk of 10%. The effects of random variability (measured by simulation analysis) in the random effect model is also presented in this figure. It was found that the effect of practice variability on predicted risks for patients cannot be fully explained by random variability, as the overall distribution (blue area) with a random effects' standard deviation of about 0.17 was much larger than the distribution due to random variability (green area) with a standard deviation for random effects of about 0.01.

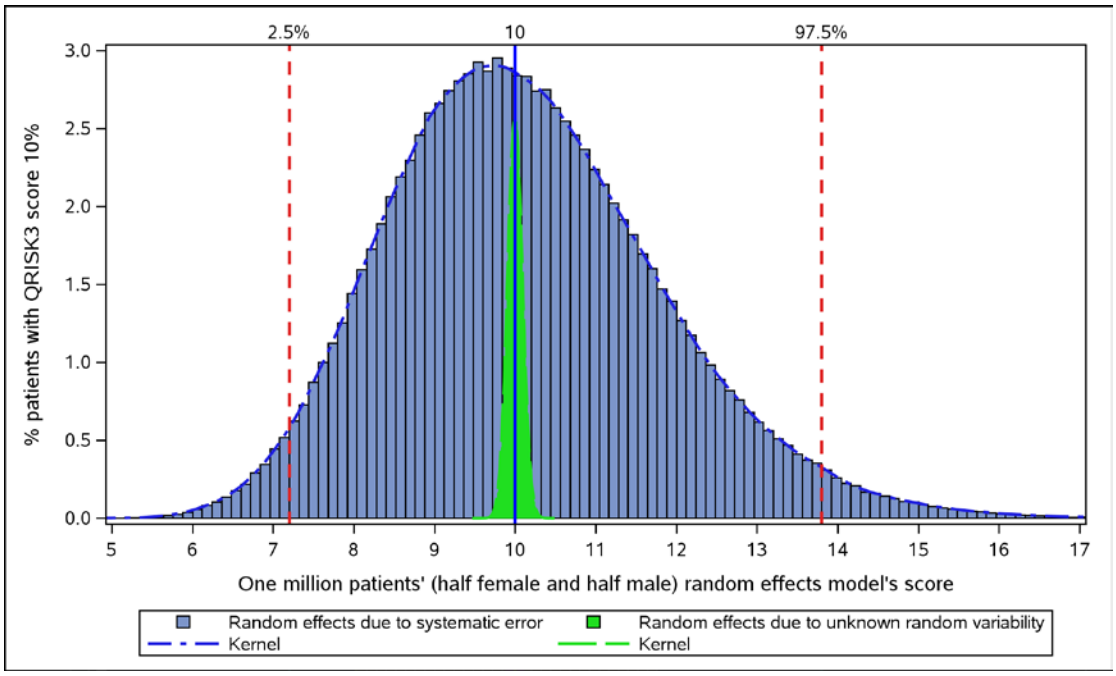


Figure 2.3: Distribution of predicted risks in the random effects model for patients with a QRISK3 predicted risk of 10% (using simulations in order to estimate the extent of random variability)

2.5 Discussion

2.5.1 Key results

This study found that incorporating practice variability in a risk prediction model substantially affected the predicted CVD risks of individual patients. The random effect model was similar to QRISK3 in terms of calibration and discrimination. Patients with a QRISK3 predicted risk of 10% had a much larger range of predicted risks after incorporating practice variability. Treatment classifications were found to be different for a substantive number of patients after considering the heterogeneity in CVD incidence between practices.

2.5.2 Limitation

There are several limitations of this study. Firstly, the observed risks had to be extrapolated for the practices with less than 10 years of follow-up to compare with QRISK3 (or random effects model) on practice level. The QRISK3 developers did not share the life table pattern of CVD risks over follow-up in QResearch. Although the validation showed that the result of extrapolation was not statistically significantly different from those practices with 10 years follow-up, the use of the actual changes in CVD risk over 10 years would have been preferable. Also, the definitions and classification of the risk factors could have been different from QRISK3 as the underlying EHR software systems vary between CPRD and QResearch (Vision and EMIS, respectively). However, the calibration and discrimination of QRISK3 in CPRD were consistent with those reported for QResearch, which suggest that the effects of differences in definitions was minimal.

2.5.3 Interpretation

Risk prediction models need to provide accurate and generalisable predictions in order to be used clinically for individual patient decision making[27]. Current guidelines for the development of risk prediction models do not include the evaluation of extent of heterogeneity in the underlying population (unaccounted for by the model) and its impact on the generalisability of the model. Conventional metrics in the evaluation of risk prediction models only include population level averages such as calibration and discrimination[28]. However, literature suggests that the risks at the population and individual levels may be determined differently[29,30]. An example of a tool with an

acceptable average measurement but unacceptable generalisability due to heterogeneity would be a blood pressure measurement that has systematic measurement errors at different times of a day. The historic treatise by Rose emphasised that the ability to predict an average risk on a population level does not always equate to the prediction of the individuals who are going to have the event soon[31]. A previous study highlighted that the Framingham and QRISK2 risk prediction models showed considerable variability in predicting high CVD risk despite comparable population-level calibration and discrimination[19]. As Briggs emphasised, risk prediction models that provide non-extreme probabilities can never empirically be proven wrong. It was also suggested, as done in the present study, to compare the impact on predictions and decision-making with different models that are statistically comparable[32]. Our study found that, the predicted CVD risks for individuals were very different after incorporating previously unmeasured variability between practices and that decisions based on the QRISK3 or random effect model could be quite different.

There may be several reasons for our finding of heterogeneity between general practices unaccounted for by QRISK3. One reason may be that the data quality of EHRs varies between general practices. A study on the EHR recording of osteoporosis reported that there was variability in inter-practice data quality with clinically important codes and with multiple ways that the same clinical concept was represented[33]. Also, different practice computer systems have different versions of clinical coding[33]. Damen et al. in their recent literature review of all CVD prediction models, pointed out that consistent codes such as ICD-9 or ICD-10 should be used in models' development and validation, as different definitions of CVD outcome lead to variation of model performance[5]. Another reason may be unmeasured heterogeneity in CVD risks in the populations of the different practices. There is substantive evidence that risks of disease are not uniformly distributed. A nation-wide study reported that there are severe inequalities in all-cause mortality between the North and South of England from 1965 to 2008[34]. A study by Langford et al. reported that region accounted for four times more variation in mortality than that explained by the classification of residential neighbourhoods by household type including socioeconomic status[35]. In order to use a risk prediction model for individual decision making, it should be established whether or not to allow these models to miss important causal predictors. If they do, this can then lead to a substantial

misclassification on an individual level.

Riley et al have proposed a statistical way to measure heterogeneity between sites by evaluating the C-statistics across practices in funnel plots with approximate 95% confidence interval based on the observed standard error observed[36]. We replicated Riley's funnel plot of QRISK2 and found similar variation of the C-statistic among practices in CPRD with QRISK3 (eFigure 2.11.4 in the Appendices). But this approach of funnel plots is limited as it does not assess the impact of heterogeneity on individual risk predictions. Random effects models are the standard approach to assess the effects of practice heterogeneity[21]. Our results highlight that it is not enough to only consider calibration and discrimination on the population level when assessing a prediction model's clinical utility on individual patients. The extent of heterogeneity in risk prediction unaccounted for by the model will need to be evaluated in addition to calibration and discrimination.

2.5.4 Implications for Research and Practice

This study found that QRISK3 has limited generalisability and accuracy in predicting individual risks in heterogeneous settings. The predictions of CVD risks of individual patients substantially changed after incorporating practice variability which could impact the clinical decisions for many patients. In order to improve the clinical utility of these risk prediction models, the level of unexplained heterogeneity in populations, disease incidence and data quality must be assessed before implementing such models for individual clinical decision making. Given the uncertainty with risk prediction models that use routinely collected EHR data, it is questionable whether these tools should be used without additional clinical interpretation and without incorporating causal risk factors that better capture the unmeasured heterogeneity between different general practices. Recently an online calculator was launched by Public Health England which allows members of the public to estimate their heart age based on a QRISK model [37]. Our study indicates that these estimates could be quite different when incorporating unmeasured heterogeneity and that the level of uncertainty with these predictions is considerable.

2.6 Funding

This study was funded by the China Scholarship Council. The funder's role was to provide salary support to Yan Li. The funder did not participate in this study and did not review results; other authors are independent of the funder.

2.7 Acknowledgements

This study is based in part on data from the Clinical Practice Research Datalink obtained under license from the UK Medicines and Healthcare products Regulatory Agency. The data is provided by patients and collected by the NHS as part of their care and support. The Office for National Statistics (ONS) is the provider of the ONS Data contained within the CPRD Data. Hospital Episode Data and the ONS Data Copyright © (2014), are re-used with the permission of The Health & Social Care Information Centre. All rights reserved. The interpretation and conclusions contained in this study are those of the authors alone.

2.8 Additional Information

There are no conflicts of interest among authors.

2.9 References

1. Phe. Action plan for cardiovascular prevention: 2017 to 2018.
2. Piepoli, M. F. *et al.* 2016 European Guidelines on cardiovascular disease prevention in clinical practice. *Eur. Heart J.* **37,** 2315–2381 (2016).
3. Cardiovascular disease prevention overview - NICE Pathways. Available at: https://pathways.nice.org.uk/pathways/cardiovascular-disease-prevention. (Accessed: 6th December 2017)
4. Prevention of Cardiovascular Disease Pocket Guidelines for Assessment and Management of Cardiovascular Risk. (2007).
5. Damen, J. A. A. G. *et al.* Prediction models for cardiovascular disease risk in the general population: systematic review. *BMJ* **353,** i2416 (2016).
6. Bitton, A. & Gaziano, T. A. The Framingham Heart Study's impact on global risk assessment. *Prog. Cardiovasc. Dis.* **53,** 68–78 (2010).
7. Hippisley-Cox, J., Coupland, C. & Brindle, P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. *Bmj* **2099,** j2099 (2017).
8. Matheny, M. *et al. Systematic Review of Cardiovascular Disease Risk Assessment Tools. Systematic Review of Cardiovascular Disease Risk Assessment Tools* (Agency for Healthcare Research and Quality (US), 2011).
9. NICE recommends wider use of statins for prevention of CVD | News and features | News | NICE.
10. Sáez, C., Robles, M., García-Gó Mez, J. M. & García-Gómez, J. M. Stability metrics for multi-source biomedical data based on simplicial projections from probability distribution distances. *Stat. Methods Med. Res.* **26,** 312–336 (2017).
11. NHS Digital. SNOMED CT implementation in primary care - NHS Digital. Available at: https://digital.nhs.uk/services/terminology-and-classifications/snomed-ct/snomed-ct-implementation-in-primary-care. (Accessed: 1st May 2018)
12. Wynants, L., Riley, R. D., Timmerman, D. & Van Calster, B. Random-effects meta-analysis of the clinical utility of tests and prediction models. *Stat. Med.* **37,** 2034–2052 (2018).
13. Herrett, E. *et al.* Data Resource Profile: Clinical Practice Research Datalink (CPRD). *Int. J. Epidemiol.* **44,** 827–836 (2015).
14. Clinical Practice Research Datalink - CPRD. Available at: https://www.cprd.com/intro.asp. (Accessed: 20th August 2017)
15. Hippisley-Cox, J., Coupland, C. & Brindle, P. The performance of seven QPrediction risk scores in an independent external sample of patients from general practice: a validation study. *BMJ Open* **4,** e005809 (2014).
16. Hippisley-Cox, J. *et al.* Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. *BMJ* **336,** 1475–82 (2008).
17. Collins, G. S. & Altman, D. G. Predicting the 10 year risk of cardiovascular disease in the United Kingdom: independent and external validation of an updated version of QRISK2. *BMJ* **344,** e4181 (2012).
18. ClinRisk Ltd. QRISK®3-2017 risk calculator. 1 (2017). Available at: https://qrisk.org/three/index.php. (Accessed: 10th November 2017)
19. van Staa, T.-P., Gulliford, M., Ng, E. S.-W., Goldacre, B. & Smeeth, L. Prediction of cardiovascular risk using Framingham, ASSIGN and QRISK2: how well do they predict individual rather than population risk? *PLoS One* **9,** e106455 (2014).
20. Kim, T. K. T test as a parametric statistic. *Korean J. Anesthesiol.* **68,** 540–6 (2015).

21. Hougaard, P. Frailty models for survival data. *Lifetime Data Anal.* **1,** 255–73 (1995).
22. Antolini, L. *et al.* A time-dependent discrimination index for survival data. *Stat. Med. Stat. Med* **24,** 3927–3944 (2005).
23. Kronek, L.-P. & Reddy, A. Logical analysis of survival data: prognostic survival models by detecting high-degree interactions in right-censored data. *Bioinformatics* **24,** i248–i253 (2008).
24. Graf, E., Schmoor, C., Sauerbrei, W. & Schumacher, M. Assessment and comparison of prognostic classification schemes for survival data. *Stat. Med.* **18,** 2529–45
25. Vickers, A. J. & Elkin, E. B. Decision curve analysis: a novel method for evaluating prediction models. doi:10.1177/0272989X06295361
26. Schafer, J. L. (Joseph L. . *Analysis of incomplete multivariate data*. (Chapman & Hall, 1997).
27. Can we really predict risk of cancer? *Cancer Epidemiol.* **37,** 349–352 (2013).
28. Siontis, G. C. M., Tzoulaki, I., Siontis, K. C. & Ioannidis, J. P. A. Comparisons of established risk prediction models for cardiovascular disease : systematic review. *BMJ* **3318,** 1–11 (2012).
29. O'flaherty, M. & Capewell, S. New perspectives on cardiovascular risk in individuals and in populations. doi:10.1136/jech-2012-201409
30. Elmore, J. G. & Fletcher, S. W. The Risk of Cancer Risk Prediction: "What Is My Risk of Getting Breast Cancer?" *JNCI J. Natl. Cancer Inst.* **98,** 1673–1675 (2006).
31. Somerville, M. Rose's Strategy of Preventive Medicine. *J. Public Health (Bangkok).* **30,** 349–349 (2008).
32. BRIGGS, W. *UNCERTAINTY : the soul of modeling, probability & statistics.* (SPRINGER, 2018).
33. de Lusignan, S. *et al.* Problems with primary care data quality: osteoporosis as an exemplar. *Inform. Prim. Care* **12,** 147–56 (2004).
34. Hacking, J. M., Muller, S. & Buchan, I. E. Trends in mortality from 1965 to 2008 across the English north-south divide: comparative observational study. *BMJ* **342,** d508 (2011).
35. Regional variations in mortality rates in England and Wales: An analysis using multi-level modelling. *Soc. Sci. Med.* **42,** 897–908 (1996).
36. Riley, R. D. *et al.* External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. *BMJ* **353,** i3140 (2016).
37. What's your heart age? - NHS. Available at: https://www.nhs.uk/conditions/nhs-health-check/check-your-heart-age-tool/. (Accessed: 27th September 2018)

2.10 Author Contribution statement

Yan Li: Designed the study; conducted all statistical analysis; produced all tables and figures; wrote the main manuscript text.

Matthew Sperrin: Designed the study; proposed the main statistical method; helped in interpretation of statistical results; reviewed all statistical results; reviewed and edited the main manuscript text.

Miguel Belmonte: Reviewed all statistical methods and results; provided technical details of statistical methods; reviewed and edited paper.

Alexander Pate: Produced the raw statistical analysis dataset; reviewed all statistical results; reviewed and edited paper.

Darren M Ashcroft: Improved the major interpretation of statistical results and discussion; reviewed and edited paper;

Tjeerd Pieter van Staa: Designed and supervised the study; Quality control of all aspects of the paper; wrote the main manuscript text;

2.11 Supplementary Online Content

eAppendix. Interpretation of appendix tables and figures.

eTable 2.11.1. Distribution across practices of the number of CVD cases, number of patients at risk and survival rate over 10 years

eFigure 2.11.1. Comparison of random effects model's score and QRISK3 score in the same group of patients (grouped by certain range (red lines) of QRISK3 score)

eFigure 2.11.2. Comparison of net benefit between QRISK3 and random effects model

eFigure 2.11.3.1. Calibration plot of QRISK3

eFigure 2.11.3.2. Calibration plot of random effects model

eFigure 2.11.4. Variation of QRISK3's C-statistic among practices—a replication of Riley's funnel plot

eTable 2.11.1 shows the distribution of number of CVD events, number of patients at risk and survival rate among practices at 10 years. The number of CVD events, number of patients at risk and survival rate of practices are generally decreasing along the 10 years. The number of CVD events and the number of patients at risk varied between 5~95% percentile practices. The survival rate of 5~95% practices has less variation. Practices which do not have survival rate in 10 years are because their longest follow-ups are less than 10 years.

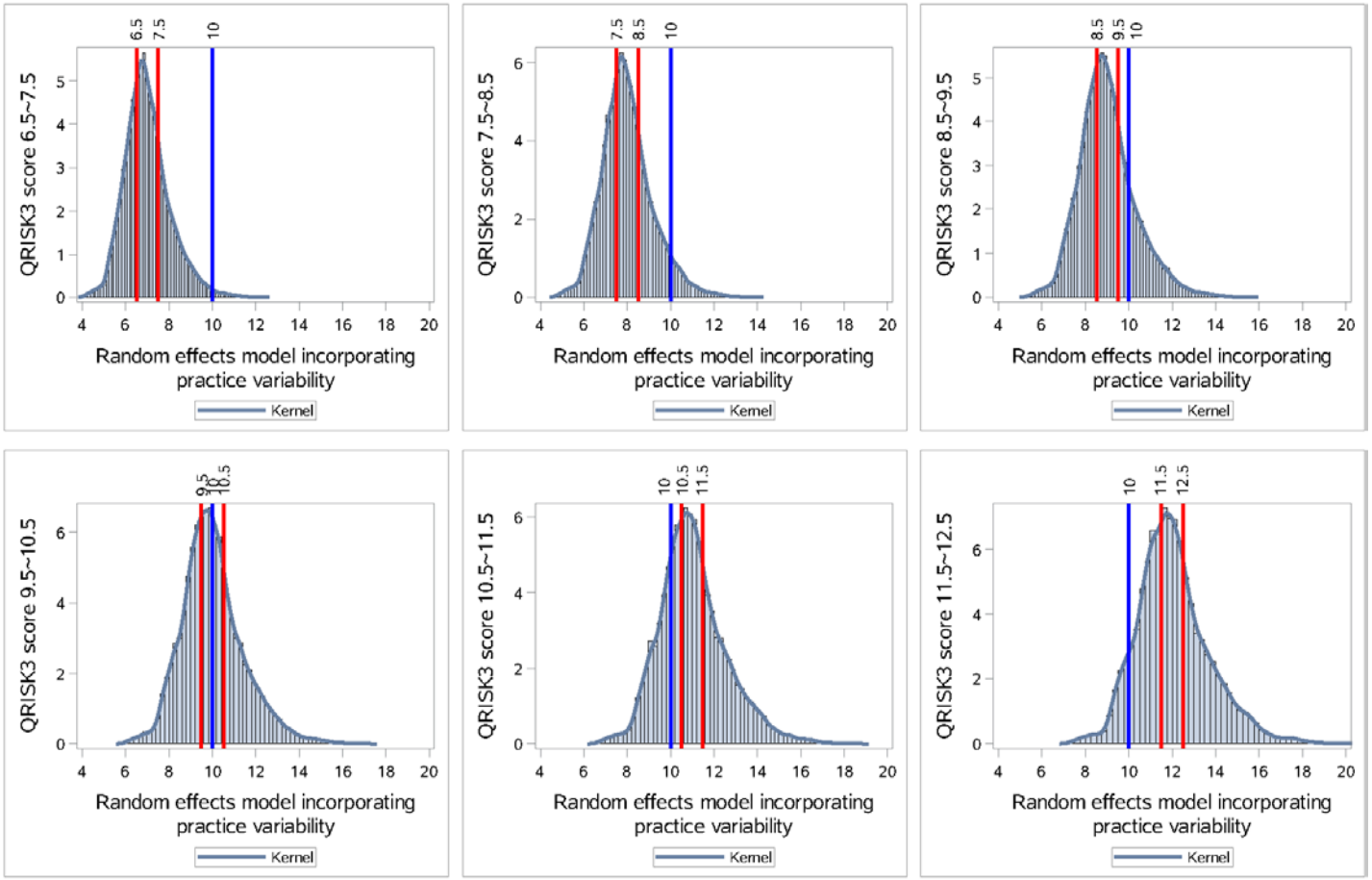
eTable 2.11.1 Distribution across practices of the number of CVD cases, number of patients at risk and survival rate over 10 years

	Number of patients with CVD events in practices (Percentile)				Number of patients at risk[a] in practices (Percentile)				Life table estimate of proportion of patients without CVD (Percentile)			
Year	**5th**	**25th**	**75th**	**95th**	**5th**	**25th**	**75th**	**95th**	**5th**	**25th**	**75th**	**95th**
1	5.0	31.0	83.5	134.0	1789.0	4358.0	10247.5	14843.0	1.00	1.00	1.00	1.00
2	2.0	21.0	60.0	93.0	575.0	3019.3	7352.0	10560.0	0.99	0.99	0.99	1.00
3	0.0	14.0	49.0	76.0	85.5	2324.8	5763.3	8329.0	0.97	0.98	0.99	0.99
4	0.0	12.0	41.0	64.0	0.0	1821.3	4671.0	6707.5	0.95	0.97	0.98	0.99
5	0.0	10.5	34.0	54.0	0.0	1456.0	3893.0	5530.5	NA[b]	0.96	0.98	0.98
6	0.0	9.0	29.5	48.0	0.0	1128.8	3225.0	4713.5	NA	0.95	0.97	0.98
7	0.0	7.0	26.0	42.0	0.0	839.0	2688.5	3995.5	NA	0.94	0.96	0.97
8	0.0	5.0	22.0	39.0	0.0	601.3	2208.0	3345.0	NA	0.93	0.95	0.97
9	0.0	4.0	18.0	32.0	0.0	420.0	1803.8	2788.0	NA	0.92	0.94	0.96
10	0.0	2.0	15.0	26.0	0.0	263.8	1426.8	2297.5	NA	0.91	0.93	0.95

a. Number of patients in the middle point of each year was used

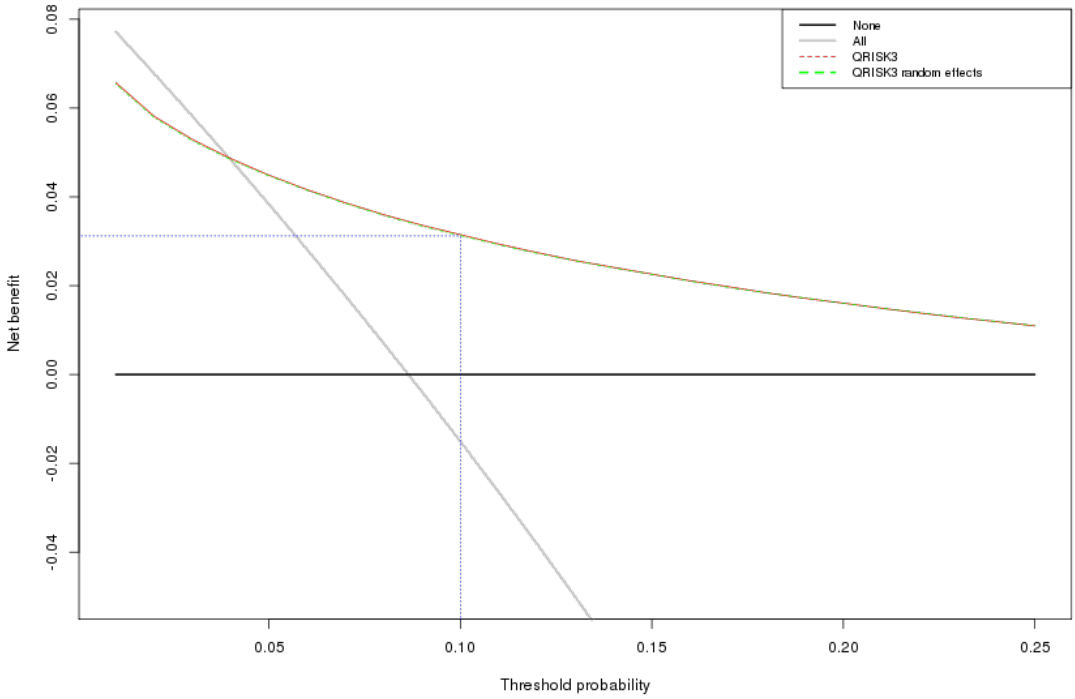
b. NA is because practices have less than 10 years follow-up data

eFigure 2.11.1 visualises the inconsistencies between the risks predicted for the same group of individual patients by QRISK3 and the random effects model. Patients with a predicted QRISK3 risk between 9.5% ~ 10.5% were found to have a much larger range of risks in the random effects model (between about 6% ~15%)



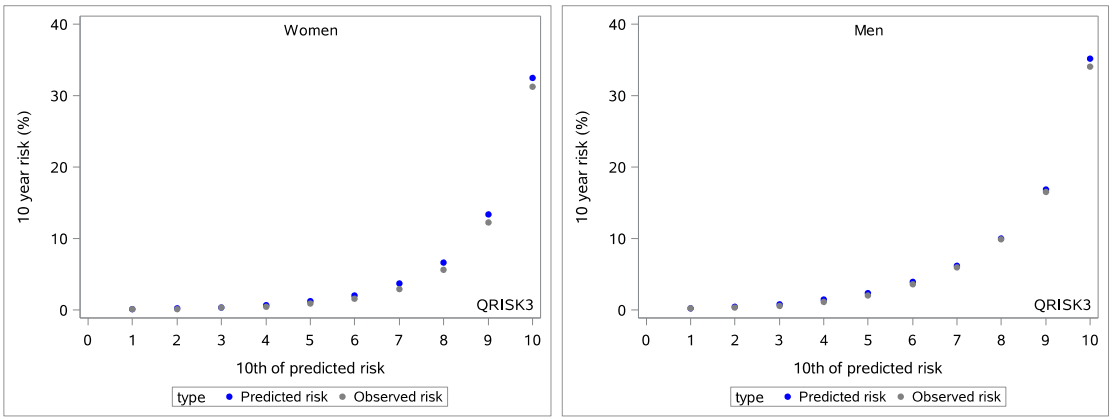
eFigure 2.11.1 Comparison of random effects model's score and QRISK3 score in the same group of patients (grouped by certain range (red lines) of QRISK3 score)

eFigure 2.11.2 shows two models' net benefit is about 3.1% at the threshold 10%, which means both of models predict about 3 true positive CVD events without adding new false positive CVD patients.

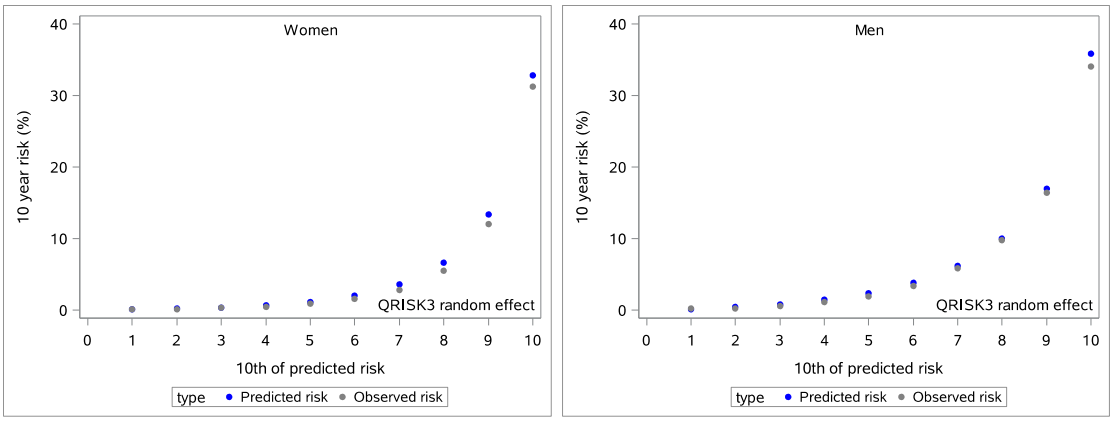


eFigure 2.11.2 Net benefit analysis on QRISK3 and random effects model

eFigure 2.11.3.1 and 2.11.3.2 show two models have similar calibration.

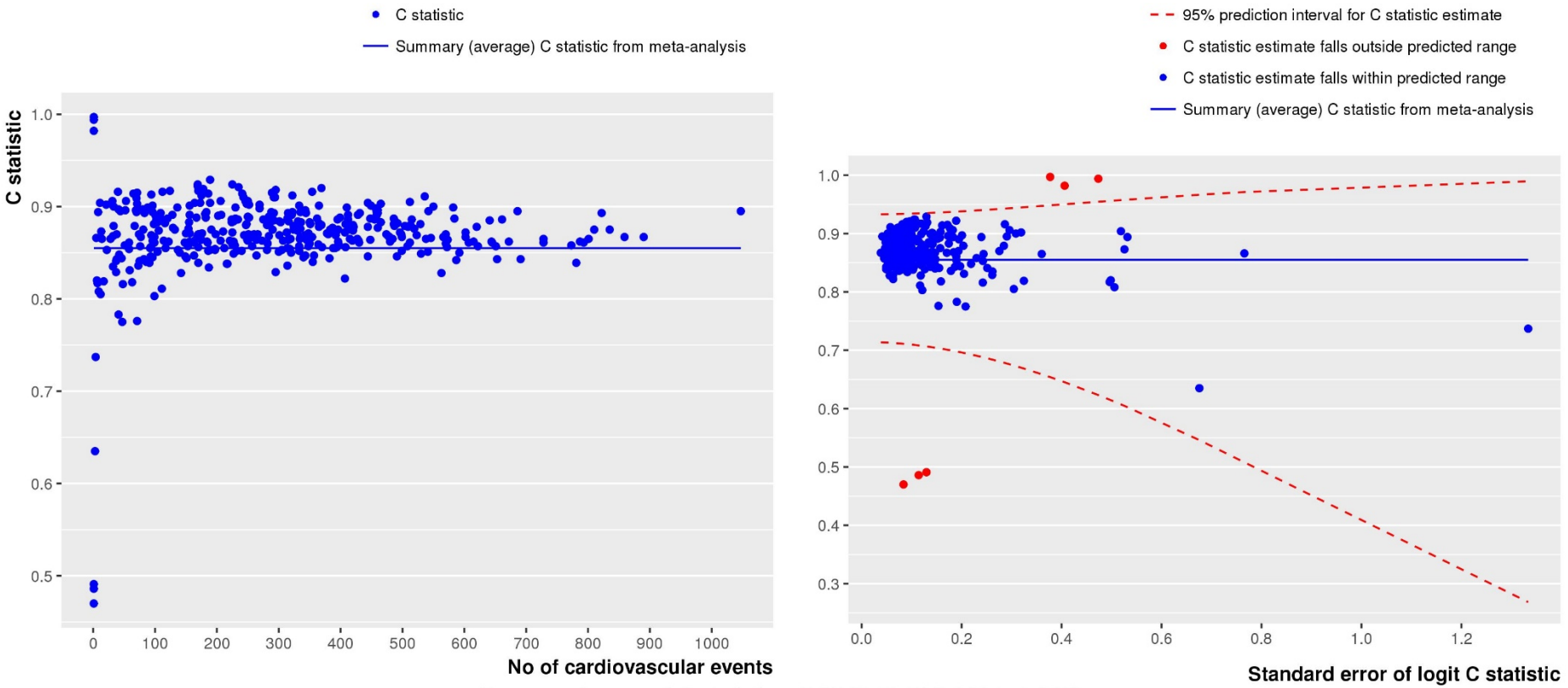


eFigure 2.11.3.1 Calibration plot of QRISK3



eFigure 2.11.3.2 Calibration plot of random effects model

eFigure 2.11.4 is a replication of Riley's [1] funnel plot. The left panel shows that QRISK3 has variation of C-statistic among practices, and the right panel performed a formal meta-analysis to identify outlier practices (those red dots outside the 95% prediction interval). The figure shows that QRISK3 performs differently on different practices, which is consistent to Riley's[1] finding on QRISK2.



eFigure 2.11.4. Variation of QRISK3's C-statistic among practices—a replication of Riley's[1] funnel plot

2.11.5 References to online-only supplement

1. Riley, R. D. *et al*. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. *BMJ* **353,** i3140 (2016).

Blank page

Chapter 3 Examining the impact of data quality and completeness of electronic health records on predictions of patients' risks of cardiovascular disease

Yan Li[1], Matthew Sperrin[1], Glen P. Martin[1], Darren M Ashcroft[2,3], Tjeerd Pieter van Staa[1,4,5]

[1]Health e-Research Centre, Farr Institute, School of Health Sciences, Faculty of Biology, Medicine and Health, the University of Manchester, Manchester Academic Health Sciences Centre (MAHSC), Oxford Road, Manchester, M13 9PL, UK
[2]Centre for Pharmacoepidemiology and Drug Safety, School of Health Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Oxford Road, Manchester, M13 9PL, UK
[3]NIHR Greater Manchester Patient Safety Translational Research Centre, School of Health Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Oxford Road, Manchester, M13 9PL, UK
[4]Utrecht Institute for Pharmaceutical Sciences, Utrecht University, Utrecht, Netherlands
[5]Alan Turing Institute, Headquartered at the British Library, London, UK

Corresponding author: Tjeerd van Staa, tjeerd.vanstaa@manchester.ac.uk

Journal title: International Journal of Medical Informatics
Doi: https://doi.org/10.1016/j.ijmedinf.2019.104033

Word count: 3188
Abstract: 244
Number of tables: 2
Number of figures: 4

3.1 Abstract

Objective

To assess the extent of variation of data quality and completeness of electronic health records and impact on the robustness of risk predictions of incident cardiovascular disease (CVD) using a risk prediction tool that is based on routinely collected data (QRISK3).

Methods

Design: Longitudinal cohort study.

Setting: 392 general practices (including 3.6 million patients) linked to hospital admission data.

Methods: Variation in data quality was assessed using Sáez's stability metrics quantifying outlyingness of each practice. Statistical frailty models evaluated whether accuracy of QRISK3 predictions on individual predictions and effects of overall risk factors (linear predictor) varied between practices.

Results

There was substantial heterogeneity between practices in CVD incidence unaccounted for by QRISK3. In the lowest quintile of statistical frailty, a QRISK3 predicted risk of 10% for female was in a range between 7.1% and 9.0% when incorporating practice variability into the statistical frailty models; for the highest quintile, this was 10.9%-16.4%. Data quality (using Saez's metrics) and completeness were comparable across different levels of statistical frailty. For example, recording of missing information on ethnicity was 55.7%, 62.7%, 57.8%, 64.8% and 62.1% for practices from lowest to highest quintiles of statistical frailty respectively. The effects of risk factors did not vary between practices with little statistical variation of beta coefficients.

Conclusions

The considerable unmeasured heterogeneity in CVD incidence between practices was not explained by variations in data quality or effects of risk factors. QRISK3 risk prediction should be supplemented with clinical judgement and evidence of additional risk factors.

Key words

Electronic health records; QRISK; practice variability; statistical frailty model; CVD risk prediction; random slope model

3.2 Introduction

Cardiovascular disease (CVD) has been the most common cause of death around the world for decades[1]. The prevention of CVD through targeting treatment to high risk patients is recommended in many international guidelines[1-4]. Risk prediction models are now an important part of CVD prevention strategies[5]. Many CVD risk prediction models have been developed around the world[6], including the Framingham risk score (FRS)[7] in the USA, QRISK3[8] in the UK and ESC HeartScore in Europe[9]. These models were developed by fitting statistical survival models (e.g. Cox model[10]) incorporating CVD risk factors on longitudinal patient cohorts. Specifically, QRISK was first developed in 2008 using routinely collected electronic health records (EHRs) from 355 general practices included in the QResearch database[8]. It considered age, sex and CVD risk factors such as body mass index (BMI) and smoking status. A recent update, QRISK3, incorporated more risk factors, such as variation in systolic blood pressure[8].

A previous study has found that QRISK3 scores that are derived from EHRs can have limited generalisability and accuracy, as they do not account for the substantive heterogeneity between different general practices[11]. Considerable changes in the individual risk estimates occurred when taking into account the heterogeneity between different general practices. Additionally, this study found that a CVD risk of 10% over 10 years as predicted by QRISK3 could change by over absolute 13% in a model that also incorporated variability between sites. Heterogeneity between sites may be related to either data quality (mainly including variation of missingness and coding[12]) or unadjusted underlying practice heterogeneities (variation of patient case mix and association between outcome and predictors[13]). However, it is unknown which of these influences contribute to the observed effects of practice variability on individual risk prediction. Therefore, the objective of this study was to assess the extent of variation of data quality and completeness of electronic health records and impact on the robustness of risk predictions of cardiovascular disease (CVD) using QRISK3. The QRISK3 model is recommended to be used in UK general practice and is now also accessible for members of the public[14, 15].

3.3 Methods

The study used data from approximately 3.6 million anonymised patient records derived from 392 general practices from the Clinical Practice Research Datalink (CPRD GOLD), which had been linked to Hospital Episode Statistics (HES), Office for National Statistics (ONS) mortality records and Townsend deprivation scores[8]. CPRD GOLD is a representative demographic sample of the UK population in terms of age, gender and ethnicity[16]. Overall, CPRD includes data on about 6.9% of the UK population. The linkages to other datasets such as HES or ONS provide additional patient information about secondary care, specific disease and cause-specific mortality[16]. CPRD includes patients' electronic health records from general practice capturing detailed information such as demographics, symptoms, tests, diagnoses, prescribed treatments, health-related behaviours and referrals to secondary care[16]. CPRD data has been widely used for public health research[17], including an external validation of the QRISK2 model[18].

The study used the same patient population as described in a previous study[11], and used similar selection criteria and risk factors to QRISK3[8]. The follow-up of patients started at the date of the patient's registration with the practice, 25th birthday, or January 1 1998 (whichever latest), and ended at the date of death or CVD outcome, the date of leaving the practice, end of study window or last date of data collection (whichever earliest). The index date for measurement of CVD risk was randomly chosen from the total period of follow-up[19]. This study used a random index date, as it captures time-relevant practice variability with a better spread of calendar time and age[11]. The use of a random index date was the only difference with the original QRISK3 studies[8]. The main inclusion criteria for the study population were aged between 25 and 84 years, with no CVD history or any statin prescription prior to the index date. Patients were censored at the date of the statin prescription if received during follow-up.

There were four analysis parts in this study. The first measured data quality and completeness in each of the different practices. Second, we evaluated the heterogeneity between practices in CVD incidence that was not taken into account in the development of QRISK3. This analysis addressed the miscalibration of QRISK3 at practice level which can be described as the closeness (accuracy) of the QRISK3 prediction to the observed CVD incidence in each practice. Unmeasured

heterogeneity between groups is also known as statistical frailty, which can be modelled in regression analyses[20]. The level of unmeasured heterogeneity in CVD incidence (statistical frailty) for each practice was used to stratify practices into quintiles. Third, we evaluated whether the effects of the QRISK3 risk factors (i.e., the overall linear predictor) varied between practices (i.e., whether the beta coefficients varied). This variation in the linear predictor between practices could occur in case of unmeasured effect moderators for CVD incidence or differences in data recording/misclassification of risk factors. Finally, we compared data quality across different levels of statistical frailty.

Several indicators of data quality were used in this study to measure the variation in coding between general practices. First, the percentages of missing records were measured for the variables ethnicity, systolic blood pressure (SBP), body mass index (BMI), cholesterol, high-density lipoprotein (HDL), ratio of cholesterol and HDL, smoking status and Townsend score for deprivation. Second, two metrics as proposed by Sáez[21] were used to measure the multidimensional variability (stability) in data quality across practices. The proposed metrics quantified the variability in the probability distribution functions of practices. Variation of coding was measured by the distribution-dissimilarity (quantified by Sáez's metrics) of CVD risk factors and their missingness among practices. Sáez's metric[21], which was based on Jensen–Shannon divergence[22] measured the distribution-dissimilarity of variables across practices. Specifically, source probabilistic outlyingness (SPO) can be thought of as a measure of how different a practice is from the average practice in terms of distribution of variables. SPO ranges from 0 to 1 measuring the extent of outlyingness of the variables' distribution. A variable with a SPO close to 1 means that the distribution of the variable in the practice is more different from the overall average indicating the outlyingness of coding. Further technical details about the Sáez's metric are provided in the eAppendix 3.9.2.

The unmeasured heterogeneity between practices in CVD incidence was evaluated by fitting a Cox proportional hazards model that included a statistical frailty term on its intercept (this type of model is also known a random intercept Cox model). The outcome of interest was the time to CVD onset. The linear predictor of QRISK3 (sum of the multiplication of beta coefficients and predictors) was used as an offset (i.e. coefficient fixed at one) to calculate the statistical frailty for each practice[20].

The variation between practices in the effects of the QRISK3 risk factors was evaluated by also adding a single frailty term to the beta coefficients of the QRISK3 linear predictor (known as a mixed effects Cox model[23]). This model calculated a random slope for the QRISK3 risk factors in each practice (assuming fixed effects and independent random effects of the QRISK3 linear predictor) in addition to the random intercept (assuming unmeasured heterogeneity in CVD incidence between practices). The random slopes and intercepts were calculated separately for each gender as QRISK3 has separate model formula for each gender[24].

The effects of practices' random slopes on individual risk prediction were visualised by estimating the difference of individual CVD risk predictions in the random slope model to that of the random intercept model. The range of individual risk predictions were calculated from the random slope model. Using a QRISK3 risk of 10%, a random slope and a random intercept were randomly drawn from a Gaussian distribution based on the variation of the random slope and random intercept calculated from this study's original cohort and the predicted risk was estimated (this was repeated one million times). The difference of the predicted CVD risk when the same patient was from practices with the same random intercept but different random slope was visualised. Two hypothetical variations in random slopes (0.03 and 0.1) were used as reference lines. The variation in random slopes of 0 indicates that there was no variation in the effects of CVD risk factors between practices. The variation of 0.03 was chosen as reference because a previous study found that this variation in the random effects of the intercept[11] resulted in large differences in individual risk predictions (a QRISK3 predicted risk of 10% would change in the random effects model to a range from 7.2% to 13.7%).

Finally, practices were grouped by quintiles of statistical frailty and data quality metrics were estimated for each quality indicator. The random intercept Cox models estimated the level of statistical frailty for a QRISK3 predicted CVD risk of 10% (over 10 years). The mean and standard deviation of each CVD risk factors were summarised. Sáez's metric for the CVD risk factors and their missingness were plotted against the percentile of practice frailty to show possible correlations and the Pearson correlation coefficients were calculated[25]. Practice statistical frailty was also plotted against the percentile of the mean (for continuous variables) or percentage (for categorical variable) of CVD risk factors at practice level and their corresponding

Sáez's metric using a Beeswarm plot[26] to identify any correlation between them, and practice statistical frailty as 1 was plotted as a reference line (red line). Beeswarm plots visualise the distribution by plotting practices as separate dots in each bin, so it has benefit to highlight individual points in distribution comparing to classical distribution-visualisation such as histograms.

The statistical software R version 3.4.2[27] with package "coxme"[28] was used to model the data; SAS 9.4 was used in data preparation, missing value imputation and visualisation. Multiple imputation using Markov chain Monte Carlo (MCMC) method with monotone style[29] was used to impute missing values before model fitting. Ten imputed datasets were created with pooling of the results based on the averages.

3.4 Results

There were 3,630,818 patients included in the study cohort, 103,350 of which had a CVD event in the 10 years after the index date. Table 3.1 shows the differences between the predictions by the QRISK3 and random intercept models (statistical frailty) for patients with a QRISK3 prediction of 10%. The practices were classified into quintiles of practice statistical frailty. Practices in the lowest quintile had predicted CVD risks at 10 years between 7.1% and 9.0% in the random intercept model for females compared to a predicted risk of 10% with QRISK3. For males, this was 6.1% and 9.0%. For practices in the highest quintile, QRISK3 predictions underestimated CVD risks compared to the random intercept model with predicted risks between 10.9% and 16.4% (for males, this was 10.9% and 15.5%). As shown in Table 1, a practice statistical frailty below 1 indicated that QRISK3 overestimated CVD risk and above 1 underestimated CVD risk compared with the random intercept models.

Table 3.1 Predicted CVD risks in random intercept models (for patients with QRISK3 predicted risk of 10%) stratified into quintiles based on the level of differences between these predictions

Quintile of practice frailty	Number of practices	Frailty	Predicted CVD risk with random intercept model (%)
Female			
0~20%	78	0.7~ 0.9	7.1~ 9.0
20~40%	78	0.9~ 1.0	9.0~ 10.0
40~60%	79	1.0~ 1.0	10.0~ 10.0
60~80%	78	1.0~ 1.1	10.0~ 10.9
80~100%	79	1.1~ 1.7	10.9~ 16.4
Male			
0~20%	78	0.6~ 0.9	6.1~ 9.0
20~40%	78	0.9~ 1.0	9.0~ 10.0
40~60%	79	1.0~ 1.0	10.0~ 10.0
60~80%	78	1.0~ 1.1	10.0~ 10.9
80~100%	79	1.1~ 1.6	10.9~ 15.5

Table 3.2 compares the baseline characteristics between practices with different levels of statistical frailty (i.e., mean difference between individual risk predictions by QRISK3 and random intercept models). For example, practices in the second quintile (20%~40%) of practice statistical frailty have on average 62.7% (standard deviation: 20.0%) patients with missing values on ethnicity and practices in the fourth quintile (60%~80%) of practice statistical frailty also have similar average 64.8% (standard deviation: 23.0%) of patients with missing values on ethnicity. There were no major differences in CVD risk factors and missing levels between practices with high and low statistical frailty. Practices with high/low statistical frailty had comparable means and standard deviations for these characteristics.

Table 3.2 Characteristics of the practices stratified by different quintiles of statistical frailty

	Male					Female				
	mean (SD))					(mean (SD))				
	Frailty (0~20%) (0.7 ~ 0.9)	**Frailty (20~40%) (0.9 ~ 1.0)**	**Frailty (40~60%) (1.0 ~ 1.0)**	**Frailty (60~80%) (1.0 ~ 1.1)**	**Frailty (80~100%) (1.1 ~ 1.7)**	**Frailty (0~20%) (0.6 ~ 0.9)**	**Frailty (20~40%) (0.9 ~ 1.0)**	**Frailty (40~60%) (1.0 ~ 1.0)**	**Frailty (60~80%) (1.0 ~ 1.1)**	**Frailty (80~100%) (1.1 ~ 1.6)**
General characteristics of practices										
Average number of CVD events in 10 years within practice strata by gender	84.5 (58.5)	133.1 (89.3)	142.0 (92.1)	181.2 (101.7)	182.4 (87.0)	89.0 (57.6)	104.0 (86.3)	122.7 (91.5)	143.4 (80.3)	159.5 (75.8)
Average age within practice	43.6 (2.9)	44.5 (3.0)	44.2 (3.1)	44.7 (2.4)	44.2 (2.0)	44.9 (4.0)	46.0 (3.8)	45.0 (3.6)	46.2 (2.7)	45.6 (2.6)
Average number of patients within practice strata by gender at index date	4528.0 (2543.6)	4680.9 (2737.6)	4330.1 (2392.9)	4930.5 (2723.2)	4112.1 (1825.2)	5415.8 (2805.9)	4590.3 (2951.9)	4578.5 (2767.0)	4819.0 (2375.8)	4341.2 (2012.4)
Number of practices	78	78	79	78	79	78	78	79	78	79
CVD risk factors										
% patients with alcohol abuse	1.5 (0.8)	1.6 (1.0)	1.8 (1.7)	2.0 (2.0)	2.4 (1.3)	0.7 (0.4)	1.0 (1.7)	0.9 (0.7)	0.8 (0.4)	1.1 (0.7)
% patients with anxiety	8.9 (3.5)	8.9 (3.0)	9.7 (3.7)	10.9 (3.4)	12.0 (5.2)	15.5 (5.9)	15.9 (5.9)	17.2 (6.7)	17.4 (6.3)	20.1 (7.6)
% patients with HIV	0.1 (0.2)	0.1 (0.1)	0.1 (0.1)	0.1 (0.1)	0.1 (0.1)	0.1 (0.1)	0.1 (0.1)	0.1 (0.2)	0.1 (0.1)	0.0 (0.1)
% patients with left ventricular hypertrophy	0.2 (0.1)	0.2 (0.1)	0.3 (0.2)	0.3 (0.1)	0.3 (0.2)	0.2 (0.1)	0.2 (0.1)	0.2 (0.2)	0.2 (0.1)	0.2 (0.2)
% patients with atrial fibrillation	0.8 (0.4)	0.9 (0.4)	0.8 (0.4)	0.9 (0.3)	0.7 (0.3)	0.6 (0.3)	0.7 (0.3)	0.6 (0.3)	0.7 (0.3)	0.6 (0.3)
% patients on atypical antipsychotic medication	0.4 (0.3)	0.4 (0.3)	0.4 (0.3)	0.4 (0.2)	0.5 (0.2)	0.4 (0.2)	0.4 (0.2)	0.4 (0.3)	0.4 (0.2)	0.5 (0.2)
% patients with chronic kidney disease (stage 3, 4 or 5)	0.8 (0.4)	0.8 (1.0)	0.8 (0.5)	0.7 (0.4)	0.6 (0.3)	1.3 (0.9)	1.5 (1.0)	1.5 (2.2)	1.3 (0.8)	1.1 (0.6)
% patients on regular steroid tablets	0.1 (0.1)	0.1 (0.1)	0.1 (0.1)	0.1 (0.1)	0.1 (0.1)	0.1 (0.1)	0.1 (0.1)	0.1 (0.1)	0.1 (0.1)	0.1 (0.1)
% patients with angina or heart attack in a 1st degree relative < 60	3.5 (2.5)	3.5 (3.1)	3.2 (2.2)	2.9 (2.8)	2.8 (2.2)	4.5 (3.8)	4.3 (3.5)	4.0 (2.7)	4.0 (4.1)	3.2 (2.4)
% patients on blood pressure treatment	5.5 (2.0)	5.9 (1.6)	6.0 (1.7)	6.0 (1.5)	5.6 (1.5)	7.4 (2.7)	8.2 (2.6)	7.6 (2.0)	8.2 (1.8)	7.8 (2.2)
% patients with migraines	3.3 (1.3)	3.3 (1.2)	3.6 (1.3)	3.7 (1.2)	3.6 (1.5)	8.7 (3.0)	8.5 (2.7)	9.6 (2.9)	9.3 (2.8)	9.7 (3.6)
% patients with rheumatoid arthritis	0.3 (0.2)	0.3 (0.2)	0.3 (0.1)	0.4 (0.1)	0.4 (0.2)	0.8 (0.3)	0.9 (0.3)	0.9 (0.4)	0.9 (0.3)	0.9 (0.3)

	Male					Female				
	mean (SD))					(mean (SD))				
	Frailty (0~20%) (0.7 ~ 0.9)	**Frailty (20~40%) (0.9 ~ 1.0)**	**Frailty (40~60%) (1.0 ~ 1.0)**	**Frailty (60~80%) (1.0 ~ 1.1)**	**Frailty (80~100%) (1.1 ~ 1.7)**	**Frailty (0~20%) (0.6 ~ 0.9)**	**Frailty (20~40%) (0.9 ~ 1.0)**	**Frailty (40~60%) (1.0 ~ 1.0)**	**Frailty (60~80%) (1.0 ~ 1.1)**	**Frailty (80~100%) (1.1 ~ 1.6)**
% patients with severe mental illness	4.5 (2.7)	5.3 (3.0)	4.8 (2.7)	6.2 (3.0)	6.4 (3.3)	8.1 (5.2)	8.8 (4.9)	10.4 (7.0)	11.1 (6.1)	12.0 (6.5)
% patients with Systemic Lupus Erythematosus	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.1 (0.1)	0.1 (0.1)	0.1 (0.1)	0.1 (0.1)	0.1 (0.1)
SBP										
Average SBP within practice	130.1 (2.6)	130.4 (2.8)	130.6 (3.2)	130.8 (2.6)	130.6 (2.5)	123.4 (3.2)	124.4 (3.2)	123.8 (3.2)	124.8 (2.7)	125.0 (2.5)
% patients with missing SBP	36.2 (9.8)	34.9 (8.4)	36.7 (8.9)	37.5 (8.4)	38.3 (9.4)	14.7 (5.8)	13.7 (5.8)	13.9 (5.3)	14.6 (6.6)	16.5 (6.5)
BMI										
Average BMI when recorded	26.5 (0.7)	26.5 (0.6)	26.7 (0.5)	26.7 (0.6)	26.7 (0.5)	25.7 (1.1)	26.1 (1.0)	26.2 (0.9)	26.3 (0.6)	26.6 (0.7)
% patients with missing BMI	45.6 (12.5)	47.3 (11.2)	47.2 (13.3)	50.2 (10.7)	50.0 (12.1)	30.4 (11.7)	29.1 (13.6)	28.9 (11.8)	31.3 (11.7)	33.6 (12.9)
Cholesterol/HDL ratio										
Average Cholesterol/HDL ratio within practice	4.3 (0.2)	4.4 (0.2)	4.4 (0.3)	4.4 (0.2)	4.4 (0.2)	3.6 (0.2)	3.6 (0.2)	3.6 (0.3)	3.7 (0.2)	3.8 (0.2)
% patients with missing Cholesterol/HDL ratio	65.6 (10.4)	66.3 (8.8)	64.8 (8.9)	69.0 (9.4)	65.5 (8.1)	63.3 (10.4)	60.1 (11.2)	61.8 (10.7)	64.9 (11.0)	62.7 (11.4)
Smoking										
% current-smokers	32.3 (8.0)	32.9 (6.1)	33.5 (6.1)	36.4 (6.9)	38.8 (6.7)	22.1 (5.5)	24.3 (6.5)	24.4 (7.4)	26.5 (5.4)	31.0 (6.7)
% patients with missing smoking status	26.2 (9.4)	28.6 (8.2)	27.7 (9.7)	31.7 (8.8)	33.1 (8.1)	17.4 (7.5)	17.8 (8.4)	16.9 (8.4)	21.0 (8.3)	22.9 (8.8)
Diabetes										
% patients with type 1 diabetes	0.2 (0.1)	0.2 (0.1)	0.2 (0.1)	0.2 (0.1)	0.3 (0.1)	0.2 (0.1)	0.2 (0.1)	0.2 (0.1)	0.2 (0.1)	0.2 (0.1)
% patients with type 2 diabetes	1.2 (0.5)	1.3 (0.4)	1.5 (0.5)	1.6 (0.4)	1.6 (0.4)	1.0 (0.4)	1.1 (0.4)	1.1 (0.4)	1.3 (0.4)	1.3 (0.5)
Ethnicity										
% white patients	83.0 (16.3)	87.4 (15.5)	83.2 (20.0)	89.7 (11.0)	90.1 (12.9)	84.6 (15.9)	86.4 (16.9)	83.9 (18.1)	88.9 (12.1)	90.3 (11.8)
% patients with missing ethnicity	55.7 (21.7)	62.7 (20.0)	57.8 (26.6)	64.8 (23.0)	62.1 (22.4)	54.2 (23.0)	54.9 (26.3)	53.2 (25.1)	61.4 (23.1)	58.9 (24.7)
Townsend (Socioeconomic Status)										
% patients with Townsend score 5 (the most deprived)	16.2 (25.8)	11.9 (21.0)	15.1 (21.9)	14.0 (19.1)	24.5 (20.1)	10.0 (19.5)	14.9 (22.4)	17.2 (24.0)	11.2 (16.4)	25.9 (21.6)

	Male					Female				
	mean (SD))					(mean (SD))				
	Frailty (0~20%) (0.7 ~ 0.9)	**Frailty (20~40%) (0.9 ~ 1.0)**	**Frailty (40~60%) (1.0 ~ 1.0)**	**Frailty (60~80%) (1.0 ~ 1.1)**	**Frailty (80~100%) (1.1 ~ 1.7)**	**Frailty (0~20%) (0.6 ~ 0.9)**	**Frailty (20~40%) (0.9 ~ 1.0)**	**Frailty (40~60%) (1.0 ~ 1.0)**	**Frailty (60~80%) (1.0 ~ 1.1)**	**Frailty (80~100%) (1.1 ~ 1.6)**
% patients with Townsend score missing	0.1 (0.1)	0.1 (0.1)	0.1 (0.1)	0.1 (0.1)	0.2 (1.4)	0.1 (0.1)	0.1 (0.1)	0.1 (0.1)	0.1 (0.1)	0.2 (1.4)

Figure 3.1 shows the relationship between risk factors' dissimilarity between practices (measured by Sáez's metric) and practice statistical frailly. Sáez's metrics for the CVD variables and for their missingness were not related to the statistical frailty of practices (blue lines), indicating that practices with high or low statistical frailty had similar distribution of these risk factors. Only a few variables (including Townsend score) were distributed differently between practices with high or low statistical frailty, but there were differences in the patterns between females and males.

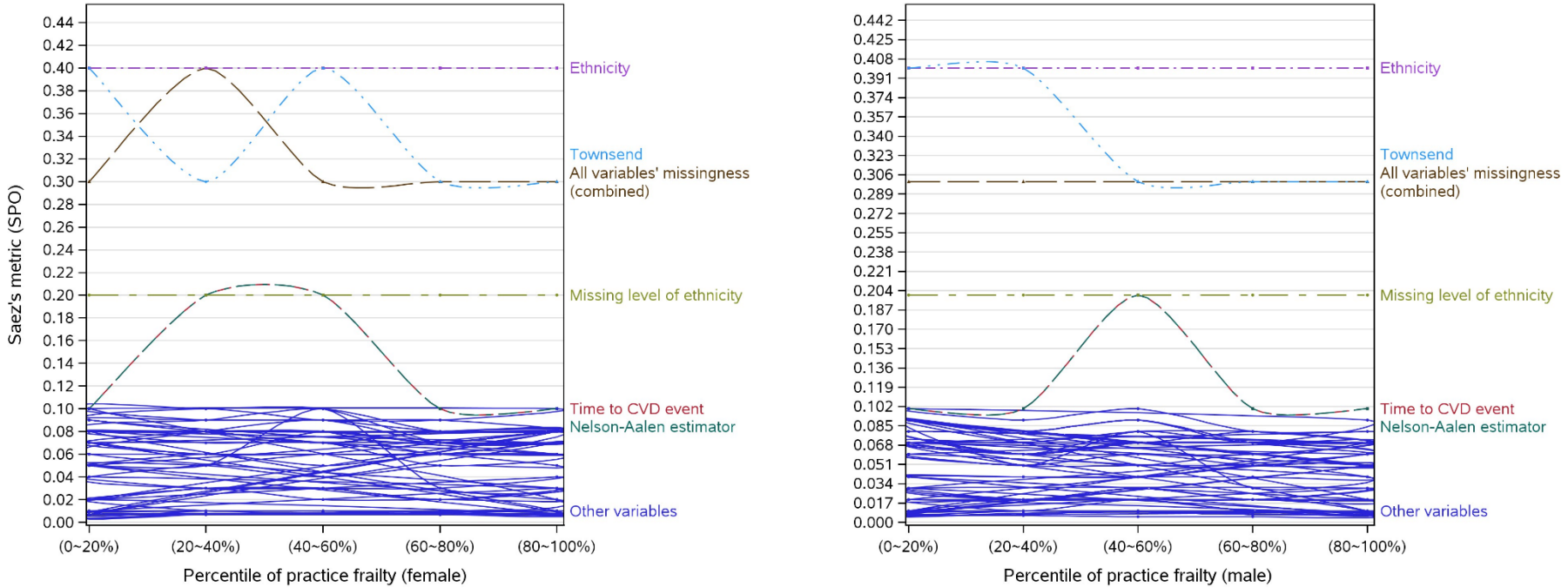


Figure 3.1 Relationship between quintiles of statistical frailty in practices and the stability metrics for QRISK3 CVD predictors and level of missingness

Figure 3.2 (Beeswarm plot) also confirms that there was no visual relationship between practice statistical frailty and most CVD risk factors and their stability metrics (only variables with non-flat trend in Figure 3.1 are shown). Practices were grouped by percentile of CVD predictors or their stability metrics. The Pearson correlation coefficients between practice frailty and practice characteristics were low. The percentage of smokers had the highest correlations of 0.46 (95% CI: 0.38, 0.54) for females and 0.35 (95%CI: 0.26, 0.44) for males.

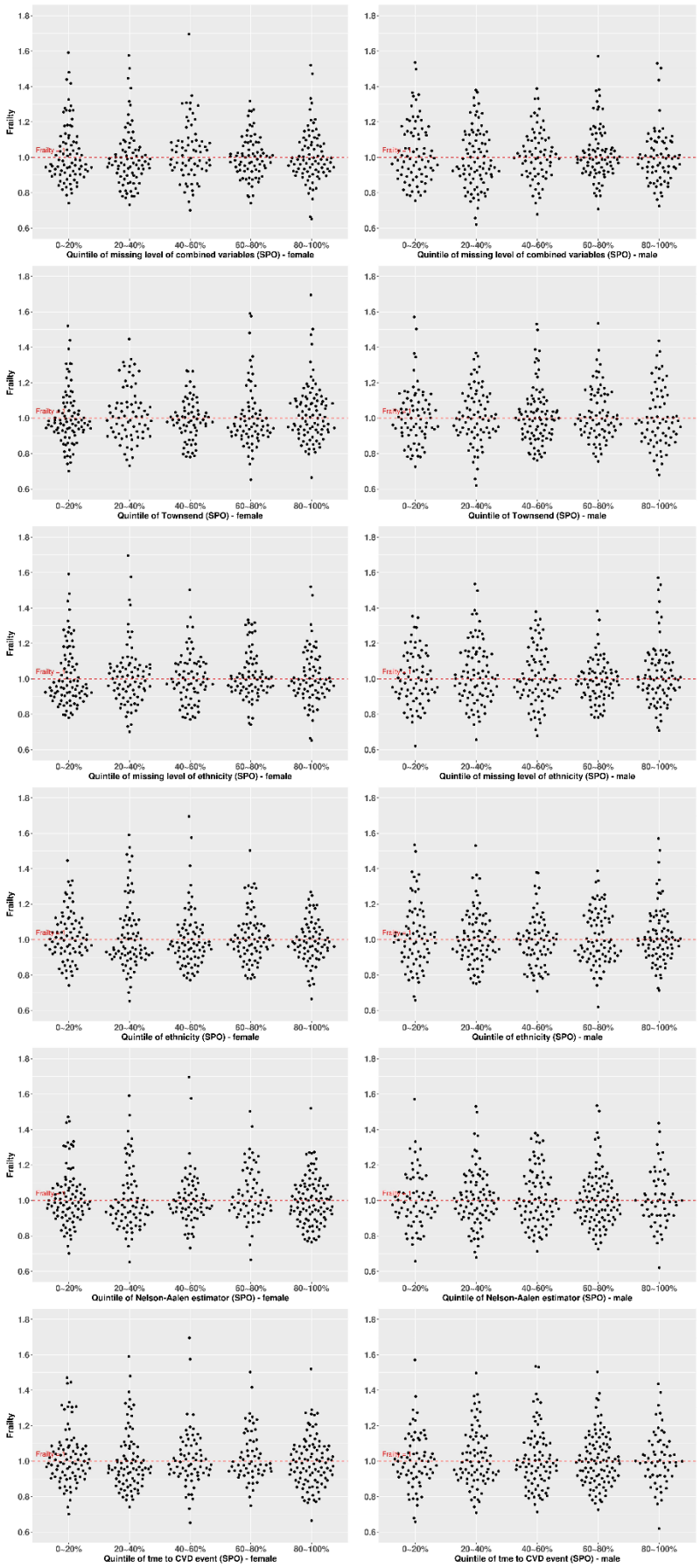


Figure 3.2 Relationship between quintiles of statistical frailty in practices and CVD risk predictors and their stability metrics (SPO) - Beeswarm plot

Figure 3.3 shows that there was no variation across practices in the effects of the risk factors on the risk of CVD outcomes, as the fixed effects of QRISK3 linear predictor was near 1 and the variation of random effects on slope among practices was near 0 (0.000111 for females and 0.000302 for males) in the random slope model. Comparing two reference variations of random slope (0.03 and 0.1), Figure 3.3 shows that there was almost no random slope, indicating similar associations between predictors and CVD outcome among practices.

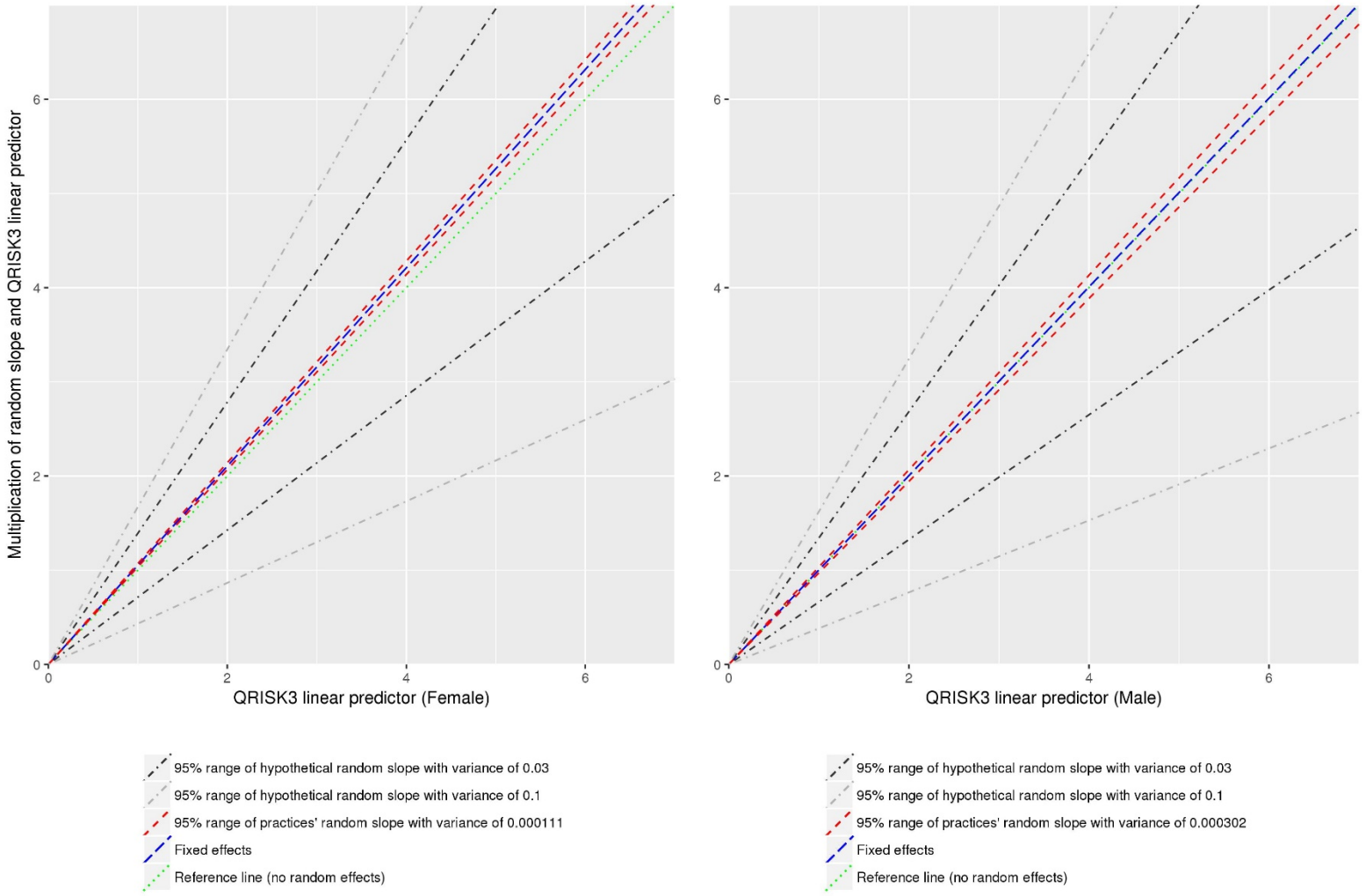


Figure 3.3 Effects of the variability between practices of the QRISK3 linear predictor (random slope)

As shown in Figure 3.4, the incorporation of random slopes into the models (i.e., varying effects of the risk factors on CVD between practices) did not increase the accuracy in individual risk prediction. The distributions of the individual risk predictions for patients with a QRISK3 of 10% were comparable between the random slope and the random intercept model. For patients with a QRISK3 predicted 10% risk, the random slope alone would only change the patients' risk by an absolute 0.6% between practices on 97.5% and 2.5% random slope percentile (eFigure3.9.3). The effects of variation of random slope on individual patients' risk was small compared with the effects of the random intercept, which could change patient's risk from 10% to a range of 5% and 17%.

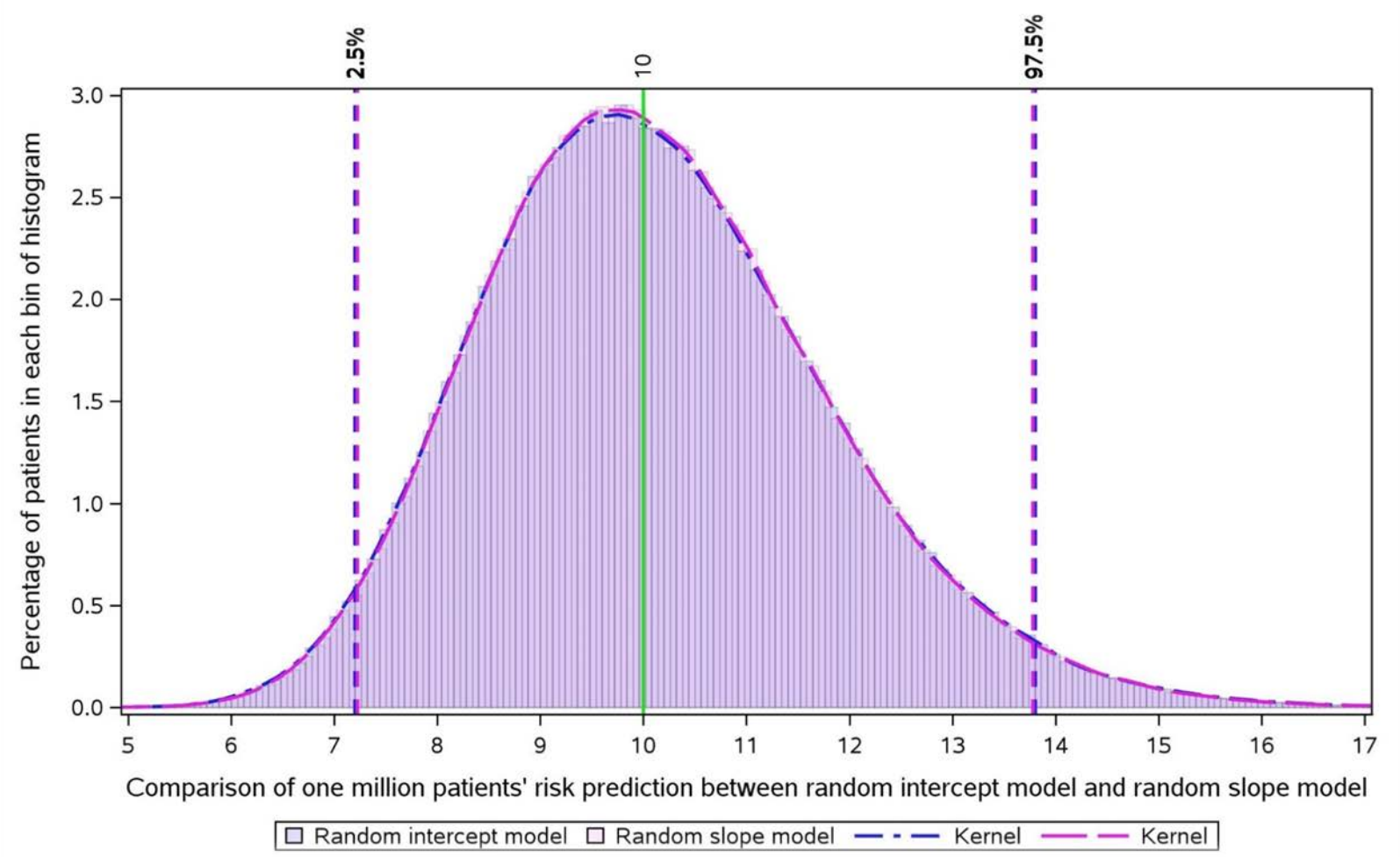


Figure 3.4 Comparison of the CVD risk predictions between the random intercept and slope models for patients with a QRISK3 risk of 10% (in a cohort of one million patients with 50% males and 50% females)

3.5 Discussion

3.5.1 Key results

This study found that the observed variation in data quality between general practices did not explain the unmeasured heterogeneity in QRISK3 risk prediction across practices (miscalibration at practice level). Specifically, practices with higher or lower statistical frailty had comparable indicators of data quality, including those based on more innovative techniques (Sáez metric). In addition, the effects of the QRISK3 predictors on CVD risk were comparable across practices despite these differences in data quality, since the random slope models found little variation of the beta coefficients across practices.

3.5.2 Strength and limitation

This study was based on a very large patient cohort It also used the innovative Sáez's metric, which could quantify the distribution-dissimilarity comparing to a test result from classical methods[30]. There are several limitations in this study. We considered several aspects of practice variability that covered important areas identified in literature[6, 13], but there may be other aspects of data quality. However, the other aspects of data quality might not have major effects, otherwise it would be reflected in the distribution-dissimilarity of CVD variables and then being captured by Sáez's metric. Sáez's metric, which measured the CVD risk factors' distribution-dissimilarity among practices, has information loss as it suffers the "curse of dimensionality"[31]. With more practices, there are more dimensions but this needs to be reduced to estimate summarised statistics resulting in loss of information. Another limitation concerned the estimation of the variation of the random slopes. One thousand bootstrap samples of 40% of the practices were used to estimate this rather than the whole dataset because the current random slope model algorithm[28] has computational difficulty to reach the converge criteria when there is only a small effect on the slopes with greater number of practices. The sensitivity analysis (eFigure 2) shows consistent results of variation of random slope among samples of 20%, 40%, 50% and 60% of total practices suggesting that the variation estimate is accurate.

3.5.3 Interpretation

Data quality is an important aspect of practice variability as it influences the performance and generalisability of a model. Damen et al.[6] discussed that data quality limits a model's generalisability and models developed from poor data would generally have poor performance. However, this study shows that although there was a large variation in data quality among practices, it did not affect the accuracy of the risk prediction on individual patients. This indicates data quality among practices were well handled by the data cleaning methods (including multiple imputation for missing values).

Wynants et al.[13] suggested that patient case mix or true variation of association between outcome and predictors might be related to the variation of a model's performance in a heterogeneous setting. Patient case mix was already adjusted in QRISK3[8], and the present study found no random slopes for beta coefficients across practices. This indicates that the effects of QRISK3 predictors on CVD risk were comparable across practices. The comparison between the random slope and random intercept models found that the effects of practice variability on individual patients was fully explained by the random intercept, i.e. the unmeasured heterogeneity in CVD incidence between practices and deviation from the baseline hazard.

A recent study[32] found that the addition of another risk factor (standard deviation of blood pressure) to QRISK3 did not improve model performance despite it being significantly related to CVD. Previous studies[11] discovered models with similar discrimination and calibration could predict the same patient differently using the current model's predictors. Therneau[33] demonstrated an example that the effects of random intercept could come from unknown covariates missed by a model. This study found that unexplained heterogeneity at practice level cannot be resolved using current measured risk factors. Therefore, this study supports the conclusion from Damen et al.[6] that current CVD models lack information on other important CVD risk factors, e.g. those that better measure the heterogeneity in incidence between different areas.

3.5.4 Implications for Research and Practice

This study found that variation between practices in data quality and effects of CVD predictors were not associated with the considerable heterogeneity in CVD incidence. This suggests that a further study might focus on determining whether the

CVD risk prediction models can be extended with new risk factors from patient level CVD risk factors (e.g. biomarkers) or practice level, which could reduce the unmeasured heterogeneity in CVD incidence across practices. Further research could consider more individual level based methods, such as a Bayesian clinical reasoning model[34] and machine learning models[35], as this study and other findings[11, 36, 37] show that Cox models with similar conventional model performance metrics (C-stat[38] and calibration) could predict inconsistent risk to the same patients. Alternatively, new statistics might be required to measure a population based model's performance on an individual level.

In conclusion, the considerable unmeasured heterogeneity in CVD incidence between practices was not explained by variations in data quality or effects of risk factors. QRISK3 risk prediction should be supplemented with clinical judgement and evidence of additional risk factors.

3.6 Funding

This study was funded by China Scholarship Council (PhD studentship of Yan Li).

3.7 Acknowledgements

This study is based on data from the Clinical Practice Research Datalink obtained under license from the UK Medicines and Healthcare products Regulatory Agency. The protocol for this work was approved by the independent scientific advisory committee for Clinical Practice Research Datalink research (No 17_125RMn2). The data are provided by patients and collected by the NHS as part of their care and support. The Office for National Statistics (ONS) is the provider of the ONS Data contained within the CPRD Data. Hospital Episode Data and the ONS Data Copyright © (2014), are re-used with the permission of The Health & Social Care Information Centre. All rights reserved. The interpretation and conclusions contained in this study are those of the authors alone. There are no conflicts of interest among the authors.

3.8 Summary points

What was already known

Risk prediction tools based on routinely collected data are used by clinicians to predict a 10-year CVD risk for individual patients.
A previous study (accepted by scientific reports) found that there was considerable variability between clinical sites in the robustness of individual risk predictions. This heterogeneity in incidence between sites is not incorporated into current risk prediction approaches.

What this study has added

There was substantial heterogeneity between practices in the incidence of cardiovascular disease (CVD) which was not explained by a commonly used risk prediction model (QRISK3).
Data quality, as measured by probabilistic indicators based on information theory and geometry, varied substantially between clinical sites.
This study adds that this unmeasured heterogeneity in CVD incidence was not explained by variations in data quality or effects of risk factors between clinical sites.

3.9 References

1. Phe. Action plan for cardiovascular prevention: 2017 to 2018. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/648190/cardiovascular_disease_prevention_action_plan_2017_to_2018.pdf. Accessed December 6, 2017.

2. Piepoli MF, Hoes AW, Agewall S, et al. 2016 European Guidelines on cardiovascular disease prevention in clinical practice. *Eur Heart J*. 2016;37(29):2315-2381. doi:10.1093/eurheartj/ehw106

3. Cardiovascular disease prevention overview - NICE Pathways. https://pathways.nice.org.uk/pathways/cardiovascular-disease-prevention. Accessed December 6, 2017.

4. Guidelines P, Risk C. Prevention of Cardiovascular Disease Prevention of Cardiovascular Disease. *World Heal Organ*. 2007:1-30. doi:10.1093/innovait/inr119

5. Karmali KN, Lloyd-Jones DM. Implementing Cardiovascular Risk Prediction in Clinical Practice: The Future Is Now. *J Am Heart Assoc*. 2017;6(4). doi:10.1161/JAHA.117.006019

6. Damen JAAG, Hooft L, Schuit E, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. *BMJ*. 2016;353:i2416. doi:10.1136/BMJ.I2416

7. Bitton A, Gaziano TA. The Framingham Heart Study's impact on global risk assessment. *Prog Cardiovasc Dis*. 2010;53(1):68-78. doi:10.1016/j.pcad.2010.04.001

8. Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. *Bmj*. 2017;2099(May):j2099. doi:10.1136/bmj.j2099

9. Conroy RM, Pyörälä K, Fitzgerald AP, et al. Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project on behalf of the SCORE project group 1. *Eur Heart J*. 2003;24:987-1003. doi:10.1016/S0195-668X(03)00114-3

10. Cox DR. Regression Models and Life-Tables. *J R Stat Soc Ser B*. 1972;34(2):187-220. http://links.jstor.org/sici?sici=0035-9246%281972%2934%3A2%3C187%3ARMAL%3E2.0.CO%3B2-6. Accessed January 15, 2018.

11. Yan Li, Matthew Sperrin, Miguel Belmonte, Alexander Pate, Darren M Ashcroft TP van S. Do population-level risk prediction models that use routinely collected health data reliably predict individual risks? *Accepted Scientific Reports*. 2019.

12. Sáez C, Robles M, García-Gó Mez JM, García-Gómez JM. Stability metrics for multi-source biomedical data based on simplicial projections from probability distribution distances. *Stat Methods Med Res*. 2017;26(1):312-336. doi:10.1177/0962280214545122

13. Wynants L, Riley RD, Timmerman D, Van Calster B. Random-effects meta-analysis of the clinical utility of tests and prediction models. *Stat Med*. 2018;37(12):2034-2052. doi:10.1002/sim.7653

14. *Surv Surveillance Report 2018-Eillance Report 2018-Cardio Cardiovascular Disease: Risk Assessment Vascular Disease: Risk Assessment and Reduction, Including Lipid and Reduction, Including Lipid Modification (2014) NICE Guideline Modification (2014) NICE Guideline CG181 CG181 Surveillance Report Contents Contents*.; 2018. https://www.nice.org.uk/terms-and-. Accessed May 20, 2019.

15. *2018/19 General Medical Services (GMS) Contract Quality and Outcomes*

Framework (QOF).; 2018. https://www.nhsemployers.org/-/media/Employers/Documents/Primary-care-contracts/QOF/2018-19/2018-19-QOF-guidance-for-stakeholders.pdf. Accessed May 20, 2019.

16. Herrett E, Gallagher AM, Bhaskaran K, et al. Data Resource Profile: Clinical Practice Research Datalink (CPRD). *Int J Epidemiol*. 2015;44(3):827-836. doi:10.1093/ije/dyv098

17. Clinical Practice Research Datalink - CPRD. https://www.cprd.com/intro.asp. Accessed August 20, 2017.

18. Hippisley-Cox J, Coupland C, Brindle P. The performance of seven QPrediction risk scores in an independent external sample of patients from general practice: a validation study. *BMJ Open*. 2014;4(8):e005809. doi:10.1136/bmjopen-2014-005809

19. van Staa T-P, Gulliford M, Ng ES-W, Goldacre B, Smeeth L. Prediction of cardiovascular risk using Framingham, ASSIGN and QRISK2: how well do they predict individual rather than population risk? *PLoS One*. 2014;9(10):e106455. doi:10.1371/journal.pone.0106455

20. Hougaard P. Frailty models for survival data. *Lifetime Data Anal*. 1995;1(3):255-273. http://www.ncbi.nlm.nih.gov/pubmed/9385105. Accessed August 23, 2018.

21. Sáez C, Zurriaga O, Pérez-Panadés J, Melchor I, Robles M, García-Gómez JM. Applying probabilistic temporal and multisite data quality control methods to a public health mortality registry in Spain: A systematic approach to quality control of repositories. *J Am Med Informatics Assoc*. 2016;23(6):1085-1095. doi:10.1093/jamia/ocw010

22. Fuglede B, Topsoe F. Jensen-Shannon divergence and Hilbert space embedding. In: *International Symposium OnInformation Theory, 2004. ISIT 2004. Proceedings.* IEEE; :30-30. doi:10.1109/ISIT.2004.1365067

23. Austin PC. A Tutorial on Multilevel Survival Analysis: Methods, Models and Applications. *Int Stat Rev*. 2017;85(2):185-203. doi:10.1111/insr.12214

24. Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. *BMJ*. 2017;357(3):j2099. doi:https://doi.org/10.1136/bmj.j2099

25. Liu J, Tang W, Chen G, Lu Y, Feng C, Tu XM. Correlation and agreement: overview and clarification of competing concepts and measures. *Shanghai Arch psychiatry*. 2016;28(2):115-120. doi:10.11919/j.issn.1002-0829.216045

26. *Package "Beeswarm."*; 2016. http://www.cbs.dtu.dk/~eklund/beeswarm/. Accessed November 28, 2018.

27. Previous releases of R for Windows. https://cran.r-project.org/bin/windows/base/old/. Accessed April 28, 2019.

28. Therneau T, Clinic M. Mixed Effects Cox Models. 2018. https://cran.r-project.org/web/packages/coxme/vignettes/coxme.pdf. Accessed April 10, 2018.

29. Schafer JL (Joseph L. *Analysis of Incomplete Multivariate Data*. Chapman & Hall; 1997.

30. Sáez C, Robles M, García-Gómez JM. Stability metrics for multi-source biomedical data based on simplicial projections from probability distribution distances. *Stat Methods Med Res*. 2017;26(1):312-336. doi:10.1177/0962280214545122

31. Aggarwal CC, Hinneburg A, Keim DA. *On the Surprising Behavior of Distance Metrics in High Dimensional Space*. https://bib.dbvis.de/uploadedFiles/155.pdf. Accessed March 4, 2019.

32. Stevens SL, McManus RJ, Stevens RJ. The utility of long-term blood pressure variability for cardiovascular risk prediction in primary care. *J Hypertens*. September 2018:1. doi:10.1097/HJH.0000000000001923

33. Therneau TM, Grambsch PM. *Modeling Survival Data: Extending the Cox Model*. New York, NY: Springer New York; 2000. doi:10.1007/978-1-4757-3294-8

34. Liu Y-M, Chen SL-S, Yen AM-F, Chen H-H. Individual risk prediction model for incident cardiovascular disease: A Bayesian clinical reasoning approach. *Int J Cardiol*. 2013;167(5):2008-2012. doi:10.1016/J.IJCARD.2012.05.016

35. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? Liu B, ed. *PLoS One*. 2017;12(4):e0174944. doi:10.1371/journal.pone.0174944

36. Dzeshka MS, Gill PS, Lip GYH. Cardiovascular risk prediction: balancing complexity against simple practicality. *Br J Gen Pract*. 2015;65(630):4-5. doi:10.3399/bjgp15X683005

37. Gray BJ, Bracken RM, Turner D, et al. Predicted 10-year risk of cardiovascular disease is influenced by the risk equation adopted: a cross-sectional analysis. *Br J Gen Pract*. 2014;64(627):e634-40. doi:10.3399/bjgp14X681805

38. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation*. 2007;115(7):928-935. doi:10.1161/CIRCULATIONAHA.106.672402

3.9 Supplementary Online Content

eAppendix 3.9.1. Interpretation of appendix tables and figures.

eAppendix 3.9.2. Saez's metric of distribution-dissimilarity

References

eTable 3.9.1. Stability metrics of all QRISK3 CVD predictors and their missing level on practice level

eFigure 3.9.1. Relationship between practice frailty and CVD risk predictors and their stability metrics (Beeswarm plot).

eFigure 3.9.2.1. Effects of practice variability on QRISK3 linear predictor (random slope) (20% of overall CPRD practices)

eFigure 3.9.2.2. Effects of practice variability on QRISK3 linear predictor (random slope) (50% of overall CPRD practices)

eFigure 3.9.2.3. Effects of practice variability on QRISK3 linear predictor (random slope) (60% of overall CPRD practices)

eFigure 3.9.3. Difference of individual patients' prediction between practice with 2.5% random slope and 97.5% slope and a random selected fixed random intercept

eAppendix 3.9.1. Interpretation of appendix tables and figures.

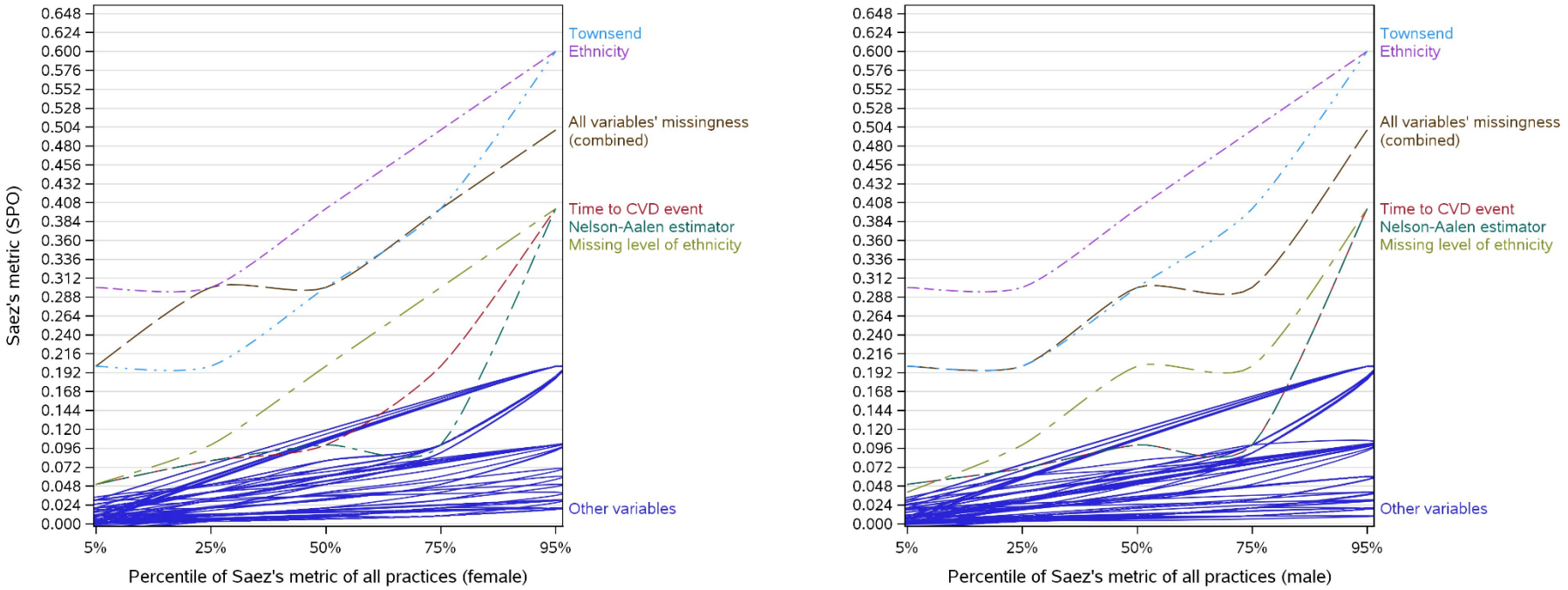
eTable 3.9.1 summarises the distribution-dissimilarity of all CVD risk factors and their missingness using Sáez's proposed metrics global probabilistic deviation (GPD) and source probabilistic outlyingness (SPO). GPD measures risk factors' overall outlyingness of practices, which ranges from 0 to 1 and the closer to 1 means there is more variation of practices. SPO measures the latent distance of risk factors' distribution between one practice to the average of overall practice, which also ranges from 0 to 1 and closer to 1 means the practice is more far away from the average. The table shows part of CVD risk factors (e.g. Rheumatoid arthritis) among practices are very stable, which means they have similar distribution among practices. Other variables, such as ethnicity, Townsend and missing level of ethnicity, were unstable among practices, which means the distribution of these variables has substantial variation among practices.

eTable 3.9.1. Stability metrics of all QRISK3 CVD predictors and their missing level on practice level

	Male							Female					
	GPD	**5th**	**25th**	**50th**	**75th**	**95th**		**GPD**	**5th**	**25th**	**50th**	**75th**	**95th**
CVD risk factors (distribution dissimilarity to the overall practice average)													
Atrial fibrillation	0.02	0.00	0.01	0.01	0.02	0.04		0.02	0.00	0.01	0.01	0.02	0.03
Whether patients on atypical antipsychotic medication	0.02	0.00	0.01	0.01	0.02	0.03		0.02	0.00	0.00	0.01	0.02	0.03
Chronic kidney disease (stage 3, 4 or 5)	0.02	0.00	0.01	0.01	0.02	0.04		0.03	0.00	0.01	0.02	0.03	0.05
CVD censors	0.08	0.02	0.02	0.03	0.05	0.11		0.08	0.02	0.02	0.03	0.05	0.10
Time to CVD event	0.19	0.05	0.07	0.10	0.14	0.40		0.20	0.05	0.08	0.11	0.15	0.42
Cholesterol	0.08	0.02	0.03	0.05	0.07	0.11		0.10	0.02	0.04	0.06	0.09	0.17
Regular steroid tablets	0.01	0.00	0.00	0.01	0.01	0.02		0.01	0.00	0.00	0.01	0.01	0.02
Erectile dysfunction	0.03	0.00	0.01	0.02	0.03	0.06		0.01	0.00	0.00	0.01	0.01	0.02
Angina or heart attack in a 1st degree relative < 60	0.06	0.01	0.02	0.03	0.06	0.10		0.07	0.01	0.02	0.04	0.06	0.11
HDL	0.11	0.03	0.06	0.07	0.10	0.15		0.12	0.03	0.05	0.08	0.11	0.17
Blood pressure treatment	0.03	0.00	0.01	0.02	0.03	0.06		0.04	0.00	0.01	0.02	0.04	0.07
Migraines	0.03	0.00	0.01	0.02	0.03	0.06		0.05	0.00	0.02	0.03	0.05	0.10
Nelson-Aalen estimator	0.18	0.05	0.07	0.10	0.14	0.36		0.19	0.05	0.08	0.11	0.14	0.36
Rheumatoid arthritis	0.01	0.00	0.00	0.01	0.01	0.02		0.02	0.00	0.00	0.01	0.01	0.03
Systemic Lupus Erythematosus	0.01	0.00	0.00	0.00	0.01	0.01		0.01	0.00	0.00	0.01	0.01	0.02
Severe mental illness (this includes schizophrenia, bipolar disorder and moderate/severe depression)	0.07	0.01	0.02	0.04	0.07	0.10		0.10	0.01	0.04	0.07	0.10	0.15
Type 1 diabetes	0.01	0.00	0.00	0.01	0.01	0.02		0.01	0.00	0.00	0.01	0.01	0.02
Type 2 diabetes	0.02	0.00	0.01	0.01	0.02	0.03		0.02	0.00	0.01	0.01	0.02	0.04

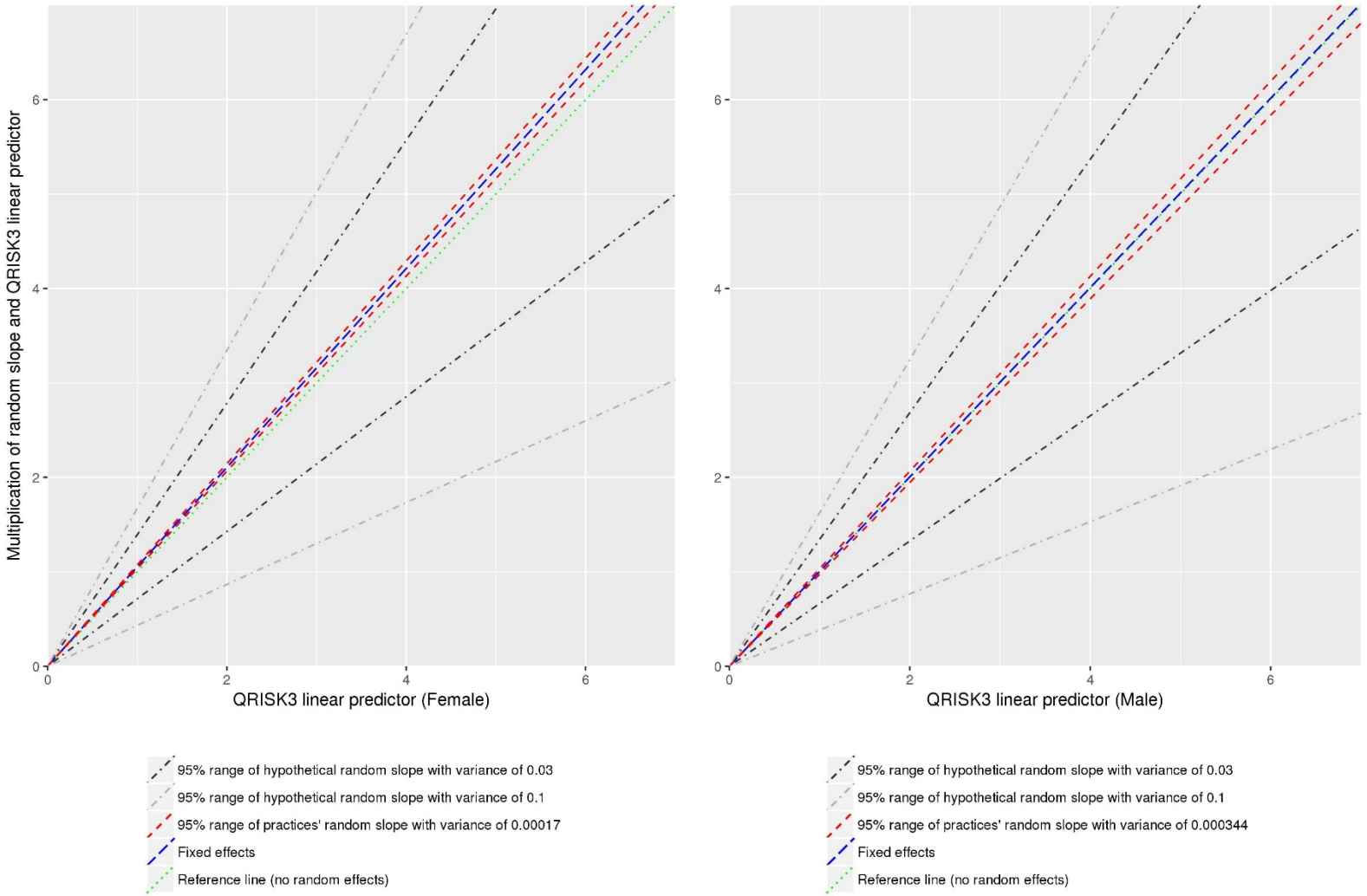
	Male							Female					
	GPD	**5th**	**25th**	**50th**	**75th**	**95th**		**GPD**	**5th**	**25th**	**50th**	**75th**	**95th**
Age	0.11	0.02	0.04	0.07	0.10	0.17		0.12	0.03	0.05	0.07	0.12	0.20
BMI	0.09	0.02	0.04	0.06	0.08	0.12		0.09	0.02	0.04	0.06	0.08	0.14
Cholesterol and HDL	0.11	0.03	0.05	0.07	0.09	0.15		0.12	0.03	0.05	0.07	0.11	0.17
Ethnicity	0.57	0.27	0.32	0.38	0.47	0.63		0.61	0.29	0.34	0.40	0.50	0.63
SBP	0.13	0.03	0.06	0.08	0.11	0.18		0.12	0.03	0.05	0.08	0.10	0.17
Standard deviation of SBP	0.09	0.02	0.04	0.06	0.08	0.13		0.08	0.02	0.04	0.05	0.07	0.12
Smoking	0.10	0.02	0.04	0.06	0.09	0.16		0.11	0.02	0.04	0.06	0.10	0.19
Townsend	0.48	0.17	0.24	0.32	0.43	0.58		0.48	0.17	0.24	0.31	0.43	0.56
Missing level (distribution dissimilarity to the overall practice average)													
Missing level of Cholesterol	0.07	0.01	0.02	0.04	0.08	0.14		0.09	0.01	0.03	0.05	0.09	0.16
Missing level of HDL	0.09	0.01	0.03	0.05	0.09	0.18		0.11	0.01	0.03	0.06	0.11	0.21
Missing level of Nelson-Aalen estimator	0.04	0.00	0.01	0.02	0.03	0.06		0.03	0.00	0.01	0.02	0.03	0.06
Missing level of BMI	0.12	0.01	0.03	0.07	0.12	0.22		0.13	0.01	0.04	0.08	0.13	0.24
Missing level of ratio of Cholesterol and HDL	0.09	0.01	0.03	0.05	0.09	0.18		0.11	0.01	0.03	0.06	0.11	0.21
Missing level of ethnicity	0.26	0.04	0.10	0.16	0.24	0.43		0.28	0.05	0.11	0.18	0.25	0.44
Missing level of SBP	0.09	0.01	0.02	0.05	0.09	0.17		0.08	0.01	0.02	0.05	0.08	0.14
Missing level of standard deviation of SBP	0.07	0.00	0.02	0.04	0.07	0.14		0.08	0.01	0.02	0.05	0.08	0.15
Missing level of smoking	0.10	0.01	0.03	0.06	0.10	0.18		0.11	0.01	0.03	0.06	0.11	0.19
Missing level of townsend	0.02	0.00	0.00	0.01	0.02	0.02		0.01	0.00	0.00	0.01	0.02	0.02

eFigure 3.9.1 visualises the distribution-dissimilarity of all CVD risk factors and their missingness using percentile of Sáez's proposed metrics SPO strata by gender. The result is consistent with eTable 3.9.1, as part of risk factors are very stable among practices (blue lines), and other variables such as Townsend and ethnicity has substantial variation among practices.

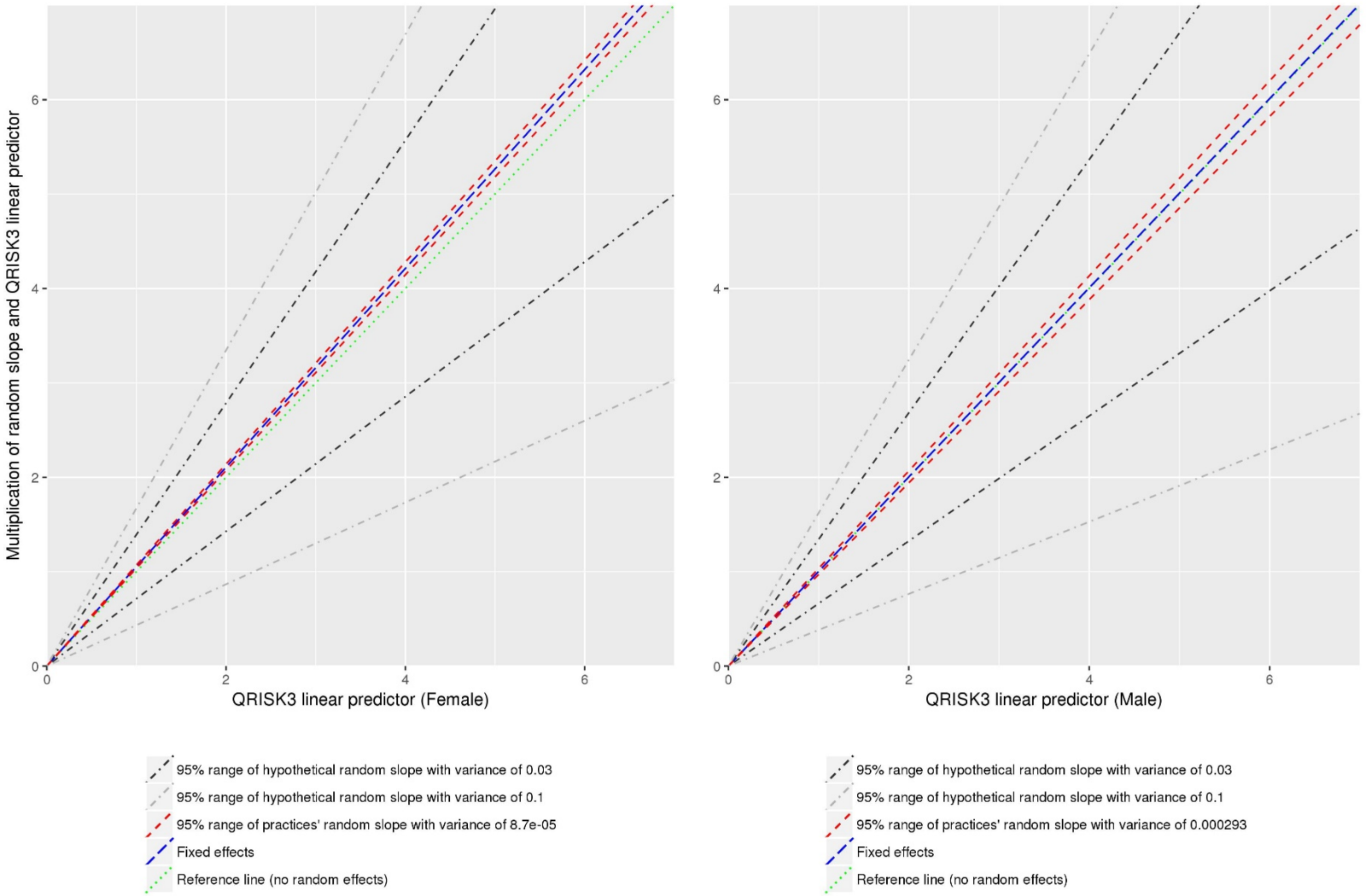


eFigure 3.9.1 Stability metrics of all QRISK3 CVD predictors and their missing level on practice level

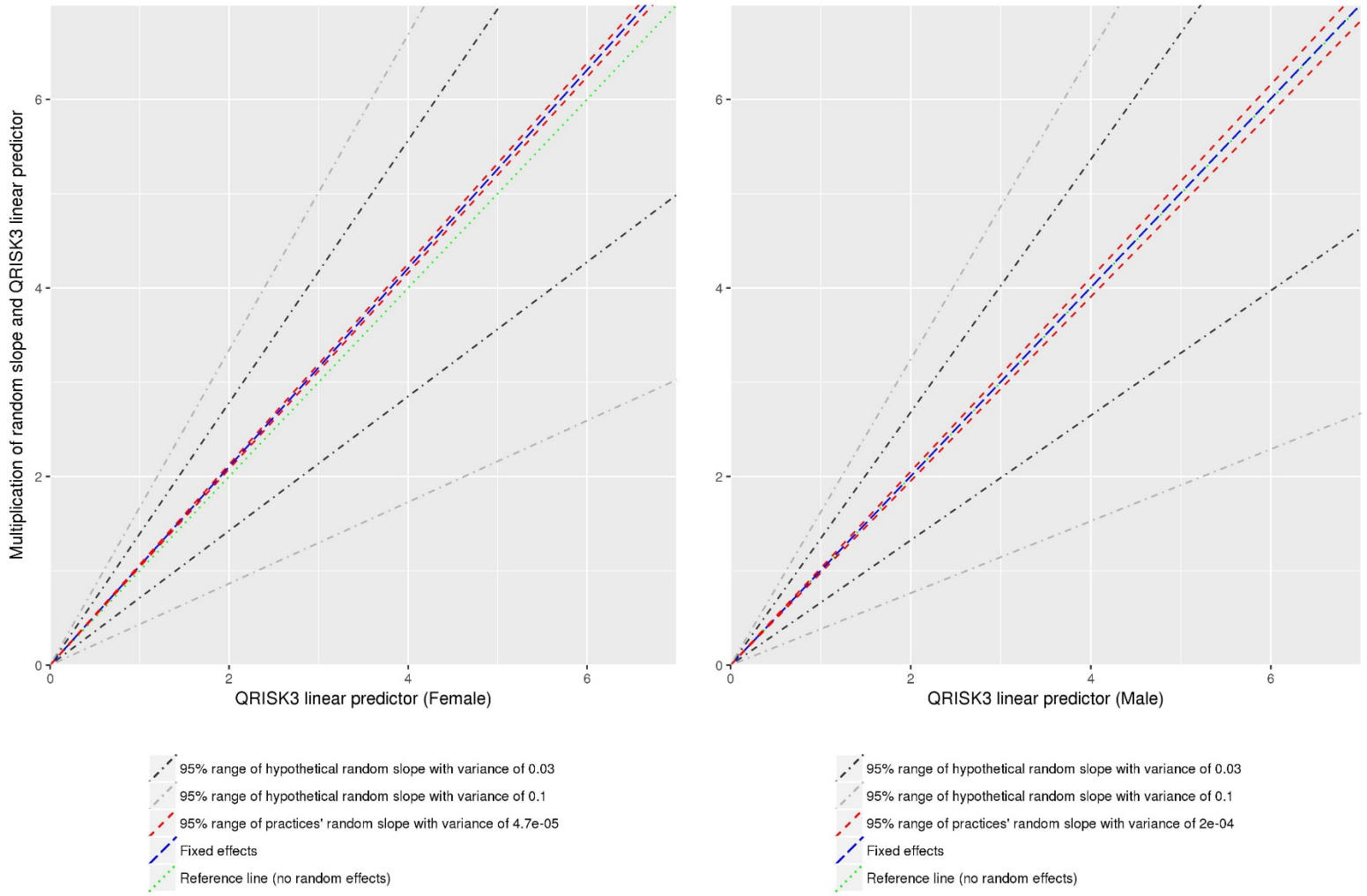
Preliminary analysis showed that the variation of random slopes of full CPRD practices was too small and current statistical package would take very long to calculate results for full practices, so the study estimated practices' variation of random slope by averaging practices' variation of random slope of 1000 random samples (each sample contained 40% practices of all CPRD practices). Sensitive analysis (**eFigure 3.9.2**) shows that there is no difference of the average variation of random slope among different sample size of practice (20%, 40%, 50% and 60% percent of full CPRD practices). All of them shows that there is no variation of random slope among practices, which suggests all practices have similar association between predictors and outcome. This also suggests all of samples are a representative sample of CPRD practices, just as CPRD is a representative sample of the whole UK practices.



eFigure 3.9.2.1 Effects of practice variability on QRISK3 linear predictor (random slope) (20% of overall CPRD practices)

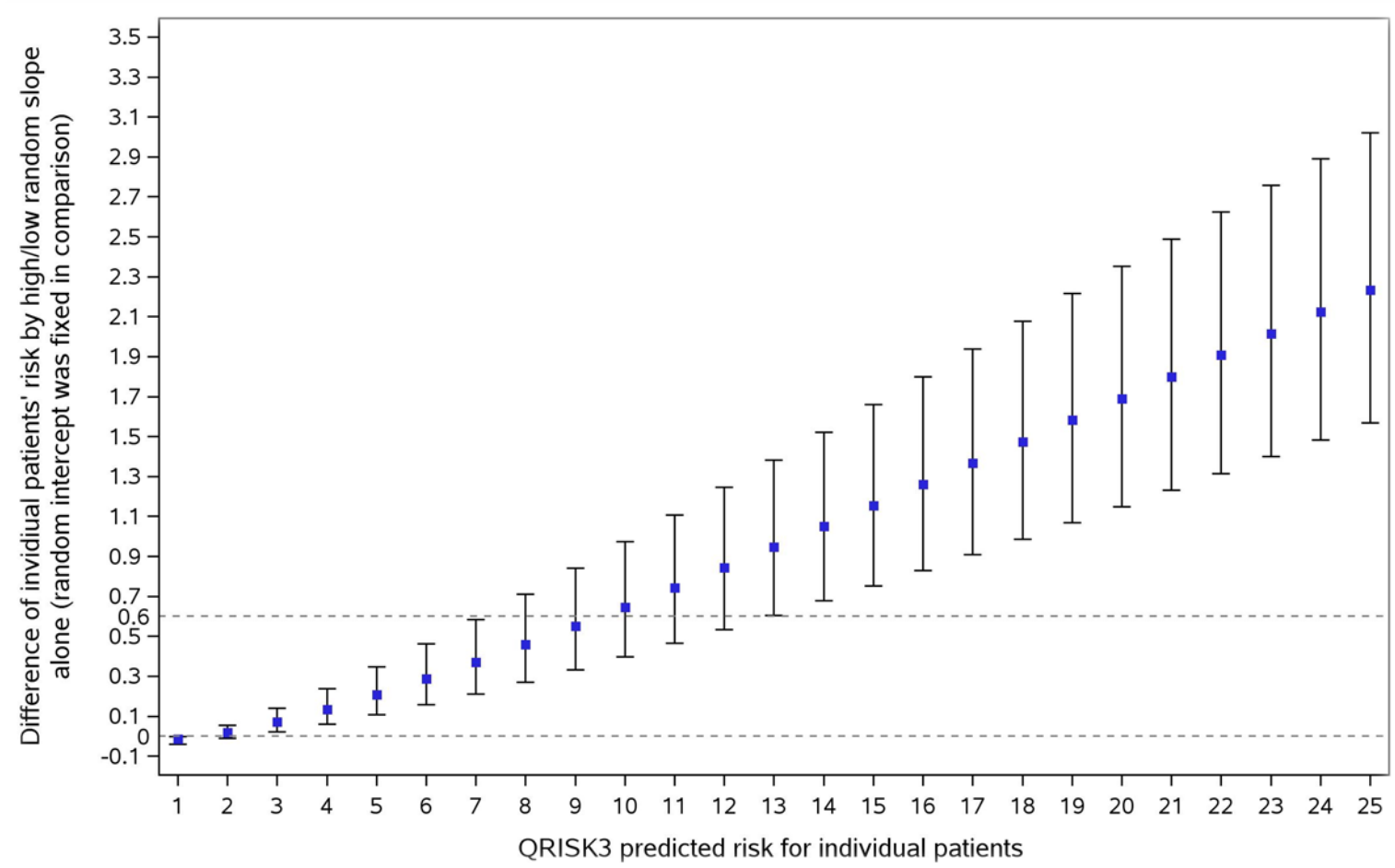


eFigure 3.9.2.2 Effects of practice variability on QRISK3 linear predictor (random slope) (50% of overall CPRD practices)



eFigure 3.9.2.3 Effects of practice variability on QRISK3 linear predictor (random slope) (60% of overall CPRD practices)

eFigure 3.9.3 shows that for patients with a QRISK3 predicted 10% risk, random slope alone would only change the patients' risk by absolute 0.6% between practices on 97.5% and 2.5% random slope percentile. The effects of variation of random slope on individual patients' risk is small comparing to that random intercept could change patient's risk from 10% a range of 5% and 17%.The effects of random slope on individual patients however increases with the increase of patients predicted risk by QRISK3, but it would not affect patients' classification at most of the time. For example, although patients' risk could change about 2% when they have 25%, but the patients would still be prescribed statin after change. Also, the larger patients' predicted risk by QRISK3 means the larger linear predictor which then enlarges the effects of random slope through exponential function from Cox model. Consider following empirical example which compares exp (10*1.05) - exp (10*1.01) = 11972.49 to exp (20*1.05) – exp (20*1.01) = 726233627. We think 20 is a linear predictor for patients with a very high risk, and 10 is for a patient with low risk. Ignoring random intercept here and think 1.01 is a sum of fixed and random slope. Say the fix slope is 1, and 97.5 random slope is 0.05 and 2.5% random slope is 0.01. We can see the larger linear predictor enlarges the differences because of exponential function here.



eFigure 3.9.3. Difference of individual patients' prediction between practice with 2.5% random slope and 97.5% slope and a random selected fixed random intercept

eAppendix 3.9.2. Technical details of Sáez's metric of distribution-dissimilarity

Sáez proposed non-parametric information theory metrics to quantify the distribution-dissimilarity of single or multiple variables among different practice (sites) [1]. Saez quantified the distribution-dissimilarity of variables using Jensen–Shannon divergence (JSD) [2]. JSD (ranges from 0 to 1) calculates an information distance (divergence) of the variable's probability distribution in different practices (sites), which measures the distribution-dissimilarity of a variable among different practices (sites). Once the information distance of a variable between all pair of practices were acquired, Euclidean embedding [3] and simplex [4] theory could be used to construct a coordinate for each practice based on the information distance among them. Based on the geometry theory, the coordinate of a latent center (centroid [4]) could be calculated by averaging all practices' coordinates. The centroid represents a latent average distribution of the variable among all practices, so the information distance between one practice to the centroid quantifies the dissimilarity of variable's distribution of one practice to the overall average. By normalising this distance (so the information distance of different variables is comparable), Sáez proposed source probabilistic outlyingness (SPO) metric, which ranges from 0 to 1 and the higher means the site is more far away from centroid, to quantify the distribution-dissimilarity of variable from one practice to the overall average. Sáez also proposed global probabilistic deviation (GPD), also ranges from 0 to 1 and the higher means the more variation of the variable among practices, to quantify variable's the overall distribution stability among practices [1]. For multiple variables, dimension reduction method such as principle component analysis (PCA) [3] and factor analysis [3], could be used to construct three or four independent principle components to represent the overall variation of multiple variables. Joint probabilities could be calculated using these principle components' distribution, and then JSD of joint probabilities could be calculated to represent the combined distribution-dissimilarity of multiple variables among practices. Take an example to calculate SPO of age for 392 practices. JSD was first calculated between each pair of practices to quantify the distribution dissimilarity of age. Once divergences (distances) of distribution of age (i.e. JSD of age) were acquired, coordinates of each practice could be acquired by Euclidean embedding with simplex theory. With the geometry theory, the coordinate of a latent centre (centroid) of age distribution could be calculated by averaging all practices' coordinates. The distance of each practice to the centroid quantifies the dissimilarity of age distribution of each practice to the average age distribution. SPO of age for each practice was then calculated by normalising this distance (so SPO of different predictors can be compared).

References

1. Sáez C, Robles M, García-Gó Mez JM, García-Gómez JM. Stability metrics for multi-source biomedical data based on simplicial projections from probability distribution distances. *Stat Methods Med Res*. 2017;26(1):312-336. doi:10.1177/0962280214545122

2. Fuglede B, Topsoe F. Jensen-Shannon divergence and Hilbert space embedding. In: *International Symposium OnInformation Theory, 2004. ISIT 2004. Proceedings.* IEEE; :30-30. doi:10.1109/ISIT.2004.1365067

3. Gower JC. Multivariate Analysis and Multidimensional Geometry. *Stat.* 1967;17(1):13. doi:10.2307/2987199

4. Weisstein EW. Simplex. http://mathworld.wolfram.com/Simplex.html. Accessed November 29, 2018.

Blank page

Chapter 4 The consistency of a variety of machine learning and statistical models in predicting clinical risks of individual patients: A Longitudinal cohort study using cardiovascular disease as exemplar

Yan Li[1], Matthew Sperrin[1], Darren M Ashcroft[2,3], Tjeerd Pieter van Staa[1,4,5,]

[1]Health e-Research Centre, School of Health Sciences, Faculty of Biology, Medicine and Health, the University of Manchester, Manchester, Oxford Road, Manchester, M13 9PL, UK
[2]Centre for Pharmacoepidemiology and Drug Safety, School of Health Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Oxford Road, Manchester, M13 9PL, UK
[3]NIHR Greater Manchester Patient Safety Translational Research Centre, School of Health Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Oxford Road, Manchester, M13 9PL, UK
[4]Utrecht Institute for Pharmaceutical Sciences, Utrecht University, Utrecht, Netherlands
[5]Alan Turing Institute, Headquartered at the British Library, London, UK

Corresponding author: Tjeerd van Staa, tjeerd.vanstaa@manchester.ac.uk

Journal title: Submitted
Doi: -
License: -

Word count: 3211
Abstract: 300
Number of tables: 4
Number of figures: 4

4.1 Abstract

Objective: To assess the consistency of machine learning and statistical techniques in predicting individual- and population-level cardiovascular disease (CVD) risks and the effects of censoring on risk predictions.

Design: Longitudinal cohort study from 1st Jan 1998 to 31st December 2018.

Setting: 3.6 million patients from the Clinical Practice Research Datalink registered at 391 general practices in England with linked hospital admission and mortality records.

Main outcome measures: Model performance including discrimination, calibration and consistency of individual risk prediction for the same patients among models with comparable model performance.

Methods: 19 different prediction techniques were applied including 12 families of machine learning models (grid searched for best models), three Cox proportional hazards models (local fitted, QRISK3 and Framingham), three parametric survival models and a logistic model.

Results: We found that the various models had similar population-level model performance (C-statistics of about 0.87 and similar calibration). However, the predictions for individual CVD risks varied widely between and within different types of machine learning and statistical models, especially in patients with higher risks. A patient with a risk of 9%~10% predicted by QRISK3 had a risk of 2.8%~9.0% in a random forest and 2.3%~6.9% in a neural network. Models that ignored censoring (i.e., censored patients assumed to be event free) substantially underestimated CVD risk. Of the 223,815 patients with a CVD risk >7.5% with QRISK3, 57.8% would be reclassified below 7.5% when using another model.

Conclusions: A variety of models predicted risks for the same patients very differently despite similar model performances. The logistic model and commonly used machine learning models should not be directly applied for the prediction of long-term risks without considering censoring; Cox models such as QRISK3 are preferable. The level of consistency within and between models should be routinely assessed prior to clinical usage to help inform treatment decisions.

4.2 Summary boxes

What is already known on this topic

Risk prediction models are widely used in clinical practice (such as QRISK or Framingham for cardiovascular disease [CVD]). Multiple techniques can be used for these predictions and recent studies claim that machine learning models can outperform models such as QRISK.

What this study adds

Nineteen different prediction techniques (including 12 machine learning and 7 statistical models) yielded similar population-level performance but CVD predictions for the same patients varied substantially between models. Models that ignored censoring (including commonly used machine learning models) yielded biased risk predictions. The level of consistency within and between models should be routinely assessed prior to clinical usage and ‘black box’ approaches may be less preferable in some settings such as CVD long-term risk prediction.

4.3 Introduction

Risk prediction models are used routinely in healthcare practice to identify high risk patients and make treatment decisions, so that appropriate healthcare resources can be allocated to those patients who most need care [1]. These risk prediction models are usually built using statistical regression techniques. Examples include the Framingham risk score (developed from a US cohort with prospectively collected data) [2] and QRISK3 (developed from a large UK cohort using retrospective electronic health records (EHR)).[3] Recently, machine learning models have gained considerable popularity. The English National Health Service has invested £250 million to further embed machine learning in health care [4]. A recent viewpoint article suggested that machine learning technology is about to start a revolution with the potential to transform the whole health care system [5]. Several studies argued that machine learning models could outperform statistical models in terms of calibration and discrimination [6 7 8 9]. However, another viewpoint expressed concern that these approaches cannot provide explainable reasons behind their predictions potentially leading to inappropriate actions [10], while a recent review found no evidence that machine learning models improved model performance compared with logistic models [11]. However, the interpretation is difficult as this review included models from mostly small sample sizes and with different outcomes and predictors. Machine learning has established strengths in image recognition which could help diagnose diseases in healthcare [12 13 14 15], but censoring (patients lost to follow-up), which is a common issue in risk prediction, does not exist in image recognition. Many commonly used machine learning models do not take into account censoring by default [16]. The objective of this study was to assess the robustness and consistency of a variety of machine learning and statistical models on individual risk prediction and the effects of censoring on risk predictions. Cardiovascular disease (CVD) was used as an exemplar. Robustness of individual risk prediction in this study was defined as the level of consistency in the prediction of risks for individual patients with models that have comparable population-level performance metrics [17 18 19].

4.4 Methods

4.4.1 Data source

The study cohort was derived from Clinical Practice Research Datalink (CPRD GOLD). The database includes data from about 6.9% of the population in England [20]. It also has been linked to Hospital Episode Statistics, Office for National Statistics mortality records and Townsend deprivation scores [3] to provide additional patient information about hospital admissions (including date and discharge diagnoses) and cause-specific mortality [20]. CPRD includes patient EHRs from general practice capturing detailed information such as demographics (age, sex and ethnicity), symptoms, tests, diagnoses, prescribed treatments, health-related behaviours and referrals to secondary care [20]. CPRD is a well-established representative cohort of UK population and thousands of studies have used CPRD [21 22 23], including a validation of the QRISK2 model [24] and an analysis of machine learning [8].

4.4.2 Study population

This study used the same selection criteria for the study population, risk factors and CVD outcomes as used for QRISK3 [3 18]. Follow-up of patients started at the date of the patient's registration with the practice, 25th birthday, or January 1 1998 (whichever latest) and ended at the date of death, incident CVD, date of leaving the practice or last date of data collection (whichever earliest). The index date for measurement of CVD risk was randomly chosen from the follow-up period in order to capture time-relevant practice variability with a better spread of calendar time and age [25]. This was different from QRISK3 which mostly used a single calendar time date [18]. The main inclusion criteria were age between 25 and 84, no CVD history or any statin prescription prior to the index date. The outcome of interest was the 10-year risk of developing CVD. The primary clinical outcome (CVD) was defined the same as QRISK3 [3], i.e. coronary heart disease, ischaemic stroke or transient ischaemic attack.

Two main cohorts were extracted from the study population, one overall cohort including all patients with at least one day of follow-up and the other removing censored patients. The cohort without censoring excluded patients who were lost to follow-up before developing CVD by year 10. The analysis of the cohort without censoring aimed to investigate the effects of ignoring censoring on patients' individual

risk prediction. This cohort mimics the methods used by some machine learning studies, i.e. only selected patients or practices with full 10 years follow-up [8].

4.4.3 CVD risk factors

The CVD risk factors at the index date included sex, age, body mass index (BMI), smoking history, cholesterol/HDL ratio, systolic blood pressure (SBP) and its standard deviation, history of prescribing of atypical antipsychotic medication, blood pressure treatment or regular oral glucocorticoids, clinical history of systemic lupus erythematosus, atrial fibrillation, chronic kidney disease (stage 3, 4 or 5), erectile dysfunction, migraine, rheumatoid arthritis, severe mental illness or type 1 or 2 diabetes mellitus, family history of angina or heart attack in a 1st degree relative aged < 60 years, ethnicity and Townsend deprivation score [3]. The same predictors from QRISK3 [3] were used for all model fitting except for Framingham [26] which used fewer and different predictors.

4.4.4 Machine learning and Cox models

The study considered 19 models including 12 families of machine learning, three Cox proportional hazards models (local fitted, QRISK3 and Framingham), three parametric survival models (assuming Weibull, Gaussian and logistic distribution) and a statistical logistic model (fitted in a statistical causal-inference framework). Machine learning models included a logistic model [27] (fitted in an automated machine learning framework), random forest [28] and neural network [29] from R package “Caret” [30]; logistic model, random forest, neural network, extra-tree model [31] and gradient boosting classifier [31] from Python package “Sklearn” [32]; logistic model, random forest, neural network and autoML [33] from Python package “h2o” [34]. The package autoML selects a best model from a broader spectrum of candidate models [33]. Details of these models are summarised in eTable 1. The study used the machine learning algorithms from different software packages with a grid search process on hyper-parameters and cross validation to acquire a series of high-performing machine learning models; this mimics the reality that different packages may subjectively be selected by practitioners for model fitting and end up with a different best model. The study treated the models from the same machine learning algorithm from different

software packages as different model families as the settings (hyper-parameters) of these packages to control the model fitting are often different, which might result in a different best performing model through the grid search process.

4.4.5 Statistical analysis

The Markov chain Monte Carlo (MCMC) method with monotone style was used to impute missing values 10 times for ethnicity (% missing in overall cohort was 54.3%), BMI (40.3%), Townsend score (0.1%), SBP (26.9%), standard deviation of SBP (53.9%), ratio of cholesterol and High-Density Lipoprotein (HDL) (65.0%) and smoking status (25.2%) [18] (only these variables have missing values). The overall cohort (which contained 10 imputations) was randomly split to an overall derivation (75%) and an overall testing (25%) cohort. A total of 1200 machine learning models with highest discrimination (C-statistic) were grid searched on hyper-parameters with two-fold cross validation estimating calibration and discrimination. They were derived from 12 model families of 100 samples with similar sample size to another machine learning study [8]. The individual CVD risk predictions (averaged for missing value imputations) and model performance of all models were then estimated using the overall testing cohort. The sample splitting and model fitting process is shown in eFigure 1.

Model performance such as sensitivity and positive predictive value (PPV) were calculated using a threshold of 7.5% for all models. The threshold was selected according to ACC/AHA Guideline on the Assessment of Cardiovascular Risk [35] and used in other machine learning studies [7] [8]. Distributions of risk predictions for the same group of individuals among models were compared. The individual risk predictions of the same patients for the other models were plotted against the reference model (QRISK3) in a scatter plot with Fieller's 95% confidence interval [36]. The differences of individual CVD risk predictions between models were plotted against the 10 deciles of CVD risk predictions for QRISK3. The differences of individual CVD risk predictions were calculated by individual risk prediction of each model minus the individual risk prediction of the reference model (QRISK3). R [30] was used to fit the models from "Caret", and Python [31] was used to fit models from "Sklearn" and "h2o". SAS procedures [37] were used to extract the raw data, create

analysis data sets and generate tables and graphs. The protocol for this work was approved by the Independent Scientific Advisory Committee for Clinical Practice Research Datalink research (protocol no 19_054R).

4.4.6 Patient and Public Involvement

No patients were involved in setting the research question or the outcome measures, nor were they involved in developing plans or implementation of the study. No patients were asked to advise on interpretation or writing up of results.

4.5 Results

The overall study population included 3.6 million patients from 391 general practices. The cohort without censoring was considerably smaller (0.4 million) than the overall cohort. Table 4.1 shows the baseline characteristics of the two study populations which were split into derivation and validation cohorts. The average age was higher in the cohort without censoring (due to younger patients leaving the practice as shown in eFigure 4.12.8).

Table 4.1: Baseline characteristics of the two study populations (patients aged 25-84 years without history of CVD or prior statin use)

	Overall cohort		Cohort without censoring	
	Derivation cohort	**Validation cohort**	**Derivation cohort**	**Validation cohort**
Number of patients	2746453	915479	335632	111868
CVD cases (N (%))	86769 (3.2)	28828 (3.1)	78826 (23.5)	26168 (23.4)
Censored patients within 10 years (N (%))	2410516 (87.8)	803916 (87.8)	NA	NA
CVD risk factors				
Females (N (%))	1406796 (51.2)	469098 (51.2)	173691 (51.8)	58169 (52.0)
Age (Mean (SD))	44.7 (15.6)	44.7 (15.7)	53.3 (16.2)	53.4 (16.2)
BMI (Mean (SD))	26.7 (5.0)	26.7 (5.0)	27.1 (4.8)	27.1 (4.8)
Cholesterol/HDL ratio (Mean (SD))	3.9 (1.3)	3.9 (1.3)	4.1 (1.3)	4.1 (1.3)
On atypical antipsychotic medication (N (%))	123060 (0.4)	40300 (0.4)	9320 (0.3)	3160 (0.3)
On blood pressure treatment (N (%))	1839640 (6.7)	619620 (6.8)	427040 (12.7)	142450 (12.7)
On regular steroid tablets (N (%))	20590 (0.1)	6940 (0.1)	2890 (0.1)	1000 (0.1)
History of Systemic Lupus Erythematosus (N (%))	18400 (0.1)	6060 (0.1)	2570 (0.1)	740 (0.1)
History of angina or heart attack in a 1st degree relative < 60 (N (%))	984550 (3.6)	326190 (3.6)	79500 (2.4)	26690 (2.4)
History of atrial fibrillation (N (%))	207780 (0.8)	69650 (0.8)	52130 (1.6)	17570 (1.6)
History of chronic kidney disease (stage 3, 4 or 5) (N (%))	301330 (1.1)	102400 (1.1)	43640 (1.3)	15140 (1.4)
History of erectile dysfunction (N (%))	396510 (1.4)	131100 (1.4)	38670 (1.2)	12870 (1.2)
History of migraines (N (%))	1774390 (6.5)	591060 (6.5)	196290 (5.8)	65930 (5.9)
History of rheumatoid arthritis (N (%))	161670 (0.6)	54590 (0.6)	30430 (0.9)	10300 (0.9)
History of severe mental illness (N (%))	2198610 (8.0)	728320 (8.0)	321900 (9.6)	106730 (9.5)
History of type 1 diabetes (N (%))	58990 (0.2)	20970 (0.2)	8200 (0.2)	2510 (0.2)
History of type 2 diabetes (N (%))	355690 (1.3)	118260 (1.3)	81340 (2.4)	26410 (2.4)
SBP (Mean (SD))	126.9 (16.7)	126.9 (16.7)	133.1 (18.3)	133.1 (18.3)
Standard deviation of each individual patients' SBP (Mean (SD))	9.9 (5.6)	9.9 (5.6)	10.7 (5.9)	10.7 (5.9)
Ethnicity				
Other ethnicity (N (%))	372240 (1.4)	125370 (1.4)	13660 (0.4)	4490 (0.4)
White or not recorded (N (%))	25731820 (93.7)	8573550 (93.7)	3287320 (97.9)	1095950 (98.0)
Smoking				
Ex-smoker (N (%))	6300299 (22.9)	2095026 (22.9)	761755 (22.7)	255109 (22.8)
Current smoker (N (%))	8066066 (29.4)	2696146 (29.5)	940730 (28.0)	312258 (27.9)
Never smoker (N (%))	13098165 (47.7)	4363618 (47.7)	1653835 (49.3)	551313 (49.3)
Townsend deprivation				
Score 1 – Least deprived (N (%))	6004107 (21.9)	1999488 (21.8)	866048 (25.8)	291496 (26.1)
Score 2 (N (%))	5947510 (21.7)	1976893 (21.6)	822072 (24.5)	272627 (24.4)
Score 3 (N (%))	5728916 (20.9)	1910200 (20.9)	698985 (20.8)	233291 (20.9)
Score 4 (N (%))	5680051 (20.7)	1895230 (20.7)	604246 (18.0)	201443 (18.0)
Score 5 – Most deprived (N (%))	4103946 (14.9)	1372979 (15.0)	364969 (10.9)	119823 (10.7)

Table 4.2 shows the model performance of the machine learning and statistical models. All models had very similar discrimination (C-statistics of about 0.87) and calibration (Brier scores of about 0.03). Details on model performances are shown in eTable 4.12.3.1-4.12.3.2 and eFigure 4.12.2.1-4.12.2.2 and 4.12.3.1-4.12.3.6).

Table 4.2: Performance indicators of machine learning and statistical models in the overall cohort

	Model performance* (95% range #)				Average absolute change of model performance (95% range)
	C-statistic (2.5% ~ 97.5%) #	**Brier score (2.5% ~ 97.5%) #**	**Recall (Sensitivity) (2.5% ~ 97.5%) #**	**Precision (PPV) (2.5% ~ 97.5%) #**	**C-statistic (2.5% ~ 97.5%) #**
Logistic (Caret)	0.879 (0.879, 0.879)	0.028 (0.028, 0.028)	0.615 (0.609, 0.620)	0.163 (0.162, 0.164)	+0.00% (-0.03%, 0.04%)
Random forest (Caret)	0.869 (0.867, 0.869)	0.028 (0.028, 0.028)	0.656 (0.620, 0.675)	0.144 (0.139, 0.153)	-1.20% (-1.33%, -1.10%)
Neural network (Caret)	0.878 (0.867, 0.880)	0.028 (0.027, 0.028)	0.670 (0.642, 0.687)	0.148 (0.141, 0.153)	-0.15% (-1.35%, 0.06%)
Statistic logistic model	0.879 (0.879, 0.879)	0.028 (0.028, 0.028)	0.614 (0.607, 0.620)	0.163 (0.162, 0.164)	+0.01% (-0.02%, 0.04%)
QRISK3	0.879	0.031	0.834	0.107	Reference model
Framingham	0.865	0.031	0.892	0.085	-1.66% (-1.66%, -1.66%)
Local Cox model	0.877 (0.877, 0.878)	0.032 (0.031, 0.032)	0.810 (0.804, 0.816)	0.112 (0.110, 0.113)	-0.22% (-0.28%, -0.17%)
Parametric survival model (Weibull)	0.877 (0.876, 0.877)	0.031 (0.031, 0.032)	0.810 (0.804, 0.815)	0.111 (0.110, 0.113)	-0.29% (-0.35%, -0.24%)
Parametric survival model (Gaussian)	0.876 (0.876, 0.877)	0.031 (0.030, 0.031)	0.834 (0.830, 0.839)	0.104 (0.103, 0.105)	-0.33% (-0.39%, -0.29%)
Parametric survival model (Logistic)	0.876 (0.875, 0.876)	0.031 (0.031, 0.032)	0.796 (0.791, 0.802)	0.114 (0.113, 0.115)	-0.36% (-0.43%, -0.31%)
Logistic (Sklearn)	0.879 (0.879, 0.879)	0.028 (0.028, 0.028)	0.615 (0.609, 0.620)	0.163 (0.161, 0.164)	0.00% (-0.05%, 0.03%)
Random forest (Sklearn)	0.872 (0.871, 0.873)	0.028 (0.028, 0.028)	0.670 (0.661, 0.679)	0.142 (0.140, 0.144)	-0.80% (-0.89%, -0.71%)
Neural network (Sklearn)	0.872 (0.832, 0.879)	0.028 (0.028, 0.029)	0.556 (0.174, 0.692)	0.163 (0.137, 0.224)	-0.85% (-5.39%, -0.03%)
Gradient boosting classifier (Sklearn)	0.878 (0.877, 0.878)	0.028 (0.028, 0.028)	0.642 (0.623, 0.657)	0.154 (0.150, 0.157)	-0.17% (-0.29%, -0.08%)
extra-trees (Sklearn)	0.863 (0.861, 0.864)	0.028 (0.028, 0.029)	0.639 (0.628, 0.650)	0.139 (0.136, 0.141)	-1.89% (-2.05%, -1.76%)
Logistic (h2o)	0.879 (0.878, 0.879)	0.028 (0.028, 0.028)	0.615 (0.608, 0.621)	0.162 (0.161, 0.164)	-0.06% (-0.10%, -0.02%)
Random forest (h2o)	0.877 (0.877, 0.878)	0.028 (0.028, 0.028)	0.646 (0.631, 0.659)	0.152 (0.149, 0.154)	-0.22% (-0.29%, -0.17%)
Neural network (h2o)	0.875 (0.870, 0.879)	0.028 (0.028, 0.031)	0.552 (0.163, 0.780)	0.169 (0.118, 0.238)	-0.45% (-1.09%, -0.04%)
autoML (h2o)	0.879 (0.879, 0.880)	0.028 (0.028, 0.028)	0.616 (0.605, 0.642)	0.162 (0.157, 0.164)	-0.01% (-0.07%, 0.06%)

*** Model performance was calculated in binaray framework. Threshold 7.5% was used to calculate precision and recall for all models**

 # 95% range (2.5% ~ 97.5%) of model performances was derived from 100 random samples

Figure 4.1a shows that models which ignored censoring substantially underestimated patients' CVD risks in the overall cohort. Patients with a predicted CVD risk between 9.5%~10.5% with QRISK3 had a median prediction of 4.1% with logistic Caret model, 5.1% with Caret neural network and 5.0% with Sklearn random forest. As shown in Figure 4.2a, multiple models fitted from the cohort without censoring substantially overestimated patients' risk compared to QRISK3. The removal of censored patients changed the magnitude but not the variability of individual CVD risk predictions (Figure 4.2b and Figure 4.1b).

Figure 4.1: Distribution of individual risk predictions with machine learning and statistical models in overall cohort

a. **For patients with predicted CVD risks of 9.5%~10.5% in QRISK3**
b. **For patients with predicted CVD risks of 7%~8% in the logistic Caret model**

X axis: predicted CVD risk

Y axis: relative frequency (estimated density value)

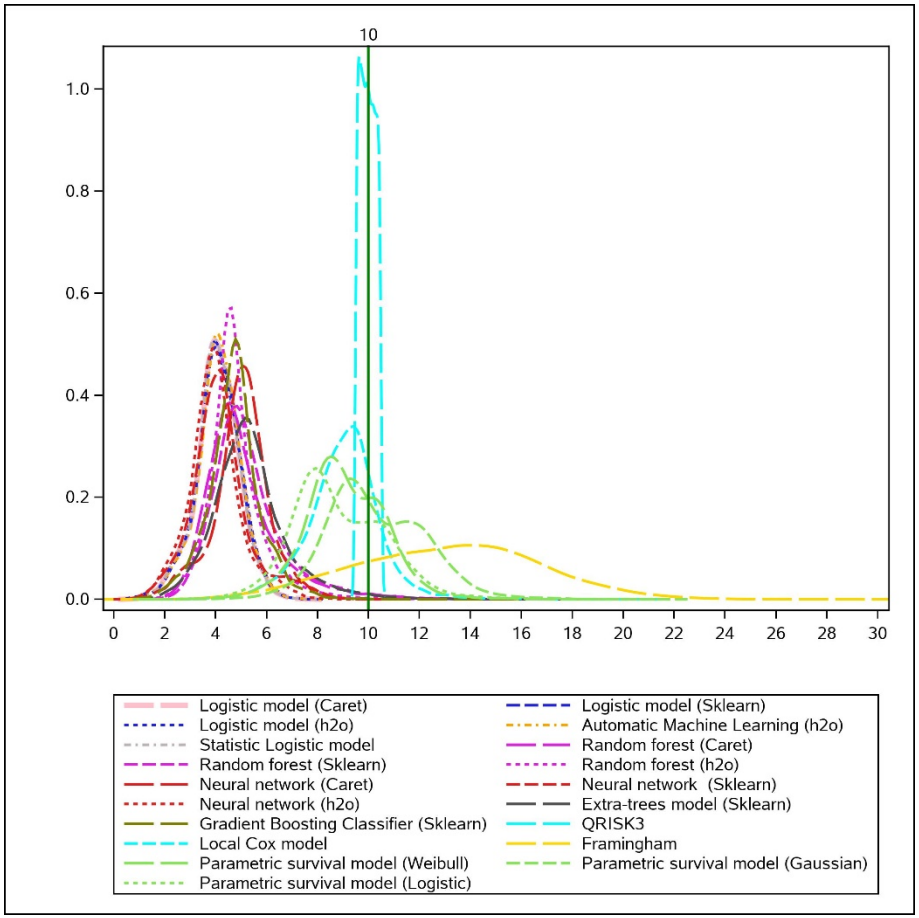


Figure 4.1a

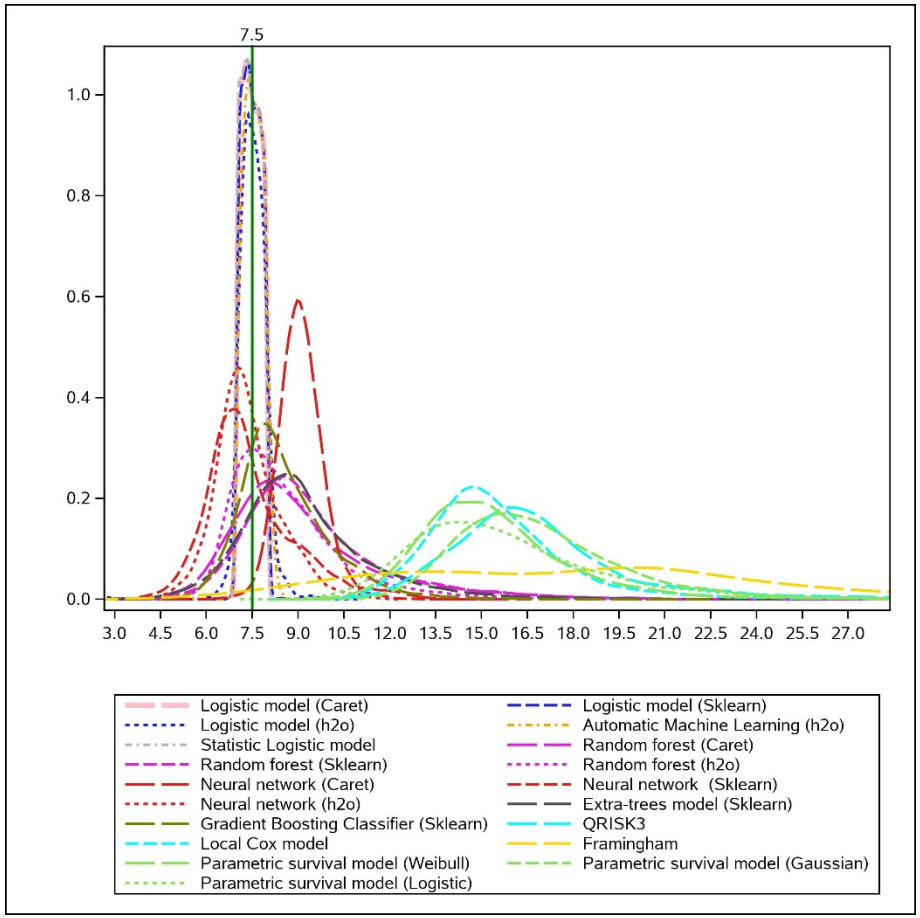


Figure 4.1b

Figure 4.2: Distribution of individual risk predictions with machine learning and statistical models in cohort without censoring

a. **For patients with predicted CVD risks of 9.5%~10.5% in QRISK3**
b. **For patients with predicted CVD risks of 7%~8% in the logistic Caret model**

X axis: predicted CVD risk

Y axis: relative frequency (estimated density value)

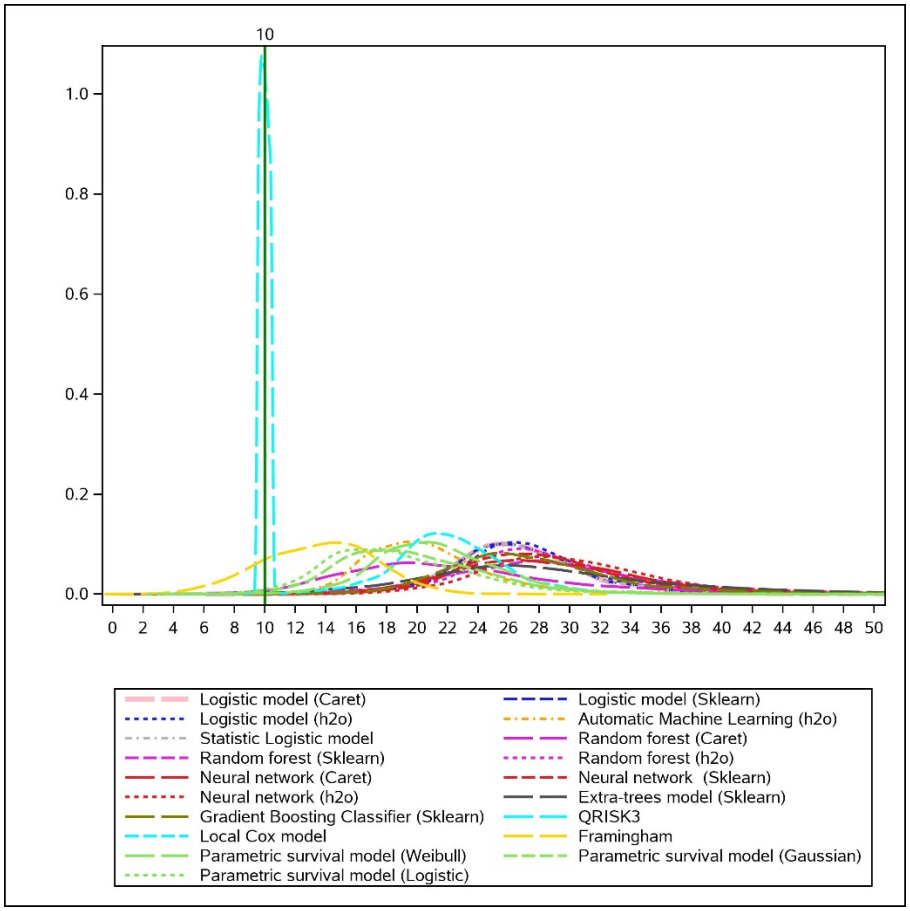


Figure 4.2a

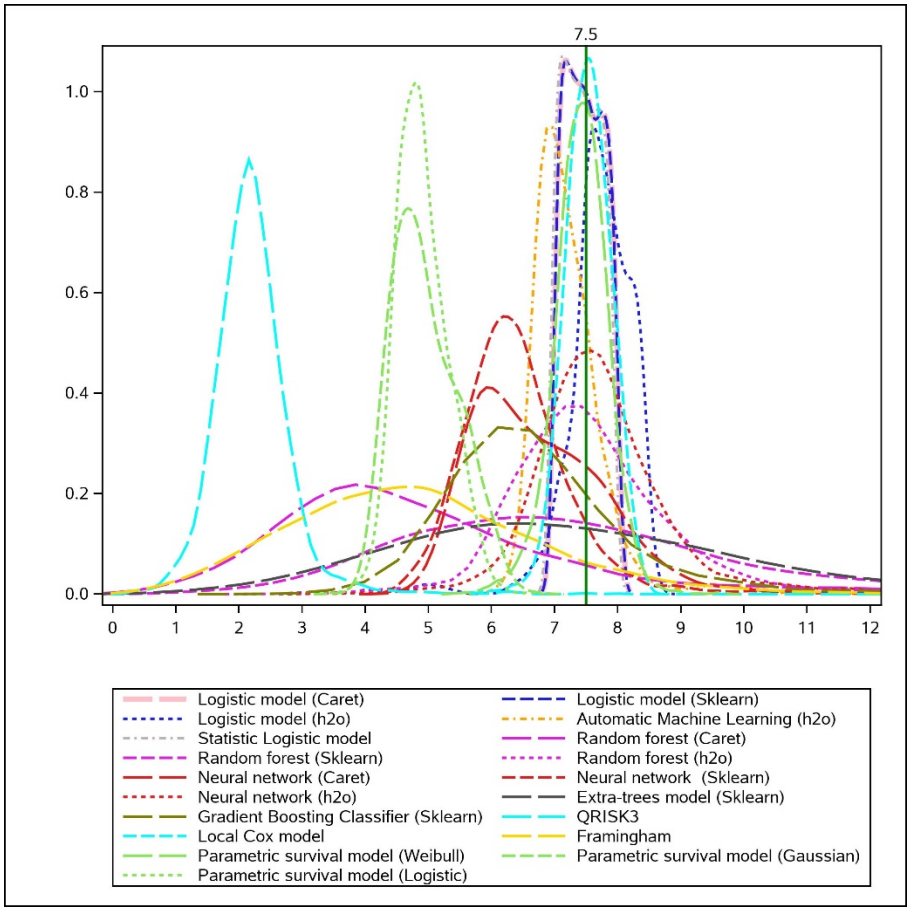


Figure 4.2b

Table 4.3 shows the range of individual CVD risk predictions across the different models. It was found that patients with a QRISK3 predicted risk of 7%~8% had predicted CVD risks between 2.0% and 7.3% in a Caret random forest, 1.7% and 5.2% in Caret neural network and 1.4% and 5.6% in Sklearn neural network (eTable 4.12.5.1). Figure 4.3 shows the variation of individual risk predictions between QRISK3 and to those generated by the other models with Fieller's confidence interval. Similar to the findings presented in Table 4.3, predictions for these same patients varied substantially between models.

Table 4.3: Comparison of individual risk predictions of machine learning and statistical models in the overall cohort and cohort without censoring

	Range of individual risk predictions (2.5^{th}~97.5^{th}) with other models compared to those of reference model								
	<6%	**6~7%**	**7~8%**	**8~9%**	**9~10%**	**10~11%**	**11~12%**	**12~13%**	**≥ 13%**
Overall cohort									
QRISK3 as reference model									
Soft voting *	0.3~3.5	2.7~5.2	3.1~6.0	3.5~6.8	3.9~7.5	4.3~8.3	4.6~9.0	5.1~9.8	7.2~36.4
All models	0.1~4.9	1.5~10.5	1.8~11.6	2.0~12.7	2.4~13.6	2.6~14.7	2.9~15.7	3.2~16.7	5.0~44.8
Logistic model (Caret) as reference model									
Soft voting	0.3~7.8	8.2~12.3	9.4~13.9	10.5~15.4	11.7~16.8	12.8~18.2	14.0~19.3	15.0~20.5	17.1~41.6
All models	0.1~9.7	5.4~20.1	6.2~22.1	7.0~24.0	7.8~26.1	8.5~28.1	9.2~29.9	9.9~31.7	12.6~53.7
Cohort without censoring									
QRISK3 as reference model									
Soft voting	1.4~16.4	12.5~25.2	14.3~28.5	15.9~30.7	17.1~33.7	18.8~36.6	20.4~38.9	21.8~41.2	28.4~80.7
All models	0.6~18.1	8.4~29.5	9.5~33.4	10.5~36.0	11.4~39.4	12.3~42.4	13.2~45.2	13.9~47.4	19.3~85.9
Logistic model (Caret) as reference model									
Soft voting	1.2~5.3	4.7~7.7	5.4~8.9	6.2~9.8	7.1~10.9	7.8~11.9	8.5~13.5	9.4~14.3	11.9~76.2
All models	0.2~6.3	1.6~9.2	2.0~10.9	2.3~12.2	2.7~14.1	3.1~15.3	3.4~17.0	3.8~18.2	8.4~82.0

***Soft voting involved averaging of predictions of all models except the reference model**

Figure 4.3: Inconsistency of individual risk predictions with machine learning and statistical models with Fieller's 95% confidence interval (each dot corresponds to an individual prediction; a random sample of these is displayed with red line enclosing 95% of the observations)

a. **Overall cohort**
b. **Cohort without censoring**

X axis: QRISK3 predicted CVD risk

Y axis: predicted CVD risk by other models

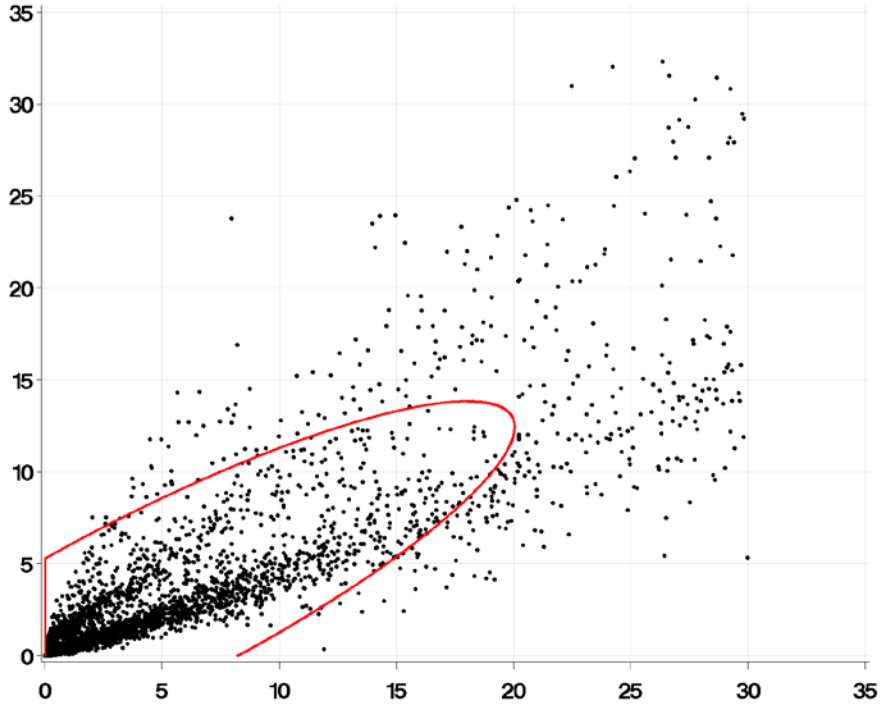


Figure 4.3a

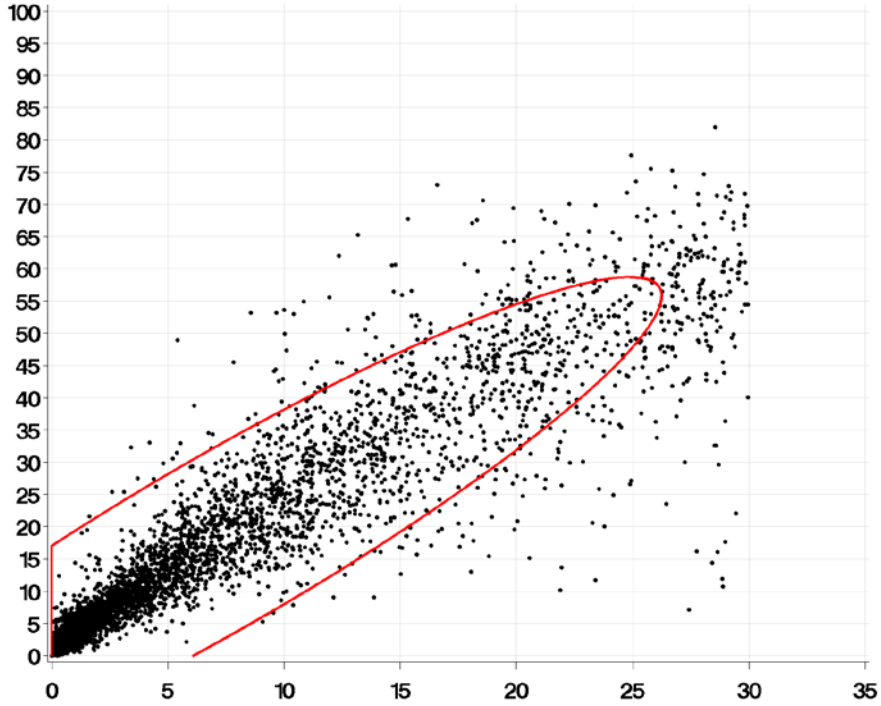


Figure 4.3b

Substantial reclassification using a threshold was found when changing between different models. Of 691,664 patients with a CVD risk ≤7.5% as predicted by QRISK3, 13.6% would be reclassified above 7.5% when using another model (Table 4.4). Of the 223,815 patients with a CVD risk >7.5%, 57.8% would be reclassified below 7.5% when using another model. High levels of reclassification were also found with a different reference model (as shown in Table 4.4) or a different threshold (eTable 4.12.7).

Table 4.4: Reclassification of individual risk predictions with machine learning and statistical models

	Reclassification in overall testing cohort*	
	Reclassified*	**Not reclassified**
Overall cohort		
QRISK3 10year risk prediction (reference model)		
Below or equal to 7.5% threshold	94186 (13.6%)	597478 (86.4%)
Above 7.5% threshold	129348 (57.8%)	94467 (42.2%)
Logistic model (Caret) 10 year risk prediction (reference model)		
Below or equal to 7.5% threshold	209221 (25.9%)	597478 (74.1%)
Above 7.5% threshold	14313 (13.2%)	94467 (86.8%)
Cohort without censoring		
QRISK3 10 year risk prediction (reference model)		
Below or equal to 7.5% threshold	34607 (54.6%)	28779 (45.4%)
Above 7.5% threshold	1248 (2.6%)	47234 (97.4%)
Logistic model (Caret) 10 year risk prediction (reference model)		
Below or equal to 7.5% threshold	6004 (17.3%)	28779 (82.7%)
Above 7.5% threshold	29851 (38.7%)	47234 (61.3%)

***Patient are re-classified if they have a risk prediction in any model which crosses the threshold compared to the prediction of the reference model**

Figure 4.4 plots the differences of individual CVD risk predictions with the different models stratified by deciles of CVD risk predictions of QRISK3. The largest range of inconsistencies in risk predictions was found in patients with highest predicted CVD risks. Low CVD risk was generally predicted consistently between and within models. Similar trends were observed using a different reference model (eFigure 4.12.4).

Figure 4.4: 95% range of individual risk predictions with machine learning and statistical models stratified by deciles of predicted CVD risks with QRISK3

a. **Overall cohort**
b. **Cohort without censoring**

X axis: 10 deciles of QRISK3 predicted CVD risk (displayed by median value within decile)

Y axis: 95% range of differences in predicted CVD risks with all other models

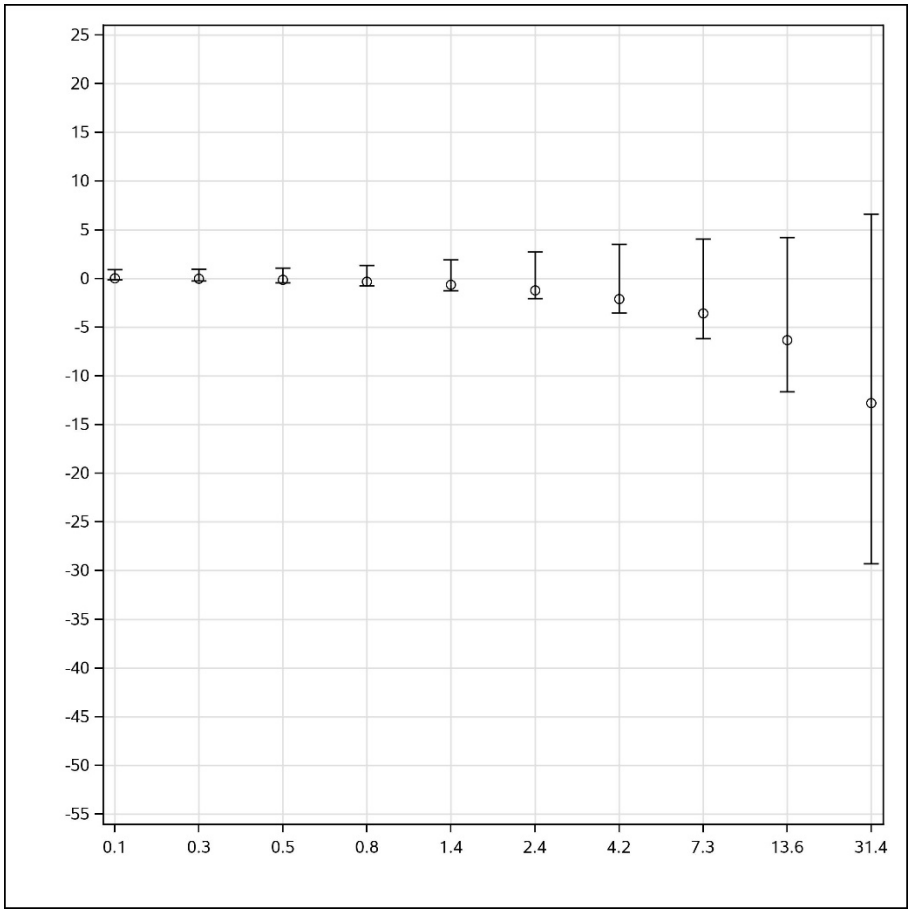


Figure 4.4a

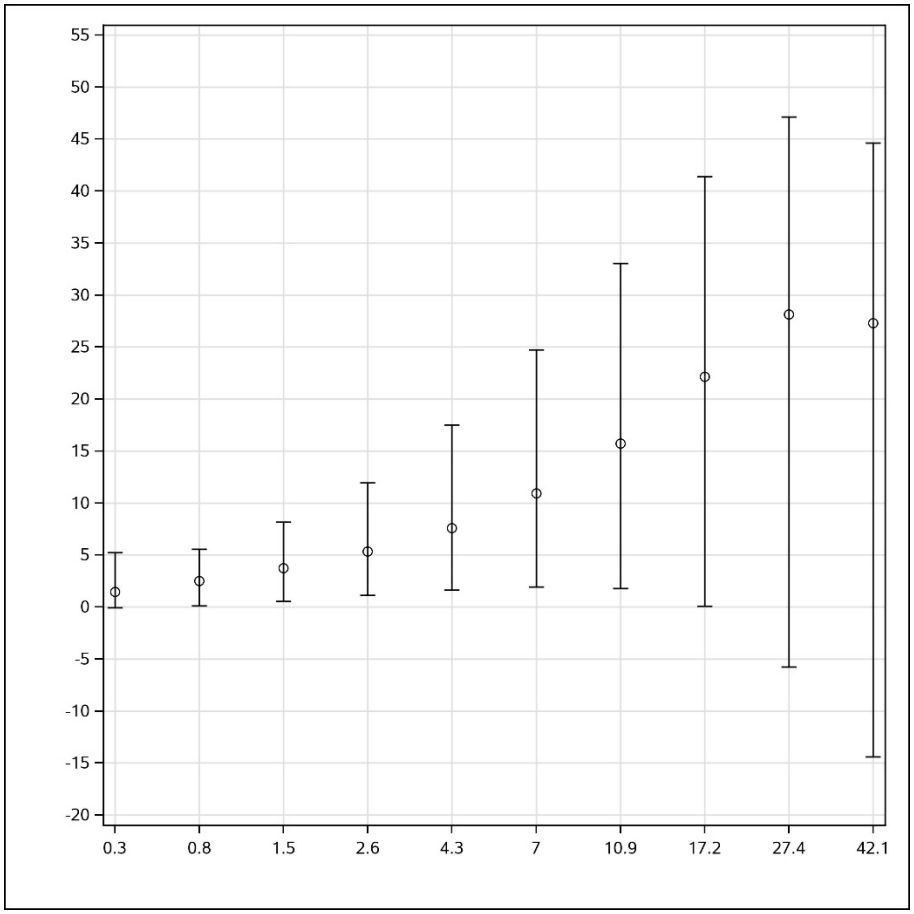


Figure 4.4b

Several sensitivity analyses were conducted with consistent findings of high levels of inconsistencies in individual risk predictions between and within models. The same machine learning algorithm with the selection of different settings (hyper parameters) from different software packages yielded different individual CVD risk predictions (e.g. Caret neural network, Sklearn neural network and h2o neural network; eTable 4.12.8 and eFigure 4.12.5). The evaluation of the effects of generalisability by developing and testing models in different regions of England showed similarly high levels of inconsistencies in CVD risk predictions (eTable 4.12.10 and eFigure 4.12.6). Changing the number of predictors did not result in lower levels of inconsistencies in CVD risk predictions with more predictors included in the models (eTable 4.12.11 and eFigure 4.12.7),

4.6 Discussion

We found that the predictions of CVD risks for individual patients varied widely between and within different types of machine learning and statistical models, especially in patients with higher risks (when using similar predictors). A statistical logistic model and the machine learning models that ignored censoring substantially underestimated CVD risk.

Despite claims that machine learning models can revolutionise risk prediction and potentially replace traditional statistical regression models in other areas [5 38 39], this study of CVD prediction in primary care found that they have similar model performance to traditional statistical methods and share similar uncertainty in individual risk predictions. Strengths of machine learning models may include their ability to automatically model non-linear associations and interactions between different risk factors [40 41]. They may also find new data patterns [31]. They have the acknowledged strength to automate model building with a better performance in specific classification tasks (e.g. image recognition) [31]. However, a critical question is whether risk prediction models provide accurate and consistent risk predictions for individual patients. Previous research has found that a traditional risk prediction model such as QRISK3 has considerable uncertainty on individual risk prediction though it has very good model performance at the population level [18 19]. This uncertainty was found to be related to unmeasured heterogeneity between clinical

sites and modelling choices such as the inclusion of secular trends [18 19]. The present study has found that machine learning models share this uncertainty as models with comparable population-level performance yielded very different individual risk predictions; consequently, different treatment decisions could be made by arbitrarily selecting another technique.

Censoring of patients is an unavoidable problem in the development and validation of prediction models for long-term risks as patients frequently move away or die. However, many popular machine learning models ignore censoring as the default framework is the analysis of a binary outcome rather than time-to-event survival outcome. A UK Biobank study of CVD risk prediction did not report how censoring was dealt with [7], like several other studies [41 42 43]. Another machine learning study incorrectly excluded censored patients [8]. Random survival forest is a machine learning model that takes into account of censoring [44]. There are also innovative techniques being developed that incorporate statistical censoring approaches into the machine learning framework [16 45]. However, to our knowledge currently there are no software packages can handle large datasets for these methods. This study shows that directly applying popular machine learning models to survival analysis without considering censoring substantially biased risk predictions. The miscalibration was large compared to observed life-table predictions. This is consistent with a recent study that reported information loss due to lack of considering censoring with random forest method [6].

The present study considered a total of 22 predictors which had been selected by the developers of QRISK on the basis on their likely causal effect on CVD [3]. Other machine learning studies have used considerably more predictors. As an example, a study using the UK Biobank included 473 predictors in the machine learning models [7]. A potentially unresolved issue in risk prediction is what type and how many variables should be included in models, as currently there is a lack of consensus and guidelines for choosing variables for risk prediction model [46]. More information incorporated into a model may increase the population model performance of risk prediction. For example, the C-statistic is related to both the effects of predictors and the variation of predictors among patients with and without events [47]. Including more predictors in a model may increase the C-statistic merely due to higher variation of predictors. On the other hand, inclusion of non-causal predictors may lower the

accuracy of the risk prediction by adding noise, increase the risk of over-fitting and suffer more data quality challenges [48]. Also, a very large number of predictors may limit the clinical utility of these machine learning models, as more predictors need to be measured before making a prediction. Further research is needed to establish whether the focus of risk prediction should be on consistently measured causal risk factors or variables that may be recorded inconsistently between clinicians or EHR systems.

Current guidelines for the development and validation of risk prediction models (called TRIPOD) focus on the assessment of population-level performance but do not consider consistencies in individual risk predictions by prediction models with comparable population-level performance [49]. Arguably, the clinical utility of risk prediction models should be based, as done with blood pressure devices for instance, on the consistent risk prediction (reliability) for a particular patient rather than broad population-level performance [50]. If models with comparable performance provide different predictions for a patient with certain risk factors, there is a need of an explanation for these discrepant predictions [51]. Explainable artificial intelligence has been described as methods and techniques in the application of artificial intelligence such that the results of the solution can be understood by human experts [52]. It contrasts with the concept of the "black box" in machine learning where predictions cannot be explained. Arguably, a prediction model which is explainable (such as QRISK3 which is based on established causal predictors) may be preferable over “black box” models that are high-dimensional (including many predictors) but that provide inconsistent results for individual patients. Better standards are needed on how to develop and test machine learning algorithms [14]

The major strength of this study was that a large number of different machine learning models with varying hyper-parameters using different packages from different programming languages were fitted to a large population-based primary care cohort. However, there are several study limitations. We only considered predictors from QRISK3 in order to compare models based on equal information, but sensitivity analyses showed similar findings of inconsistencies in CVD risk prediction independent of the number of predictors. Furthermore, more hyper-parameters in the machine learning models could have been considered in the grid search process. However, the fitted models already achieved reasonable high model performance

which indicates that the main hyper-parameters had been covered in the grid search process. There are several machine learning algorithms that were not included in this study such as support vector machine or survival random forest. The reason is due to the difficulty that software packages handling large datasets [53 54 55 56]. Another limitation is that this study concerned CVD risk prediction in primary care and findings may not be generalisable to other outcomes or settings. However, the robustness of individual risk predictions within and between models with comparable population-level performance is rarely, if ever, evaluated. Our findings do indicate the importance of assessing this.

In conclusion, a variety of models predicted CVD risks for the same patients very differently despite similar model performances. Using the logistic model and commonly used machine learning models without considering censoring in survival analysis results substantially biased risk prediction and have limited usefulness in the prediction of long-term risks. The level of consistency within and between models should be assessed prior to clinical usage for treatment decisions and considered in TRIPOD guidelines.

4.7 Funding

This study was funded by the China Scholarship Council (to cover costs of PhD studentship of Yan Li at the University of Manchester). The funder did not participate in the research or review any details of this study; the other authors are independent of the funder.

4.8 Acknowledgements

This study is based on data from the Clinical Practice Research Datalink obtained under license from the UK Medicines and Healthcare products Regulatory Agency. The protocol for this work was approved by the independent scientific advisory committee for Clinical Practice Research Datalink research (No 19_054R). The data are provided by patients and collected by the NHS as part of their care and support. The Office for National Statistics (ONS) is the provider of the ONS Data contained within the CPRD Data. Hospital Episode Data and the ONS Data Copyright © (2014),

are re-used with the permission of The Health & Social Care Information Centre. All rights reserved. The interpretation and conclusions contained in this study are those of the authors alone. There are no conflicts of interest among the authors. No additional data available.

4.9 Additional information

"The Corresponding Author has the right to grant on behalf of all authors and does grant on behalf of all authors, a worldwide licence to the Publishers and its licensees in perpetuity, in all forms, formats and media (whether known now or created in the future), to i) publish, reproduce, distribute, display and store the Contribution, ii) translate the Contribution into other languages, create adaptations, reprints, include within collections and create summaries, extracts and/or, abstracts of the Contribution, iii) create any other derivative work(s) based on the Contribution, iv) to exploit all subsidiary rights in the Contribution, v) the inclusion of electronic links from the Contribution to third party material where-ever it may be located; and, vi) licence any third party to do any or all of the above."

Prof. Tjeerd van Staa affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

The default licence, a CC BY NC licence, is needed

4.10 Author Contribution statement

Yan Li: Designed the study; conducted all statistical analysis; produced all tables and figures; wrote the main manuscript text and Appendix.
Matthew Sperrin: Supervised the study; Quality control on statistical analysis; reviewed all statistical results; reviewed and edited the main manuscript text.
Darren M Ashcroft: Reviewed and edited the main manuscript text and Appendix.
Tjeerd Pieter van Staa: Designed and supervised the study; Quality control of all aspects of the paper; wrote the main manuscript text.

4.11 References

1. NICE recommends wider use of statins for prevention of CVD | News and features | News | NICE. https://www.nice.org.uk/news/article/nice-recommends-wider-use-of-statins-for-prevention-of-cvd. Accessed April 30, 2018.

2. Mahmood SS, Levy D, Vasan RS, Wang TJ. The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective. *Lancet (London, England)*. 2014;383(9921):999-1008. doi:10.1016/S0140-6736(13)61752-3

3. Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. *Bmj*. 2017;2099(May):j2099. doi:10.1136/bmj.j2099

4. GOV.UK. Health Secretary announces £250 million investment in artificial intelligence - GOV.UK. https://www.gov.uk/government/news/health-secretary-announces-250-million-investment-in-artificial-intelligence. Accessed August 28, 2019.

5. Hinton G. Deep Learning—A Technology With the Potential to Transform Health Care. *JAMA*. 2018;320(11):1101. doi:10.1001/jama.2018.11100

6. Steele AJ, Denaxas SC, Shah AD, Hemingway H, Luscombe NM. Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease. 2018. doi:10.1371/journal.pone.0202344

7. Alaa AM, Bolton T, Di Angelantonio E, Rudd JHF, van der Schaar M. Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants. Aalto-Setala K, ed. *PLoS One*. 2019;14(5):e0213653. doi:10.1371/journal.pone.0213653

8. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? Liu B, ed. *PLoS One*. 2017;12(4):e0174944. doi:10.1371/journal.pone.0174944

9. Al'Aref SJ, Anchouche K, Singh G, et al. Clinical applications of machine learning in cardiovascular disease and its relevance to cardiac imaging. *Eur Heart J*. 2019;40(24):1975-1986. doi:10.1093/eurheartj/ehy404

10. Price WN, Gerke S, Cohen IG. Potential Liability for Physicians Using Artificial Intelligence. *JAMA*. October 2019. doi:10.1001/jama.2019.15064

11. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol*. 2019;110. doi:10.1016/j.jclinepi.2019.02.004

12. Carin L, Pencina MJ. On deep learning for medical image analysis. *JAMA - J Am Med Assoc*. 2018;320(11):1192-1193. doi:10.1001/jama.2018.13316

13. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep

learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA - J Am Med Assoc*. 2016;316(22):2402-2410. doi:10.1001/jama.2016.17216

14. Maddox TM, Rumsfeld JS, Payne PRO. Questions for Artificial Intelligence in Health Care. *JAMA - J Am Med Assoc*. 2019;321(1):31-32. doi:10.1001/jama.2018.18932

15. Nsoesie EO. Evaluating Artificial Intelligence Applications in Clinical Settings. *JAMA Netw Open*. 2018;1(5):e182658. doi:10.1001/jamanetworkopen.2018.2658

16. Vock DM, Wolfson J, Bandyopadhyay S, et al. Adapting machine learning techniques to censored time-to-event health record data: A general-purpose approach using inverse probability of censoring weighting. *J Biomed Inform*. 2016;61:119-131. doi:10.1016/j.jbi.2016.03.009

17. BRIGGS W. *UNCERTAINTY : The Soul of Modeling, Probability & Statistics.* SPRINGER; 2018.

18. Li Y, Sperrin M, Belmonte M, Pate A, Ashcroft DM, van Staa TP. Do population-level risk prediction models that use routinely collected health data reliably predict individual risks? *Sci Rep*. 2019;9(1):11222. doi:10.1038/s41598-019-47712-5

19. Pate A, Emsley R, Ashcroft DM, Brown B, van Staa T. The uncertainty with using risk prediction models for individual decision making: an exemplar cohort study examining the prediction of cardiovascular disease in English primary care. *BMC Med*. 2019;17(1):134. doi:10.1186/s12916-019-1368-8

20. Herrett E, Gallagher AM, Bhaskaran K, et al. Data Resource Profile: Clinical Practice Research Datalink (CPRD). *Int J Epidemiol*. 2015;44(3):827-836. doi:10.1093/ije/dyv098

21. Clinical Practice Research Datalink - CPRD. https://www.cprd.com/intro.asp. Accessed August 20, 2017.

22. Clinical Practice Research Datalink | CPRD. https://www.cprd.com/. Accessed November 28, 2019.

23. Danese MD, Gleeson M, Griffiths RI, Catterick D, Kutikova L. Methods for estimating costs in patients with hyperlipidemia experiencing their first cardiovascular event in the United Kingdom. *J Med Econ*. 2017;20(9):931-937. doi:10.1080/13696998.2017.1345747

24. Hippisley-Cox J, Coupland C, Brindle P. The performance of seven QPrediction risk scores in an independent external sample of patients from general practice: a validation study. *BMJ Open*. 2014;4(8):e005809. doi:10.1136/bmjopen-2014-005809

25. van Staa T-P, Gulliford M, Ng ES-W, Goldacre B, Smeeth L. Prediction of cardiovascular risk using Framingham, ASSIGN and QRISK2: how well do they predict individual rather than population risk? *PLoS One*. 2014;9(10):e106455. doi:10.1371/journal.pone.0106455

26. Anderson KM, Wilson PW, Odell PM, Kannel WB. An updated coronary risk

profile. A statement for health professionals. *Circulation*. 1991;83(1):356-362. doi:10.1161/01.CIR.83.1.356

27. Nelder JA, Wedderburn RWM. Generalized Linear Models. *J R Stat Soc Ser A*. 1972;135(3):370. doi:10.2307/2344614

28. Breiman L. *RANDOM FORESTS*.; 2001. https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf. Accessed August 22, 2019.

29. Demuth H, De Jesús B. *Neural Network Design 2nd Edition*. https://hagan.okstate.edu/NNDesign.pdf. Accessed August 22, 2019.

30. Max Kuhn. The caret Package. http://topepo.github.io/caret/index.html. Accessed September 10, 2019.

31. Géron A. *Hands-on Machine Learning with Scikit-Learn and TensorFlow : Concepts, Tools, and Techniques to Build Intelligent Systems*.

32. About us — scikit-learn 0.21.3 documentation. https://scikit-learn.org/stable/about.html. Accessed September 10, 2019.

33. h2o. AutoML: Automatic Machine Learning — H2O 3.26.0.3 documentation. http://docs.h2o.ai/h2o/latest-stable/h2o-docs/automl.html?highlight=automl. Accessed September 7, 2019.

34. The H2O Python Module — H2O documentation. http://docs.h2o.ai/h2o/latest-stable/h2o-py/docs/intro.html#what-is-h2o. Accessed September 10, 2019.

35. Goff DC, Lloyd-Jones DM, Bennett G, et al. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: A report of the American college of cardiology/American heart association task force on practice guidelines. *J Am Coll Cardiol*. 2014;63(25 PART B):2935-2959. doi:10.1016/j.jacc.2013.11.005

36. Fieller EC. Some Problems in Interval Estimation. *J R Stat Soc Ser B*. 1954;16:175-185. doi:10.2307/2984043

37. SAS® 9.4 Statements: Reference, Fifth Edition. http://support.sas.com/documentation/cdl/en/lestmtsref/69738/HTML/default/viewer.htm#n1i8w2bwu1fn5kn1gpxj18xttbb0.htm. Accessed August 20, 2017.

38. The Alan Turing Institute. Turing Lecture: Transforming medicine through AI-enabled healthcare pathways - YouTube. https://www.youtube.com/watch?v=TWI-WIoWvfk&feature=youtu.be&_cldee=dGplZXJkLnZhbnN0YWFAbWFuY2hlc3Rlci5hYy51aw%3D%3D&recipientid=contact-d2c6e6742b58e811812370106faae7f1-40027ba23fa146c189b8a3077e37916a&esid=2d3fc1d1-6593-e911-a98b-002248014cd6. Accessed September 24, 2019.

39. Makridakis S, Spiliotis E, Assimakopoulos V. Statistical and Machine Learning forecasting methods: Concerns and ways forward. Hernandez Montoya AR, ed. *PLoS One*. 2018;13(3):e0194889. doi:10.1371/journal.pone.0194889

40. Steyerberg EW. *Clinical Prediction Models : A Practical Approach to Development, Validation, and Updating*. Springer; 2009.

41. Kattan MW, Hess KR, Beck JR. Experiments to determine whether recursive partitioning (CART) or an artificial neural network overcomes theoretical limitations of cox proportional hazards regression. *Comput Biomed Res*. 1998;31(5):363-373. doi:10.1006/cbmr.1998.1488

42. Desai RJ, Wang S V., Vaduganathan M, Evers T, Schneeweiss S. Comparison of Machine Learning Methods With Traditional Models for Use of Administrative Claims With Electronic Medical Records to Predict Heart Failure Outcomes. *JAMA Netw open*. 2020;3(1):e1918962. doi:10.1001/jamanetworkopen.2019.18962

43. Kakadiaris IA, Vrigkas M, Yen AA, Kuznetsova T, Budoff M, Naghavi M. Machine Learning Outperforms ACC/AHA CVD Risk Calculator in MESA. doi:10.1161/JAHA.118.009476

44. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. RANDOM SURVIVAL FORESTS 1. *Ann Appl Stat*. 2008;2(3):841-860. doi:10.1214/08-AOAS169

45. Kvamme H, Borgan Ø, Scheel I. *Time-to-Event Prediction with Neural Networks and Cox Regression*. Vol 20.; 2019. http://jmlr.org/papers/v20/18-424.html. Accessed November 29, 2019.

46. Lee YH, Bang H, Kim DJ. How to Establish Clinical Prediction Models. *Endocrinol Metab (Seoul, Korea)*. 2016;31(1):38-44. doi:10.3803/EnM.2016.31.1.38

47. Austin PC, Steyerberg EW. Interpreting the concordance statistic of a logistic regression model: Relation to the variance and odds ratio of a continuous explanatory variable. *BMC Med Res Methodol*. 2012;12. doi:10.1186/1471-2288-12-82

48. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. *JAMA Intern Med*. 2018;178(11):1544-1547. doi:10.1001/jamainternmed.2018.3763

49. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD Statement. *Eur Urol*. 2015;67(6):1142-1151. doi:10.1016/j.eururo.2014.11.025

50. Kerr KF, Janes H. First things first: Risk model performance metrics should reflect the clinical application. *Stat Med*. 2017;36(28):4503. doi:10.1002/SIM.7341

51. Vollmer S, Mateen BA, Bohner G, et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *BMJ*. 2020;368. doi:10.1136/bmj.l6927

52. Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, et al. Explainable Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf Fusion*. 2020;58:82-115. doi:10.1016/j.inffus.2019.12.012

53. Menon AK. *Large-Scale Support Vector Machines: Algorithms and Theory*.

54. Type Package Title A Fast Implementation of Random Forests. 2020.

doi:10.1080/10618600.2014.983641

55. CRAN - Package randomForestSRC. https://cran.r-project.org/web/packages/randomForestSRC/index.html. Accessed April 25, 2020.

56. *Package "Rpart."*; 2019. https://cran.r-project.org/package=rpart. Accessed April 25, 2020.

4.12 Supplementary Online Content

eTable 4.12.1. Description of the machine learning and statistical models included in this study and the key parameters

eTable 4.12.2. Performance indicators of machine learning and statistical models in overall cohort with logistic caret model as reference model

eTable 4.12.3.1. Performance indicators of machine learning and statistical models in cohort without censoring with QRISK3 model as reference model

eTable 4.12.3.2. Performance indicators of machine learning and statistical models in cohort without censoring with logistic caret model as reference model

eTable 4.12.4.1. More performance indicators of machine learning and statistical models

eTable 4.12.4.2. More performance indicators of machine learning and statistical models in cohort without censoring

eTable 4.12.5.1. Comparison of individual risk predictions of machine learning and statistical models in overall cohort (with as reference the risk predictions of the QRISK3)

eTable 4.12.5.2. Comparison of individual risk predictions of machine learning and statistical models in overall cohort (with as reference the risk predictions of the logistic Caret model)

eTable 4.12.5.3. Comparison of the individual risk predictions of machine learning and statistical models in cohort without censoring (with as reference the risk predictions of the QRISK3 model)

eTable 4.12.5.4: Comparison of the individual risk predictions of machine learning and statistical models in cohort without censoring (with as reference the risk predictions of the Logistic Caret model)

eTable 4.12.6. Spearman correlations of machine learning models and statistical models in risk groups (logistic Caret predicted risk between 7%~8%)

eTable 4.12.7. Reclassification of individual risk predictions of machine learning and statistical models with 10% as threshold

eTable 4.12.8. Reclassification of individual risk predictions of Caret neural network models with different hyperparameters

eTable 4.12.9. Inconsistency of individual risk prediction between machine learning models derived from overall cohort and cohort without censoring

eTable 4.12.10. Performance indicators of machine learning and statistical models developed in South and validated in North England

eTable 4.12.11. Performance indicators of machine learning and statistical models with lower number of predictors

eFigure 4.12.1. Flow chart of sample splitting and model fitting process

eFigure 4.12.2.1. Calibration slope of machine learning models and statistical models in overall cohort

c. Survival framework (Observed events consider censoring)
d. Binary framework (Observed events did not consider censoring)

eFigure 4.12.2.2. Calibration slope of machine learning models and statistical models in cohort without censoring

a. Survival framework
b. Binary framework

eFigure 4.12.3.1. Calibration plots in machine learning models of Caret in overall cohort and cohort without censoring

eFigure 4.12.3.2. Calibration plots in statistical logistic models in overall cohort and cohort without censoring

eFigure 4.12.3.3. Calibration plots in Cox proportional hazard models in overall cohort and cohort without censoring

eFigure 4.12.3.4. Calibration plots in parametric survival models in overall cohort and cohort without censoring

eFigure 4.12.3.5. Calibration plots in machine learning models of Sklearn in overall cohort and cohort without censoring

eFigure 4.12.3.6. Calibration plots in machine learning models of h2o in overall cohort and cohort without censoring

eFigure 4.12.4. 95% range of individual risk predictions with machine learning and statistical models stratified by deciles of predicted risks with Caret logistic model

a. Overall cohort
b. Cohort without censoring

eFigure 4.12.5. 95% range of individual risk predictions with Caret neural network models with different grid searched best hyperparameters stratified by deciles of predicted risks with models with the most frequent selected hyperparameters

eFigure 4.12.6. Distribution of individual risk predictions with machine learning and statistical models developed in practices from South and tested in practices from North England

eFigure 4.12.7. Distribution of individual risk predictions with machine learning and statistical models developed with predictors of age and sex plus 1/3, 1/2, 2/3 of all predictors

eFigure 4.12.8. Distribution of age among removed patients due to censoring (death patients excluded)

eTable 4.12.1 describes the 19 model families used in main study and selected hyperparameters in grid search process.

eTable 4.12.2 shows the model performance among all models with logistic model Caret as reference. Like the main study, most models had similar model performance.

eTable 4.12.3.1-4.12.3.2 shows the model performance of machine learning and statistical models in the cohort without censoring. Though all models generally had a lower C-statistic than the models from the cohort with censoring (Table 4.2 in the main manuscript), the performance of these models was comparable.

eTable 4.12.4.1-4.12.4.2 shows more model performance measures including F1 score, balanced accuracy, negative predictive value (NPV) and specificity with threshold as 7.5% in binary framework in overall cohort and cohort without censoring. In general, all models had a few slightly better measures than others but also had a few slightly worse measures than other models. This was because these measures are a trade-off and being influenced by the selected threshold (i.e. a different threshold say 10% rather than 7.5% would change values of these measures).

eTable 4.12.5.1-4.12.5.4 shows the inconsistencies of range of individual CVD risk predictions for different strata of predictions with the QRISK3 and logistic Caret model as reference in the overall cohort and cohort without censoring. Logistic models and machine learning models which ignore censoring substantially underestimated patient risks (eTable 4.12.5.1), predictions for same individual patients varied substantially (eTable 4.12.5.2). Removing patients with censoring makes models overestimated patients risk compared to QRISK3 (eTable 4.12.5.3) and it did not change the magnitude of inconsistency of individual risk prediction (eTable 4.12.5.4).

eTable 4.12.6 shows the low correlation between individual risk prediction among different machine learning models. The results were consistent with the Figure 4.1b: machine learning models with similar model performance predicted the same patients differently.

eTable 4.12.7 is a similar reclassification table as Table 4.4 except using 10% rather than 7.5% threshold. Similar reclassification was found as in the main study. Of the 735,474 patients with a CVD risk $\leqslant$10%, 10% were reclassified when using another model. Of the 180,005 patients with a CVD risk >10%, 62.9% were reclassified when using another model.

eTable 4.12.8 shows the reclassification effects of choosing different hyperparameters for the same machine learning model family on individual risk prediction. Neural network Caret with hyperparameters of size (number of neurons) and decay (parameter to control the overfitting) was used as an exemplar. From 100 best models grid searched from random samples, there were 17 groups of best selected hyperparameters. Using the average risk predictions from the most frequent group as reference (in this case the biggest model group has 17 neural network models with size=3 and decay=3.5 from grid search process), risk predictions of the same patients from the same model with different best hyperparameters were compared in **eFigure 4.12.5**. The reclassification **eTable 4.12.8** shows that among 129,991 patients over threshold 7.5%, 11% of them would be reclassified if a different best hyperparameter was chosen. However, in the main study, the variation of individual risk predictions within the same model family was eliminated by model ensembling with soft voting (averaging). This additional analysis shows the inconsistency of individual risk prediction among machine learning models could be worse considering variation of individual risk prediction among the same model family, and current approach to find the best hyperparameters is data-driven and in lack of a principal way to determine what hyperparameters were more proper before fitting the model.

eTable 4.12.9 used the machine learning models from the cohort without censoring to calculate risk for the full cohort (the cohort with censoring), and then the risk prediction of the same model of the same risk group of patients was compared. Even within the same model, the risk predictions for the same patients did not agree to each other. Machine learning models derived from a cohort without censoring predicted larger range of risk than the same machine learning models derived from a cohort with censoring on the same patients. The same applied to the fitted statistical models, as both models were fitted on a biased cohort, i.e. patients censoring were artificially removed. This indicates that models developed from a censored removed cohort (this has been done in several machine learning papers) should not be used in a cohort with censoring, as ignoring censoring introduces bias (mis-calibration) to individual risk prediction.

eTable 4.12.10 shows the model performance of models developed from practices from South England and validated in practices from North England. It also shows similar inconsistency of individual risk prediction was found as in main manuscript (**eFigure 4.12.6**). As expected, models have similar model performance in the North and South England, which is consistent to the main study. However, models developed from practices from South generally have lower model performance in practices from North English practices compared to practices from South England in either machine learning models or statistical models. Previous study using the random effects model showed that there was practice variability among UK practices with an effect on individual risk prediction [1]. In this study, this sensitive analysis showed that machine learning models did not automatically capture this variability in coding.

eTable 4.12.11 compares the model performance of models with age and sex plus 1/3, 1/2 and 2/3 randomly sampled predictors. Except random forest caret, all models have similar model performance among each other as in the main study and similar inconsistency of individual risk prediction was found as in main manuscript (eFigure 4.12.7). Random forest Caret model was underfitted with 1/3 predictors as the final model grid searched the best parameter mtry (number of predictors used to grow branches) as the same number of available predictors. As expected, model performance of random forest Caret improved with the increase of number of predictors. Random forest Caret model has the similar model performance as other models with full predictors indicate enough predictors were considered in the study.

eFigure 4.12.1 visualises the workflow of sample splitting and model fitting process.

eFigure 4.12.2.1 -4.12.2.2 shows the calibration slopes of all machine learning models and statistical models in overall cohort and cohort without censoring. It shows that both models were well calibrated in their own framework (i.e. Cox models in survival framework and logistic models and machine learning models in binary framework). However, Logistic models and machine learning models ignoring censoring were mis-calibrated (eFigure 4.12.2.1a) in survival framework (i.e. cohorts with censoring). It appears that models were well calibrated in both framework in cohort without censoring, as survival framework and binary framework were similar once patients with censoring were removed. However, artificially removing patients with censoring makes the cohort non-representative as censoring often occurs over time. Figure 4.2a in main manuscript has shown that logistic models and machine learning models developed from cohort without censoring over-estimated patient risks in overall cohort compared to QRISK3.

eFigure 4.12.3.1 – 4.12.3.6 shows the calibration plots (same information as eFigure4.12.2.1-4.12.2.2 but with different visualisation) of all machine learning models and statistical models. Machine learning models had good calibration in a binary framework (i.e. treating the patients with censoring as non-events) irrespective of the cohort with censoring or cohort without censoring. However, the calibration figures of the cohort with censoring showed that machine learning models had poor calibration in the survival framework (i.e. considering the effects of censoring). Once censoring was removed, the calibration of machine learning models improved (shown in calibration plot from cohort without out censoring). This suggests that although machine learning models which ignore patient-censoring can have good calibration in a binary framework, it was poor calibrated in survival framework. Cox models including QRISK3 and Framingham have very good calibration in a survival framework but very poor calibration in a binary framework as they considered censoring with time-to-events outcome. However censoring is very common with long-term risks and should not be ignored.

eFigure 4.12.4a-4.12.4b is a similar plot as **Figure 4.4a-4.4b** in main manuscript except it used logistic Caret model as reference rather than QRISK3. It shows similar finding that inconsistency of individual risk prediction among models were mainly in higher risk group patients.

eFigure 4.12.8 shows that among patients who were censored (death excluded), younger patients were the majority. This indicates the reason that average age in the cohort without censoring is higher than the cohort with censoring is the effects of younger patients transferred out from practices that compensated the effects of older patients who died during the follow-up.

eTable 4.12.1: Description of the machine learning and statistical models included in this study and the key parameters

	Package description	Model description	Key parameters selected by analyst for grid search
Caret			
Logistic	Classification And REgression Training (Caret) is a R package which has a series of functions to create predictive models in a structural and organized way. It contains functions which can be used to split data, pre-process data, select predictors, tune model and resample [2]	Logistic model is a type of generalised linear model with a binary variable as outcome [3]	None
Random forest		Random forest is an ensemble machine learning model which combines the predictions from multiple decision-trees where each tree grows from an independent sample of predictors [4]	**mtry:** number of randomly selected predictors as candidates at each split **ntree:** number of trees
Neural network		Neural network is a machine learning model whose model-structure mimics the structure of animal brain using hidden layers and neurons in those hidden layer [5]	**size:** number of units in hidden layer (neural network in Caret only fits one hidden layer) **decay:** a regularization parameter to control over-fitting (higher decay means less chance of over-fitting)
Statistical logistic model			
Logistic model	Standard statistical way to fit logistic model with glm() function from basic R library [6]	Logistic model fitted with standard statistical approach	None
Cox proportional hazards model			
QRISK3	Effects of predictors on hazard ratio are assumed to be multiplicative. Cox models take into account censoring [7]	QRISK3 was derived from a UK cohort [8]	None
Framingham		Framingham model was derived from a US cohort [9]	None
Local Cox model		Re-fitted Cox model using the same training cohorts and validation cohorts as machine learning models.	None
Parametric survival model			
Weibull distribution	Parametric survival models are alternatives of Cox model in survival analysis. It assumes survival time follows a known parametric distribution (e.g. Weibull distribution). Parametric survival models also take into account censoring naturally [10].	Assume survival time follows Weibull distribution	None
Gaussian distribution		Assume survival time follows Gaussian distribution	None
Logistic distribution		Assume survival time follows Logistic distribution	None
Sklearn			
Logistic	Scikit-learn (Sklearn) is a free machine learning library written in Python. It supports different machine learning algorithms including classification, regression and clustering tasks [11]	Using the same mathematical algorithm as Caret but written by different computer language (Python rather than R)	**penalty:** L1 Lasso regression or L2 Ridge regression (penalty term adds to loss function to increase model-generalizability) **C**: Inverse of regularization strength to control over-fitting (smaller value means stronger regularization and more-generalizability)
Random forest		Using the same mathematical algorithm as Caret but memory-optimisation and language advantage allow python version to fit more trees than Caret version.	**n_estimators**: number of trees **max_features**: number of predictors to consider when searching for the best split

	Package description	Model description	Key parameters selected by analyst for grid search
Neural network		Using the same mathematical algorithm as Caret but additional options provided by Sklearn to further control the structure and fitting process of neural network	**hidden_layer_sizes**: control number of hidden layers and number of neurons in each hidden layer **activation:** activation function for the hidden layer calculation. **solver:** different methods to optimise weights (beta) estimation
Gradient boosting classifier		Gradient boosting is a machine learning boosting method to train model by adding new predictor to the residual error of previous predictor rather than using all predictors at once [12]	**n_estimators:** the number of boosting stages to perform (larger means better performance but higher risk of overfitting) **learning_rate**: shrinks the contribution of each tree (a trade-off to n_estimators to control overfitting) **max_features**: number of predictors to consider when searching for the best split
extra-trees		Extra-trees model is similar to random forest except that random forest grows decision-trees by searching for the best splitting while Extra-trees uses random split for each decision-tree [12]	**n_estimators:** number of trees **max_features**: number of predictors to consider when searching for the best split
h2o			
Logistic	h2o is a Java-based machine learning library which has been implanted to both of R and python. Its main strength consists of memory allocation and the ability to distributed and paralleled machine learning process (which accelerates the model creating process) [13]	Using the same mathematical algorithm as Caret and Sklearn but being optimised for better memory allocation and parallelizing	None
Random forest		Using the same mathematical algorithm as Caret and Sklearn, but further memory-optimization and parallelizing allow to fit even more trees than Sklearn.	**max_depth:** maximum tree depth **mtries**: number of randomly selected predictors as candidates at each split **ntrees:** number of trees
Neural network		Using the same mathematical algorithm as Caret and Sklearn, but parallelising makes it possible to fit neural networks with large number of hidden layers with a more complex structure (deep learning)	**hidden:** control number of hidden layers and number of neurons in each hidden layer
autoML		AutoML is an automatic machine learning model training algorithm provided by h2o, which choose a best model among several candidate machine learning models such as gradient boosting machine, deep neural net and extremely randomised forest [14]	None

eTable 4.12.2: Performance indicators of machine learning and statistical models in overall cohort with logistic caret model as reference model

	Model performance* (95% range #)				Average absolute change of model performance (95% range)
	C-statistic (2.5% ~ 97.5%) #	**Brier score (2.5% ~ 97.5%) #**	**Recall (Sensitivity) (2.5% ~ 97.5%) #**	**Precision (PPV) (2.5% ~ 97.5%) #**	**C-statistic (2.5% ~ 97.5%) #**
Logistic (Caret)	0.879 (0.879, 0.879)	0.028 (0.028, 0.028)	0.615 (0.609, 0.620)	0.163 (0.162, 0.164)	Reference model
Random forest (Caret)	0.869 (0.867, 0.869)	0.028 (0.028, 0.028)	0.656 (0.620, 0.675)	0.144 (0.139, 0.153)	-1.21% (-1.35%, -1.10%)
Neural network (Caret)	0.878 (0.867, 0.880)	0.028 (0.027, 0.028)	0.670 (0.642, 0.687)	0.148 (0.141, 0.153)	-0.16% (-1.34%, 0.05%)
Statistical logistic model	0.879 (0.879, 0.879)	0.028 (0.028, 0.028)	0.614 (0.607, 0.620)	0.163 (0.162, 0.164)	+0.00% (0.00%, 0.00%)
QRISK3	0.879	0.031	0.834	0.107	0.00% (-0.04%, 0.03%)
Framingham	0.865	0.031	0.892	0.085	-1.66% (-1.69%, -1.63%)
Local Cox model	0.877 (0.877, 0.878)	0.032 (0.031, 0.032)	0.810 (0.804, 0.816)	0.112 (0.110, 0.113)	-0.22% (-0.26%, -0.18%)
Parametric survival model (Weibull)	0.877 (0.876, 0.877)	0.031 (0.031, 0.032)	0.810 (0.804, 0.815)	0.111 (0.110, 0.113)	-0.30% (-0.34%, -0.26%)
Parametric survival model (Gaussian)	0.876 (0.876, 0.877)	0.031 (0.030, 0.031)	0.834 (0.830, 0.839)	0.104 (0.103, 0.105)	-0.34% (-0.38%, -0.30%)
Parametric survival model (Logistic)	0.876 (0.875, 0.876)	0.031 (0.031, 0.032)	0.796 (0.791, 0.802)	0.114 (0.113, 0.115)	-0.37% (-0.41%, -0.33%)
Logistic (Sklearn)	0.879 (0.879, 0.879)	0.028 (0.028, 0.028)	0.615 (0.609, 0.620)	0.163 (0.161, 0.164)	-0.01% (-0.04%, 0.00%)
Random forest (Sklearn)	0.872 (0.871, 0.873)	0.028 (0.028, 0.028)	0.670 (0.661, 0.679)	0.142 (0.140, 0.144)	-0.81% (-0.91%, -0.71%)
Neural network (Sklearn)	0.872 (0.832, 0.879)	0.028 (0.028, 0.029)	0.556 (0.174, 0.692)	0.163 (0.137, 0.224)	-0.85% (-5.41%, -0.06%)
Gradient boosting classifier (Sklearn)	0.878 (0.877, 0.878)	0.028 (0.028, 0.028)	0.642 (0.623, 0.657)	0.154 (0.150, 0.157)	-0.17% (-0.28%, -0.09%)
extra-trees (Sklearn)	0.863 (0.861, 0.864)	0.028 (0.028, 0.029)	0.639 (0.628, 0.650)	0.139 (0.136, 0.141)	-1.89% (-2.05%, -1.77%)
Logistic (h2o)	0.879 (0.878, 0.879)	0.028 (0.028, 0.028)	0.615 (0.608, 0.621)	0.162 (0.161, 0.164)	-0.06% (-0.09%, -0.04%)
Random forest (h2o)	0.877 (0.877, 0.878)	0.028 (0.028, 0.028)	0.646 (0.631, 0.659)	0.152 (0.149, 0.154)	-0.23% (-0.29%, -0.16%)
Neural network (h2o)	0.875 (0.870, 0.879)	0.028 (0.028, 0.031)	0.552 (0.163, 0.780)	0.169 (0.118, 0.238)	-0.45% (-1.10%, -0.05%)
autoML (h2o)	0.879 (0.879, 0.880)	0.028 (0.028, 0.028)	0.616 (0.605, 0.642)	0.162 (0.157, 0.164)	-0.02% (-0.05%, 0.03%)

*** Model performance was calculated in binaray framework. Threshold 7.5% was used to calculate precision and recall for all models.**

 # 95% range (2.5% ~ 97.5%) of model performance was derived from 100 random samples.

eTable 4.12.3.1: Performance indicators of machine learning and statistical models in cohort without censoring with QRISK3 as reference model

	Model performance*				Average absolute change of model performance
	C-statistic	**Brier score**	**Recall (Sensitivity)**	**Precision (PPV)**	**C-statistic**
Logistic (Caret)	0.851	0.125	0.957	0.346	+0.49%
Random forest (Caret)	0.849	0.126	0.926	0.384	+0.34%
Neural network (Caret)	0.849	0.126	0.953	0.354	+0.25%
Statistical logistic model	0.851	0.125	0.957	0.346	+0.49%
QRISK3	0.847	0.150	0.844	0.455	Reference
Framingham	0.815	0.161	0.899	0.385	-3.74%
Local Cox model	0.850	0.126	0.968	0.330	+0.39%
Parametric survival model (Weibull)	0.849	0.128	0.955	0.347	+0.25%
Parametric survival model (Gaussian)	0.848	0.130	0.932	0.379	+0.23%
Parametric survival model (Logistic)	0.848	0.129	0.925	0.386	+0.20%
Logistic (Sklearn)	0.851	0.125	0.957	0.346	+0.49%
Random forest (Sklearn)	0.849	0.126	0.957	0.346	+0.30%
Neural network (Sklearn)	0.852	0.125	0.965	0.336	+0.63%
Gradient boosting classifier (Sklearn)	0.853	0.124	0.953	0.354	+0.74%
extra-trees (Sklearn)	0.845	0.127	0.954	0.345	-0.17%
Logistic (h2o)	0.849	0.126	0.957	0.343	+0.28%
Random forest (h2o)	0.851	0.125	0.960	0.344	+0.52%
Neural network (h2o)	0.852	0.126	0.927	0.386	+0.65%
autoML (h2o)	0.853	0.125	0.952	0.356	+0.71%

*** Model performance was calculated in binaray framework. Threshold 7.5% was used to calculate precision and recall for all models.**

eTable 4.12.3.2: Performance indicators of machine learning and statistical models in cohort without censoring with logistic caret model as reference model

	Model performance*				Average absolute change of model performance
	C-statistic	**Brier score**	**Recall (Sensitivity)**	**Precision (PPV)**	**C-statistic**
Logistic (Caret)	0.851	0.125	0.957	0.346	Reference
Random forest (Caret)	0.849	0.126	0.926	0.384	-0.15%
Neural network (Caret)	0.849	0.126	0.953	0.354	-0.24%
Statistical logistic model	0.851	0.125	0.957	0.346	0.00%
QRISK3	0.847	0.150	0.844	0.455	-0.48%
Framingham	0.815	0.161	0.899	0.385	-4.21%
Local Cox model	0.850	0.126	0.968	0.330	-0.10%
Parametric survival model (Weibull)	0.849	0.128	0.955	0.347	-0.24%
Parametric survival model (Gaussian)	0.848	0.130	0.932	0.379	-0.25%
Parametric survival model (Logistic)	0.848	0.129	0.925	0.386	-0.28%
Logistic (Sklearn)	0.851	0.125	0.957	0.346	+0.00%
Random forest (Sklearn)	0.849	0.126	0.957	0.346	-0.18%
Neural network (Sklearn)	0.852	0.125	0.965	0.336	+0.15%
Gradient boosting classifier (Sklearn)	0.853	0.124	0.953	0.354	+0.25%
extra-trees (Sklearn)	0.845	0.127	0.954	0.345	-0.66%
Logistic (h2o)	0.849	0.126	0.957	0.343	-0.21%
Random forest (h2o)	0.851	0.125	0.960	0.344	+0.03%
Neural network (h2o)	0.852	0.126	0.927	0.386	+0.16%
autoML (h2o)	0.853	0.125	0.952	0.356	+0.22%

*** Model performance was calculated in binaray framework. Threshold 7.5% was used to calculate precision and recall for all models.**

eTable 4.12.4.1: More performance indicators of machine learning and statistical models

	Model performance* (95% range #)			
	F1 score (2.5% ~ 97.5%) #	**Balanced accuracy (2.5% ~ 97.5%) #**	**NPV (2.5% ~ 97.5%) #**	**Specificity (2.5% ~ 97.5%) #**
Logistic (Caret)	0.258 (0.256, 0.259)	0.756 (0.754, 0.758)	0.986 (0.986, 0.986)	0.897 (0.895, 0.899)
Random forest (Caret)	0.236 (0.230, 0.245)	0.765 (0.754, 0.770)	0.987 (0.986, 0.988)	0.873 (0.864, 0.888)
Neural network (Caret)	0.242 (0.234, 0.248)	0.772 (0.752, 0.777)	0.988 (0.987, 0.988)	0.874 (0.864, 0.885)
Statistical logistic model	0.258 (0.256, 0.259)	0.756 (0.754, 0.758)	0.986 (0.986, 0.986)	0.898 (0.896, 0.900)
QRISK3	0.190	0.804	0.993	0.775
Framingham	0.155	0.790	0.995	0.688
Local Cox model	0.197 (0.194, 0.199)	0.800 (0.799, 0.801)	0.992 (0.992, 0.992)	0.791 (0.786, 0.796)
Parametric survival model (Weibull)	0.196 (0.194, 0.198)	0.800 (0.799, 0.800)	0.992 (0.992, 0.992)	0.789 (0.785, 0.794)
Parametric survival model (Gaussian)	0.185 (0.183, 0.187)	0.800 (0.800, 0.801)	0.993 (0.993, 0.993)	0.766 (0.762, 0.771)
Parametric survival model (Logistic)	0.199 (0.197, 0.201)	0.797 (0.796, 0.798)	0.992 (0.992, 0.992)	0.798 (0.795, 0.802)
Logistic (Sklearn)	0.258 (0.256, 0.259)	0.756 (0.754, 0.758)	0.986 (0.986, 0.986)	0.897 (0.895, 0.899)
Random forest (Sklearn)	0.235 (0.232, 0.237)	0.769 (0.766, 0.772)	0.988 (0.988, 0.988)	0.869 (0.864, 0.872)
Neural network (Sklearn)	0.240 (0.191, 0.272)	0.728 (0.576, 0.777)	0.984 (0.973, 0.988)	0.901 (0.858, 0.979)
Gradient boosting classifier (Sklearn)	0.248 (0.244, 0.251)	0.763 (0.757, 0.768)	0.987 (0.986, 0.987)	0.885 (0.880, 0.890)
extra-trees (Sklearn)	0.228 (0.225, 0.231)	0.755 (0.752, 0.758)	0.987 (0.986, 0.987)	0.871 (0.867, 0.875)
Logistic (h2o)	0.257 (0.255, 0.258)	0.756 (0.753, 0.758)	0.986 (0.986, 0.986)	0.897 (0.895, 0.899)
Random forest (h2o)	0.246 (0.243, 0.248)	0.764 (0.760, 0.768)	0.987 (0.987, 0.988)	0.883 (0.878, 0.887)
Neural network (h2o)	0.246 (0.172, 0.273)	0.728 (0.573, 0.795)	0.984 (0.973, 0.991)	0.904 (0.811, 0.983)
autoML (h2o)	0.257 (0.252, 0.259)	0.756 (0.753, 0.765)	0.986 (0.986, 0.987)	0.897 (0.888, 0.900)

*** Model performance was calculated in binaray framework. Threshold 7.5% was used to calculate precision and recall for all models.**

 # 95% range (2.5% ~ 97.5%) of model performance was derived from 100 random samples.

eTable 4.12.4.2: More performance indicators of machine learning and statistical models in cohort without censoring

	Model performance* (95% range #)			
	F1 score (2.5% ~ 97.5%) #	**Balanced accuracy (2.5% ~ 97.5%) #**	**NPV (2.5% ~ 97.5%) #**	**Specificity (2.5% ~ 97.5%) #**
Logistic (Caret)	0.508	0.702	0.972	0.447
Random forest (Caret)	0.543	0.736	0.960	0.547
Neural network (Caret)	0.516	0.711	0.970	0.470
Statistical logistic model	0.509	0.703	0.971	0.449
QRISK3	0.592	0.768	0.936	0.692
Framingham	0.539	0.730	0.948	0.561
Local Cox model	0.492	0.683	0.976	0.399
Parametric survival model (Weibull)	0.510	0.704	0.971	0.452
Parametric survival model (Gaussian)	0.539	0.733	0.962	0.534
Parametric survival model (Logistic)	0.545	0.738	0.960	0.552
Logistic (Sklearn)	0.508	0.702	0.972	0.447
Random forest (Sklearn)	0.508	0.702	0.971	0.448
Neural network (Sklearn)	0.498	0.691	0.975	0.417
Gradient boosting classifier (Sklearn)	0.516	0.711	0.971	0.468
extra-trees (Sklearn)	0.507	0.701	0.969	0.447
Logistic (h2o)	0.505	0.698	0.971	0.439
Random forest (h2o)	0.506	0.700	0.973	0.441
Neural network (h2o)	0.545	0.739	0.961	0.551
autoML (h2o)	0.518	0.713	0.970	0.475

*** Model performance was calculated in binaray framework. Threshold 7.5% was used to calculate precision and recall for all models.**

 # 95% range (2.5% ~ 97.5%) of model performance was derived from 100 random samples.

eTable 4.12.5.1: Comparison of individual risk predictions of machine learning and statistical models in overall cohort (with as reference the risk predictions of the QRISK3)

	Range of individual risk predictions (2.5^{th}~97.5^{th}) with other models compared to those from QRISK3 model								
	<6%	**6~7%**	**7~8%**	**8~9%**	**9~10%**	**10~11%**	**11~12%**	**12~13%**	**≥ 13%**
Caret									
Logistic	0.1~2.3	1.4~3.7	1.6~4.3	1.8~4.9	2.1~5.5	2.3~6.1	2.5~6.7	2.7~7.4	4.2~35.7
Random forest	0.0~2.9	1.7~6.3	2.0~7.3	2.4~8.1	2.8~9.0	3.1~9.9	3.5~10.9	3.9~11.5	5.5~30.4
Neural network	0.1~2.5	1.4~4.4	1.7~5.2	1.9~6.1	2.3~6.9	2.5~7.6	2.8~8.4	3.2~9.2	5.3~24.6
Statistical logistic model									
Statistical logistic model	0.1~2.3	1.4~3.7	1.6~4.3	1.8~4.9	2.1~5.5	2.3~6.1	2.5~6.6	2.7~7.3	4.2~35.5
Cox model									
QRISK3	Reference	Reference	Reference	Reference	Reference	Reference	Reference	Reference	Reference
Framingham	0.0~10.0	4.5~15.3	5.0~16.7	5.3~18.1	5.9~19.4	6.2~20.8	6.2~22.2	6.7~23.2	8.2~49.7
Local Cox model	0.4~5.4	4.1~8.2	4.7~9.2	5.2~10.4	5.7~11.3	6.2~12.4	6.7~13.7	7.2~14.7	10.3~61.1
Parametric survival model									
Parametric survival model (Weibull)	1.0~5.5	4.2~8.3	4.7~9.4	5.2~10.7	5.7~11.7	6.2~12.8	6.7~14.0	7.2~15.1	10.1~59.2
Parametric survival model (Gaussian)	1.0~6.2	4.7~9.7	5.3~11.1	6.0~12.4	6.6~13.6	7.2~14.9	7.7~16.1	8.3~17.3	11.4~49.4
Parametric survival model (Logistic)	1.0~5.1	3.8~8.2	4.3~9.5	4.8~10.8	5.3~12.0	5.9~13.3	6.3~14.7	6.9~15.9	9.8~57.5
Sklearn									
Logistic	0.1~2.3	1.4~3.7	1.6~4.3	1.8~4.9	2.1~5.5	2.3~6.1	2.5~6.7	2.7~7.3	4.3~35.5
Random forest	0.0~3.1	1.8~6.4	2.1~7.4	2.5~8.2	2.9~9.0	3.4~9.9	3.6~11.0	4.1~11.6	5.8~30.6
Neural network	0.1~2.4	1.2~4.8	1.4~5.6	1.6~6.4	1.9~7.1	2.1~7.6	2.3~8.3	2.5~8.9	4.1~23.9
Gradient boosting classifier	0.1~2.5	1.4~4.4	1.7~5.2	2.0~5.8	2.3~6.7	2.7~7.3	2.9~8.3	3.3~9.2	5.2~30.5
extra-trees	0.0~3.2	1.6~6.2	2.0~7.1	2.3~7.9	2.6~8.7	3.0~9.5	3.2~10.5	3.7~11.2	5.6~29.4
h2o									

	Range of individual risk predictions (2.5^{th}~97.5^{th}) with other models compared to those from QRISK3 model								
	<6%	**6~7%**	**7~8%**	**8~9%**	**9~10%**	**10~11%**	**11~12%**	**12~13%**	**≥ 13%**
Logistic	0.1~2.4	1.4~3.8	1.6~4.4	1.8~5.0	2.0~5.6	2.3~6.1	2.4~6.7	2.7~7.4	4.3~35.0
Random forest	0.1~2.8	1.8~5.1	2.0~6.0	2.4~6.7	2.7~7.3	3.1~8.2	3.4~9.1	3.8~9.6	5.4~28.9
Neural network	0.1~2.2	1.2~4.0	1.4~4.7	1.6~5.4	1.9~6.1	2.1~6.7	2.2~7.5	2.4~8.2	4.0~29.2
autoML	0.1~2.3	1.4~3.8	1.6~4.3	1.8~4.9	2.1~5.5	2.3~6.1	2.5~6.7	2.8~7.4	4.3~35.0
Overall									
Soft voting *	0.3~3.5	2.7~5.2	3.1~6.0	3.5~6.8	3.9~7.5	4.3~8.3	4.6~9.0	5.1~9.8	7.2~36.4
All models #	0.1~4.9	1.5~10.5	1.8~11.6	2.0~12.7	2.4~13.6	2.6~14.7	2.9~15.7	3.2~16.7	5.0~44.8

*** 95% range of individual risk prediction from soft voting (averaging) of all models except the reference model**

 # 95% range of individual risk prediction from all models except the reference model

eTable 4.12.5.2: Comparison of individual risk predictions of machine learning and statistical models in overall cohort (with as reference the risk predictions of the logistic Caret model)

	Range of individual risk predictions (2.5^{th}~97.5^{th}) with other models compared to those from logistic Caret model								
	<6%	**6~7%**	**7~8%**	**8~9%**	**9~10%**	**10~11%**	**11~12%**	**12~13%**	**≥ 13%**
Caret									
Logistic	Reference	Reference	Reference	Reference	Reference	Reference	Reference	Reference	Reference
Random forest	0.0~6.7	4.9~14.5	5.8~16.1	6.6~17.7	7.4~19.1	8.2~20.0	9.1~20.9	10.0~21.8	12.5~33.6
Neural network	0.1~6.3	6.7~9.9	7.8~11.4	8.8~12.6	9.8~13.9	10.6~14.9	11.5~15.7	12.2~16.6	13.8~26.2
Statistical logistic model									
Statistical logistic model	0.1~5.0	6.0~6.9	7.0~7.9	8.0~8.9	9.0~9.9	10.0~10.9	11.0~11.9	12.0~12.9	13.2~40.9
Cox model									
QRISK3	0.1~12.5	10.9~23.4	12.4~26.2	13.9~28.6	15.4~31.3	16.7~33.5	18.2~35.2	19.5~36.6	23.7~59.6
Framingham	0.1~17.1	7.1~28.3	7.3~30.3	7.6~32.0	7.6~34.5	8.1~35.7	7.9~36.6	8.5~38.5	10.5~57.7
Local Cox model	0.4~11.0	10.5~19.0	11.9~21.7	13.4~24.0	14.8~26.9	16.1~29.5	17.5~31.5	18.9~33.3	23.1~69.0
Parametric survival model									
Parametric survival model (Weibull)	1.0~11.1	10.3~19.0	11.7~21.8	13.1~24.6	14.3~27.6	15.5~30.0	16.8~32.5	18.2~34.4	22.2~66.6
Parametric survival model (Gaussian)	1.0~12.6	11.3~21.2	12.8~23.8	14.1~26.0	15.4~28.4	16.7~30.4	17.9~32.4	19.0~33.6	22.6~54.0
Parametric survival model (Logistic)	1.0~11.0	9.7~20.8	11.2~23.9	12.6~27.0	13.9~30.3	15.2~33.0	16.5~35.4	17.9~37.1	22.3~63.1
Sklearn									
Logistic	0.1~5.1	6.0~7.0	7.0~8.0	8.0~9.0	9.0~10.0	10.0~11.0	11.0~12.0	12.0~13.0	13.2~40.9
Random forest	0.0~6.9	5.2~14.4	6.1~15.9	7.0~17.7	7.8~19.1	8.6~19.8	9.5~20.7	10.5~21.8	13.0~33.6
Neural network	0.1~5.3	4.5~9.4	5.1~10.5	5.8~11.4	6.4~12.3	7.1~13.1	7.7~13.9	8.3~14.7	10.2~26.6
Gradient boosting classifier	0.1~6.0	5.3~10.7	6.2~12.2	7.1~13.5	8.0~14.9	8.8~16.3	9.5~17.4	10.4~18.8	12.5~34.3
extra-trees	0.0~6.9	5.0~13.5	5.8~14.8	6.6~16.2	7.4~17.5	8.1~18.5	9.1~19.6	9.8~20.4	12.5~32.4
h2o									
Logistic	0.1~5.1	5.9~7.2	6.8~8.4	7.7~9.5	8.7~10.6	9.6~11.6	10.5~12.8	11.5~13.9	13.2~40.0

	Range of individual risk predictions (2.5^{th}~97.5^{th}) with other models compared to those from logistic Caret model								
	<6%	**6~7%**	**7~8%**	**8~9%**	**9~10%**	**10~11%**	**11~12%**	**12~13%**	**≥ 13%**
Random forest	0.1~5.9	5.5~12.2	6.3~13.7	7.0~15.4	7.8~16.7	8.6~17.5	9.4~18.6	10.4~19.6	12.8~32.1
Neural network	0.1~5.0	4.7~8.6	5.4~9.8	6.3~10.9	7.0~12.0	7.8~13.0	8.6~14.0	9.4~14.9	11.6~33.0
autoML	0.1~5.1	6.0~7.1	7.0~8.1	8.0~9.1	9.0~10.2	10.0~11.2	11.0~12.2	11.9~13.2	13.3~40.1
Overall									
Soft voting *	0.3~7.8	8.2~12.3	9.4~13.9	10.5~15.4	11.7~16.8	12.8~18.2	14.0~19.3	15.0~20.5	17.1~41.6
All models #	0.1~9.7	5.4~20.1	6.2~22.1	7.0~24.0	7.8~26.1	8.5~28.1	9.2~29.9	9.9~31.7	12.6~53.7

*** 95% range of individual risk prediction from soft voting (averaging) of all models except the reference model**

 # 95% range of individual risk prediction from all models except the reference model

eTable 4.12.5.3: Comparison of the individual risk predictions of machine learning and statistical models in cohort without censoring (with as reference the risk predictions of the QRISK3 model)

	Range of individual risk predictions (2.5^{th}~97.5^{th}) with other models compared to those from QRISK3 model								
	<6%	**6~7%**	**7~8%**	**8~9%**	**9~10%**	**10~11%**	**11~12%**	**12~13%**	**≥ 13%**
Caret									
Logistic	1.2~17.8	12.4~26.3	14.4~30.5	15.6~32.7	17.3~35.7	18.9~38.8	20.3~42.3	21.7~44.8	30.1~88.6
Random forest	0.2~17.9	6.5~31.4	7.9~35.8	9.5~37.9	10.7~40.4	12.6~44.7	13.5~46.9	14.9~48.9	24.1~85.8
Neural network	2.8~18.0	11.3~29.5	13.4~34.8	15.4~37.9	16.9~41.7	19.0~43.8	21.6~48.0	23.0~50.5	31.8~75.1
Statistical logistic model									
Statistical logistic model	1.2~17.7	12.3~26.2	14.3~30.4	15.5~32.5	17.2~35.6	18.7~38.7	20.2~42.2	21.6~44.7	30.0~88.6
Cox model									
QRISK3	Reference	Reference	Reference	Reference	Reference	Reference	Reference	Reference	Reference
Framingham	0.1~11.1	4.8~15.7	5.4~16.9	5.9~18.8	5.9~19.9	6.6~21.4	7.2~22.1	7.2~23.9	9.0~53.6
Local Cox model	1.8~15.6	10.7~22.1	12.2~25.2	13.1~27.1	14.3~29.8	15.6~32.5	16.4~35.1	17.3~37.4	24.2~90.2
Parametric survival model									
Parametric survival model (Weibull)	1.6~15.5	10.6~22.1	11.9~25.3	12.7~27.0	13.7~29.7	14.4~32.1	15.6~34.4	16.3~37.3	22.5~87.8
Parametric survival model (Gaussian)	1.0~13.8	8.1~21.4	9.5~24.3	10.3~26.2	11.3~28.9	12.5~30.8	13.6~33.2	14.4~36.4	20.4~76.5
Parametric survival model (Logistic)	1.0~12.7	7.6~20.1	9.0~23.4	9.6~25.1	10.6~28.1	11.7~30.1	12.6~32.8	13.5~36.5	19.4~80.3
Sklearn									
Logistic	1.2~17.8	12.4~26.3	14.4~30.5	15.6~32.7	17.3~35.7	18.8~38.8	20.3~42.3	21.7~44.8	30.1~88.6
Random forest	0.4~22.9	10.5~35.8	12.3~41.0	14.2~42.4	15.4~45.3	17.2~48.0	19.0~49.8	21.0~51.7	29.7~83.4
Neural network	0.3~18.8	12.5~30.1	14.5~34.4	16.1~36.8	17.6~41.0	19.6~43.9	21.6~47.1	22.5~50.1	30.8~84.1
Gradient boosting classifier	1.0~19.2	12.4~31.6	14.4~35.7	15.9~38.3	17.5~42.2	19.0~47.2	21.0~48.2	22.6~53.1	30.1~87.2
extra-trees	0.3~24.0	9.1~38.0	10.8~42.4	12.7~45.8	13.6~48.7	15.0~51.2	16.5~53.6	17.7~55.5	26.8~86.3
h2o									
Logistic	1.3~17.9	12.2~25.6	14.3~29.1	15.6~31.5	16.9~34.6	18.0~37.6	20.3~40.4	21.3~42.8	30.2~87.2

	Range of individual risk predictions (2.5^{th}~97.5^{th}) with other models compared to those from QRISK3 model								
	<6%	**6~7%**	**7~8%**	**8~9%**	**9~10%**	**10~11%**	**11~12%**	**12~13%**	**≥ 13%**
Random forest	1.6~19.5	13.4~30.6	15.1~33.8	16.8~35.8	18.1~39.3	19.5~42.0	21.5~45.0	22.7~46.7	30.7~83.1
Neural network	0.5~20.4	13.4~32.3	15.9~37.2	17.4~39.6	19.1~44.3	21.1~47.4	23.3~48.8	24.6~52.2	33.7~87.2
autoML	5.3~13.6	10.2~23.0	11.4~27.6	12.3~30.0	13.6~34.4	14.8~36.5	16.0~40.0	17.3~43.8	23.8~87.7
Overall									
Soft voting *	1.4~16.4	12.5~25.2	14.3~28.5	15.9~30.7	17.1~33.7	18.8~36.6	20.4~38.9	21.8~41.2	28.4~80.7
All models #	0.6~18.1	8.4~29.5	9.5~33.4	10.5~36.0	11.4~39.4	12.3~42.4	13.2~45.2	13.9~47.4	19.3~85.9

*** 95% range of individual risk prediction from soft voting (averaging) of all models except the reference model**

 # 95% range of individual risk prediction from all models except the reference model

eTable 4.12.5.4: Comparison of the individual risk predictions of machine learning and statistical models in cohort without censoring (with as reference the risk predictions of the Caret Logistic model)

	Range of individual risk predictions (2.5^{th}~97.5^{th}) with other models compared to those from Caret Logistic model								
	<6%	**6~7%**	**7~8%**	**8~9%**	**9~10%**	**10~11%**	**11~12%**	**12~13%**	**≥ 13%**
Caret									
Logistic	Reference	Reference	Reference	Reference	Reference	Reference	Reference	Reference	Reference
Random forest	0.0~7.1	1.2~12.0	1.6~14.4	2.1~15.1	2.4~17.4	2.7~18.6	3.2~20.9	3.9~21.8	7.7~83.6
Neural network	2.8~6.0	4.7~7.8	5.2~9.1	5.8~10.4	6.4~11.9	7.1~13.1	7.8~14.9	8.5~16.1	11.9~74.2
Statistical logistic model									
Statistical logistic model	1.0~5.8	6.0~6.9	7.0~7.9	8.0~8.9	9.0~9.9	10.0~10.9	11.0~11.9	12.0~12.9	13.7~85.7
Cox model									
QRISK3	0.1~1.8	0.9~2.8	1.2~3.5	1.4~4.1	1.5~4.6	1.9~5.2	2.1~5.9	2.4~6.4	4.2~52.5
Framingham	0.1~5.1	1.1~8.0	1.5~9.2	1.7~10.4	2.0~11.3	2.2~12.6	2.6~13.6	2.7~14.3	5.1~47.8
Local Cox model	1.5~6.2	6.0~7.3	6.8~8.2	7.6~9.0	8.3~9.9	9.0~10.7	9.7~11.5	10.4~12.3	12.4~86.3
Parametric survival model									
Parametric survival model (Weibull)	1.3~6.2	5.8~7.4	6.5~8.2	7.3~9.2	7.9~10.1	8.6~11.0	9.2~12.0	9.8~12.8	11.8~83.0
Parametric survival model (Gaussian)	1.0~3.8	3.5~5.2	4.1~6.1	4.8~7.1	5.4~8.0	6.1~9.1	6.8~10.1	7.3~11.0	9.3~72.2
Parametric survival model (Logistic)	1.0~3.9	3.6~5.0	4.2~5.8	4.8~6.7	5.3~7.5	5.9~8.4	6.5~9.3	7.0~10.1	8.7~76.4
Sklearn									
Logistic	1.0~5.8	6.0~7.0	7.0~8.0	8.0~9.0	9.0~10.0	10.0~11.0	11.0~12.0	12.0~13.0	13.7~85.7
Random forest	0.3~9.4	2.4~15.5	3.0~18.5	3.7~19.1	4.3~21.8	4.9~22.5	5.6~24.9	6.3~26.3	11.8~80.8
Neural network	0.3~4.8	4.0~7.1	5.0~8.6	6.0~9.8	7.1~11.6	8.2~12.6	9.2~14.4	10.4~15.5	13.4~81.3
Gradient boosting classifier	0.9~5.8	3.6~9.5	4.5~10.7	5.3~12.8	6.1~14.9	7.0~15.5	7.8~18.4	8.7~19.6	12.7~84.1
extra-trees	0.2~10.2	2.1~16.5	2.6~18.6	3.1~20.1	3.8~22.6	4.4~23.6	4.9~27.2	5.5~28.2	11.2~82.9
h2o									
Logistic	1.1~6.1	5.5~7.4	6.4~8.4	7.4~9.5	8.3~10.6	8.2~11.6	10.0~12.6	10.8~13.7	13.9~84.4

	Range of individual risk predictions (2.5^{th}~97.5^{th}) with other models compared to those from Caret Logistic model								
	<6%	**6~7%**	**7~8%**	**8~9%**	**9~10%**	**10~11%**	**11~12%**	**12~13%**	**≥ 13%**
Random forest	1.5~7.1	4.7~11.9	5.5~13.7	6.3~14.1	7.2~16.8	7.7~17.5	8.5~20.1	9.2~21.4	13.1~79.7
Neural network	0.4~6.1	4.6~8.8	5.9~10.6	7.0~11.7	8.2~13.4	9.2~14.5	10.4~16.1	11.3~17.6	14.6~84.0
autoML	5.3~6.8	6.2~8.0	6.5~9.3	6.8~10.1	7.2~11.3	7.5~12.1	7.8~13.7	8.3~14.5	10.3~85.3
Overall									
Soft voting *	1.2~5.3	4.7~7.7	5.4~8.9	6.2~9.8	7.1~10.9	7.8~11.9	8.5~13.5	9.4~14.3	11.9~76.2
All models #	0.2~6.3	1.6~9.2	2.0~10.9	2.3~12.2	2.7~14.1	3.1~15.3	3.4~17.0	3.8~18.2	8.4~82.0

*** 95% range of individual risk prediction from soft voting (averaging) of all models except the reference model**

 # 95% range of individual risk prediction from all models except the reference model

eTable 4.12.6: SPEARMAN correlations of Machine learning models and statistical models in risk groups with logistic (Caret) predicted 7%~8%

	SPEARMAN Correlation																		
	Caret			Statistical model							Sklearn					H2o			
	Logit *	RF	NN	Logit	QRISK3	Framingham	Cox	Parametric Weibull	Parametric Gaussian	Parametric Logistic	Logit	RF	NN	GBC	extra-trees	Logit	RF	NN	auto ML
Caret																			
Logistic	1.00	0.11	0.37	1.00	0.15	0.05	0.21	0.19	0.17	0.16	0.98	0.11	0.18	0.20	0.12	0.60	0.15	0.26	0.85
Random forest	0.11	1.00	0.16	0.11	0.44	-0.06	0.35	0.38	0.38	0.38	0.13	0.99	0.15	0.48	0.65	0.16	0.90	0.02	0.26
Neural network	0.37	0.16	1.00	0.38	0.30	0.09	0.32	0.15	0.17	0.13	0.42	0.15	0.67	0.38	0.38	0.14	0.21	0.71	0.36
Statistical logistic model																			
Statistical logistic model	1.00	0.11	0.38	1.00	0.15	0.06	0.21	0.19	0.17	0.16	0.98	0.11	0.18	0.20	0.12	0.60	0.14	0.26	0.85
Cox model																			
QRISK3	0.15	0.44	0.30	0.15	1.00	0.32	0.60	0.50	0.48	0.49	0.17	0.43	0.14	0.37	0.35	0.20	0.43	0.11	0.23
Framingham	0.05	-0.06	0.09	0.06	0.32	1.00	-0.04	-0.32	-0.34	-0.36	0.01	-0.05	-0.30	0.21	-0.02	0.03	-0.23	-0.06	0.06
Local Cox model	0.21	0.35	0.32	0.21	0.60	-0.04	1.00	0.85	0.79	0.80	0.28	0.33	0.32	0.13	0.25	0.30	0.33	0.15	0.25
Parametric survival model																			
Parametric survival model (Weibull)	0.19	0.38	0.15	0.19	0.50	-0.32	0.85	1.00	0.97	0.99	0.26	0.36	0.23	0.04	0.22	0.32	0.44	0.03	0.26
Parametric survival model (Gaussian)	0.17	0.38	0.17	0.17	0.48	-0.34	0.79	0.97	1.00	0.99	0.23	0.36	0.27	0.03	0.24	0.24	0.44	0.10	0.23
Parametric survival model (Logistic)	0.16	0.38	0.13	0.16	0.49	-0.36	0.80	0.99	0.99	1.00	0.23	0.36	0.25	0.03	0.22	0.28	0.45	0.06	0.24
Sklearn																			
Logistic	0.98	0.13	0.42	0.98	0.17	0.01	0.28	0.26	0.23	0.23	1.00	0.13	0.27	0.21	0.14	0.67	0.18	0.29	0.88
Random forest	0.11	0.99	0.15	0.11	0.43	-0.05	0.33	0.36	0.36	0.36	0.13	1.00	0.12	0.50	0.68	0.17	0.89	-0.00	0.27
Neural network	0.18	0.15	0.67	0.18	0.14	-0.30	0.32	0.23	0.27	0.25	0.27	0.12	1.00	0.19	0.24	0.17	0.21	0.68	0.22

	SPEARMAN Correlation																		
	Caret			Statistical model							Sklearn					H2o			
	Logit *	RF	NN	Logit	QRISK3	Framingham	Cox	Parametric Weibull	Parametric Gaussian	Parametric Logistic	Logit	RF	NN	GBC	extra-trees	Logit	RF	NN	auto ML
Gradient boosting classifier	0.20	0.48	0.38	0.20	0.37	0.21	0.13	0.04	0.03	0.03	0.21	0.50	0.19	1.00	0.45	0.14	0.52	0.26	0.36
extra-trees	0.12	0.65	0.38	0.12	0.35	-0.02	0.25	0.22	0.24	0.22	0.14	0.68	0.24	0.45	1.00	0.05	0.65	0.27	0.21
h2o																			
Logistic	0.60	0.16	0.14	0.60	0.20	0.03	0.30	0.32	0.24	0.28	0.67	0.17	0.17	0.14	0.05	1.00	0.23	0.12	0.86
Random forest	0.15	0.90	0.21	0.14	0.43	-0.23	0.33	0.44	0.44	0.45	0.18	0.89	0.21	0.52	0.65	0.23	1.00	0.12	0.34
Neural network	0.26	0.02	0.71	0.26	0.11	-0.06	0.15	0.03	0.10	0.06	0.29	-0.00	0.68	0.26	0.27	0.12	0.12	1.00	0.28
autoML	0.85	0.26	0.36	0.85	0.23	0.06	0.25	0.26	0.23	0.24	0.88	0.27	0.22	0.36	0.21	0.86	0.34	0.28	1.00

*** Abbreviation: Logit - Logistic model, RF - Random forest, NN - Neural network, Cox - Cox proportional hazard model, GBC - Gradient boosting classifier**

eTable 4.12.7: Reclassification of individual risk predictions of machine learning and statistical models with 10% as threshold

	Reclassification in overall testing cohort	
	Reclassified*	**Not reclassified**
Overall cohort		
QRISK3 as reference model		
Below or equal to the threshold (≤10%)	73871 (10.0%)	661603 (90.0%)
Above the threshold (>10%)	113260 (62.9%)	66745 (37.1%)
Logistic model (Caret) as reference model		
Below or equal to the threshold (≤10%)	170983 (20.5%)	661603 (79.5%)
Above the threshold (>10%)	16148 (19.5%)	66745 (80.5%)
Cohort without censoring		
QRISK3 as reference model		
Below or equal to the threshold (≤10%)	34691 (49.1%)	35891 (50.9%)
Above the threshold (>10%)	2269 (5.5%)	39017 (94.5%)
Logistic model (Caret) as reference model		
Below or equal to the threshold (≤10%)	6872 (16.1%)	35891 (83.9%)
Above the threshold (>10%)	30088 (43.5%)	39017 (56.5%)

*** For patients who are below or equal to the threshold, they are re-classified if they have prediction above the threshold in any model.**
For patients who are above the threshold, they are re-classified if they have prediction below or equalt to the threshold in any model.

eTable 4.12.8: Reclassification of individual risk predictions of Caret neural network models with different hyperparameters

	Reclassification in overall testing cohort	
	Reclassified*	**Not reclassified**
Overall cohort		
Models with the most frequent selected hyperparameters as reference model		
Below or equal to the threshold (<=7.5%)	12016 (1.5%)	773472 (98.5%)
Above the threshold (>7.5%)	14987 (11.5%)	115004 (88.5%)

*** For patients who are below or equal to the threshold, they are re-classified if they have prediction above the threshold in any model.**
For patients who are above the threshold, they are re-classified if they have prediction below or equalt to the threshold in any model.

eTable 4.12.9: Inconsistency of individual risk prediction between machine learning models derived from overall cohort and cohort without censoring

	Range of individual risk predictions (2.5^{th}~97.5^{th}) for the same group of patients *								
	<6%	**6~7%**	**7~8%**	**8~9%**	**9~10%**	**10~11%**	**11~12%**	**12~13%**	**≥ 13%**
Caret									
Logistic	0.8~31.9	21.6~61.2	24.2~65.5	27.2~69.4	29.5~73.7	31.8~75.4	34.1~77.4	36.2~79.6	44.2~90.2
Random forest	0.2~26.8	10.5~52.5	11.6~57.9	13.5~61.7	15.3~65.3	16.8~69.5	18.8~73.7	19.4~76.8	29.1~87.9
Neural network	0.8~29.7	19.7~54.4	22.7~57.5	25.3~60.7	28.0~64.3	31.0~68.0	33.3~70.2	35.7~72.8	42.2~76.2
Cox model									
QRISK3	0.1~5.4	6.0~7.0	7.0~8.0	8.0~9.0	9.0~10.0	10.0~11.0	11.0~12.0	12.0~13.0	13.3~54.0
Framingham	0.0~5.7	6.0~7.0	7.0~8.0	8.0~9.0	9.0~10.0	10.0~11.0	11.0~12.0	12.0~13.0	13.2~48.2
Local Cox model	2.5~28.0	24.1~42.9	27.1~47.7	29.9~52.2	32.7~56.2	35.4~59.7	37.8~62.9	40.4~66.3	48.4~99.2
Sklearn									
Logistic	0.8~31.8	21.6~61.1	24.3~65.0	27.3~69.9	29.6~73.2	32.0~75.3	34.1~77.5	36.3~79.1	44.3~90.2
Random forest	0.4~32.1	12.7~56.8	13.7~60.4	15.1~64.8	16.1~68.8	18.4~72.1	19.0~74.7	19.5~77.5	27.1~87.2
Neural network	1.0~35.5	20.9~59.8	23.7~63.5	26.8~67.4	30.3~70.5	33.9~73.5	37.4~75.0	40.8~76.3	49.1~83.0
Gradient boosting classifier	0.8~33.2	17.7~62.1	19.8~65.9	22.0~71.5	23.4~76.9	25.2~81.0	27.9~83.3	29.7~85.2	38.1~89.2
extra-trees	0.1~33.0	6.8~64.9	7.6~71.2	6.0~74.2	6.5~80.1	5.1~85.1	10.1~86.6	11.2~88.5	15.0~97.1
h2o									
Logistic	0.8~30.9	21.9~55.3	24.6~59.8	27.4~63.4	29.8~66.7	32.0~70.9	34.2~72.4	36.4~74.2	44.4~88.4
Random forest	1.4~30.5	18.2~51.0	20.1~54.8	22.3~57.8	24.4~61.1	26.4~65.4	28.6~67.9	31.1~71.4	39.2~84.9
Neural network	0.1~31.8	19.6~69.9	22.5~75.2	25.4~80.0	28.1~82.5	31.1~84.7	33.2~86.8	36.4~88.0	48.4~93.5
autoML	5.0~29.2	17.5~63.5	19.9~69.1	22.6~76.3	24.9~80.7	27.2~81.8	29.9~83.9	31.3~85.7	40.4~91.3

*** 95% range of individual risk prediction of the same risk-group patients predicted by model derived from cohort without censoring comparing to the same model dervived from overall cohort (risk-group displayed in the second line of the table title)**

eTable 4.12.10: Performance indicators of machine learning and Cox models developed in South and validated in North

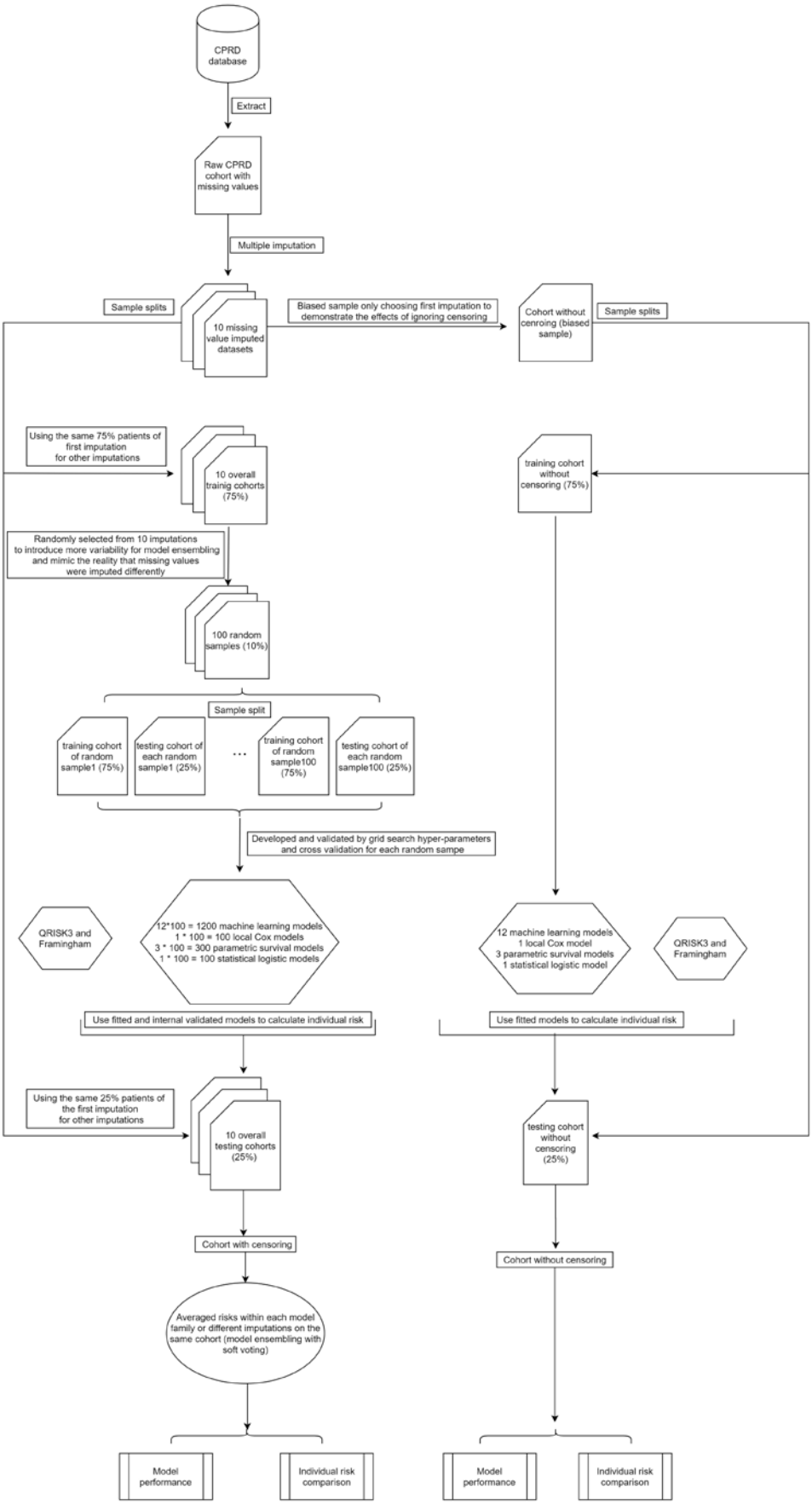
	Model performance*				Average absolute change of model performance
	C-statistic	**Brier score**	**Recall (Sensitivity)**	**Precision (PPV)**	**C-statistic**
North#					
Logistic (Caret)	0.871	0.032	0.575	0.179	Reference
Neural network (Caret)	0.871	0.032	0.631	0.167	-0.02%
Local Cox model	0.869	0.036	0.798	0.124	-0.21%
South$					
Logistic (Caret)	0.877	0.028	0.607	0.164	Reference
Neural network (Caret)	0.877	0.028	0.659	0.151	+0.01%
Local Cox model	0.875	0.031	0.803	0.112	-0.21%

*** Model performance was calculated in binaray framework. Threshold 7.5% was used to calculate precision and recall for all models.**
 # Testing cohort only including practices from North of UK which was different from development cohort (i.e. practices from south)
$ Testing cohort only including practices from South of UK which was similar to development cohort

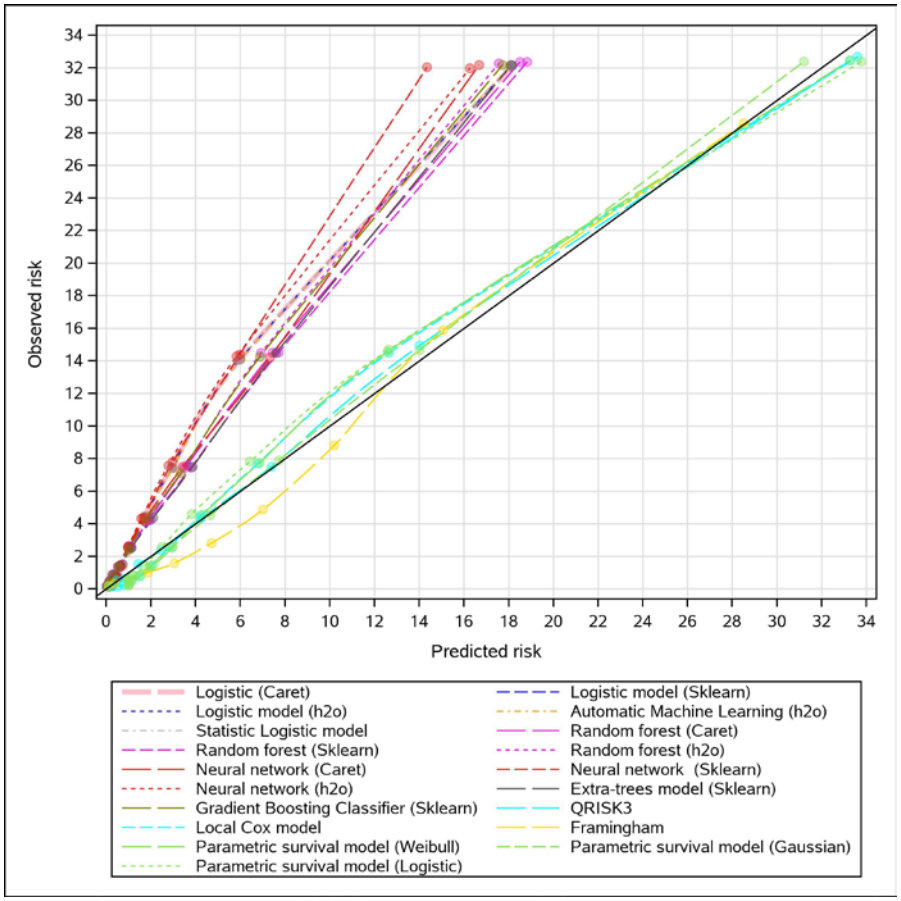
eTable 4.12.11: Performance indicators of machine learning and Cox models with lower number of predictors

	Model performance*				Average absolute change of model performance
	C-statistic	**Brier score**	**Recall (Sensitivity)**	**Precision (PPV)**	**C-statistic**
Using the same 1/3 random predictors #					
Logistic (Caret)	0.870	0.028	0.591	0.157	Reference
Random forest (Caret)	0.705	0.036	0.302	0.125	-18.95%
Neural network (Caret)	0.870	0.028	0.655	0.142	+0.01%
Local Cox model	0.869	0.032	0.801	0.108	-0.08%
Using the same 1/2 random predictors					
Logistic (Caret)	0.875	0.028	0.602	0.160	Reference
Random forest (Caret)	0.832	0.029	0.594	0.132	-4.96%
Neural network (Caret)	0.876	0.028	0.669	0.145	+0.03%
Local Cox model	0.875	0.031	0.809	0.110	-0.07%
Using the same 2/3 random predictors					
Logistic (Caret)	0.878	0.028	0.610	0.162	Reference
Random forest (Caret)	0.858	0.028	0.621	0.143	-2.27%
Neural network (Caret)	0.878	0.028	0.665	0.149	+0.02%
Local Cox model	0.876	0.031	0.810	0.111	-0.22%

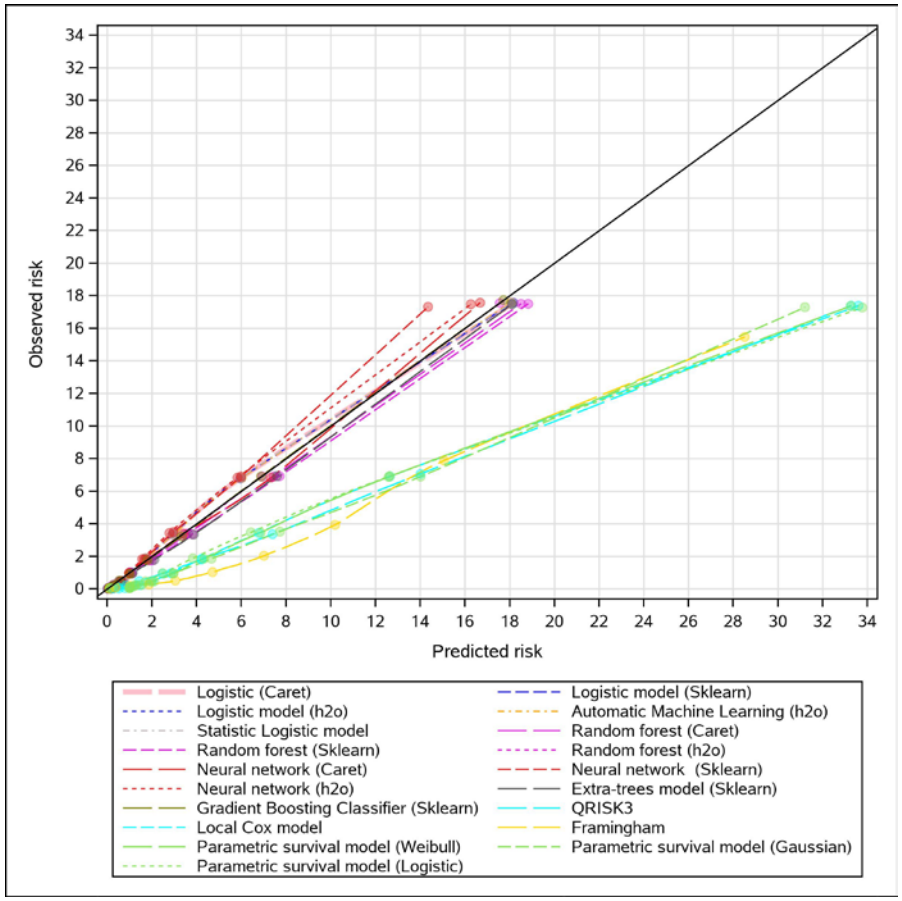
*** Model performance was calculated in binaray framework. Threshold 7.5% was used to calculate precision and recall for all models.**
 # Age and gender were always included as predictors in all scenarios.



eFigure 4.12.1. Workflow of sample splitting and model fitting process

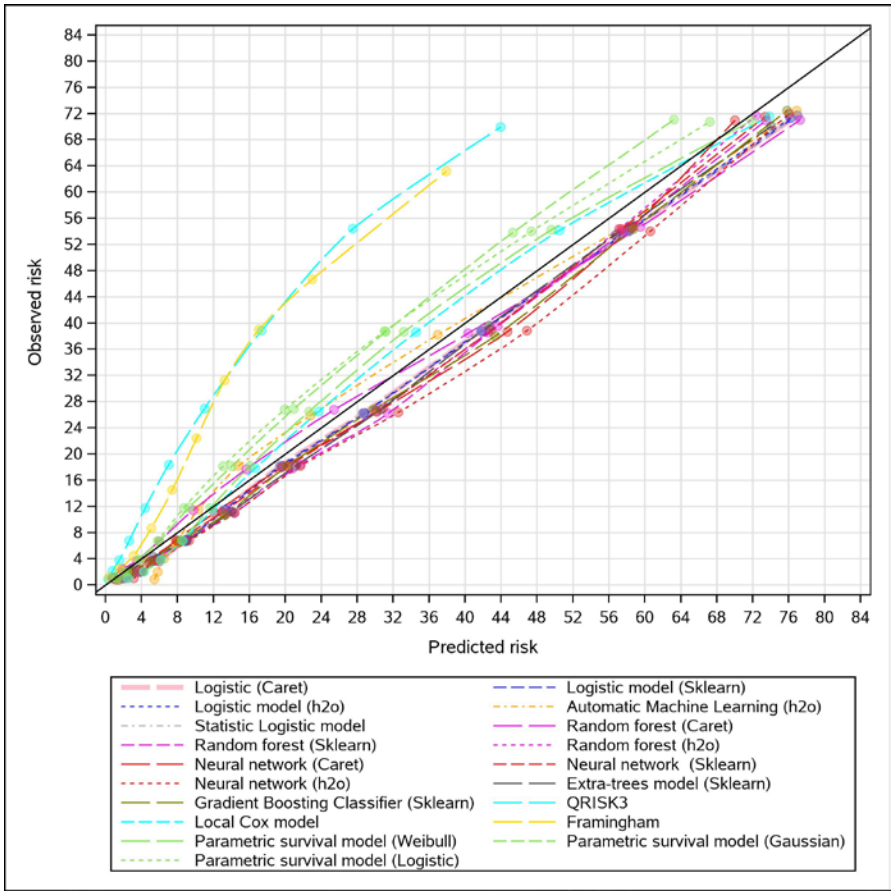


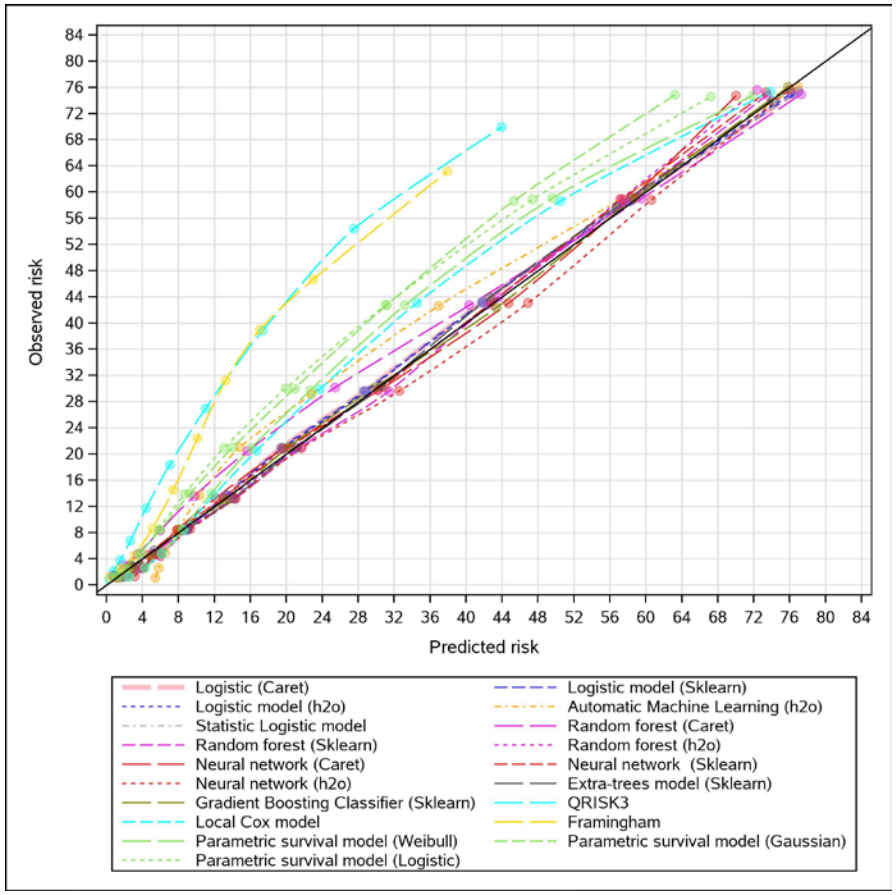
eFigure 4.12.2.1a



eFigure 4.12.2.1b

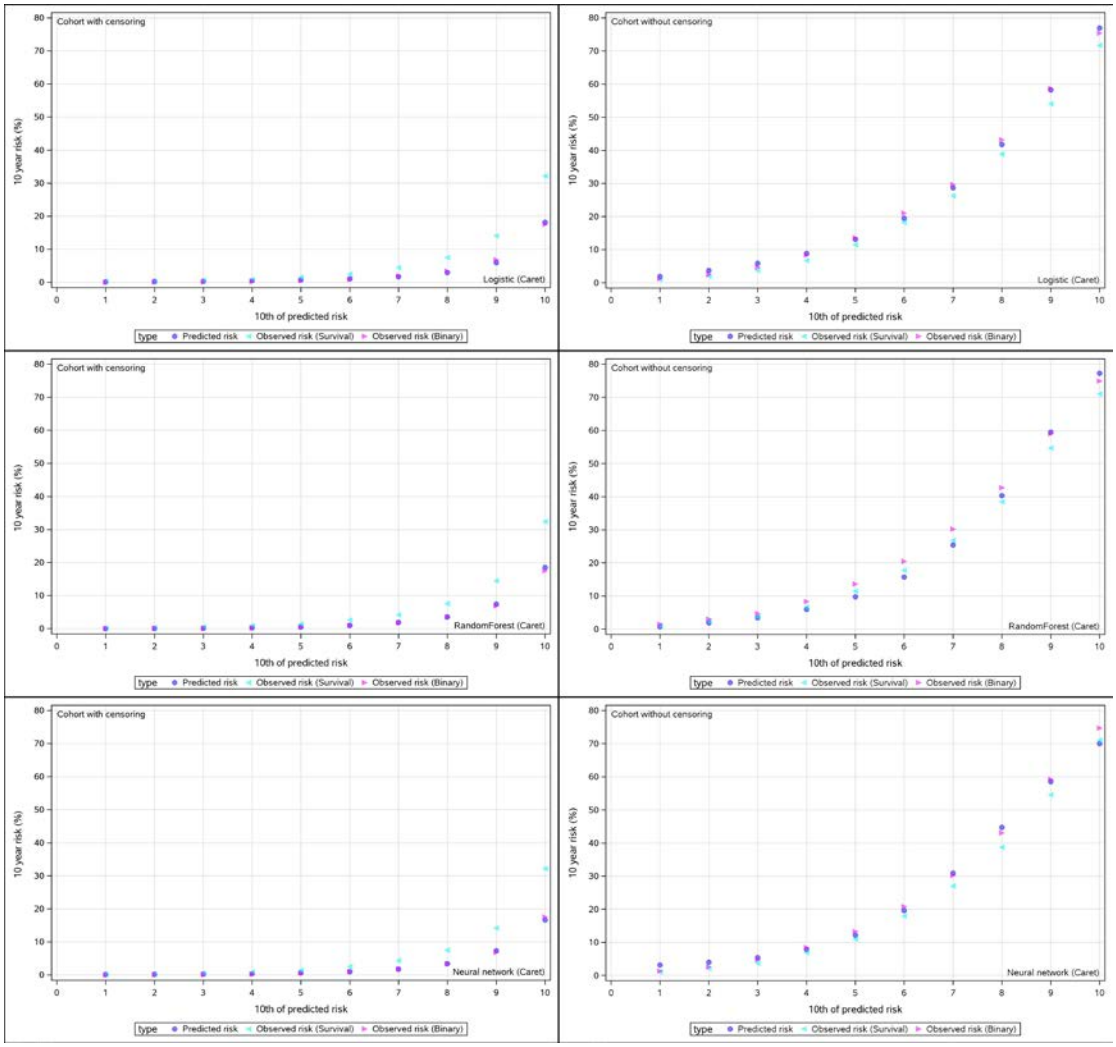
eFigure 4.12.2.1. Calibration slope of machine learning models and statistical models in overall cohort



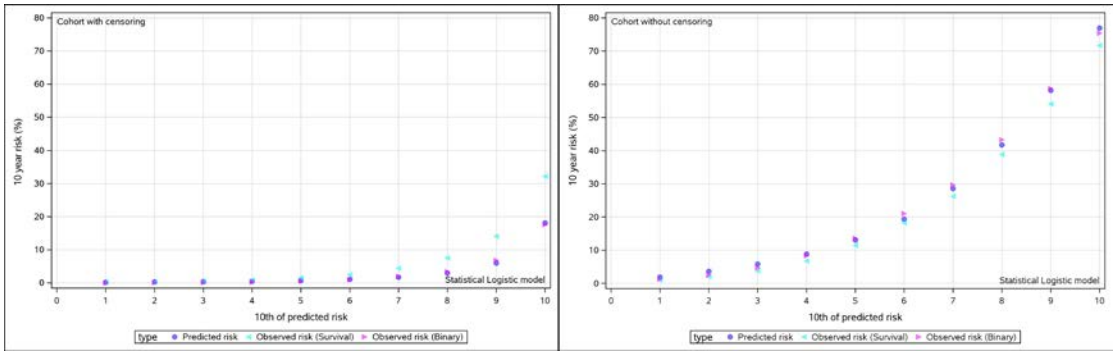


eFigure 4.12.2.2a **eFigure 4.12.2.2b**

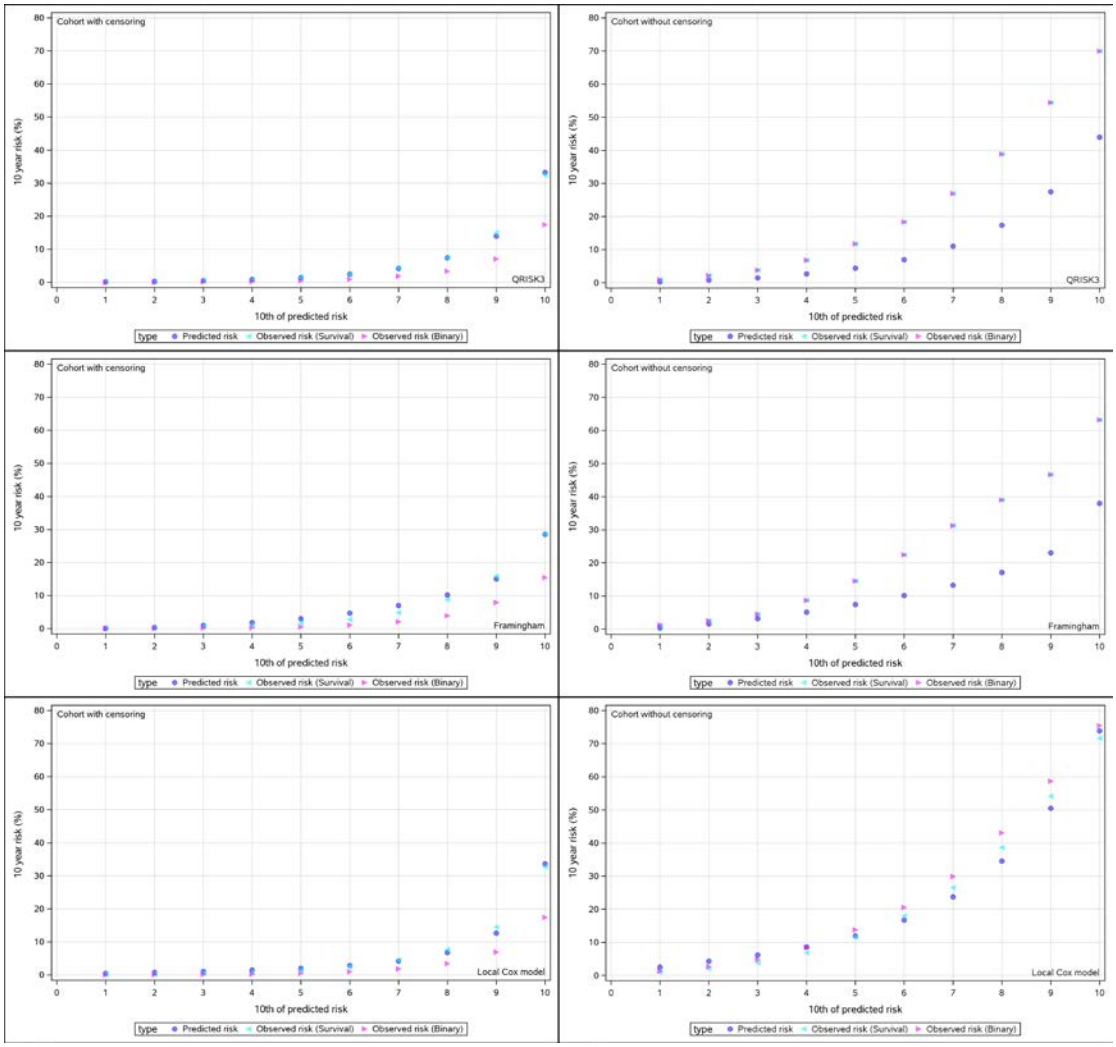
eFigure 4.12.2.2. Calibration slope of machine learning models and statistical models in cohort without censoring



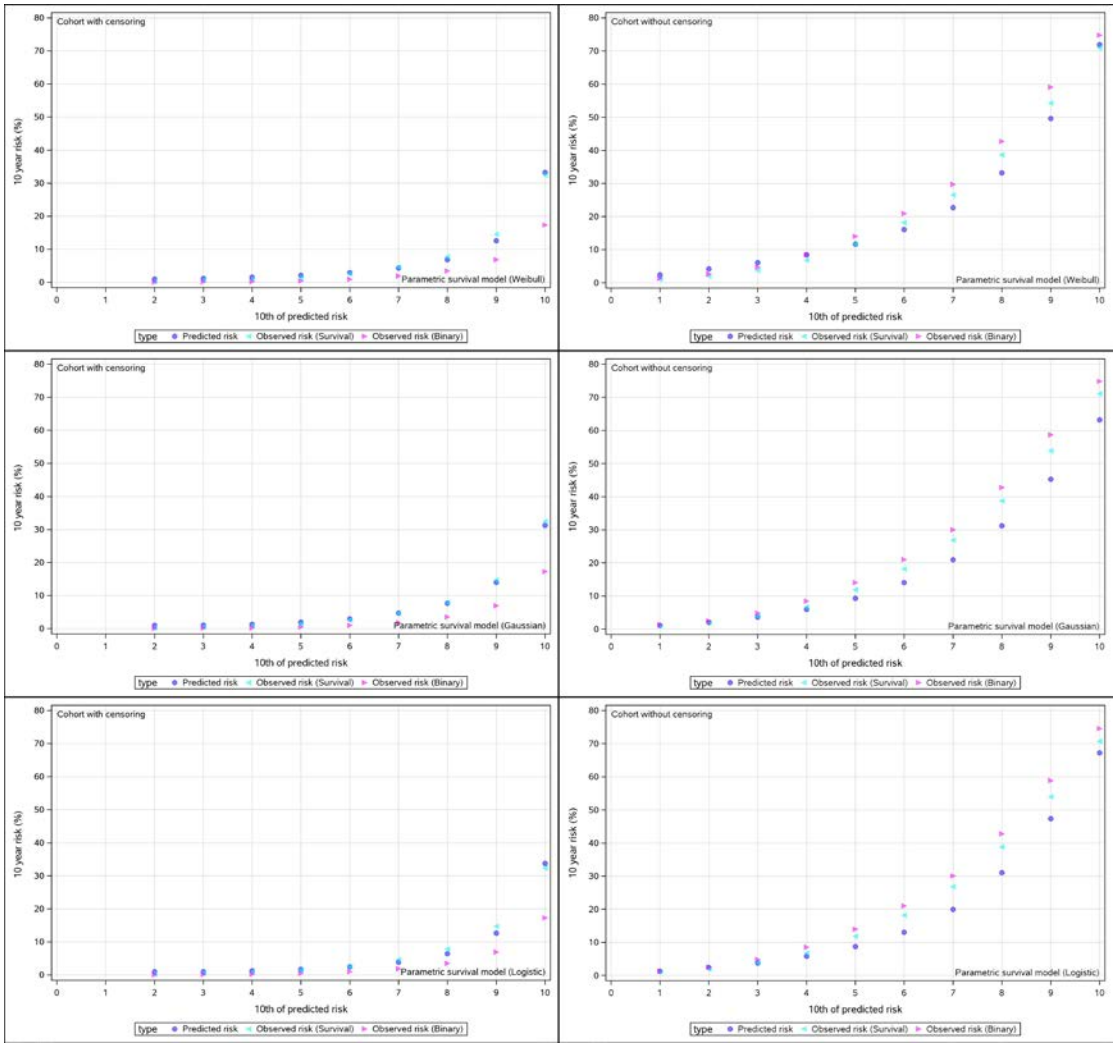
eFigure 4.12.3.1. Calibration plots in machine learning models of Caret in overall cohort and cohort without censoring



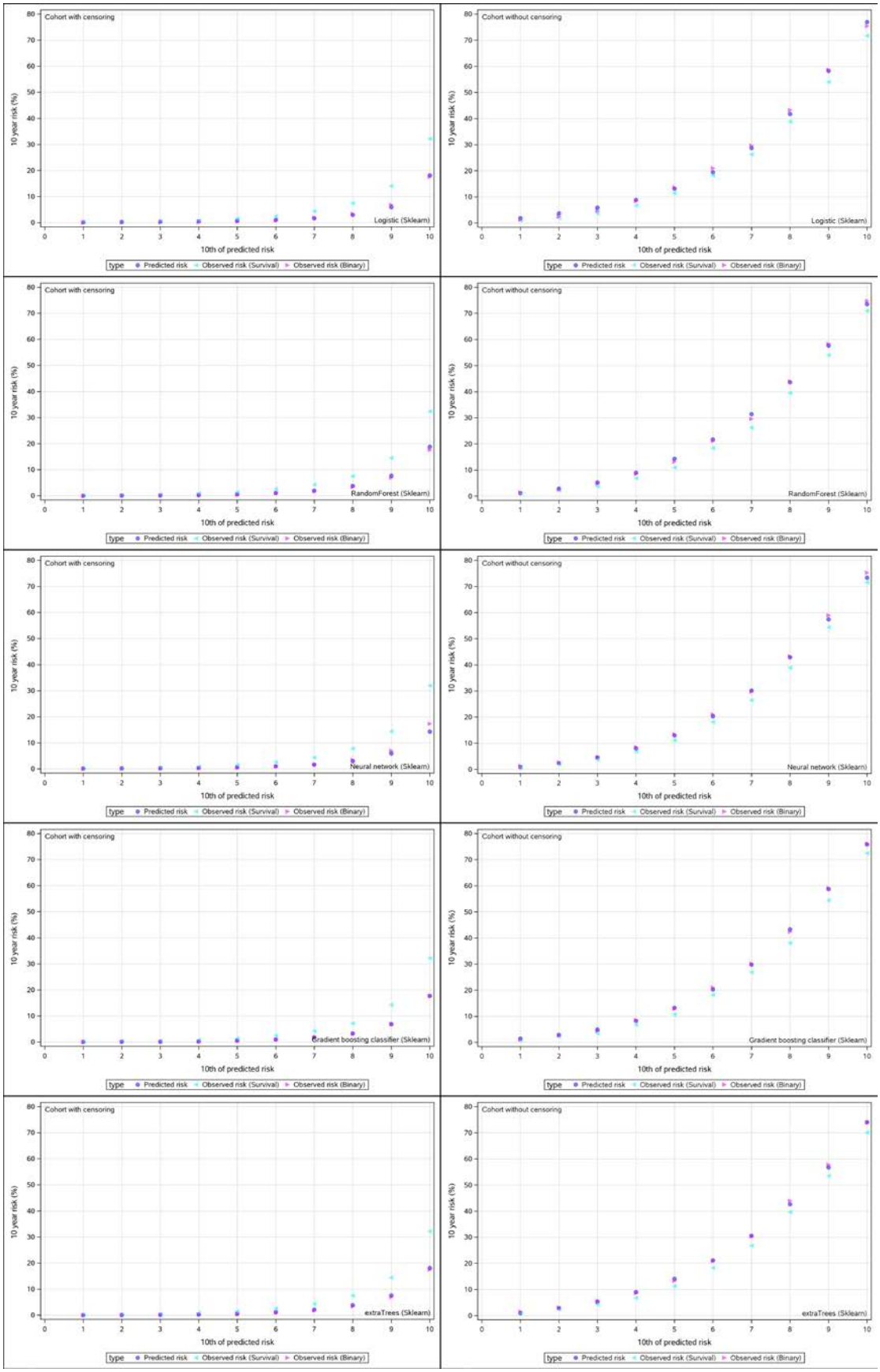
eFigure 4.12.3.2. Calibration plots in statistical logistic models in overall cohort and cohort without censoring



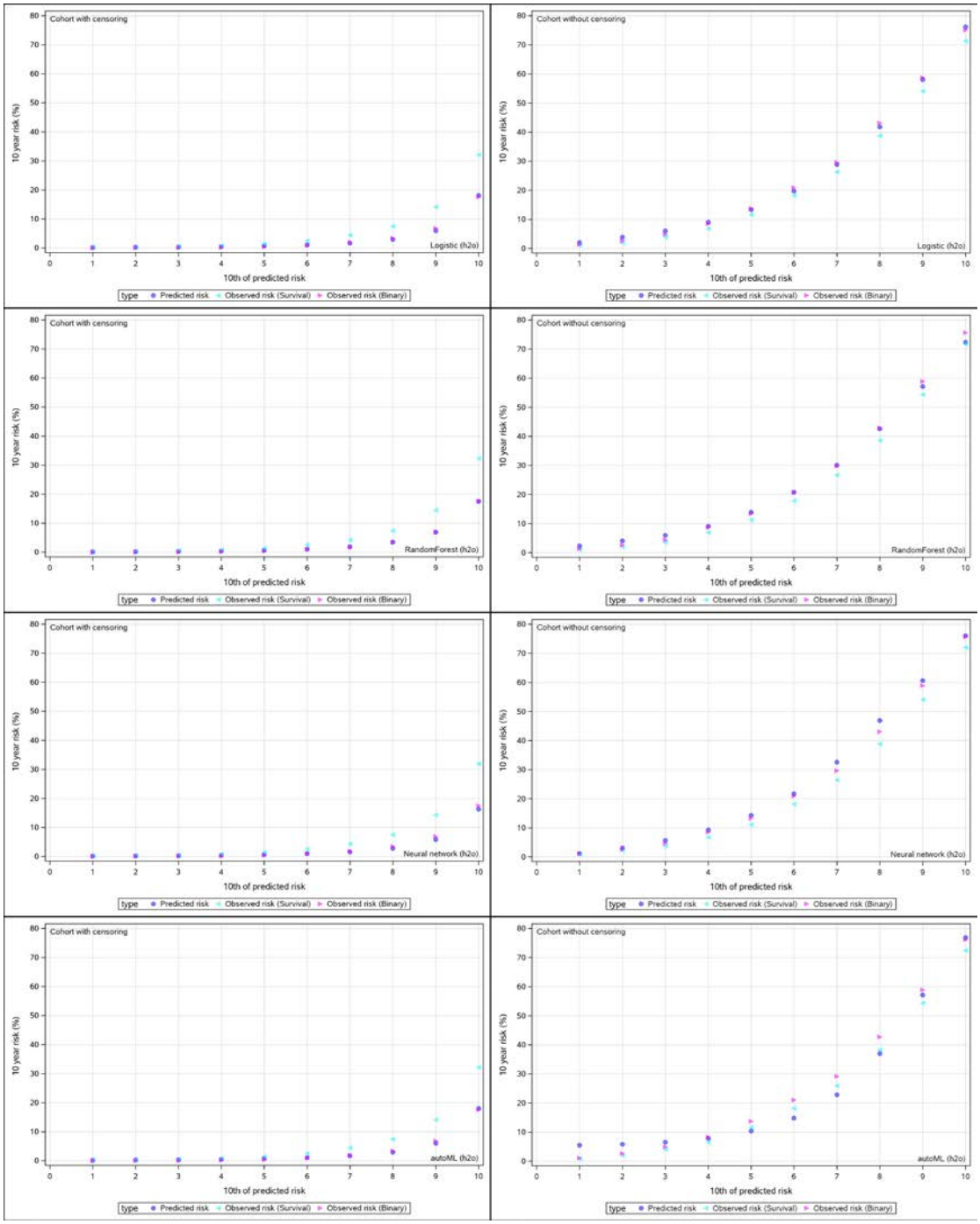
eFigure 4.12.3.3. Calibration plots in Cox proportional hazard models in overall cohort and cohort without censoring



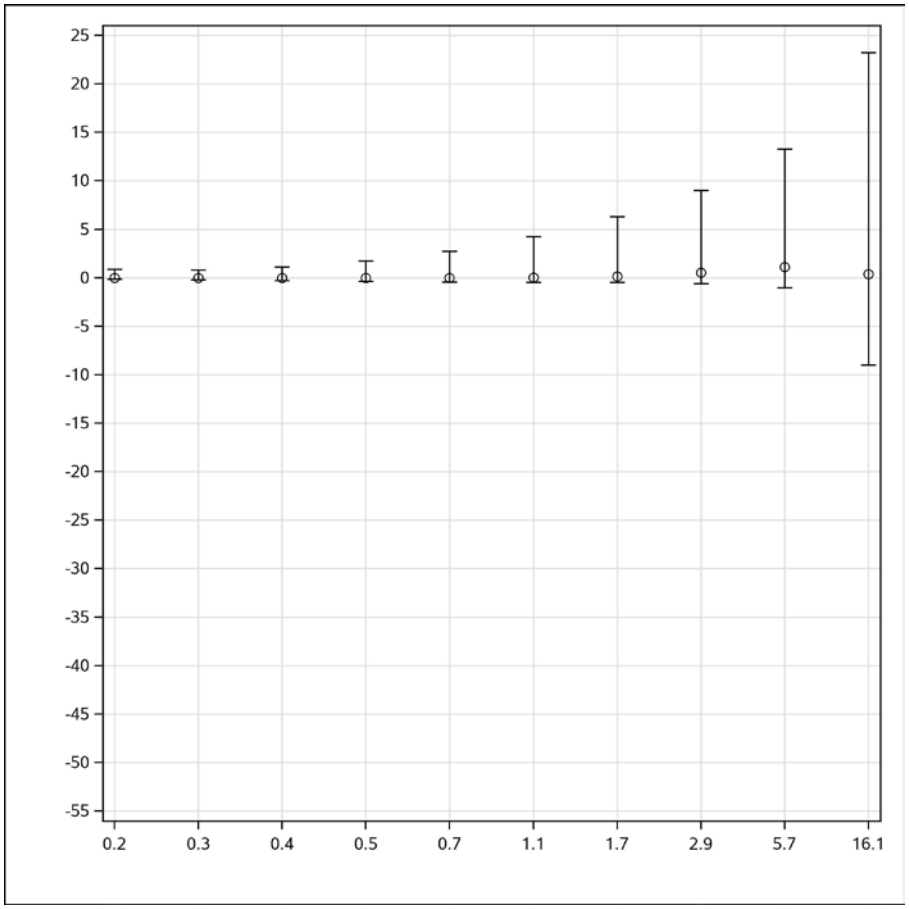
eFigure 4.12.3.4. Calibration plots in parametric survival models in overall cohort and cohort without censoring

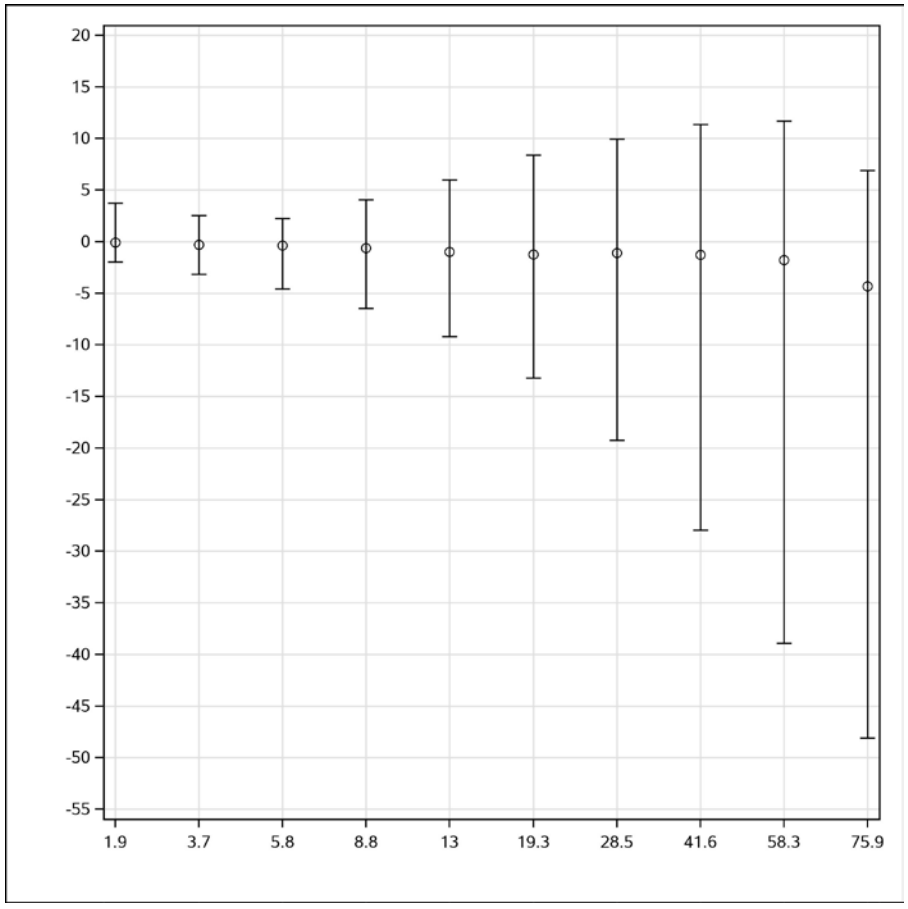


eFigure 4.12.3.5. Calibration plots in machine learning models of Sklearn in overall cohort and cohort without censoring



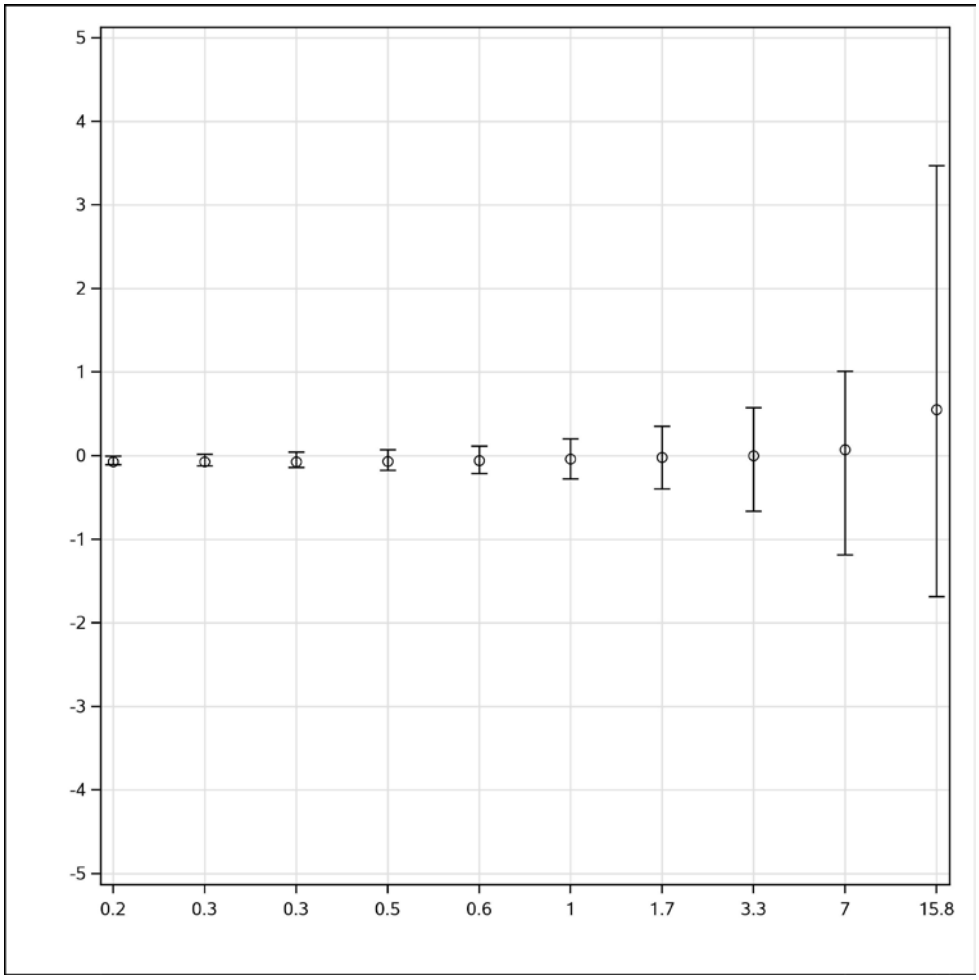
eFigure 4.12.3.6. Calibration plots in machine learning models of h2o in overall cohort and cohort without censoring



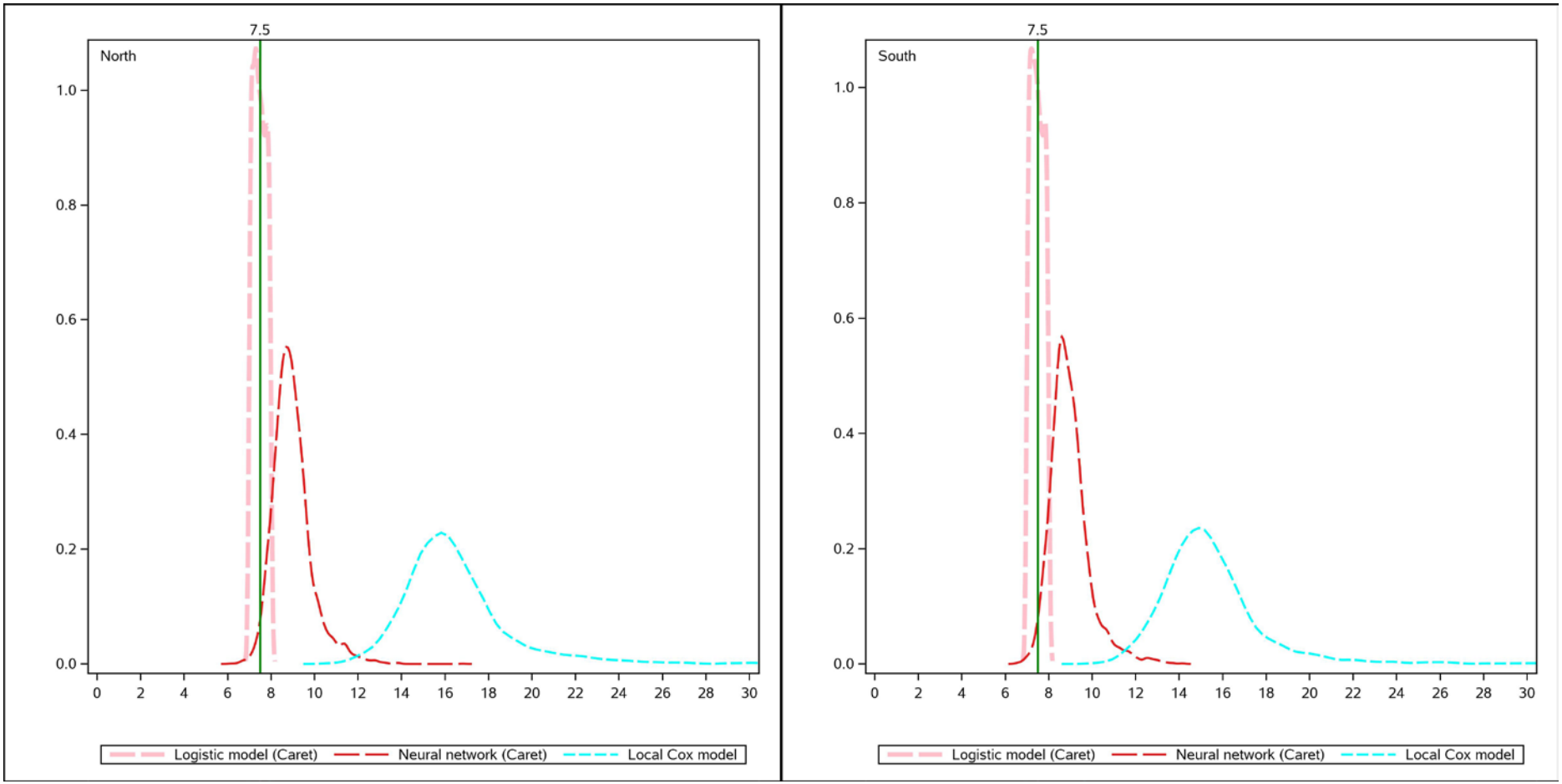


eFigure 4.12.4a **eFigure 4.12.4b**

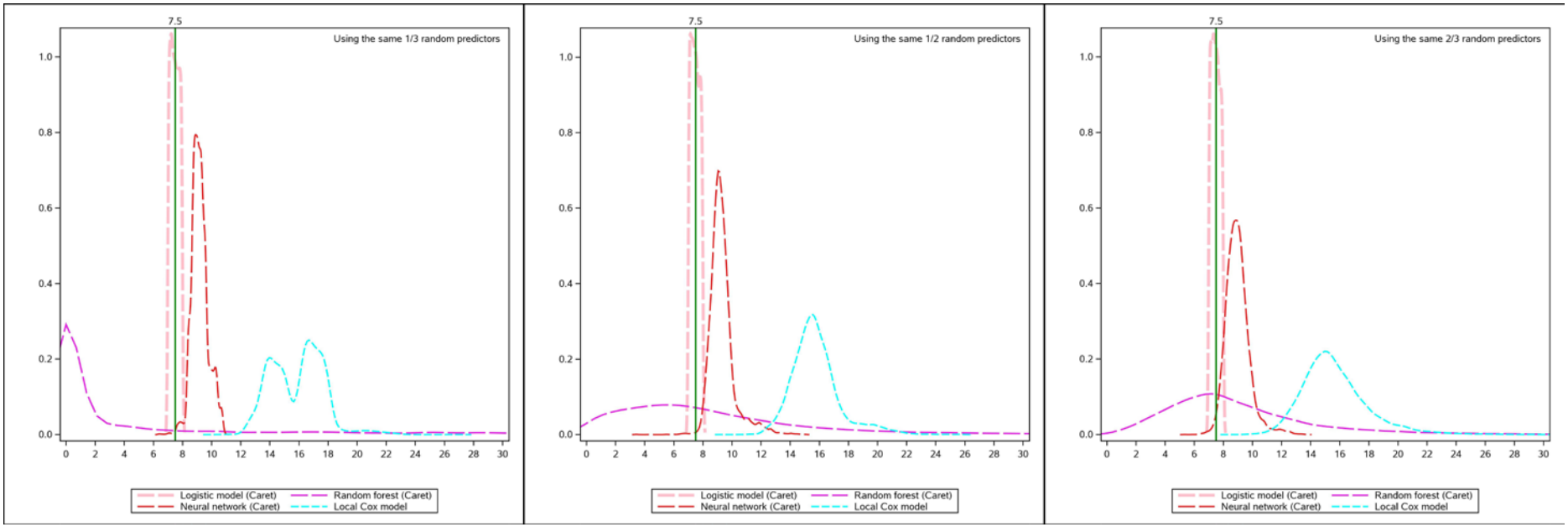
eFigure 4.12.4. 95% range of individual risk predictions with machine learning and statistical models stratified by deciles of predicted risks with Caret logistic model



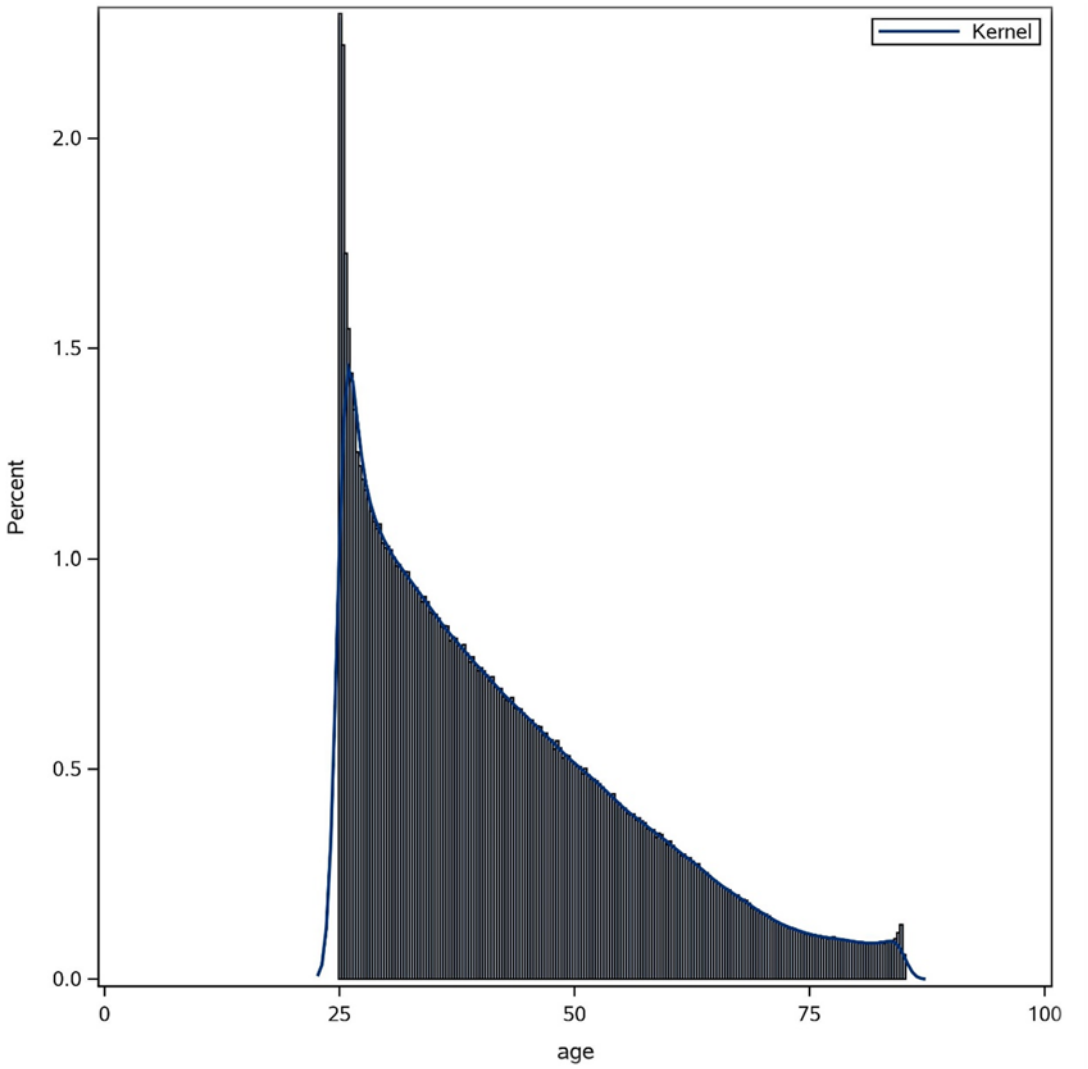
eFigure 4.12.5. 95% range of individual risk predictions with Caret neural network models with different grid searched best hyperparameters stratified by deciles of predicted risks with models with the most frequent selected hyperparameters



eFigure 4.12.6. Distribution of individual risk predictions with machine learning and statistical models developed in practices from South and tested in practices from North



eFigure 4.12.7. Distribution of individual risk predictions with machine learning and statistical models developed with predictors of age and sex plus 1/3, 1/2, 2/3 of all predictors



eFigure 4.12.8 Distribution of age among removed patients due to censoring (death patients excluded)

Blank page

Chapter 5 R package "QRISK3": an unofficial research purposed implementation of ClinRisk's QRISK3 algorithm into R

Yan Li[1], Matthew Sperrin[1], Tjeerd Pieter van Staa[1,2,3]

[1]Health e-Research Centre, School of Health Sciences, Faculty of Biology, Medicine and Health, the University of Manchester, Manchester Academic Health Sciences Centre (MAHSC), Oxford Road, Manchester, M13 9PL, UK

[2]Utrecht Institute for Pharmaceutical Sciences, Utrecht University, Utrecht, Netherlands

[3]Alan Turing Institute, Headquartered at the British Library, London, UK

Corresponding author: Tjeerd van Staa, tjeerd.vanstaa@manchester.ac.uk

Journal title: F1000Research
Doi: https://doi.org/10.12688/f1000research.21679.1

Word count: 2423
Abstract:109
Number of tables: 2
Number of figures: 0

5.1 Abstract

Cardiovascular disease has been the leading cause of death for decades. Risk prediction models are used to identify high risk patients; the most common model used in the UK is ClinRisk's QRISK3. In this paper we describe the implementation of the QRISK3 algorithm into an R package. The package was successfully validated by the open sourced QRISK3 algorithm and QRISK3 SAS program. We provide detailed examples of the use of the package, including assigning QRISK3 scores for a large cohort of patients. This R package could help the research community to better understand risk prediction scores and improve future risk prediction models. The package is available from CRAN: https://cran.r-project.org/web/packages/QRISK3/index.html.

5.2 Introduction

Cardiovascular disease (CVD) was responsible for 17.9 million deaths in 2016, which represents 31% of overall global deaths, and over 75% of these deaths happened in low/middle-income countries [1]. People who are at high risk of CVD need to be identified and treated early [1]. Risk prediction models that use risk factors to calculate the probability of patients developing diseases are often used to identify high risk patients [2]. QRISK3 is the most popular risk prediction model for CVD developed in the UK. It calculates risk of patients developing CVD in the next 10 years and has been incorporated into the electronic health records (EHRs) system in the UK in order to detect high risk CVD patients and help clinicians make treatment decisions [3 4]. NICE guidelines recommend clinicians to consider prescribing statins to patients with a risk over 10% identified from QRISK3 [5]. QRISK3 was developed from historical patients' EHR data using Cox proportional hazard model [6] and has been well validated at population level corresponding to discrimination and calibration [3 4 7].

The implementation of QRISK3 into R would not only benefit researchers to improve future risk prediction but also enable them to use QRISK3 scores to identify patients at certain risk levels, e.g. for clinical trial recruitment. There is also scope to improve these risk predictions; it has been found that QRISK3 has uncertainty on individual risk prediction [7 8] due to unmeasured heterogeneity between practices, which was not captured. A follow-up study suggests that QRISK3 may need to include additional causal risk factors as this uncertainty on individual risk prediction was not related to data quality and variation of association between disease and outcome [9]. The current QRISK3 can only be accessed through an online web calculator or specialised commercial software [10] and its original algorithm was written by C, which is a low level programming language appealing to software engineering rather than data science [11]. R is the most popular statistical programming language in the data science field due to its great advantage as free and open-source, with fast computing and a well-supported community [12]. This paper explains the incorporation of the QRISK3 algorithm into R for ease of research concerning QRISK3 and how the package was developed and validated. The package aims to help researchers to improve risk prediction models and better detect high risk CVD patients.

5.3 Methods

5.3.1 Extraction of the QRISK3 algorithm

The original QRISK3 algorithm was written in C by ClinRisk under a GNU Lesser General Public License [11]. Their previously published QRISK3 paper was used to understand the original algorithm and the associations between variables used in the original algorithm and risk factors of QRISK3 [3].

5.3.2 Development and validation of the QRISK3 R package

The QRISK3 algorithm was written in both R (3.4.2) and SAS (9.4) [13] independently, in order to mimic double programming, with a plan to use the SAS implementation to validate the R package. An additional C program, which could directly call the original QRISK3 algorithm to calculate risk, was written for validation. Two validation datasets (QRISK3_2017_test and QRISK3_2019_test) were then created and included in the R package. Dataset QRISK3_2017_test was created by manually recording the calculated QRISK3 risk score from the original QRISK3 algorithm for a group of simulated patients. The simulated patient groups were generated by changing each risk factor sequentially covering the changes of all QRISK3 risk factors. For example, patient 1 in QRISK3_2017_test does not have any positive CVD risk factors, patient 2 is similar to patient 1 expect he has atrial fibrillation, patient ID 3 is similar to patient 2 except he is on atypical antipsychotic medication rather than atrial fibrillation and so on until all the change of CVD predictors are covered. Therefore, each patient is similar to the previous patient except change of one CVD predictor. QRISK3_2019_test was the version recorded using the original QRISK3 algorithm with different value changes for each risk factor. Risk scores of the same simulated patient groups (QRISK3_2017_test and QRISK3_2019_test) was compared among different versions of QRISK3, including QRISK3 R package, QRISK3 SAS macro and QRISK3 C function for validation. The R package was created using R CMD tool [14] with several useful online tutorials [15] [16] [17] [18].

5.4 Results

5.4.1 Implementation

The QRISK3 package can be directly installed from CRAN [19] using "install(QRISK3)" or GitHub respiratory [20] with "install_github("YanLiUK/QRISK3") ". The package contains one function (QRISK3_2017) to calculate the risk of patients developing CVD in the next 10 years using the QRISK3 algorithm [11] and the two datasets for testing.

Variables used by the QRISK3 package were summarised and compared to the original algorithm in Table 5.1. All variables have the same definition as the QRISK3 paper [3], most of variables were coded into numeric variables similar to the original algorithm. The coding of ethnicity and smoking was different from the original algorithm (written in C), as the C index starts from 0 but R's index starts from 1.

Table 5.1: Description of QRISK3 variables.

Parameters in QRISK3 R package	Meaning of variables	Variables in original algorithm
age	Specify the age of the patient in year (e.g. 64 years-old)	age
atrial_fibrillation	Atrial fibrillation? (0: No, 1: Yes)	b_AF
atypical_antipsy	On atypical antipsychotic medication? (0: No, 1: Yes)	b_atypicalantipsy
regular_steroid_tablets	On regular steroid tablets? (0: No, 1: Yes)	b_corticosteroids
erectile_disfunction	A diagnosis of or treatment for erectile disfunction? (0: No, 1: Yes)	b_impotence2 (only for men)
migraine	Do patients have migraines? (0: No, 1: Yes)	b_migraine
rheumatoid_arthritis	Rheumatoid arthritis? (0: No, 1: Yes)	b_ra
chronic_kidney_disease	Chronic kidney disease (stage 3, 4 or 5)? (0: No, 1: Yes)	b_renal
severe_mental_illness	Severe mental illness? (0: No, 1: Yes)	b_semi
systemic_lupus_erythematosis	Systemic lupus erythematosis (SLE)? (0: No, 1: Yes)	b_sle
blood_pressure_treatment	On blood pressure treatment? (0: No, 1: Yes)	b_treatedhyp
diabetes1	Diabetes status: type 1? (0: No, 1: Yes)	b_type1
diabetes2	Diabetes status: type 2? (0: No, 1: Yes)	b_type2
weight (kg)	Weight	Not available
height (cm)	Height	Not available
weight (m) / (height (cm) /100)^2	Body mass index (BMI)	bmi

Parameters in QRISK3 R package	Meaning of variables	Variables in original algorithm
ethnicity	1 White or not stated 2 Indian 3 Pakistani 4 Bangladeshi 5 Other Asian 6 Black Caribbean 7 Black African 8 Chinese 9 Other ethnic group	ethrisk: 0, --not stated 1, --white 2, --inidan 3, --Pakistani 4,--Bangladeshi 5,--Other Asian 6,--Black Caribbean 7,--Black African 8,--Chinese 9--Other ethnic group
heart_attack_relative	Angina or heart attack in a 1st degree relative < 60? (0: No, 1: Yes)	fh_cvd
cholesterol_HDL_ratio	Cholesterol/HDL ratio? (range from 1 to 11, e.g. 4)	rati
systolic_blood_pressure	Systolic blood pressure (mmHg, e.g. 180 mmHg)	sbp
std_systolic_blood_pressure	Standard deviation of at least two most recent systolic blood pressure readings(mmHg)	sbps5
smoke	1 non-smoker 2 ex-smoker 3 light smoker (less than 10) 4 moderate smoker (10 to 19) 5 heavy smoker (20 or over)	smoke_cat: 0 non-smoker 1 ex-smoker 2 light smoker (less than 10) 3 moderate smoker (10 to 19) 4 heavy smoker (20 or over)
townsend	Townsend deprivation scores	town

5.4.2 Validation

The two datasets QRISK3_2017_test and QRISK3_2019_test were used for validation. Risk scores calculated from this QRISK3 package, the original algorithm and the SAS version on the same group of patients were exactly the same. The external validation of this QRISK3 package in a big CPRD cohorts with 3.6 million patients shows a good and similar discrimination (C statistic: 0.85) and calibration to a previous study [7] compared to the original QRISK3 paper [3].

5.4.3 Usage and features

A patient cohort with anonymous patient identifiers and CVD risk factors should first extracted and coded similarly to QRISK3 by the user. Missing values in the dataset should be handled (e.g. multiple imputation) before using this package. Column names of CVD risk factors (e.g. “age”) should then be specified correctly to the QRISK3_2017 function. The function returns calculated risk scores through a dataset with three columns, including patient identifier, calculated QRISK3 score and calculated QRISK3 score with one digit. It also reminds users to double check whether the definition of their variables was the same as the definition of QRISK3. The package also automatically detects whether all variables were coded as numeric and whether age of patients was ranged between 25 and 84, if not an error message returns (explained in Table 5.2).

Table 5.2: Description of error message in the QRISK3 R package.

Error message	Conditions	Explanation
"Variables including XXX, XXX must be coded as numeric (0/1) variable."	When at least one of variables in dataset are not numeric	QRISK3 algorithm needs numeric variable (0/1) to calculate risk
"Age of patients must be between 25 and 84."	When at least one patient in the dataset has age below 25 or above 84	QRISK3 algorithm was developed from a population with age between 25 and 84
"Variables including XXX, XXX has missing values."	When at least one of variables in dataset has missing value	Missing values must be handled before using this QRISK3 algorithm

5.5 Workflow

5.5.1. Set path and read data from CSV file

```
dataPath <- "yourPath"
dataName <- "yourDataName.csv"

setwd(dataPath)
myData <- read.csv(dataName, check.names=FALSE)
```

5.5.2. See the data structure and other information

```
#See data structure

str(myData)
```

```
## 'data.frame':    48 obs. of  26 variables:
##  $ QRISK_C_algorithm_score  : num  17.2 36 21.6 24.1 17.2 19.1 20.9 22.3 19.3 23.5 ...
##  $ age            : int  64 64 64 64 64 64 64 64 64 64 ...
##  $ gender         : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ b_AF           : int  0 1 0 0 0 0 0 0 0 0 ...
##  $ b_atypicalantipsy: int  0 0 1 0 0 0 0 0 0 0 ...
##  $ b_corticosteroids: int  0 0 0 1 0 0 0 0 0 0 ...
##  $ b_impotence2     : int  0 0 0 0 1 0 0 0 0 0 ...
##  $ b_migraine       : int  0 0 0 0 0 1 0 0 0 0 ...
##  $ b_ra           : int  0 0 0 0 0 0 1 0 0 0 ...
##  $ b_renal        : int  0 0 0 0 0 0 0 1 0 0 ...
##  $ b_semi         : int  0 0 0 0 0 0 0 0 1 0 ...
##  $ b_sle          : int  0 0 0 0 0 0 0 0 0 1 ...
##  $ b_treatedhyp     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ b_type1        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ b_type2        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ weight         : int  70 70 70 70 70 70 70 70 70 70 ...
##  $ height         : int  180 180 180 180 180 180 180 180 180 180 ...
##  $ ethrisk        : int  2 2 2 2 2 2 2 2 2 2 ...
##  $ fh_cvd         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ rati           : int  4 4 4 4 4 4 4 4 4 4 ...
##  $ sbp            : int  180 180 180 180 180 180 180 180 180 180 ...
##  $ sbps5          : int  20 20 20 20 20 20 20 20 20 20 ...
##  $ smoke_cat        : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ surv           : int  10 10 10 10 10 10 10 10 10 10 ...
##  $ town           : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ ID             : int  1 2 3 4 5 6 7 8 9 10 ...


#See missing value
# summary(myData)

#If there is any missing value
#please use methods (e.g. multiple imputation) to impute missing value

#Once there is no missing value
#Get all variable names in your data
# colnames(myData)
```

```
#Use help of this package to map your variable to QRISK3 variables
# ?QRISK3_2017
```

5.5.3. Call the QRISK3 function to calculate risk score

```
test_all_rst <- QRISK3_2017(data= myData, patid="ID", gender="gender",age="age",
atrial_fibrillation="b_AF", atypical_antipsy="b_atypicalantipsy",
regular_steroid_tablets="b_corticosteroids", erectile_disfunction="b_impotence2",
migraine="b_migraine", rheumatoid_arthritis="b_ra",
chronic_kidney_disease="b_renal", severe_mental_illness="b_semi",
systemic_lupus_erythematosis="b_sle",
blood_pressure_treatment="b_treatedhyp", diabetes1="b_type1",
diabetes2="b_type2", weight="weight", height="height",
ethiniciy="ethrisk", heart_attack_relative="fh_cvd",
cholesterol_HDL_ratio="rati", systolic_blood_pressure="sbp",
std_systolic_blood_pressure="sbps5", smoke="smoke_cat", townsend="town")

##
## This R package was based on open-sourced original QRISK3-2017 algorithm.

## <https://qrisk.org/three/src.php> Copyright 2017 ClinRisk Ltd.

##
## The risk score calculated from this R package can only be used for  research purpose.

##
## Please refer to QRISK3 website for more information

## <https://qrisk.org/three/index.php>

##
## Important: Please double check whether your variables are coded the same as the
QRISK3 calculator

##
## Height should have unit as (cm)

## Weight should have unit as (kg)

##
## Ethnicity should be coded as:

##   Ethnicity_category Ethnicity
## 1 White or not stated         1
## 2            Indian         2
## 3          Pakistani         3
## 4        Bangladeshi         4
## 5        Other Asian         5
## 6    Black Caribbean         6

##
## Smoke should be coded as:

##              Smoke_category Smoke
## 1              non-smoker     1
## 2              ex-smoker     2
## 3 light smoker (less than 10)     3
```

```
## 4  moderate smoker (10 to 19)     4
## 5   heavy smoker (20 or over)     5

##
## The head of result in all patients is:

##   ID QRISK3_2017 QRISK3_2017_1digit
## 1  1    17.22985              17.2
## 2  2    17.89260              17.9
## 3  3    36.02081              36.0
## 4  4    21.60346              21.6
## 5  5    24.06195              24.1
## 6  6    17.22985              17.2
```

5.6 Use Case

Users first need to create a statistical analysis dataset similar to the provided test dataset (e.g. QRISK3_2019_test) which contains information of patients' identifier and QRISK3 risk factors and mimic QRISK3's training cohort [3]. The structure of this statistical analysis dataset would be each row (observation) represents one individual patient and each column represents one of QRISK3 predictors. The exact definition of all QRISK3 predictors could be found from Box 1 of original QRISK3 paper [3]. Variables used by QRISK3 can be extracted from EHR databases, such as CPRD [21] or QResearch [22]. Code lists (Read code) for the outcome variable (CVD) can be obtained from the supplementary materials of QRISK3 paper [3]. Code lists for variables included in QRISK2 can be extracted from a previous study [23]. Code lists for other variables including anxiety, alcohol abuse, atypical anti-psychotic medication, erectile dysfunction, HIV/AIDS, left ventricular hypertrophy, migraine and systemic lupus erythematosus could be found from CPRD [24] or clinical codes website [25]. All CVD risk factors should be coded as numeric, binary variables should be coded as 0 or 1, categorical variables such as smoking status should be coded as the same as this package. Any differences between users' variables and QRISK3 predictors (e.g. different criteria to define smoking status) should be mentioned in users' final report. Once the analysis dataset was extracted, it is recommended to compare the distribution of users' analysis dataset to Qresearch's cohort using their baseline table (Table 1) [3 26]. Missing value should be imputed with multiple imputation [27]. Finally, users follow the above workflow and carefully match their variable names to pre-defined QRISK3 predictors to calculate risk score, the function would return a dataset with patient identifier, calculated score and calculated score with 1 digit.

5.7 Discussion

This R package successfully implements the QRISK3 algorithm into R, which allows researchers to calculate CVD risk of patients in the next 10 years. The R package was validated by the original algorithm and a SAS version. This is also the first R implementation of the QRISK3 algorithm at the date of writing.

Though QRISK3 was already published and released from the online website, it is time consuming for researchers to calculate QRISK3 risk score, as the online calculator cannot be used as a service to obtain QRISK3 scores for a large cohort, and the original algorithm is written in C rather than a well-established data science language such as R. This package bridges this gap. It allows researchers to obtain QRISK3 scores for large cohorts, which could help to improve model accuracy of QRISK3 and help with any more applied tasks that require knowing CVD risk at a patient level.

Although it is easy to use this R function to calculate a risk score, researchers should carefully check whether their variables are coded the same as the original QRISK3 cohort, otherwise the calculated score might not be the correct risk of the patient in the cohort. For example, a patient who is a smoker is coded as "1" in the variable "smoking" would be in conflict with the definition of the QRISK3 algorithm ("smoking" equals 1 in this R package means non-smoker). Since QRISK is updated annually every spring, researchers who are interested in the latest work should refer to their website [10].

In conclusion, we developed this R package to allow researchers to obtain QRISK3 scores for large cohorts. It allows the research community to better understand and apply a currently used risk prediction model for CVD risk.

5.8 Software availability

Package available from CRAN: https://cran.r-project.org/web/packages/QRISK3/index.html

Source code available from: https://github.com/YanLiUK/QRISK3

Archived source code as at time of publication:

https://doi.org/10.5281/zenodo.3570682

Source code of validation available from:

https://github.com/YanLiUK/QRISK3_valid

C code and SAS version for validation at time of publication:

https://doi.org/10.5281/zenodo.3571304

License: GPL-3

Data availability

Underlying data

Original QRISK3 algorithm: https://qrisk.org/three/src.php

5.9 References

1. Cardiovascular diseases (CVDs). https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds). Accessed November 16, 2019.
2. Grant SW, Collins GS, Nashef SAM. Statistical Primer: developing and validating a risk prediction model†. *Eur J Cardio-Thoracic Surg*. 2018;54(2):203-208. doi:10.1093/ejcts/ezy180
3. Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. *Bmj*. 2017;2099(May):j2099. doi:10.1136/bmj.j2099
4. Collins GS, Altman DG. An independent and external validation of QRISK2 cardiovascular disease risk score: a prospective open cohort study. *BMJ*. 2010;340(July):c2442. doi:10.1136/bmj.c2442
5. CVD risk assessment and management - NICE CKS. https://cks.nice.org.uk/cvd-risk-assessment-and-management#!scenario:2. Accessed November 16, 2019.
6. Cox DR. *Regression Models and Life-Tables*. Vol 34.; 1972. http://www.stat.cmu.edu/~ryantibs/journalclub/cox_1972.pdf. Accessed February 21, 2019.
7. Li Y, Sperrin M, Belmonte M, Pate A, Ashcroft DM, van Staa TP. Do population-level risk prediction models that use routinely collected health data reliably predict individual risks? *Sci Rep*. 2019;9(1):11222. doi:10.1038/s41598-019-47712-5
8. Pate A, Emsley R, Ashcroft D, Brown B, Staa T Van. The uncertainty with using risk prediction models for individual decision making: an exemplar cohort study examining the prediction of cardiovascular disease in English primary care. *BMC Med*. June 2019.
9. Li Y, Sperrin M, Martin GP, Ashcroft DM, van Staa TP. Examining the impact of data quality and completeness of electronic health records on predictions of patients' risks of cardiovascular disease. *Int J Med Inform*. November 2019:104033. doi:10.1016/j.ijmedinf.2019.104033
10. QRISK3. https://qrisk.org/three/. Accessed November 16, 2019.
11. (No Title). https://qrisk.org/three/src.php. Accessed November 16, 2019.
12. R: The R Project for Statistical Computing. https://www.r-project.org/. Accessed April 28, 2019.
13. SAS® 9.4 Statements: Reference, Fifth Edition. http://support.sas.com/documentation/cdl/en/lestmtsref/69738/HTML/default/viewer.htm#n1i8w2bwu1fn5kn1gpxj18xttbb0.htm. Accessed August 20, 2017.
14. R Installation and Administration. https://cran.r-project.org/doc/manuals/r-release/R-admin.html. Accessed November 17, 2019.
15. Submitting your first package to CRAN, my experience | R-bloggers. https://www.r-bloggers.com/submitting-your-first-package-to-cran-my-experience/. Accessed November 17, 2019.

16. Writing an R package from scratch | Not So Standard Deviations. https://hilaryparker.com/2014/04/29/writing-an-r-package-from-scratch/. Accessed November 17, 2019.

17. R package primer. http://kbroman.org/pkg_primer/. Accessed November 17, 2019.

18. Collins D, Lee J, Bobrovitz N, Koshiaris C, Ward A, Heneghan C. whoishRisk - an R package to calculate WHO/ISH cardiovascular risk scores for all epidemiological subregions of the world. *F1000Research*. 2017;5. doi:10.12688/f1000research.9742.2

19. CRAN - Package QRISK3. https://cran.r-project.org/web/packages/QRISK3/index.html. Accessed December 8, 2019.

20. YanLiUK/QRISK3: A QRISK3 R package implements QRISK3 algorithm into R. https://github.com/YanLiUK/QRISK3. Accessed December 12, 2019.

21. Clinical Practice Research Datalink - CPRD. https://www.cprd.com/intro.asp. Accessed August 20, 2017.

22. Home - QResearch. https://www.qresearch.org/. Accessed December 8, 2019.

23. van Staa T-P, Gulliford M, Ng ES-W, Goldacre B, Smeeth L. Prediction of cardiovascular risk using Framingham, ASSIGN and QRISK2: how well do they predict individual rather than population risk? *PLoS One*. 2014;9(10):e106455. doi:10.1371/journal.pone.0106455

24. CPRD @ Cambridge - Code Lists - Primary Care Unit. http://www.phpc.cam.ac.uk/pcu/cprd_cam/codelists/. Accessed November 18, 2019.

25. ClinicalCodes Repository. https://clinicalcodes.rss.mhs.man.ac.uk/. Accessed November 18, 2019.

26. Pate A, Emsley R, Ashcroft DM, Brown B, van Staa T. The uncertainty with using risk prediction models for individual decision making: an exemplar cohort study examining the prediction of cardiovascular disease in English primary care. *BMC Med*. 2019;17(1):134. doi:10.1186/s12916-019-1368-8

27. BMJ. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. 2009. doi:10.1136/bmj.b2393

Blank page

Chapter 6 The instability of machine learning and statistical models in predicting individual patient risks:

an approach to improve the clinical utility of these models

Yan Li[1], Matthew Sperrin[1], Darren M Ashcroft[2,3], Tjeerd Pieter van Staa[1,4,5,7]

[1]Health e-Research Centre, School of Health Sciences, Faculty of Biology, Medicine and Health, the University of Manchester, Manchester, Oxford Road, Manchester, M13 9PL, UK
[2]Centre for Pharmacoepidemiology and Drug Safety, School of Health Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Oxford Road, Manchester, M13 9PL, UK
[3]NIHR Greater Manchester Patient Safety Translational Research Centre, School of Health Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Oxford Road, Manchester, M13 9PL, UK
[4]Utrecht Institute for Pharmaceutical Sciences, Utrecht University, Utrecht, Netherlands
[5]Alan Turing Institute, Headquartered at the British Library, London, UK

Corresponding author: Tjeerd van Staa, tjeerd.vanstaa@manchester.ac.uk

Journal title: Ready to submit
Doi:

Word count: 2955
Abstract: 313
Number of tables: 1
Number of figures: 4

6.1 Abstract

Objective: Previous work found that different machine learning models and statistical models can give inconsistent absolute risk predictions for patients even with similar model performances. This study aims to evaluate whether ranking of individual risk predictions may improve consistency between different prediction models.

Design: Longitudinal cohort study from 1st Jan 1998 to Jan 2019.

Setting: 3.6 million patients from the Clinical Practice Research Datalink in primary care.

Main outcome measures: Consistency of individual rank and absolute risk prediction for the same patients among models with comparable model performance.

Methods: 15 different prediction techniques including 12 families of machine learning model and 3 families of Cox model from previous study were considered. Model performance of ensembled machine learning models and Cox models were compared in the same test cohort. Individual rank was derived by ranking individual risk and percentage of rank was derived by individual rank over number of patients. Distribution of individual rank and individual patients were compared among different models for the same patients.

Results: All ensembled machine learning models and Cox models had similar population-level model performance (C statistics and calibration). The study found that ranking of risk predictions improved the consistency between different machine learning and statistical prediction models compared to absolute risks, For patients in the highest risk group, the 25% percentile (Q1) to 75% percentile (Q3) of differences of absolute predicted risk between models was -18.8% ~ -9.0%. The Q1 to Q3 when ranking risks was -0.6% ~ 1.0%. There was larger variability between models in ranking for patients in the medium risk group.

Conclusions: The clinical utility of risk prediction model could be improved by supplying percentage of patient ranks with their individual risk prediction from

multiple models in clinical practice. Treatment decision based on risk prediction model for patients especially for medium risk groups should be made in conjunction with additional clinical testing and clinical judgment.

6.2 Introduction

Risk prediction models aim to assist clinicians in screening for high risk patients and making evidence-based clinical decision quickly [1]. Traditionally, they were developed from statistical models such as Cox proportional hazard model with time to patients' disease or status of disease as outcome variable and risk factors of disease as predictors. Recently, with the increase of computing power and breakthrough of deep learning [2], machine learning models have started to show their strength in image recognition [3] and are also being used to develop model to predict individual patient risks of clinical outcomes. However, literatures have shown that though these models are well validated on population level in terms of good discrimination (ability to discriminate high/low risk patients) and calibration (agreement between predicted risk and observed risk), they have great uncertainty on individual risk prediction [4 5 6]. Models with similar population level model performance but developed from a heterogeneous setting [4], with different model choices [5] or using different model algorithms [6] predict risks inconsistently for same patients. A patient treatment decision could thus be strongly influenced by the arbitrary choice of a modelling technique or arbitrary design choice. A previous study found that the inconsistency of individual risk predictions among models with similar model performance was larger in high risk patients, which may strongly limit the clinical utility of risk prediction model in identifying high risk patients [6]. The ability of models to discriminate high/low risk patients is often measured by C-statistic [7 8], which is a proportion of pairs of patients whose risks were correctly ranked (i.e. a true high risk patient would be ranked as higher risk than a true low risk patient) by the model among all random pairs of patients in the cohort. This study aims to evaluate whether ranking of predicted patient risks, rather than estimating their absolute predicted risks, improves consistency between different machine learning and statistical prediction models thus improving their clinical utility and reducing dependency on arbitrary model choices. The prediction of incident cardiovascular disease (CVD) was used as exemplar. Various statistical CVD risk prediction models are being recommended for use in

treatment guidelines such as Framingham for US [9], QRISK for UK [10] and ESC score for Europe [11] and machine learning models have been proposed to considerably improve the performance in CVD risk prediction [12 13 14 15] .

6.3 Method

6.3.1 Data source

Clinical Practice Research Datalink (CPRD GOLD) was used to derive the study cohort. CPRD database includes patients' electronic health records (EHR) of about 6.9% of the population in England [16]. Patients' EHRs were extracted from general practices which include detailed information such as demographics (age, gender and ethnicity), symptoms, tests, diagnoses, prescribed treatments, health-related behaviours and referrals to secondary care [16]. CPRD has also been linked to Hospital Episode Statistics, Office for National Statistics mortality records and Townsend deprivation scores to acquire additional patient information about hospital admissions (e.g. date and discharge diagnoses), cause-specific mortality and deprivation [16]. CPRD is a well-recognised representative cohort of UK population and has been used to develop and validate clinical risk prediction models, thousands of studies [17] have used CPRD including a validation of a popular UK risk prediction model (QRISK2) [18] and several analysis of machine learning [19 20].

6.3.2 Study population

The same selection criteria as QRISK3 [10] was used in this study to derive study population, risk factors and CVD outcome [21 22]. The main inclusion criteria of patients including age between 25 and 84, no CVD history or any statin prescription to the index date. The index date was randomly selected from the patients' follow-up period. Follow-up was defined as

started at the date of the patient's registration with the practice, 25th birthday, or January 1 1998 (whichever latest) and ended at the date of death, incident CVD, the date of leaving the practice or last date of data collection (whichever earliest) [6]. Random index date was used rather than a single calendar time date (as in QRISK3) aims to capture time-relevant practice variability and CVD risk factors with a better spread of calendar time and age [23]. The primary clinical outcome of interest of CVD (including coronary heart disease, ischaemic stroke, and transient ischaemic attack) was defined similarly to QRISK3 [10].

6.3.3 CVD risk factors

This study considered the same predictors selected for QRISK3, including gender, age, body mass index (BMI), smoking history, cholesterol/HDL ratio, systolic blood pressure (SBP) and its standard deviation, history of prescribing of atypical antipsychotic medication, blood pressure treatment or regular oral glucocorticoids, clinical history of systemic lupus erythematosus, atrial fibrillation, chronic kidney disease (stage 3, 4 or 5), erectile dysfunction, migraine, rheumatoid arthritis, severe mental illness or type 1 or 2 diabetes mellitus, family history of angina or heart attack in a 1st degree relative aged < 60 years, ethnicity and Townsend socioeconomic score [10]. All models except Framingham in this study considered all these predictors, Framingham used fewer and different predictors [6 24].

6.3.4 Risk prediction models, model development and validation

The study used the same risk prediction models as developed in a previous study [6], including 12 machine learning models (ensembled from 1200 machine learning models from 12 families of machine learning models) and three Cox models. The 12 families of machine learning models included logistic model [25], random forest [26] and neural network [27] from R package "Caret" [28]; logistic model, random forest, neural network, extra-tree model [29] and gradient boosting classifier [29] from Python package "Sklearn" [30]; logistic model, random forest, neural network and autoML [31] from Python package "h2o" [32]. Models of the same

machine learning algorithm but different software were treated as different model family in the study. This is because the differences in the settings (hyper-parameters) to control model fitting might result a different best performed model through model fitting process. The Cox models in this study included QRISK3 [10], Framingham [33] and a local fitted Cox model.

The detailed model fitting process is described in the previous study [6]. The variables with missing values including ethnicity (% missing in overall cohort was 54.3%), BMI (40.3%), Townsend score (0.1%), SBP (26.9%), standard deviation of SBP (53.9%), ratio of cholesterol and High-Density Lipoprotein (HDL) (65.0%) and smoking status (25.2%) were imputed 10 times with Markov chain Monte Carlo method with monotone style. The missing value imputed overall cohort were randomly split into overall training set (75%) and overall testing set (25%). 100 random samples of the overall training set were used to train machine learning models with grid searched on selected hyper-parameters and two-fold cross validation estimating calibration and discrimination [6]. A total of 1200 machine learning models were first fitted and validated in its own testing cohort with high discrimination and calibration, then model performance and individual risk predictions of these models on the overall testing cohort were calculated. Individual risk predictions of 12 ensemble machine learning models were averaged from individual risk predictions of machine learning models from each model family (model ensemble with soft voting). The model performance of 12 ensemble machine learning models were calculated and presented in this study. Threshold of 7.5% (according to ACC/AHA Guideline on the Assessment of Cardiovascular Risk [34]) was used to calculate sensitivity and positive predictive value (PPV). The refitted local Cox model followed the same process as machine learning in terms of derivation and validation except that there was no tuning hyper-parameters process. This study considered QRISK3 [10 35] and Framingham [33] by their published model formula, as they were both internal and external validated [10 21 36]. The individual risk predictions among 12 ensembled machine learning models and three Cox models were calculated and compared in the same overall testing cohort [6].

6.3.5 Main Statistical analysis

Fifteen individual ranks for the same patients were derived from 12 ensembled machine learning models and three Cox models. Ranking of individual risk predictions was defined by the descending order of model predictions. The percentage of rank was derived by rank over total number of ranks (total number of patients). The distribution of patient percentages of rank with machine learning and Cox models was plotted and compared for the same group of patients with a predicted risk between 7% and 8% from two different reference models (logistic Caret model or local Cox model). This investigated whether different models rank patients similarly. The logistic Caret model was used as reference model as it is a neutral model between machine learning and traditional statistic models (can be fitted in both ways) and being used as reference in the previous study [6]. The local Cox model was selected as reference model as previous study showed that both Cox and machine learning models have comparable model performance while Cox models additionally consider patient censoring (i.e., the effects of patients drop out early). Boxplots of differences of individual ranks among models were plotted against deciles of predicted absolute risks of the local Cox model. This investigated whether the differences of individual ranks came from lower or higher risk groups.

6.4 Results

There were 3.6 million patients from 391 general practices in the study population. Table 6.1 summarises the baseline characteristics of the study population showing that the derivation and validation cohort had similar characteristics.

Table 6.1: Baseline characteristics of the study population (patients aged 25-84 years without history of CVD or prior statin use at study entry

	Derivation cohort	Validation cohort
General characteristics		
Number of patients	2746453	915479
Number of CVD cases (%)	86769 (3.2)	28828 (3.1)
Number of female patients (%)	1406796 (51.2)	469098 (51.2)
CVD risk factors		
Age (Mean (SD))	44.7 (15.6)	44.7 (15.7)
BMI (Mean (SD))	26.7 (5.0)	26.7 (5.0)
Cholesterol/HDL ratio (Mean (SD))	3.9 (1.3)	3.9 (1.3)
Number of patients on atypical antipsychotic medication (%)	123060 (0.4)	40300 (0.4)
Number of patients on blood pressure treatment (%)	1839640 (6.7)	619620 (6.8)
Number of patients on regular steroid tablets (%)	20590 (0.1)	6940 (0.1)
Number of patients with Systemic Lupus Erythematosus (%)	18400 (0.1)	6060 (0.1)
Number of patients with angina or heart attack in a 1st degree relative < 60 (%)	984550 (3.6)	326190 (3.6)
Number of patients with atrial fibrillation (%)	207780 (0.8)	69650 (0.8)
Number of patients with chronic kidney disease (stage 3, 4 or 5) (%)	301330 (1.1)	102400 (1.1)
Number of patients with erectile dysfunction (%)	396510 (1.4)	131100 (1.4)
Number of patients with migraines (%)	1774390 (6.5)	591060 (6.5)
Number of patients with rheumatoid arthritis (%)	161670 (0.6)	54590 (0.6)
Number of patients with severe mental illness (this includes schizophrenia, bipolar disorder and moderate/severe depression) (%)	2198610 (8.0)	728320 (8.0)
Number of patients with type 1 diabetes (%)	58990 (0.2)	20970 (0.2)
Number of patients with type 2 diabetes (%)	355690 (1.3)	118260 (1.3)
SBP (Mean (SD))	126.9 (16.7)	126.9 (16.7)
Standard deviation of each individual patients' SBP (Mean (SD))	9.9 (5.6)	9.9 (5.6)
Ethnicity		
Number of patients with other ethnicity (%)	372240 (1.4)	125370 (1.4)
White or not recorded (%)	25731820 (93.7)	8573550 (93.7)
Smoking		
Number ex-smokers (%)	6300299 (22.9)	2095026 (22.9)
Number of current-smokers (%)	8066066 (29.4)	2696146 (29.5)
Number of patients who never smoked (%)	13098165 (47.7)	4363618 (47.7)

	Derivation cohort	Validation cohort
Townsend		
Number of patients with Townsend - 1.the least deprived (%)	6004107 (21.9)	1999488 (21.8)
Number of patients with Townsend - 2.less deprived (%)	5947510 (21.7)	1976893 (21.6)
Number of patients with Townsend - 3.deprived (%)	5728916 (20.9)	1910200 (20.9)
Number of patients with Townsend - 4.more deprived (%)	5680051 (20.7)	1895230 (20.7)
Number of patients with Townsend - 5.the most deprived (%)	4103946 (14.9)	1372979 (15.0)

All models had similar model performance in terms of high discrimination (C statistic about 0.88) and calibration (similar low brier score and good calibration plots in their own framework [6]). Similar model performance of the 12 ensembled machine learning models and three Cox models were also showed in eTable 6.9.1.

Figure 6.1a (using logistic Caret model as reference model) and Figure 6.1b (using local Cox model as reference model) plotted the distribution of ranking of predicted CVD risks with different models for the same group of patients. Figure 6.1a shows that patients with absolute risk of 7%~8% from logistic Caret model had 25% percentile (Q1) to 75% percentile (Q3) of percentage of rank of 11.6%~12.3% in the logistic Caret model, 10.9%~14.3% in a random forest model, 11.4%~12.6% in a neural network, 11.2%~13.6% in QRISK3 and 11.1%~13.2% in Local Cox model. However, previous study [6] showed that patients with an absolute risk of 7%~8% from logistic model had a risk of 5.8%~16.1% in a random forest and 4.5%~9.4% in a neural network. It shows that there was variation (inconsistency) of ranking for the same patients among different models, but that this inconsistency was smaller than the inconsistency of magnitude of individual risk predictions. This statement holds when changing reference model to the local Cox model. Figure 6.1b shows that patients with risk of 7%~8% from Local Cox model had percentage of rank of 22.3%~23.5% in Local Cox model, 21.2%~24.3% in a logistic model, 21.0%~26.6% in a random forest model, 21.4%~24.0 in a neural network and 21.5%~24.8% in QRISK3.

Legends to figures

Figure 6.1: Distribution of percentage of individual patients' rank (rank was defined by decreasing order of individual risk predictions and percentage of rank was derived by dividing rank by number of patients) with machine learning and Cox models for patients with predicted risks of 7~8% in reference model

a. **Use logistic Caret model as the reference model**
b. **Use local Cox model as the reference model**

X axis: percentage of patients' rank

Y axis: relative frequency (estimated density value)

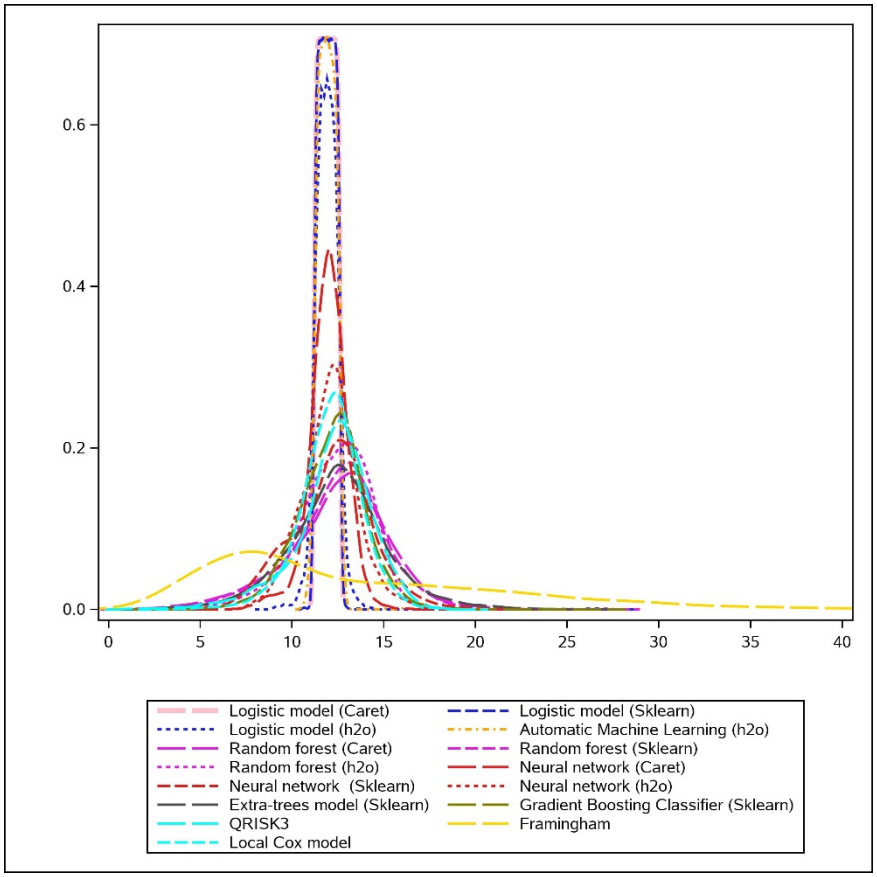


Figure 6.1a (Legends)

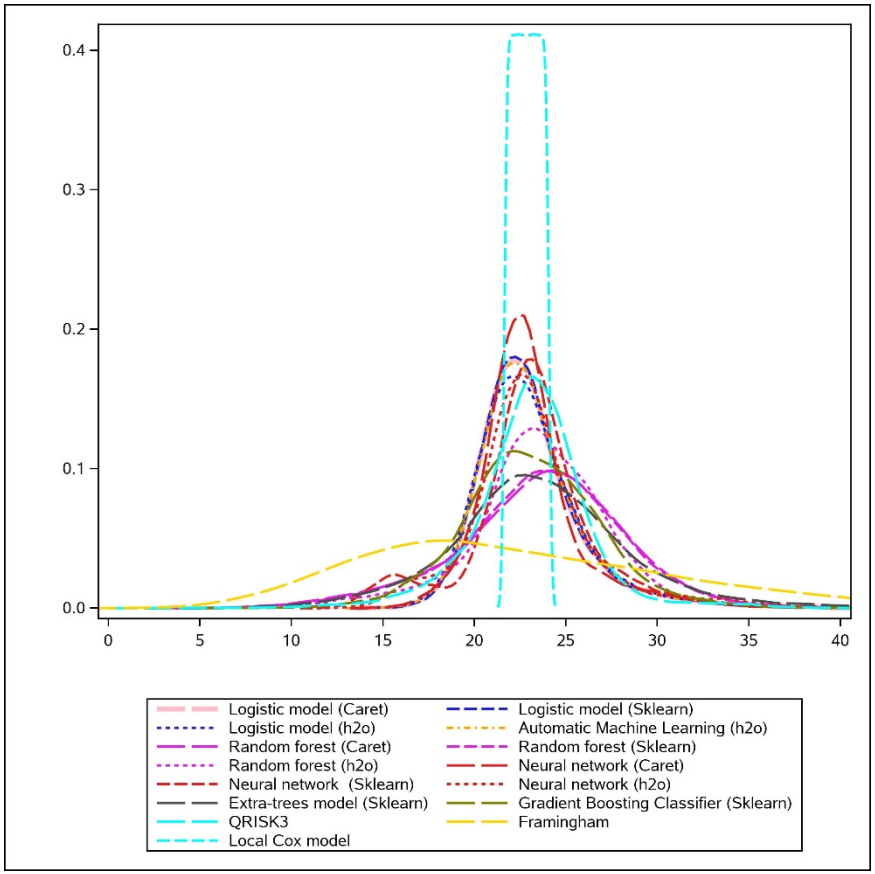


Figure 6.1b (Legends)

X axis: percentage of patients' rank

Y axis: relative frequency (estimated density value)

Figure 6.2a and 6.2b plotted the differences in individual risk and risk prediction ranking between the different models stratified by decile of absolute risk as estimated by the local Cox model. For patients in the highest risk group, the Q1 to Q3 of differences of predicted risk among models was (-18.8% ~ -9.0%). This was (-1.5% ~ -1.1%) for patients with medium predicted risk, and (-0.4% ~ -0.3%) for patients with lower predicted risk. However, for percentage of rank among models, patients in the highest risk group has an Q1 to Q3 of differences of percentage of rank as (-0.6% ~ 1.0%). This was (-2.7% ~ 3.4%) for patients in the medium risk group and (-2.9% ~ -0.6%) for patients in lower risk group. It shows that the differences in patient ranking had lower variation in higher risk groups but larger in medium risk group.

Figure 6.2: : Boxplot of differences of individual risk prediction ranks with machine learning and Cox models stratified by deciles of absolute predicted risks with local Cox model (reference model)

a. **Boxplot of differences of individual absolute risk predictions**
b. **Boxplot of differences of individual risk prediction ranks**

X axis: decile of absolute predicted risk with local Cox model

Y axis: differences with other models

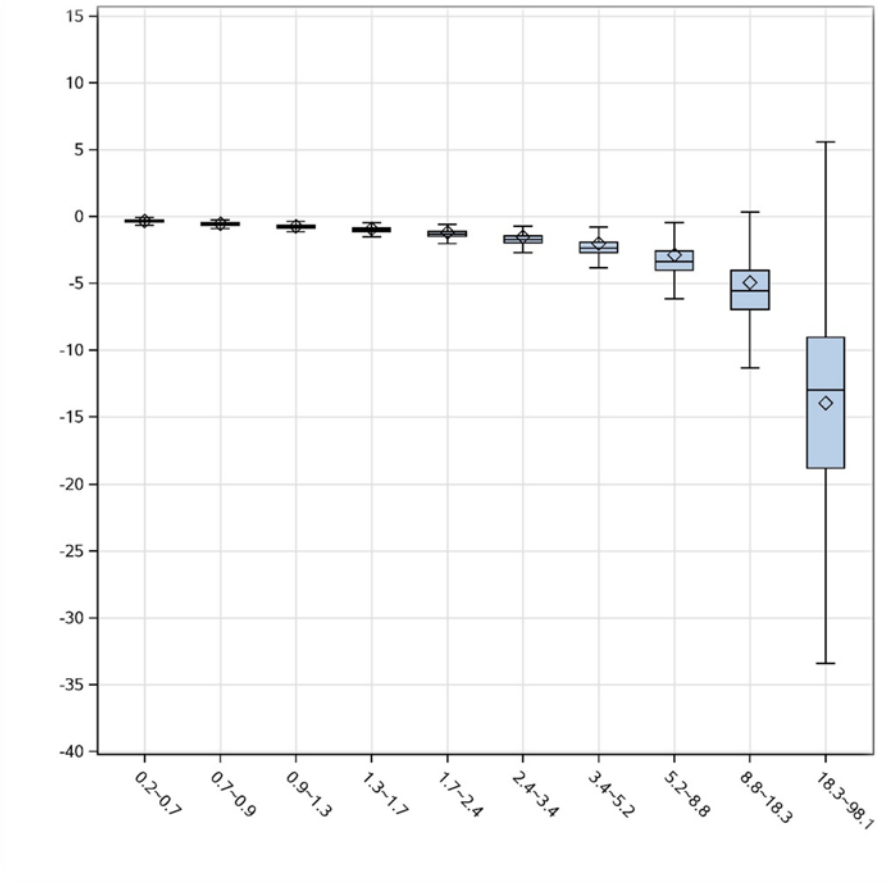


Figure 6.2a

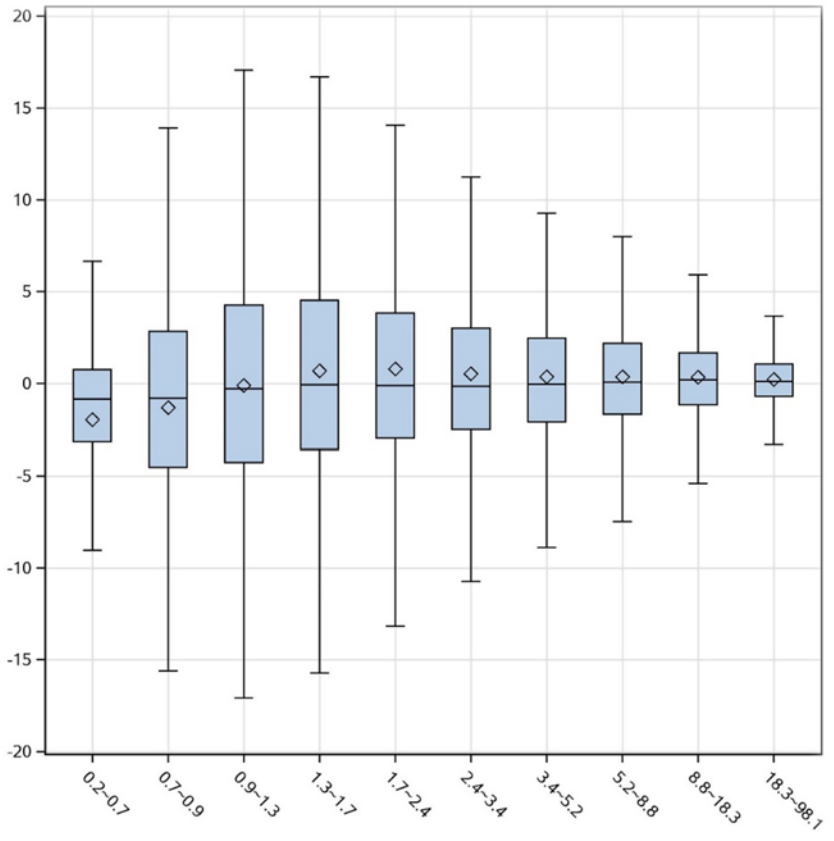


Figure 6.2b

Figure 6.3a plots the boxplot of differences in magnitude of patient risks against the decile of risk of local Cox model grouped by each model. Figure 6.3b was plotted similarly to Figure 6.3a except that it used differences of percentage of patient ranks rather than predicted risk. Figure 6.3a shows that for individual risk prediction, higher risk group patients have less consistency of risk among different models, while Figure 6.3b shows that for individual ranking, higher risk group patients had more consistency of ranking among different models.

Figure 6.3: Boxplot of differences of percentage of individual patients' rank or individual risk predictions with machine learning and Cox models stratified by deciles of predicted risks with local Cox model

a. **Boxplot of differences of individual risk predictions**
b. **Boxplot of differences of percentage of individual patients' rank**

X axis: decile of predicted risk displayed as the actual value of each decile with local Cox model

Y axis: differences in percentage of patients' rank (Figure 3a) or predicted risks with other models (Figure 3b)

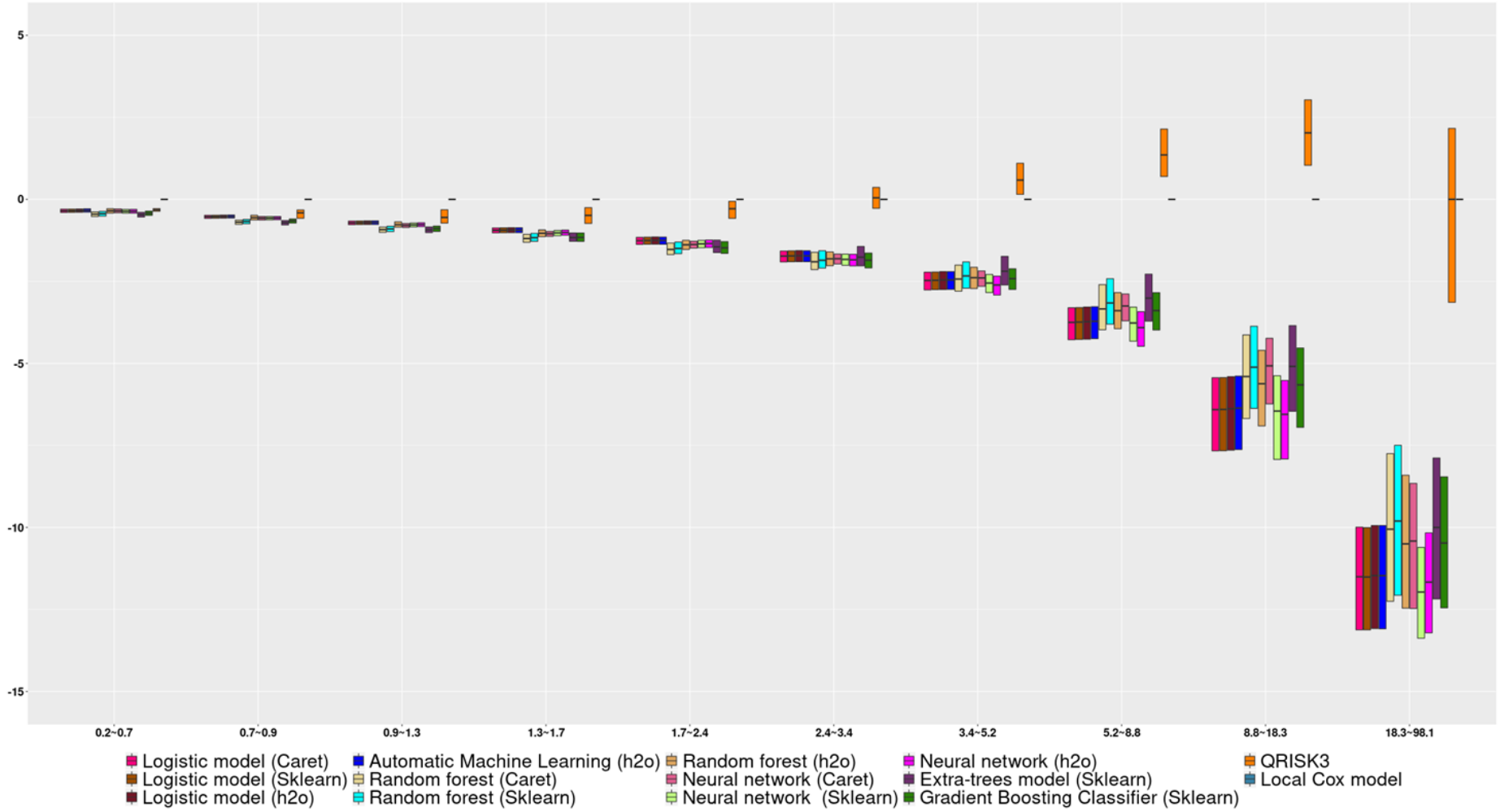


Figure 6.3a (Legends)

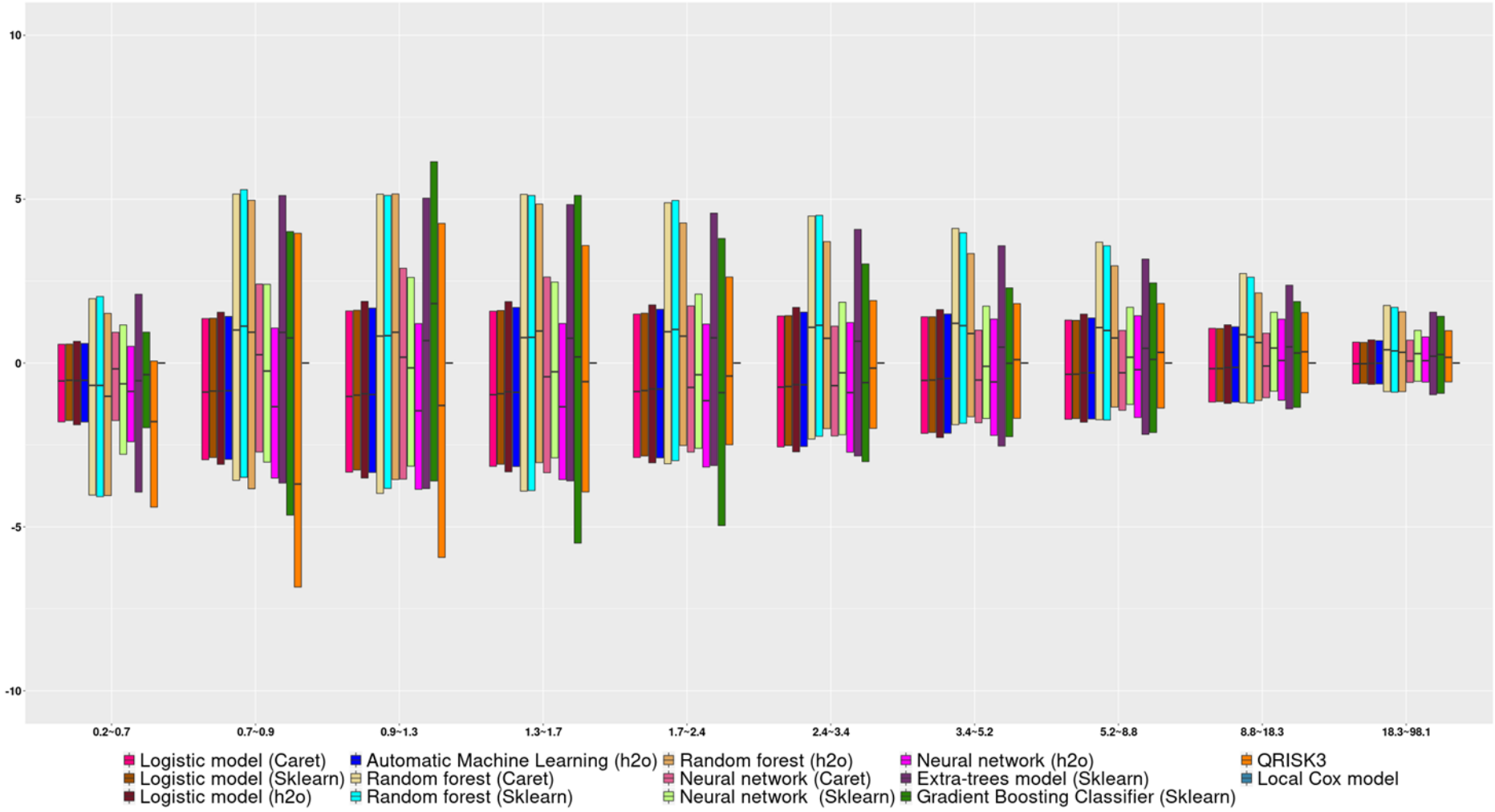


Figure 6.3b (Legends)

Figure 6.4 is a smoothed plot showing the variation of differences of percentage of patients' ranking (including Q1 to Q3, mean and median) among models of the same group of patients who have a predicted risk above different treatment thresholds in X-axis. It shows that different models rank the patients more consistently in higher risk group, which is consistent to the Figure 6.2b and Figure 6.3b. The Q1 to Q3 of percentage of differences of patients' ranking could narrow down from (-0.8%, 1.2%) to (-0.6%, 1.0%) if treatment threshold changed from 10% predicted risk to a higher 18.3% of local Cox model.

Figure 6.4: Smoothed 95% range of differences of percentage of individual patients' rank with machine learning and Cox models (Local Cox model as reference model) in patients who have the predicted risk above the selected threshold of probability (X-axis)

X axis: potential threshold of probability (with Local Cox model as reference model)

Y axis: differences of percentage of individual patients' rank with other models

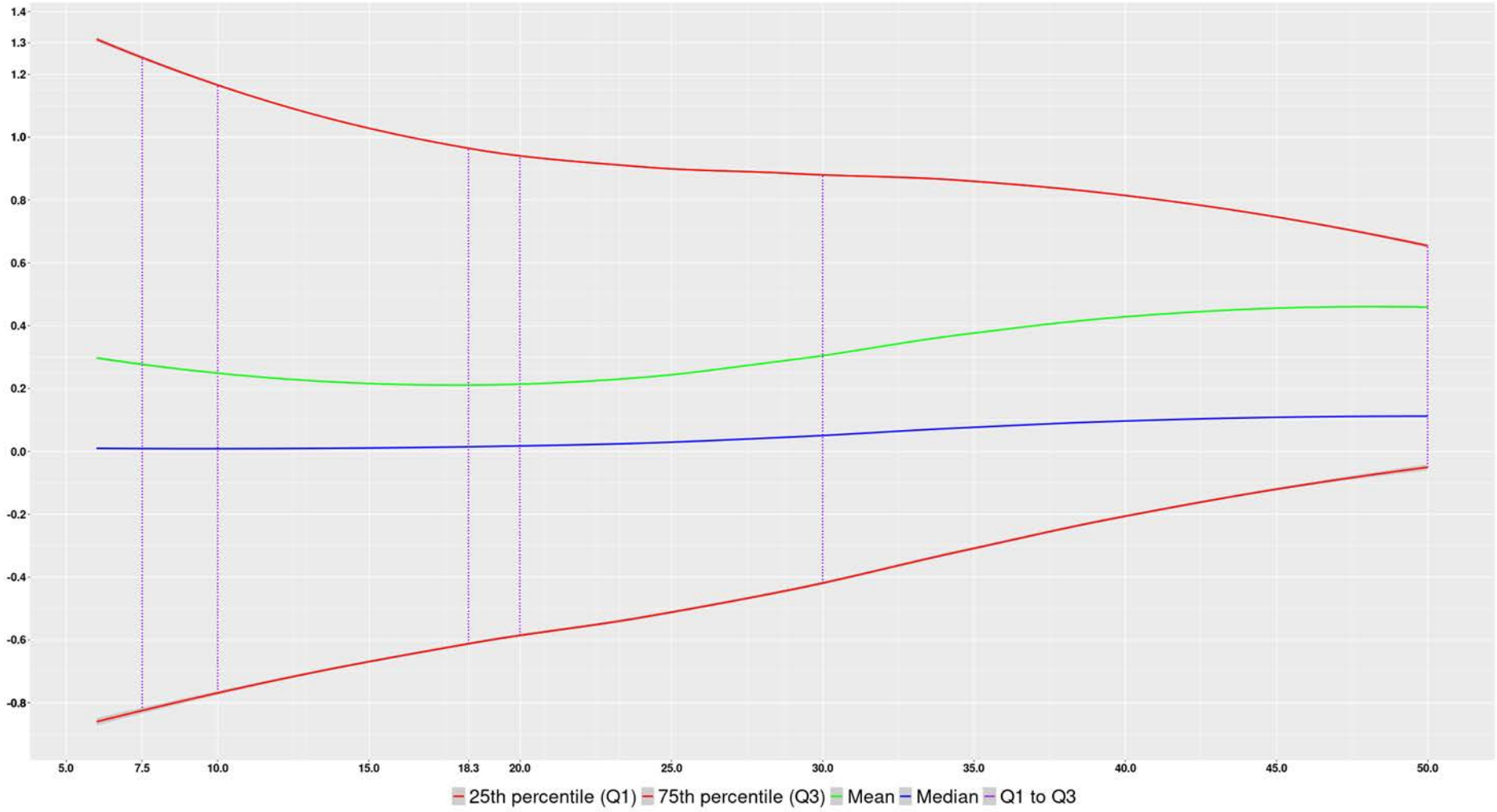


Figure 6.4 (Legends)

Figure 6.5 plotted differences of percentage of patient ranking stratified by different predictors. Smokers or patients who have ethnicity as white or unknown had less variation of ranking than patients with other characteristics, although the magnitudes of the differences comparing to patients with other predictors were relatively small.

Figure 6.5: Boxplot of differences of percentage of individual patients' rank with machine learning and Cox models (Local Cox model as reference model) for patients with different characteristics

X axis: patients with different CVD predictors (only select patients' CVD predictors with at least 1000 patients in their group)

Y axis: differences in percentage of patients' rank with other models

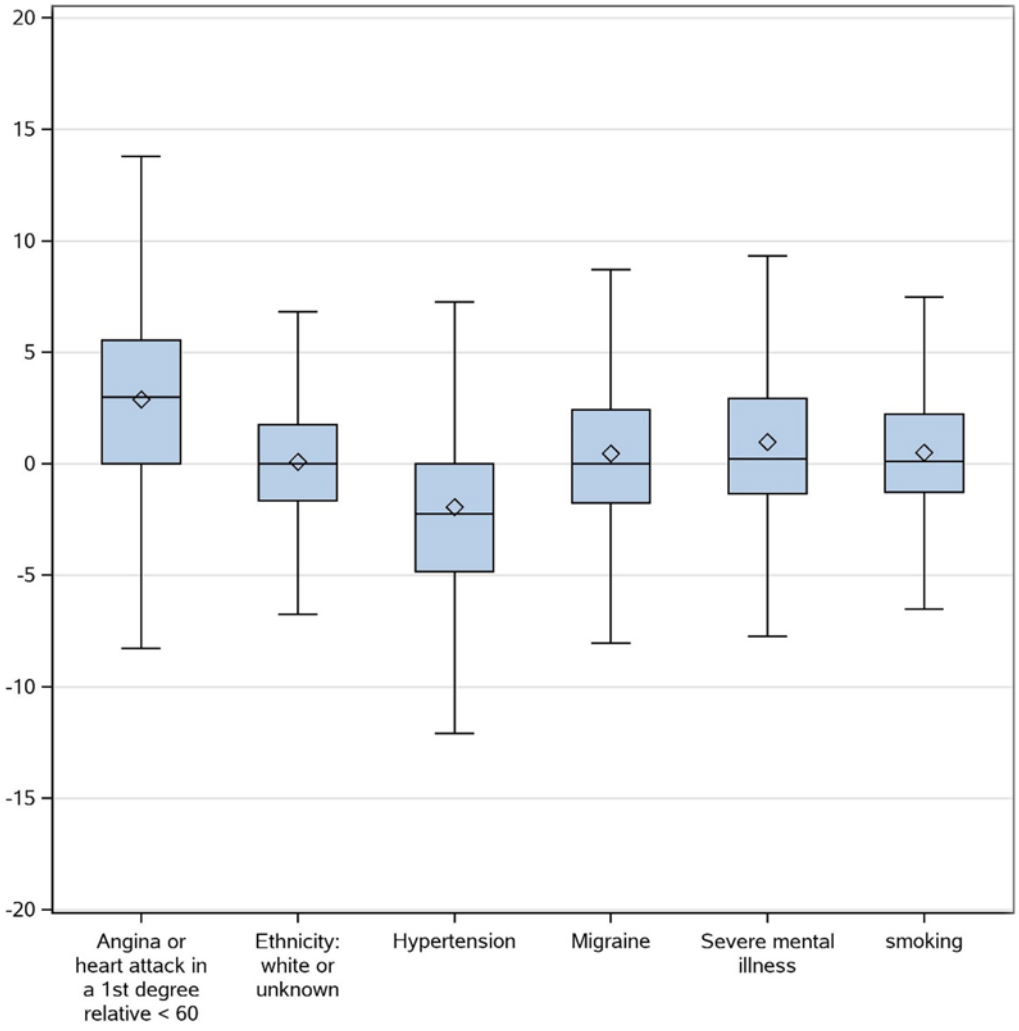


Figure 6.5 (Legends)

X axis: patients with different predictor (only select patients' predictors with at least 1000 patients in its group)

Y axis: differences of percentage of patients rank with other models

6.5 Discussion

This study found that machine learning and Cox models with comparable model performance rank higher risk group patients more consistently than medium/lower risk group for CVD risk prediction. Although patients from higher risk group had large inconsistency of individual absolute risk predictions, they were more likely ranked similarly among different models.

Previous studies used CVD risk prediction as an exemplar shows that individual risk would change much (potentially affecting treatment decisions) if considering heterogeneity between sites [21], model design of choice [37] and different types of model [6]. In this case, Absolute risk (probability) alone might not be enough for clinical usage on individual level giving the inconsistencies in prediction between models and dependency on modelling choices [6 21 37]. However, this study found that ranking of risks had less uncertainty on individual level among different models, compared to absolute risks, as there are finite possibilities of rank for a fixed cohort, the validity of individual rank was considered by model's discrimination ability and it could be derived from directly comparing linear predictor (a sum of multiplication of beta coefficients and predictors) of model alone. This study shows that though models have great uncertainty especially on higher risk patients, they rank higher risk patients more consistently and thus individual ranking of patients could be used in clinical practice to help identify patients who are a true high-risk patient. To our knowledge, this is the first study to consider using individual rank rather than individual risk to improve clinical usage of risk prediction model.

Our results show that the current threshold including 7.5% from ACC/AHA Guideline or 10% from NICE guidelines were not evidenced enough for using risk predictions on individual level, as there was large uncertainty on both of individual risk and rank for patients around these risks which could affect their treatment decisions. The results of this study support the concern from clinicians that using the new lower risk threshold 10% from NICE guideline for prescribing statin would over-treat healthy patients [38], as models have large inconsistency of risk prediction and ranking for patients above this threshold. This study also adds that the higher the threshold means the more certainty of high-risk patients on individual as different model would rank these high-risk patients more consistently than patients in lower risk group.

Future model development will need to focus on discriminating patients with medium risk on individual level due to uncertainty of both individual risk and rank for these patients.

Practically, a patient with very high or very low risk might be easily identified by clinicians through clinical judgment without the need of risk prediction model. A new statistic may be required measuring consistency of models in predicting individual risks and lack of effect of arbitrary modelling choices. Current model performance measurements including calibration and discrimination merely measure population level characteristics. A previous study [21 37 22] compared distribution of individual predictions for the same group of patients and this study compared distribution of individual rank for patients, a new statistic might be proposed to quantify uncertainty of individual patients' risk or rank. Individual rank seems to have better property than individual risk given its consistency on high risk patients, new type of model might focus on estimating patients' individual rank or linear predictor. Though individual rank does not directly reflect calibration of model, supplying with individual risk might resolve this. A future statistical and clinical useful model should predict accurate and robust individual risk prediction on both of population and individual level.

There are several limitations of this study. Models in this study considered 22 predictors as provided by QRISK3 developer [10]. A limitation was that more predictors could have been considered. However, the models with the 22 predictors already achieved a high discrimination and good calibration and more predictors could limit model's clinical utility as one needs to measure more predictors to make predictions with a higher chance of data quality issue (e.g. missingness of predictor). Another limitation in the analysis of specific CVD predictors was that not all specific predictors could be considered due to small sample sizes.

In conclusion, the clinical utility of risk prediction models could be improved by supplying ranking of individual risk predictions from multiple models in clinical practice. Consistency in ranking of risks between different models could give more confidence that a higher rank individual patient is a true high-risk patient who needs care. This may be preferable on arbitrarily picking one prediction model out of series of possible models. For patients who have uncertainty on both risk and rank (for the patients with medium risk), additional clinical testing and clinical judgement is needed to make treatment decisions.

6.6 Funding

This study was funded by China Scholarship Council (PhD studentship of Yan Li).

6.7 Acknowledgements

This study is based on data from the Clinical Practice Research Datalink obtained under license from the UK Medicines and Healthcare products Regulatory Agency. The protocol for this work was approved by the independent scientific advisory committee for Clinical Practice Research Datalink research (No 19_054R). The data are provided by patients and collected by the NHS as part of their care and support. The Office for National Statistics (ONS) is the provider of the ONS Data contained within the CPRD Data. Hospital Episode Data and the ONS Data Copyright © (2014), are re-used with the permission of The Health & Social Care Information Centre. All rights reserved. The interpretation and conclusions contained in this study are those of the authors alone. There are no conflicts of interest among the authors.

6.8 References

1. NICE recommends wider use of statins for prevention of CVD | News and features | News | NICE. https://www.nice.org.uk/news/article/nice-recommends-wider-use-of-statins-for-prevention-of-cvd. Accessed April 30, 2018.

2. Glorot X, Bengio Y. *Understanding the Difficulty of Training Deep Feedforward Neural Networks*. http://www.iro.umontreal. Accessed January 19, 2020.

3. FDA permits marketing of artificial intelligence-based device to detect certain diabetes-related eye problems | FDA. https://www.fda.gov/news-events/press-announcements/fda-permits-marketing-artificial-intelligence-based-device-detect-certain-diabetes-related-eye. Accessed January 19, 2020.

4. Yan Li, Matthew Sperrin, Miguel Belmonte, Alexander Pate, Darren M Ashcroft TP van S. Do population-level risk prediction models that use routinely collected health data reliably predict individual risks? *Submitted(TBD)*. 2019.

5. Pate A, Emsley R, Ashcroft DM, Brown B, van Staa T. The uncertainty with using risk prediction models for individual decision making: an exemplar cohort study examining the prediction of cardiovascular disease in English primary care. *BMC Med*. 2019;17(1):134. doi:10.1186/s12916-019-1368-8

6. Yan Li, Matthew Sperrin, Darren M Ashcroft TP van S. Does machine learning improve the accuracy of clinical risk predictions? An exemplar examining risk of cardiovascular disease.

7. Pencina MJ, D'Agostino RB, Song L. Quantifying discrimination of Framingham risk functions with different survival C statistics. *Stat Med*. 2012;31(15):1543-1553. doi:10.1002/sim.4508

8. Steyerberg EW. *Clinical Prediction Models : A Practical Approach to Development, Validation, and Updating*. Springer; 2009.

9. Bitton A, Gaziano TA. The Framingham Heart Study's impact on global risk assessment. *Prog Cardiovasc Dis*. 2010;53(1):68-78. doi:10.1016/j.pcad.2010.04.001

10. Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. *Bmj*. 2017;2099(May):j2099. doi:10.1136/bmj.j2099

11. Piepoli MF, Hoes AW, Agewall S, et al. 2016 European Guidelines on cardiovascular disease prevention in clinical practice. *Eur Heart J*. 2016;37(29):2315-2381. doi:10.1093/eurheartj/ehw106

12. Steele AJ, Denaxas SC, Shah AD, Hemingway H, Luscombe NM. Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease. 2018. doi:10.1371/journal.pone.0202344

13. Alaa AM, Bolton T, Di Angelantonio E, Rudd JHF, van der Schaar M. Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants. Aalto-Setala K, ed. *PLoS One*. 2019;14(5):e0213653. doi:10.1371/journal.pone.0213653

14. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? Liu B, ed. *PLoS One*. 2017;12(4):e0174944. doi:10.1371/journal.pone.0174944

15. Al'Aref SJ, Anchouche K, Singh G, et al. Clinical applications of machine learning in

cardiovascular disease and its relevance to cardiac imaging. *Eur Heart J*. 2019;40(24):1975-1986. doi:10.1093/eurheartj/ehy404

16. Herrett E, Gallagher AM, Bhaskaran K, et al. Data Resource Profile: Clinical Practice Research Datalink (CPRD). *Int J Epidemiol*. 2015;44(3):827-836. doi:10.1093/ije/dyv098

17. Clinical Practice Research Datalink | CPRD. https://www.cprd.com/. Accessed November 28, 2019.

18. Hippisley-Cox J, Coupland C, Brindle P. The performance of seven QPrediction risk scores in an independent external sample of patients from general practice: a validation study. *BMJ Open*. 2014;4(8):e005809. doi:10.1136/bmjopen-2014-005809

19. Hill NR, Ayoubkhani D, McEwan P, et al. Predicting atrial fibrillation in primary care using machine learning. *PLoS One*. 2019;14(11):e0224582. doi:10.1371/JOURNAL.PONE.0224582

20. Ford E, Rooney P, Oliver S, et al. Identifying undetected dementia in UK primary care patients: A retrospective case-control study comparing machine-learning and standard epidemiological approaches. *BMC Med Inform Decis Mak*. 2019;19(1). doi:10.1186/s12911-019-0991-9

21. Li Y, Sperrin M, Belmonte M, Pate A, Ashcroft DM, van Staa TP. Do population-level risk prediction models that use routinely collected health data reliably predict individual risks? *Sci Rep*. 2019;9(1):11222. doi:10.1038/s41598-019-47712-5

22. Li Y, Sperrin M, Martin GP, Ashcroft DM, van Staa TP. Examining the impact of data quality and completeness of electronic health records on predictions of patients' risks of cardiovascular disease. *Int J Med Inform*. November 2019:104033. doi:10.1016/j.ijmedinf.2019.104033

23. van Staa T-P, Gulliford M, Ng ES-W, Goldacre B, Smeeth L. Prediction of cardiovascular risk using Framingham, ASSIGN and QRISK2: how well do they predict individual rather than population risk? *PLoS One*. 2014;9(10):e106455. doi:10.1371/journal.pone.0106455

24. Anderson KM, Wilson PW, Odell PM, Kannel WB. An updated coronary risk profile. A statement for health professionals. *Circulation*. 1991;83(1):356-362. doi:10.1161/01.CIR.83.1.356

25. Nelder JA, Wedderburn RWM. Generalized Linear Models. *J R Stat Soc Ser A*. 1972;135(3):370. doi:10.2307/2344614

26. Breiman L. *RANDOM FORESTS*.; 2001. https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf. Accessed August 22, 2019.

27. Demuth H, De Jesús B. *Neural Network Design 2nd Edition*. https://hagan.okstate.edu/NNDesign.pdf. Accessed August 22, 2019.

28. Max Kuhn. The caret Package. http://topepo.github.io/caret/index.html. Accessed September 10, 2019.

29. Géron A. *Hands-on Machine Learning with Scikit-Learn and TensorFlow : Concepts, Tools, and Techniques to Build Intelligent Systems*.

30. About us — scikit-learn 0.21.3 documentation. https://scikit-learn.org/stable/about.html. Accessed September 10, 2019.

31. h2o. AutoML: Automatic Machine Learning — H2O 3.26.0.3 documentation.

http://docs.h2o.ai/h2o/latest-stable/h2o-docs/automl.html?highlight=automl. Accessed September 7, 2019.

32. The H2O Python Module — H2O documentation. http://docs.h2o.ai/h2o/latest-stable/h2o-py/docs/intro.html#what-is-h2o. Accessed September 10, 2019.

33. Mahmood SS, Levy D, Vasan RS, Wang TJ. The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective. *Lancet (London, England)*. 2014;383(9921):999-1008. doi:10.1016/S0140-6736(13)61752-3

34. Goff DC, Lloyd-Jones DM, Bennett G, et al. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: A report of the American college of cardiology/American heart association task force on practice guidelines. *J Am Coll Cardiol*. 2014;63(25 PART B):2935-2959. doi:10.1016/j.jacc.2013.11.005

35. Li Y, Sperrin M, van Staa T. R package "QRISK3": an unofficial research purposed implementation of ClinRisk's QRISK3 algorithm into R. *F1000Research*. 2019;8:2139. doi:10.12688/f1000research.21679.1

36. Collins GS, Altman DG. An independent and external validation of QRISK2 cardiovascular disease risk score: a prospective open cohort study. *BMJ*. 2010;340(July):c2442. doi:10.1136/bmj.c2442

37. Pate A, Emsley R, Ashcroft DM, Brown B, van Staa T. The uncertainty with using risk prediction models for individual decision making: an exemplar cohort study examining the prediction of cardiovascular disease in English primary care. *BMC Med*. 2019;17(1):134. doi:10.1186/s12916-019-1368-8

38. *Concerns about the Latest NICE Draft Guidance on Statins Introduction.*; 2014.

6.9 Supplementary Online Content

The instability of machine learning and statistical models in predicting individual patient risks: an approach to improve the clinical utility of these models

Published online XXX
Doi: XXXX

eTable 6.9.1: Performance indicators of ensembled machine learning and Cox models

eFigure 6.9.1. Boxplot of differences of individual risk prediction ranks and absolute risk with machine learning and Cox models stratified by 5 percentiles of predicted risks with local Cox model

e. **Boxplot of differences of individual risk predictions**
f. **Boxplot of differences of percentage of individual patients' rank**

X axis: decile of absolute predicted risk with local Cox model

Y axis: differences with other models

eFigure 6.9.2. Boxplot of differences of individual risk prediction ranks and absolute risk with machine learning and Cox models stratified by 20 percentiles of predicted risks with local Cox model

a. **Boxplot of differences of individual risk predictions**
b. **Boxplot of differences of percentage of individual patients' rank**

X axis: decile of absolute predicted risk with local Cox model

Y axis: differences with other models

eFigure 6.9.3. Boxplot of differences of individual risk prediction ranks and absolute risk with machine learning and Cox models stratified by 25 percentiles of predicted risks with local Cox model

a. **Boxplot of differences of individual risk predictions**
b. **Boxplot of differences of percentage of individual patients' rank**

X axis: decile of absolute predicted risk with local Cox model

Y axis: differences with other models

eTable 6.9.1 shows the model performance of the 12 ensembled machine learning models and three Cox models. All models had similar model performance in terms of high discrimination (C statistic about 0.88) and calibration (similar low brier score and good calibration plots in their own framework [6]).

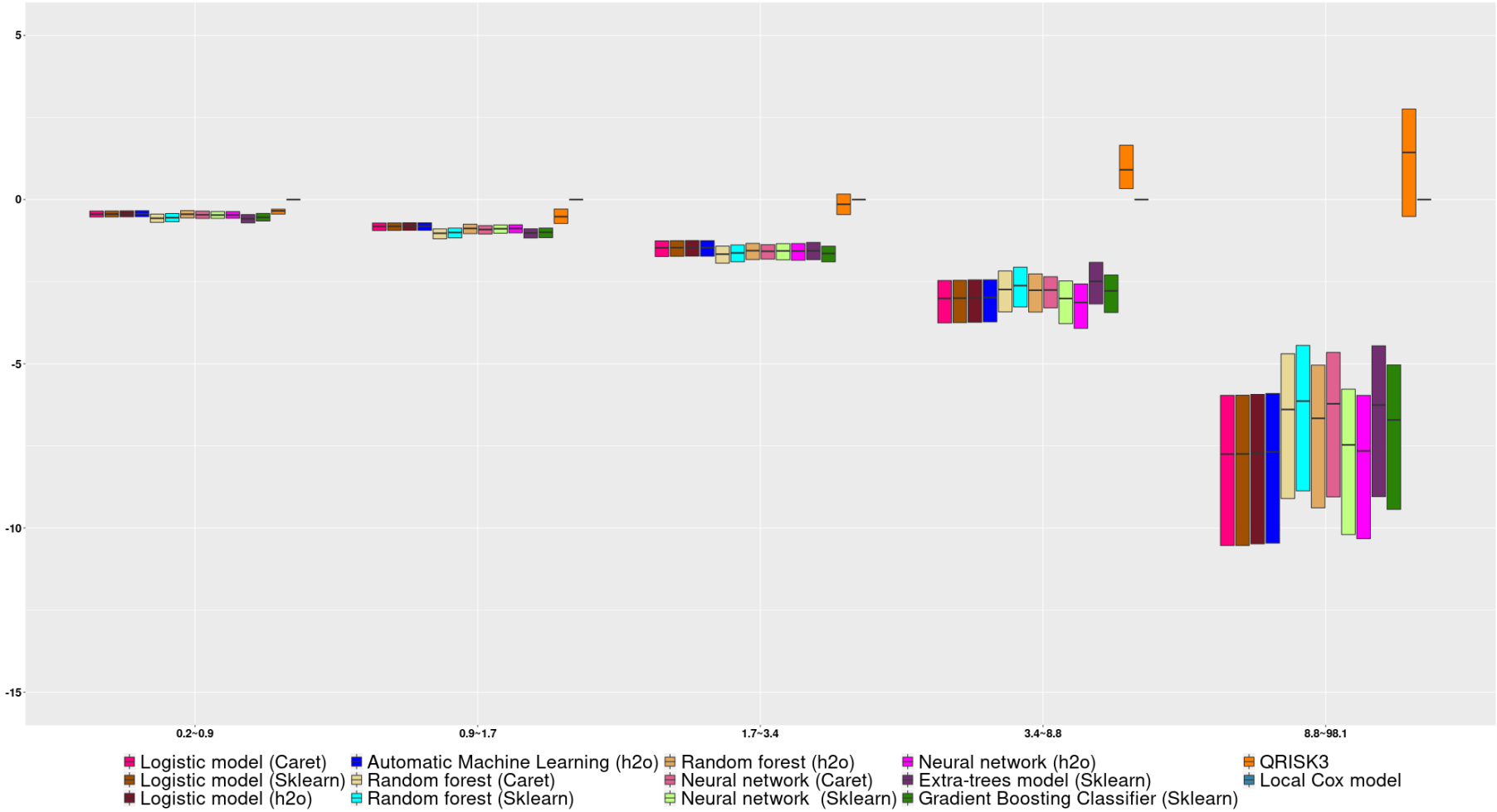
eFigure 6.9.1-6.9.3 plots similar boxplot as Figure3 in the main manuscript**.** **eFigure 6.9.1-6.9.3a** plot the boxplot of differences in magnitude of patients risks against the 5 percentiles, 20 percentiles and 25 percentiles of risk of local Cox model grouped by each model. Similarly, **eFigure 6.9.1-6.9.3b** plot the boxplot of differences of percentage of patient ranks against percentiles of risk of local Cox model. These plots show the similar results as in **Figure 6.3**, i.e. the predicted risk has larger variation or inconsistency in the higher risk group, but the predicted rank has smaller inconsistency in the higher risk group. The result of this trend was not influenced by whether we group patients by decile or percentiles of predicted risk.

eTable 6.9.1: Performance indicators of machine learning and Cox models

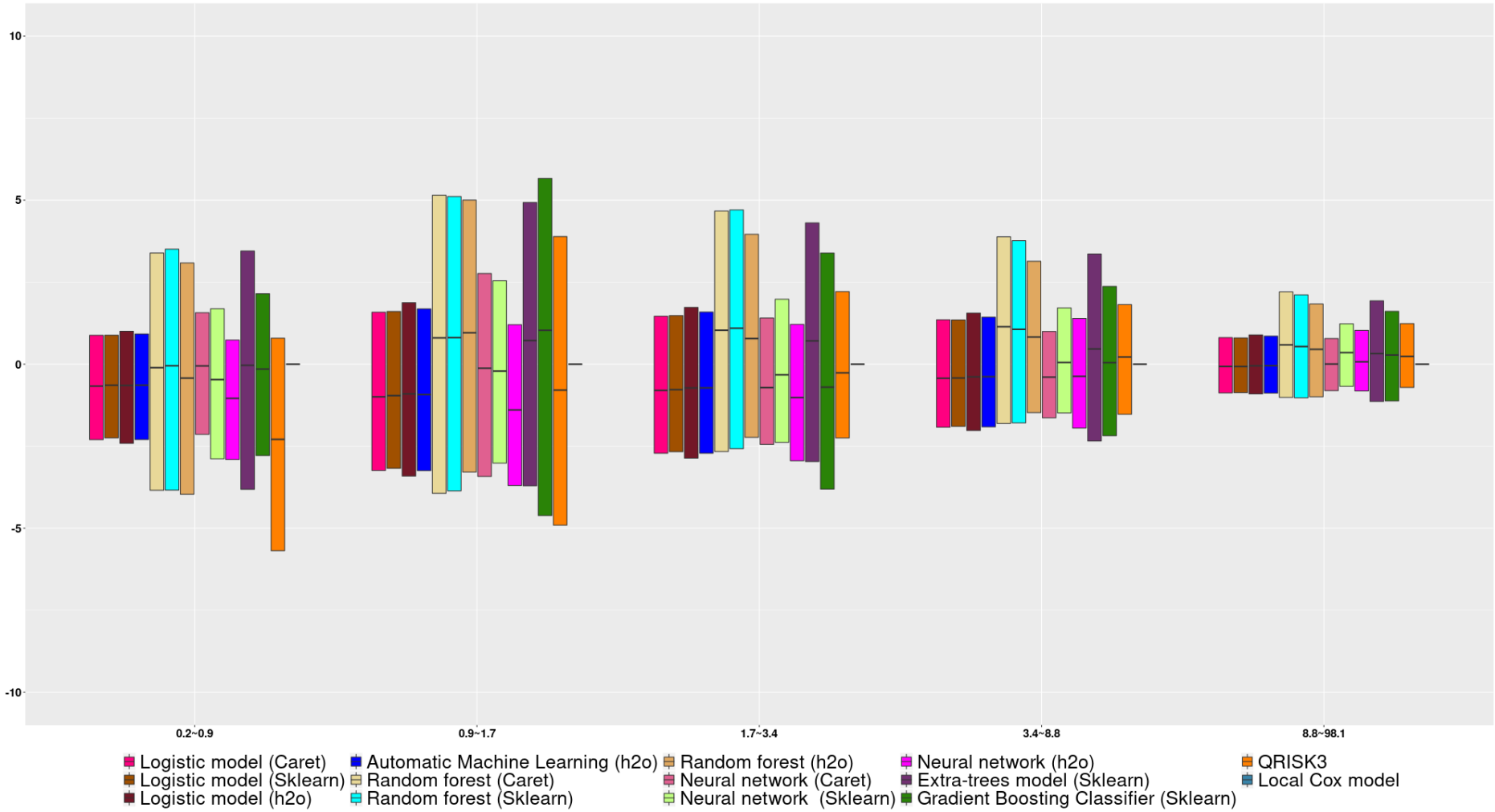
	Model performance* (95% range #)			
	C-statistic (2.5% ~ 97.5%) #	**Brier score (2.5% ~ 97.5%) #**	**Sensitivity (Recall) (2.5% ~ 97.5%) #**	**PPV (Precision) (2.5% ~ 97.5%) #**
Caret				
Logistic	0.879 (0.878, 0.881)	0.028	0.615	0.163
Random forest	0.879 (0.877, 0.880)	0.028	0.672	0.147
Neural network	0.880 (0.878, 0.881)	0.028	0.673	0.148
Cox model				
QRISK3	0.879 (0.878, 0.881)	0.032	0.858	0.101
Framingham	0.865 (0.863, 0.866)	0.031	0.892	0.085
Local Cox model	0.877 (0.876, 0.879)	0.032	0.810	0.112
Sklearn				
Logistic	0.879 (0.878, 0.881)	0.028	0.616	0.163
Random forest	0.879 (0.877, 0.881)	0.028	0.683	0.145
Neural network	0.877 (0.875, 0.878)	0.028	0.600	0.161
Gradient boosting classifier	0.881 (0.880, 0.883)	0.027	0.659	0.154
extra-trees	0.879 (0.877, 0.881)	0.028	0.679	0.146
h2o				
Logistic	0.879 (0.877, 0.880)	0.028	0.616	0.163
Random forest	0.879 (0.878, 0.881)	0.027	0.652	0.152
Neural network	0.879 (0.877, 0.880)	0.028	0.605	0.163
autoML	0.879 (0.878, 0.881)	0.028	0.617	0.163

*** Model performance was calculated in binaray framework. Threshold 7.5% was used to calculate precision and recall for all models.**

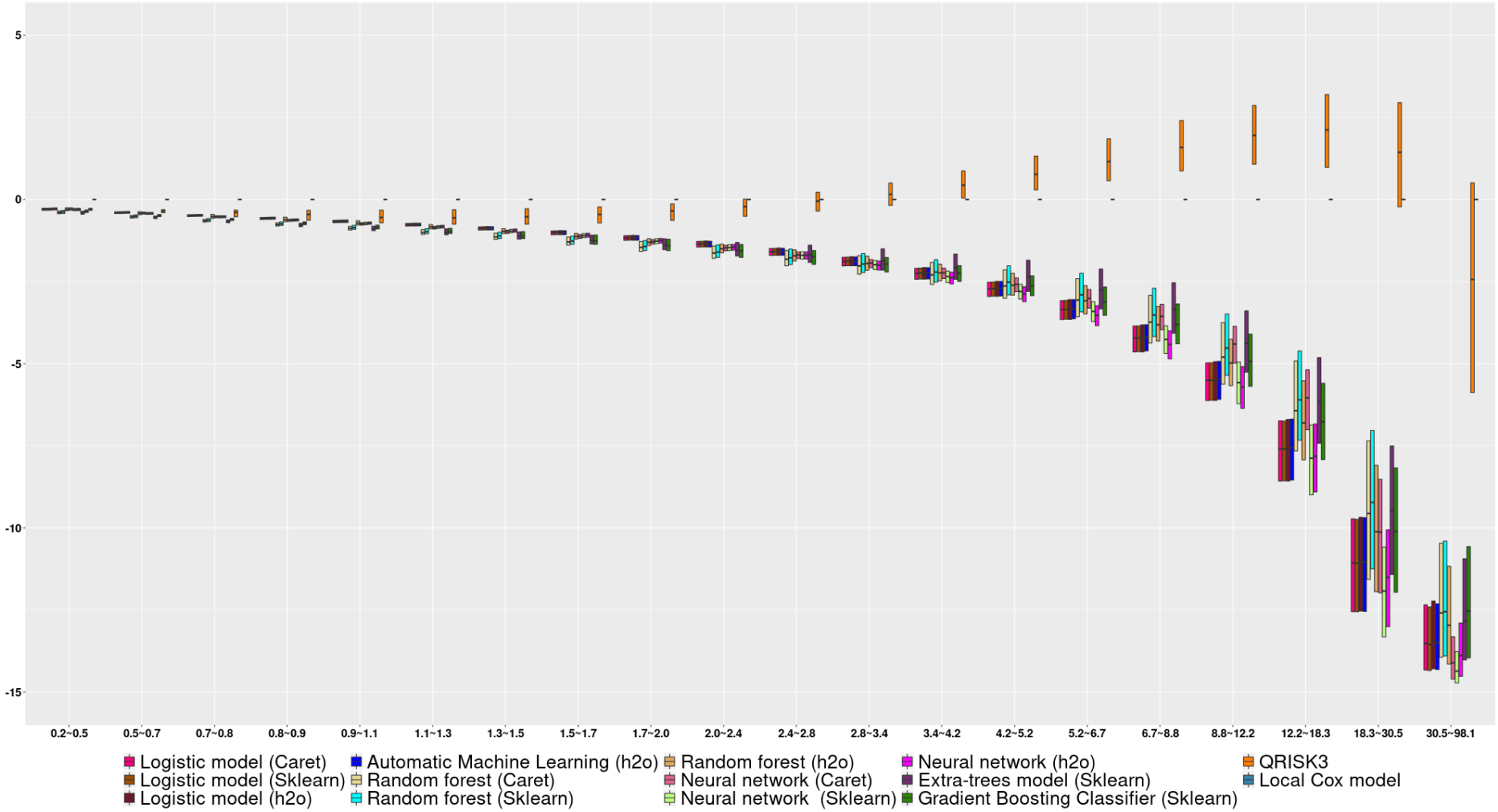
Appendix Figures



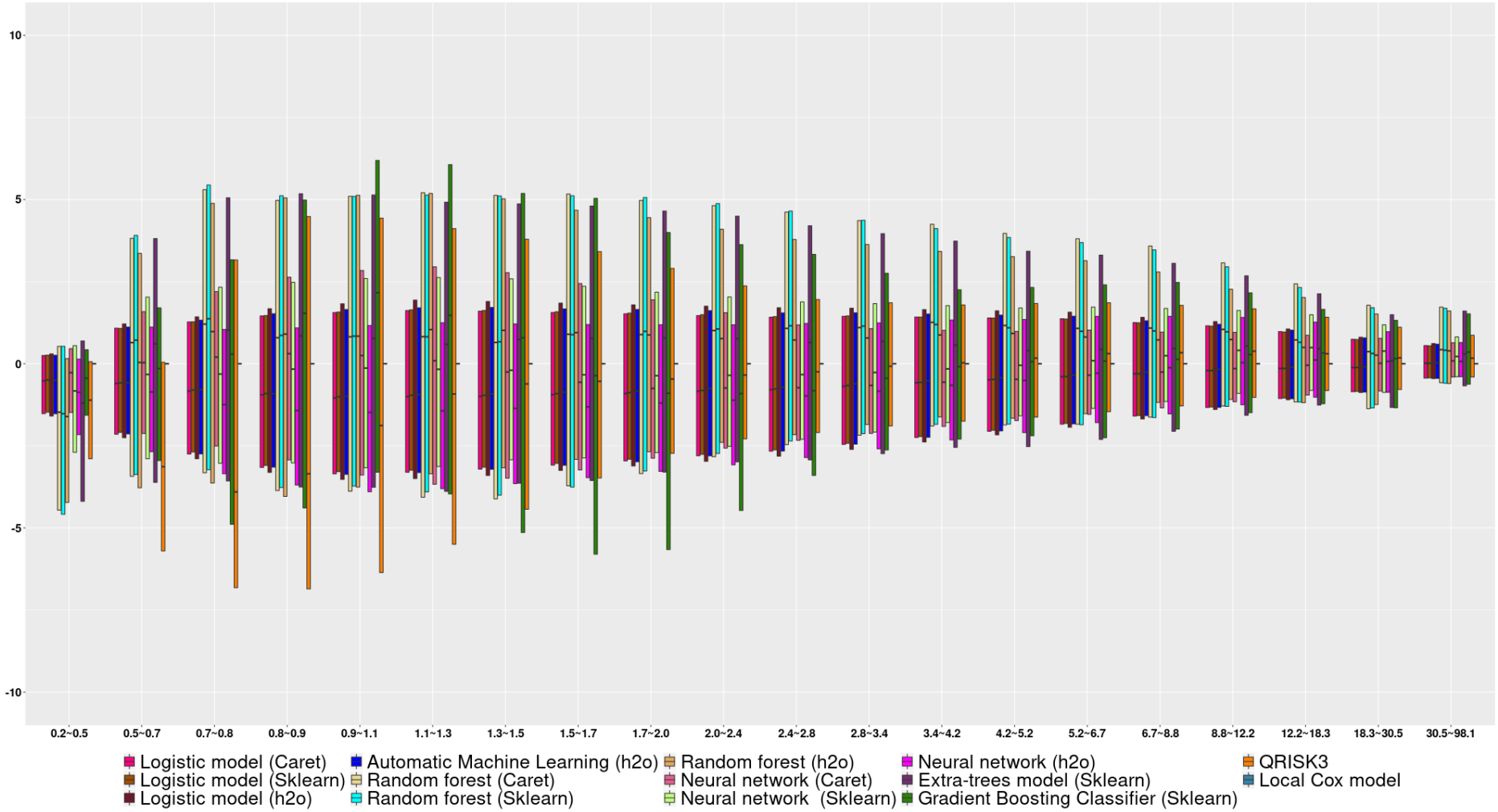
eFigure 6.9.1a



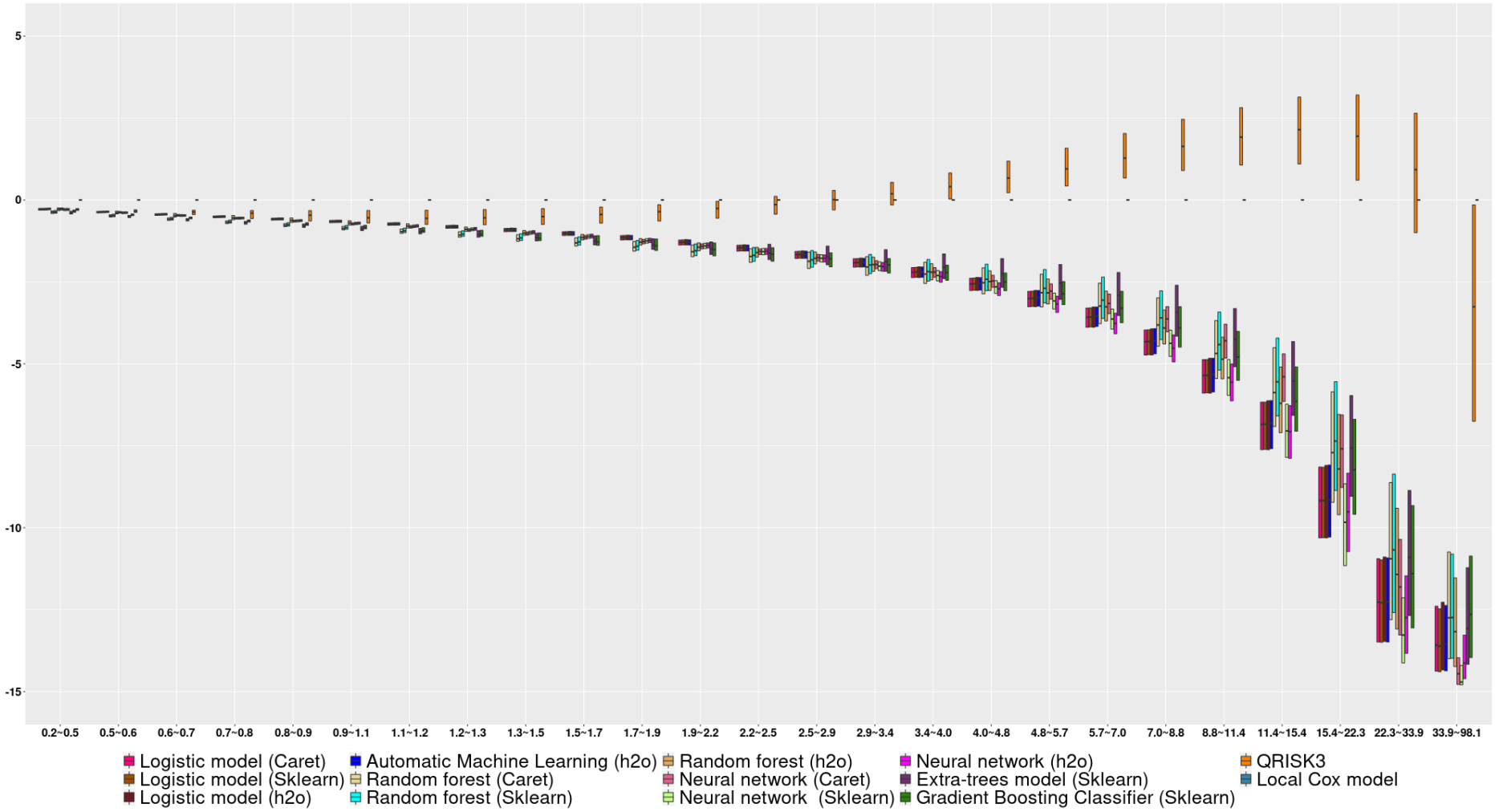
eFigure 6.9.1b



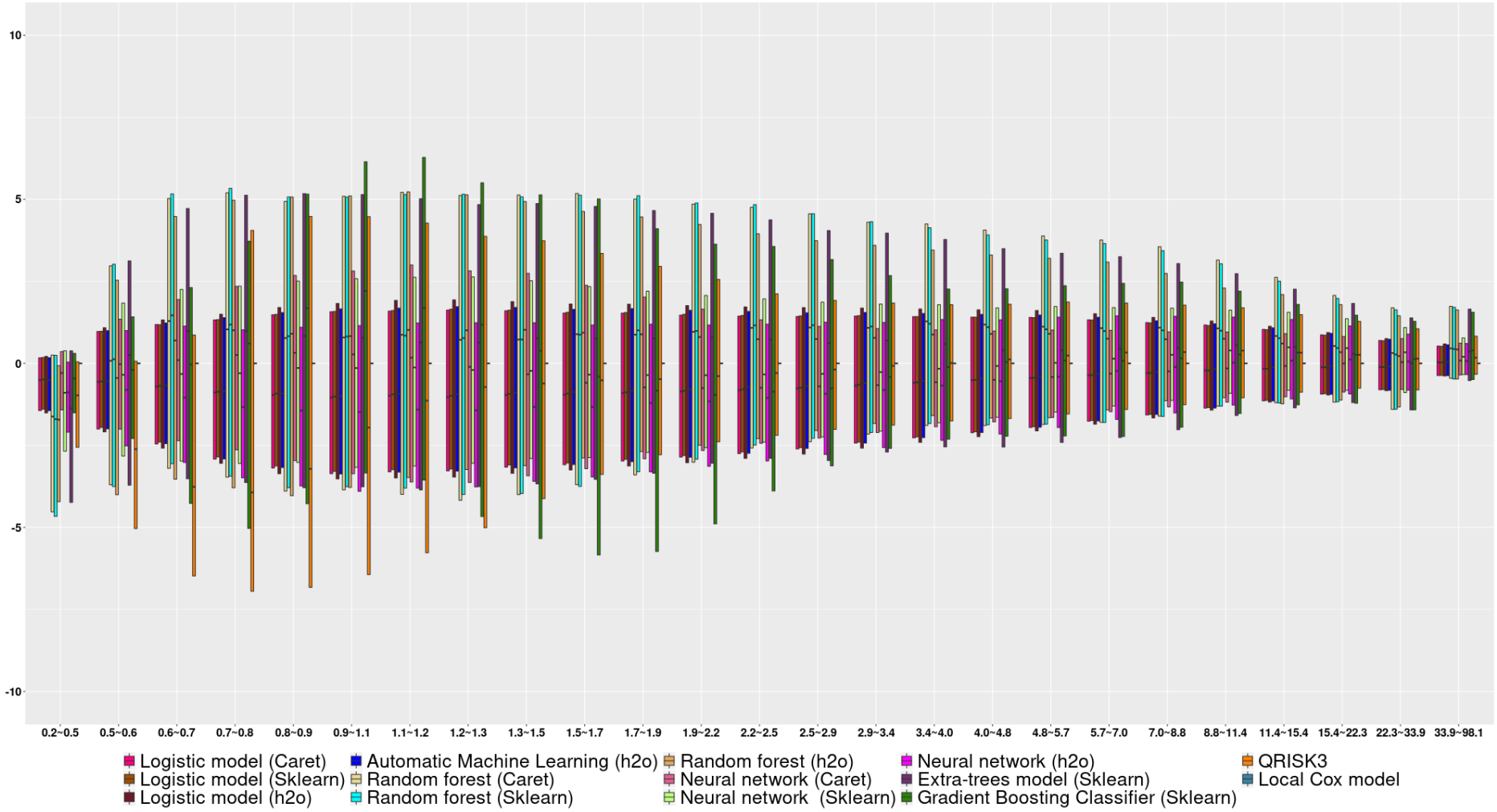
eFigure 6.9.2a



eFigure 6.9.2b



eFigure 6.9.3a



eFigure 6.9.3b

eAppendix 6.9.4 Statistical interpretation of uncertainty of individual risk

Current individual risk (probability) estimated by risk prediction model is pre-defined and conditioning on the predictors considered in the model, model assumptions and how model defines similarity of patients based on all selected predictors, i.e. P(outcome | considered conditions) [1]. However, a clinical robust individual risk should be defined as a risk based on all conditions including considered conditions and unknown conditions, i.e. P(outcome | All conditions) or P(outcome | considered conditions + unknown conditions). All of individual risk prediction of current risk prediction models use P(outcome | considered conditions) to approximate P(outcome | all conditions) for clinical usage. Ideally, the statistical estimated risk should be close to the risk which are robust enough for clinical usage on individual. This means the considered conditions already capture the main variation of disease outcome and unknown conditions would have small effects on estimated probability (i.e. adding new conditions would not change probability much and would not affect clinical decision of patients [1]). Previous studies have shown that CVD risk prediction could be changed completely (affects clinical decision of patients) if additional unknown conditions are considered including practice heterogeneity [2], model design of choice [3] and different types of model [4]. This indicates that the P(CVD | considered conditions) is not close to P(CVD | All conditions), and using different P(CVD | considered specific conditions) as P(CVD | All conditions) would result great uncertainty of individual risk of patients. This then suggests new individual level measurement might be needed addition to individual risk.

6.9.5 References

1. BRIGGS, W. *UNCERTAINTY : the soul of modeling, probability & statistics.* (SPRINGER, 2018).

2. Li, Y. *et al.* Do population-level risk prediction models that use routinely collected health data reliably predict individual risks? *Sci. Rep.* **9**, 11222 (2019).

3. Pate, A., Emsley, R., Ashcroft, D. M., Brown, B. & van Staa, T. The uncertainty with using risk prediction models for individual decision making: an exemplar cohort study examining the prediction of cardiovascular disease in English primary care. *BMC Med.* **17**, 134 (2019).

4. Yan Li, Matthew Sperrin, Darren M Ashcroft, T. P. van S. Does machine learning improve the accuracy of clinical risk predictions? An exemplar examining risk of cardiovascular disease.

Blank page

Chapter 7 Clinical risk prediction model using routinely collected electronic health records or longitudinal cohort in daily practice: Are they robust enough for clinical decision making?

Yan Li[1], Matthew Sperrin[1], Darren M Ashcroft[2,3], Tjeerd Pieter van Staa[1,4,5,7]

[1]Health e-Research Centre, School of Health Sciences, Faculty of Biology, Medicine and Health, the University of Manchester, Manchester, Oxford Road, Manchester, M13 9PL, UK
[2]Centre for Pharmacoepidemiology and Drug Safety, School of Health Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Oxford Road, Manchester, M13 9PL, UK
[3]NIHR Greater Manchester Patient Safety Translational Research Centre, School of Health Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Oxford Road, Manchester, M13 9PL, UK
[4]Utrecht Institute for Pharmaceutical Sciences, Utrecht University, Utrecht, Netherlands
[5]Alan Turing Institute, Headquartered at the British Library, London, UK

Corresponding author: Tjeerd van Staa, tjeerd.vanstaa@manchester.ac.uk

Journal title: Close to submit
Doi:
License:

Word count: 2796
Abstract: Not available due to paper type
Number of tables: 2
Number of figures: 1

7.1 Key message

Risk prediction models that may have good statistical performance on population level (such as C statistics) can have limited generalisability and clinical utility in predicting individual patient risks. The reason is that prediction models based on different techniques or modelling decisions can yield inconsistent individual results. Risk prediction models should be used in conjunction with additional clinical tests and clinical judgement.

7.2 Brief history of risk prediction models

Risk prediction models are mathematical formulas that use predictors such as risk factors of disease to calculate the risk (probability) of individual patients developing a disease in the future [1]. Historically, they were developed from large prospective cohorts with pre-defined inclusion criterion, outcomes and predictors (e.g. Framingham study [2]). More recently, risk prediction models are also being developed from routinely collected electronic health records (EHR) due to their advantage in large sample size, more available predictors, multiple time points, frequently updated data and better representativeness due to inclusion of patients from daily practice rather than volunteers in prospective cohort studies [3]. Examples of risk prediction models are the UK QRISK models [3] predicting cardiovascular disease (CVD) [4], diabetes mellitus [5] and fracture [6], US Framingham model [2] predicting CVD, European ESC score predicting CVD and Michael et al.'s model [7] predicting acute kidney injury. These models are now recommended by clinical guidelines for disease prevention [8] and implemented into healthcare system to assist clinicians to make treatment decision for individual patients. For example, NICE guideline recommends to prescribe a statin to patients if they have QRISK predicted CVD risk of over 10% [9], ESC score is used in European guidelines on CVD prevention in clinical practice [10] and Framingham model is recommended by PBS guideline [11]. These models were transparently reported following the model development guideline of TRIPOD [1]. Recently, machine learning models have started to show their strength in image

recognition [12] due to increase of computing power and breakthrough in deep learning [13]. Multiple studies reported that machine learning models also provided promising predictions comparing to traditional statistical models thus should be considered for clinical practices [14] [15]. This paper aims to analyse whether risk prediction models as developed from EHR or prospective cohort studies are robust enough for clinical decision making for individual patients.

7.3 User case

Mr. Jonathan walked into Doctor Nice's practice. "With the information such as your age, gender, whether you smoke or any relatives have heart attack before, we could predict your risk having CVD in next 10 years". Says Doctor Nice. "Look, your risk was 7.5% predicted by a logistic model, 17% according to QRISK3, 12% according to Framingham model, 9.5% according to Cambridge's Autoprognosis, 6.5% according to a neural network model and 9% according to another neural network model..." "Wait what? Which one should I believe?", Mr. Jonathan was confused…

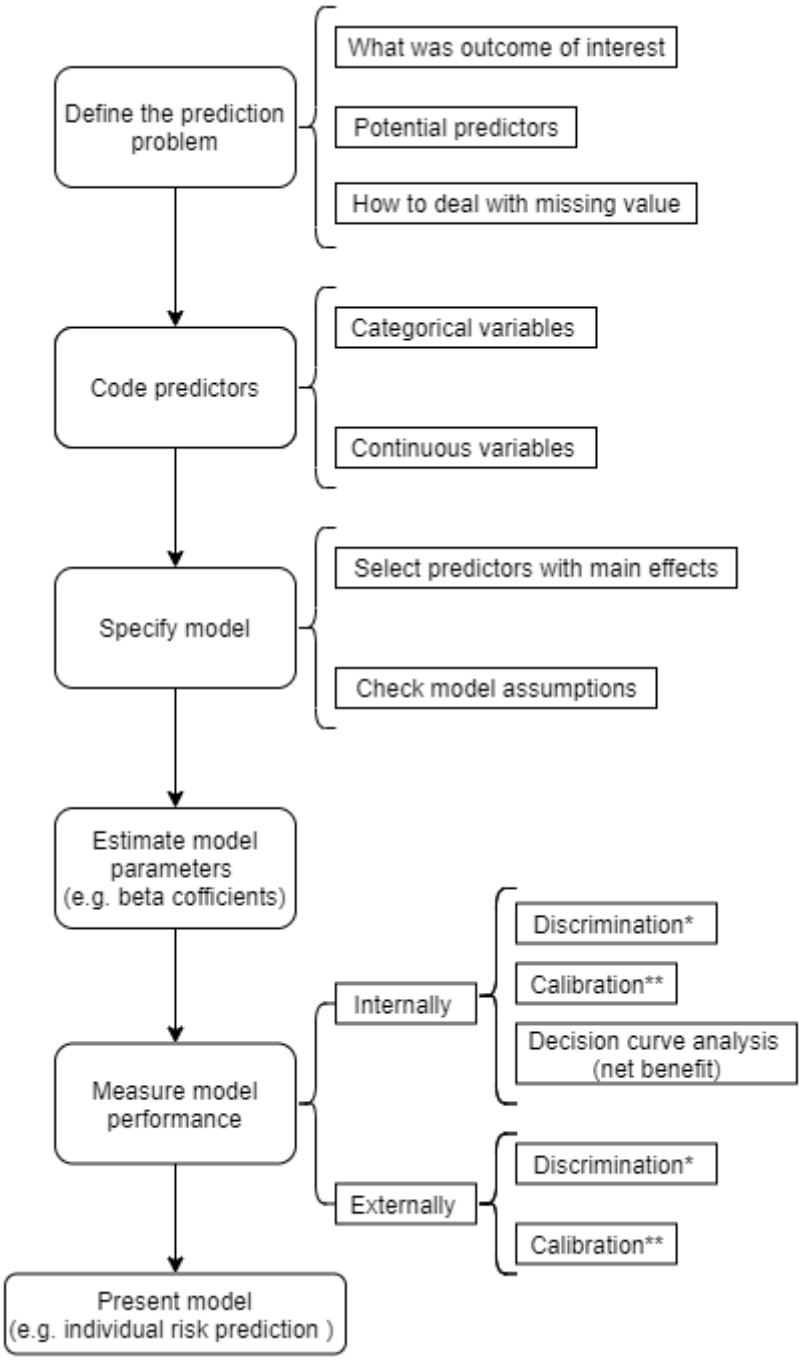
7.4 Current model fitting process

Several steps are required to develop a statistically validated model [16] (Figure 7.1). The first step to consider is the prediction problem, i.e. how to define the outcome of interest, potential predictors and how to deal with any missing values. The next step to consider is how to code predictors (including categorical variables and continuous variables) to best capture predictive information for model. Then one needs to consider how to specify a model, i.e. select predictors with main effects (causal predictors) and check model assumptions. With specified model formula, one could estimate model parameters (e.g. beta coefficients or intercept). Once obtained the fitted model, one needs to measure its model performance (i.e. discrimination and calibration) and clinical usefulness (net benefit analysis) internally (within the model

development setting) and ideally externally (external setting outside the model development sitting). The final step is to present the model results of individual risk predictions [16]. Machine learning models are generally similar to statistical model in the model fitting process except that machine learning models have more complex model structure with a higher focus on recognising data pattern, tuning model with hyperparameters (control model structure and model fitting criteria) and automation (automate model developing, validating and updating process) [17].

The performance of risk prediction models is assessed with discrimination measures (i.e., the ability to discriminate between high/low risk patients) and calibration measures (i.e., agreement between average predicted events and observed events). There are internationally accepted guidelines for the development and validation of risk prediction models (TRIPOD [1]). These guidelines include determine whether the study develop and/or validate model, identify target population, clarify the outcome of interest, definition of outcome and any blind procedure to measure outcome, provide drivers of developing a new model, references to existing model, describe source of data and study design, specify key dates of start, end and follow-up of the study, specify study setting such as primary or secondary care and number and location of centres, describe inclusion criteria of patients, define considered predictors and any blind procedure when collecting predictors and how predictors were used in analysis, report how the sample size was derived, how missing data was handled, explain what type of model was used and process of model fitting and internal validation process, select measures to assess model performance and compare to multiple models if possible, report any model updating procedure, describe whether and how risk groups were created if relevant, describe general differences of data such as setting or criteria between development and validation, report baseline information, show a distribution comparison between development and validation for important variables, present model to predict individual patients risk and explain how to use the model, discuss the limitation, discuss the differences of model performance between

development cohort and validation, give an overall interpretation of model, discuss the clinical usage of model, provide supplementary resources for model and provide funder of the study.



***:** Discrimination measures the ability of model to discriminate high/low risk patients
****:** Calibration measures the agreement between average predicted events and observed events

Figure 7.1 model fitting and validating process [16]

7.5 Challenges of current risk prediction approach

Table 7.1 identifies key challenges with developing risk prediction models that are currently ether not or minimally included in the TRIPOD guidelines. These include unmeasured heterogeneity between different clinical sites (i.e., predicted risks were affected by heterogeneity between clinical sites), variation of the associations between predictors and outcome, data quality, causal factors not included in the model, choice of model technique, design choices and settings within the model.

Table 7.1: Challenges in developing risk prediction models

Challenges	Examples
Some causal factors not included in the model	A study found that CVD risk prediction models incorporating different predictors such as secular trend have similar model performance but predicted risks for same patients differently [21]. Another study found that adding a new predictor (C-Reactive Protein) to CVD models did not improve model performance but the change of individual risk prediction improved the risk classification in women [22]. A recent study reported that machine learning model outperformed traditional statistic model incorporating all the possible predictors (causal and non-causal) from database into model [23].
Unmeasured heterogeneity between clinical sites	A study found that QRISK3 has considerable uncertainty in predicting individual CVD risks. A patient with a predicted 10-year risk of 10% could have predicted risks ranging from 7.2% to 13.7% when incorporating random effects of each site into the model. This random effects model (of intercepts) found considerable unmeasured site heterogeneity in CVD risk leading to over- and under-treatment depending on the site [18].
Data quality	Models developed from data with poor data quality generally have poor generalisability [16]. However, a study found data quality did not explain the effects of unmeasured practice heterogeneity on individual risk prediction [19]. In the study, data quality was measured by stability metrics of predictors and missingness for each practice which quantifies the distance/diversity of distribution of predictors of each practice to a latent average. Study found practices with high/low random effects have comparable stability metrics and missing percentages, vice versa. This indicates data quality was not related to effects of practice heterogeneity on individual risk prediction.
Variation of association between predictors and outcome	A study pointed out that the heterogeneity of model performance in different practices could reflect true variations between predictors and outcome [20]. A recent study, using a random effect model, found that variation of association between predictors and outcome did not explain the effects of unmeasured practice heterogeneity on individual risk prediction [19]. This random effect model (of slopes) tested whether the relative rates of outcomes with predictors statistically varied between sites [19].
Choice of model type and structure and underlying model assumptions	Studies have shown that different model choices [21] and types [24] can yield models with comparable statistical population-level performance but inconsistent prediction of individual risks. This includes machine learning models that predicted individual risks differently (even for models developed from the same algorithms and structure) [24].

Risk prediction models typically use data collected from different clinical sites but the variability in the accuracy of predictions between sites is often not considered. As an example, QRISK3 was developed based on data from 1309 general practices. A recent study found that this site variability substantially affected the estimates of risks for individual patients and potential treatment decisions. A model that incorporated site variability by estimating random effects of each site using random effects model found that a QRISK3 predicted risk of 10% of developing CVD over 10 years changed to predicted risks ranging between 7.2% and 13.7% in the random effects model [18].

Data quality, which is a particular challenge for EHRs as data are routinely being collected as part of clinical care rather than research, is an important aspect needs to be considered by risk prediction modelling. Models developed from poor quality data generally have limited generalisability [16]. There are few studies that have systematically assessed the effects of data quality on risk predictions; data quality is rarely considered in the development of risk prediction models (most studies only handled missing value). However, a recent study found that data quality did not explain the effects of unmeasured practice heterogeneity when missing values were imputed and distribution of predictors were similarly among different settings [19].
The associations between predictors and outcome could also vary between clinical sites (for example, the effects of diabetes mellitus on CVD risk could be different between sites). One study reported that this was responsible for the heterogeneity of model performance between clinical sites [20]. Another study found that the association between predictors and outcome did not explain the effects of unmeasured practice heterogeneity (using a random effect with random slope model) [19].
Whether models captured all the main predictors, what are the effects of unmeasured predictors and what predictors should/should not include in model is another challenge for risk prediction modelling. One study found that incorporating different predictors such as secular trend end up with models with similar model performance but completely different individual risk prediction [21]. Another study added a new

predictor (C-Reactive Protein) to the model and found that model performance was similar after adding the predictor but that the risk classification in women improved due to the change of individual risk prediction [22]. A recent study claimed that machine learning model outperformed traditional statistic model with almost all the possible predictors from dataset [23], but this "all in" approach strongly limits clinical utility (needs more predictors to make predictions) and interpretability (no clue how predictions were made) of models and makes model more prone to issues such as data quality issue and overfitting.

Type of both statistical models and machine learning models was related to prediction task and whether model could achieve a good model performance, but there is only empirical way (no general principle) to decide which model, what model structure was the most suitable for specific prediction task and whether all model assumptions were met. Recent research mimicked the practical model developing process by comparing 12 family of machine learning models to 3 family of Cox models considering different representative samples, different model fitting software, model types, hyper-parameters to control model structures. The study found all models have similar model performance but predicts the same patients differently even for models developed from the same algorithm, thus the treatment decision of patients was decided by what model was used in clinical practice [24].

Inconsistency of predicted risks from statistical comparable models for the same patients directly limits clinical utility of risk prediction models, as the above story goes that patients and clinicians would have no clue which predicted risk should be used. Studies have shown that models with similar model performance but differ in whether consider practice heterogeneity [18], model choices [21] and model types [24] predict inconsistent risks for the same patients.

7.6 Solutions

The ultimate goal is to improve the generalisability and clinical utility of current risk prediction models. Table 7.2 summarises potential specific solutions including

assessing effects of practice heterogeneity with random effects model, developing new metrics to measure effects of unconsidered predictors from application environment, dealing with missing value using multiple imputation and assessing data quality by comparing distribution of predictors, assessing variation of association between predictors and outcome with random slope model, continuing exploring more predictors and developing standard to incorporate predictors into model, improving current guideline to standardise model fitting process, developing new statistic model performance measurements to measure model performance on individual level and new individual level measurement for patients and use current model with a combination of clinical testing and clinical judgement.

Table 7.2: Possible solutions in overcoming challenges in risk prediction models

Challenges	Solutions
Some causal factors not included in the model	Unmeasured effects of predictors could be partially assessed by random effects model. New systematic statistical approach incorporating current available statistics such as R-square was needed to assess whether models captured all the main predictors and effects of unmeasured predictors on risk prediction. Predictors used in the current implemented risk prediction model were mainly selected by experts due to their likely causal relationship to outcome and large improvement of model performance [16]. Though incorporating large number of non-casual predictors into model provides more predictive information, it limits clinical utility (needs more predictors to make predictions) and interpretability (no clue how predictions were made) of model and makes model more prone to issues such as data quality issue and overfitting. What predictors to include/not include in risk prediction model should be carefully discussed by experts and systematically studied. New principle should be added in the current risk prediction model guideline about what predictors could add considering causal relationship, model performance, generalisability and clinical utility.
Unmeasured heterogeneity between clinical sites	Assess effects of practice heterogeneity with random effects model. Alternatively using fixed effects approach with linear predictor of model as offset and practices as categorical variable to compare the fixed effects of each practice [16]. Develop new metrics to measure clinical utility of models for environment of clinical application [29].
Data quality	Deal with missing value using multiple imputation [25]. Indirectly measure effects of data quality by comparing distribution of predictors with traditional statistical approach such as N-way ANOVA or innovative stability metric [26] in different settings (e.g. in development setting or validation setting or in different practices).
Variation of association between predictors and outcome	Assess variation of association between predictors and outcome with random slope model.

Challenges	**Solutions**
Choice of model type and structure and underlying model assumptions	Statistical model field should refer to machine learning model field to create open access model collections (model encyclopaedia) [31] about what model was used for specific diseases with transparent records of model type, structures, model assumptions, predictors, outcomes and test data, so the consensus of risk prediction modelling could be established for each specific prediction task. Machine learning field should refer to current statistical model guideline focusing on reaching consensus of choosing model type and how hyper-parameters and model structure should be considered in a principal way rather than the current ad-hoc approach.
Inconsistency of predicted risk estimated by statistical comparable models	New statistic is needed to measure model performance on individual level. Additional individual level measurement such as percentage of rank [30] and other individual level metrics could be considered to improve clinical utility of risk prediction model. Treatment decisions for patients should be recommended by risk predictions in conjunction with clinical tests and clinical judgement.

As done previously [18], models developed from homogenous setting may not be generalisable to heterogeneity setting (i.e. treatment decision of patients would be influenced), so the effects of practice heterogeneity should first be assessed with random effects model approach. Afterwards, model should be updated with model updating approaches for multiple settings (practices) [16]. Besides of practice heterogeneity, future study should continue exploring effects of unmeasured predictors from clinical application environment and new metrics were needed to assess clinical utility of models in these application environments.

Data quality is always an issue of any task using data. Missing value could be dealt with multiple imputations [25]. Other effects of data quality could be assessed by comparing distribution of predictors using statistical appoches such as stability metric [26] among different settings, as the effects of data quality would reflect on distribution of predictors.

Variation of association between predictors and outcome may exist in development, validation and application setting. Random slope model and model updating approach (i.e. re-calibrate in new setting) [16] could be used to assess and deal with the miss-calibration problem.

Risk predictors are usually acquired by consulting the expert of the disease and decided by their possible causal relationship to the disease [16]. Currently, there is no general principal or consensus of how many and what predictors should be included. Including too many predictors could enrich much information that model could use and increase its model performance but limit its clinical utility as more predictors needs to be measured before making prediction and more predictors means higher probability of missing value. Also, the improvement of model performance might be merely due to higher variation of predictors rather than their predictive effects. New predictors need to be assessed carefully whether they have potential causal relationship with disease of outcome before adding them into model, otherwise noise predictors like missingness of BMI could be found predictive to the model [27]. What predictors to include/not include in risk prediction model should be carefully

discussed by experts and systematically studied.

For a given prediction task, different approaches including statistical model and machine learning model could be considered for model fitting. However, there is a lack of consensus which model was most suitable for a specific scenario. As stated in study, treatment decision of patients was depend on which model was used [18]. An open access model collection (model encyclopaedia) with transparent records of model type, structures, model assumptions, predictors, outcomes and test data should be created to establish consensus for each specific prediction task. Machine learning model field should develop similar guideline as statistical model guideline focusing on control model type and structure with a principal way rather than ad-hoc. Currently, it is hard to distinguish which model is better if only use discrimination and calibration alone as model performance measurements. New statistics should be proposed to measure performance of model on individual level including effects of unmeasured predictors on individual risk prediction. One example is random effects [28] could measure unconsidered effects from practice heterogeneity, as if there is no practice heterogeneity (all practices were similar to each other) there would be no random effects. This proposal supports the suggestion that performance metrics of risk prediction model should reflect the clinical application [29], as the applied clinical setting could have conditions that model did not consider in development such as practice heterogeneity. If such new statistic was proposed and could distinguish performance of model on individual level, then models with best individual performance should be used.

Individual risk prediction (probability) was limited as it was defined on population level, so alternative individual level measurement could be considered. A recent study found that though higher risk patients have larger inconsistency among statistical comparable models than lower risk patients, they were ranked more similarly comparing to lower risk group [30]. Therefore, individual risk prediction supplying with individual rank could improve clinical utility of risk prediction model as patients with higher risk in one model could be ranked similar high in other models

which provide more confidence this was a truly high-risk patient. Other individual level measurement such as prediction score [16] may be further studied.

Though current risk prediction models could have more certainty on higher risk patients with individual risk and rank, they still have uncertainty on patients with medium risk (those who were near the threshold). Rather than use risk predictions alone to make treatment decision for patients, integrate multiple statistical validated risk prediction models into a large clinical decision system in combination with lab testing and clinical judgment might be preferred. Clinical decision of individual patients should be made on in a conjunction with individual risk prediction, additional lab tests and clinical judgment.

Overall, risk prediction models based on current guideline could have good performance on population level but with limited generalisability and clinical utility especially on individual level. Future guideline should consider reporting whether model captured all the main causal predictors, the effects of unmeasured causal predictors, consistency of estimated probability from statistical comparable models on the same individual patients and reasons of these inconsistencies. Future research could consider new statistic to measure model performance on individual level, new individual level measurement, adding more causal predictors and using model in conjunction with additional lab test and clinical judgement.

7.7 Use case ending

7.7.1 Ending1.

“Well, though these models have inconsistency in your risk prediction, it appears that all of them rank you as the top 2% high risk patients among the overall patients, so now we have confidence that you need a statin.” Says Dr. Nice. “Now I understand”. Mr. Jonathan seems to be relieved.

7.7.2 Ending2.

“You made your point, Mr. Jonathan, these models also rank you differently

among all patients. In this case we need to run more clinical tests.” Says Dr. Nice. After a while, “According to results from tests, your healthy conditions and my experience on patients with similar condition as you, it would be better if you take a statin.” Says Dr. Nice. Mr. “Now I understand” Jonathan replies.

7.8 References

1. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. M. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD Statement. *Eur. Urol.* **67**, 1142–1151 (2015).
2. Wilson, P. W. F. *et al.* Prediction of Coronary Heart Disease Using Risk Factor Categories. *Circulation* **97**, 1837–1847 (1998).
3. Goldstein, B. A., Navar, A. M., Pencina, M. J. & Ioannidis, J. P. A. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *J. Am. Med. Informatics Assoc.* **24**, 198–208 (2017).
4. Hippisley-Cox, J., Coupland, C. & Brindle, P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. *BMJ* **357**, j2099 (2017).
5. Hippisley-Cox, J. & Coupland, C. Development and validation of QDiabetes-2018 risk prediction algorithm to estimate future risk of type 2 diabetes: cohort study. *BMJ* **359**, j5019 (2017).
6. Derivation and validation of updated QFracture algorithm to predict risk of osteoporotic fracture in primary care in the United Kingdom: prospective open cohort study. doi:10.1136/bmj.e3427.
7. Matheny, M. E. *et al.* Development of inpatient risk stratification models of acute kidney injury for use in electronic health records. *Med. Decis. Mak.* **30**, 639–650 (2010).
8. Betts, M. B. *et al.* Comparison of Recommendations and Use of Cardiovascular Risk Equations by Health Technology Assessment Agencies and Clinical Guidelines. *Heal. Policy Anal.* (2019) doi:10.1016/j.jval.2018.08.003.
9. CVD risk assessment and management - NICE CKS. https://cks.nice.org.uk/cvd-risk-assessment-and-management#!scenario:2.
10. Piepoli, M. F. *et al.* 2016 European Guidelines on cardiovascular disease prevention in clinical practice. *Eur. Heart J.* **37**, 2315–2381 (2016).
11. Pharmaceutical Benefits Scheme (PBS) | Nicotinic Acid, tablets (prolonged release), 500 mg, 750 mg and 1 g, Niaspan®, July 2006. https://www.pbs.gov.au/pbs/industry/listing/elements/pbac-meetings/psd/2006-07/nicotinic.
12. FDA permits marketing of artificial intelligence-based device to detect certain diabetes-related eye problems | FDA. https://www.fda.gov/news-events/press-announcements/fda-permits-marketing-artificial-intelligence-based-device-detect-certain-diabetes-related-eye.
13. Glorot, X. & Bengio, Y. *Understanding the difficulty of training deep feedforward neural networks*. http://www.iro.umontreal.
14. Steele, A. J., Denaxas, S. C., Shah, A. D., Hemingway, H. & Luscombe, N. M. Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease. *PLoS One* **13**, e0202344 (2018).

15. Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M. & Qureshi, N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS One* **12**, e0174944 (2017).
16. Steyerberg, E. W. *Clinical prediction models : a practical approach to development, validation, and updating*. (Springer, 2009).
17. Géron, A. *Hands-on machine learning with Scikit-Learn and TensorFlow : concepts, tools, and techniques to build intelligent systems*.
18. Li, Y. *et al.* Do population-level risk prediction models that use routinely collected health data reliably predict individual risks? *Sci. Rep.* **9**, 11222 (2019).
19. Li, Y., Sperrin, M., Martin, G. P., Ashcroft, D. M. & van Staa, T. P. Examining the impact of data quality and completeness of electronic health records on predictions of patients' risks of cardiovascular disease. *Int. J. Med. Inform.* 104033 (2019) doi:10.1016/j.ijmedinf.2019.104033.
20. Wynants, L., Riley, R. D., Timmerman, D. & Van Calster, B. Random-effects meta-analysis of the clinical utility of tests and prediction models. *Stat. Med.* **37**, 2034–2052 (2018).
21. Pate, A., Emsley, R., Ashcroft, D. M., Brown, B. & van Staa, T. The uncertainty with using risk prediction models for individual decision making: an exemplar cohort study examining the prediction of cardiovascular disease in English primary care. *BMC Med.* **17**, 134 (2019).
22. Cook, N. R., Buring, J. E. & Ridker, P. M. The effect of including C-reactive protein in cardiovascular risk prediction models for women. *Ann. Intern. Med.* **145**, 21–29 (2006).
23. Alaa, A. M., Bolton, T., Di Angelantonio, E., Rudd, J. H. F. & van der Schaar, M. Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants. *PLoS One* **14**, e0213653 (2019).
24. Yan Li, Matthew Sperrin, Darren M Ashcroft, T. P. van S. Does machine learning improve the accuracy of clinical risk predictions? An exemplar examining risk of cardiovascular disease.
25. Rubin, D. B. *Multiple imputation for nonresponse in surveys*. (Wiley-Interscience, 2004).
26. Sáez, C., Robles, M. & García-Gómez, J. M. Stability metrics for multi-source biomedical data based on simplicial projections from probability distribution distances. *Stat. Methods Med. Res.* **26**, 312–336 (2017).
27. Schampion. *Do we need machine-learning for cardiovascular risk prediction in clinical practice?* https://journals.plos.org/plosone/article/comment?id=10.1371/annotation/e4766ec3-0bcb-406b-a320-b2b43fcd2e4b (2017).
28. Therneau, T. & Clinic, M. Mixed Effects Cox Models. (2018).
29. Kerr, K. F. & Janes, H. First things first: Risk model performance metrics should reflect the clinical application. *Stat. Med.* **36**, 4503 (2017).
30. Yan Li; Matthew Sperrin; Darren M Ashcroft; Tjeerd Pieter van. *The instability of machine learning and statistical models in predicting individual patient risks: an approach to improve the clinical utility of these models*.

31. tensorflow/models: Models and examples built with TensorFlow. https://github.com/tensorflow/models.

Blank page

Chapter 8 Overall discussion

8.1 Summary of each chapter

Chapter 1 points out this PhD aims to assess the generalisability and clinical utility of risk prediction models in different settings especially on accurately predicting high risk patients who are missed by the model, with CVD risk prediction as exemplar.

Chapter 2 found that incorporating practice variability in a risk prediction model substantially affected the predicted CVD risks of individual patients. Clinicians and patients need to understand that risk prediction models based on routinely collected health data perform well for populations but with great uncertainty for individuals.

Chapter 3 found that variations in data quality or effects of risk factors cannot explain the considerable unmeasured heterogeneity in CVD incidence between practices. QRISK3 risk prediction should be supplemented with clinical judgement and evidence of additional risk factors.

Chapter 4 found a variety of models predicted risks for the same patients very differently despite similar model performances. The logistic model and machine learning models should not be directly applied for the prediction of long-term risk without considering censoring. The level of consistency within and between models should be routinely assessed prior to clinical usage to help inform treatment decisions.

Chapter 5 implemented QRISK3 into R package to help research community to improve future risk prediction modelling based on a used risk prediction model.

Chapter 6 found supplying percentage of patient ranks with their individual risk prediction from multiple models in clinical practice could help improve clinical utility of models. Treatment decision based on risk prediction model for patients especially

for medium risk groups should be made in conjunction with additional clinical testing and clinical judgment.

Chapter 7 discussed all the identified challenges and possible solutions for the current risk prediction model. Future guideline should consider reporting whether model captured all the main causal predictors, the effects of unmeasured causal predictors, consistency of estimated probability from statistical comparable models on the same individual patients and reasons of these inconsistencies.

8.2 Overall implications of the thesis – for clinical practice and research

An example is provided to demonstrate how to implicate finding of this research, i.e. what specific generalisability and clinical issues could be considered if trying to generalise current UK QRISK3 model to Chinese population. The first issue would be the heterogeneity of sites in China would be larger than the heterogeneity in UK, as larger variations of deprivations were expected in provinces of China, so random effects model should be used to measure these effects on individual patients. Since China did not establish large EHR cohort as UK, the data quality must be assessed with outlier assessment analysis and distribution of predictors should be compared to UK population. Missing value should be dealt with. The association between predictors and outcome is expected different between UK population and Chinese population due to the heterogeneity of nationality. Model updating approach such as re-calibration should be used to re-calibrate UK model to Chinese population. Random slope model could be used to assess whether association between predictors and outcome varied in different provinces of China. Larger population of China than UK indicates that more predictors could be considered without the fear of losing degree of freedom (the more degree of freedom means higher generalisability of model). However, these predictors might be collected differently from UK since China only has secondary care (hospitals). Whether similar UK predictors could be collected, what differences these predictors might have comparing to UK and what additional predictors are available should be considered. Due to the inconsistency of individual risk prediction among statistical comparable models, whether developing one model for the whole China or re-calibrated models for each province should be

studied. Distribution of individual risk prediction and rank should be compared among statistical comparable models to assess the robustness of decision making for individual patients. Decision making for individual patients based on risk prediction should be done with clinical testing and clinical judgement especially for patients with medium risk.

Clinicians and patients should be aware of this uncertainty issue of individual risk prediction of current population-level validated risk prediction models. Risk prediction models such as QRISK3 should be supplemented with clinical judgement and evidence of additional risk factors for clinical use.

8.3 Planned future research

Overall, current risk prediction models developed from routinely collected electronic health records or longitudinal cohorts could have very good performance on population level but with limited generalisability and clinical utility especially on individual level. Generalisability of model on population level in new clinical setting could be improved by model updating approach such as model shrinkage and re-calibration, and current guideline should require reporting the generalisability of model. To increase generalisability and clinical utility on individual level, planned future research would focus on developing new metrics to measure model performance on individual level, new individual level measurement for patients and effects of unmeasured predictors should be considered in future model development. Current model is useful in “no model” scenario (model versus nothing), future model development would move on solving the challenges from “model versus model” (e.g. inconsistency of individual risk prediction among statistical comparable model in table 7.1).

8.4 Strengths and weaknesses of the overall approach

The main strength in this PhD is that substantial quantitative analysis using representative UK cohorts with large sample size along with different statistic models, different statistical methods including those innovative ones and multiple machine learning models were conducted to assess the generalisability and clinical utility of

risk prediction model. To our knowledge, this is also the first PhD work starts to consider performance of model on both population level and individual level. There are spaces for improvement. Large cohorts from different databases such as biobank with a different disease of outcome rather than CVD could be used to further assess the generalisability and clinical utility of risk prediction model. Though CPRD is already known as a representative cohort for UK population and CVD is a regular disease of outcome for risk prediction, analysis in different settings with different outcome might be helpful to generalise the finding of this PhD. There are more statistical models and machine learning models could be considered, but most common models were covered in this PhD. Overall, future research could consider assess generalisability and clinical utility of risk prediction model with different cohorts, disease of outcome and other models.

Blank page

Chapter 9 Appendices

9.1 Literature review on statistical methods to compare sites, identify outliers and quantify practice variability

9.1.1 Abstract

Generalisability reflects the ability of model to be transferred in new clinical settings. There are different aspects of generalisability. Literatures identified that model performance of QRISK3 has heterogeneity among practices. In order to investigate on the effects of practice variability on both of population level and individual level for risk prediction model like QRISK3. One needs statistical methods to identify outlier practices which QRISK3 predicts poorly and quantify the effects of practice variability. The aim of this literature review was to review existing methods which statistically compare sites and identify outliers (outlier assessment).

9.1.2 Introduction

Generalisability (Transferability/Transportability/reproducibility) of a clinical prediction model reflects the ability (availability/applicability) of model to be applied in a different setting. Models with good generalisability would have good performance on population level and accurate risk prediction on individual patients level in new setting [1]. There are different aspects of generalisability, and external validation was the most common way to consider the generalisability of model during current model development [1234]. Different aspects of generalisability were different definitions of risk and outcome variables, different predictor effects due to location and time, strict inclusion and exclusion criteria, overfitting of the model, lack of internal validation, unreasonable to generalise due to the big differences between two populations, small sample size, different ways to handle missing data, between-practice heterogeneity (site variability), differences of CVD incidence between two populations, between-country heterogeneity [1]. Research [4] points out that external validation of CVD prediction models can only at the best evaluate the generalisability of the model in one new setting, population and time. It appears that in order to measure the whole generalisability of risk prediction models, one needs to conduct external validation on all the different setting, population and time for the model, which seems implausible practically.

This PhD first considered generalisability in the context of site variability, measuring the effects of developing a model in one set of sites and then using it in a different set of sites. The effects of site heterogeneity on risk prediction of model were tested. Site variability means the differences between sites, and sites or institutions could be GP practices, hospitals, regions of countries. Heterogeneity is a common term often used in meta-analysis study to describe true inherent differences between studies not due to chance [5]. Between-sites heterogeneity means the true differences between sites not due to chance; the differences could be different means or distributions of risk factors, different percentage of missing values (data quality). QRISK3 has different model performance (discrimination) among practices [6]. This indicates part of practices might have higher or lower predicted CVD risk than others. To investigate on this research question, one needs statistical methods to identify outlier practices which QRISK3 predicts poorly. The aim of this literature was to review existing methods which statistically compare sites and identify outliers (outlier assessment).

9.1.2 Methods

The first method used in the literature review was "snowballing", which was a method to find more papers based on the references from key sources. The key source was identified based on a small literature research around outlier assessment and discussion with professionals. The first key paper identified was the HQIP report which included guidelines and suggestions for clinical audits to detect and manage outliers. HQIP report introduced statistical methods, performance indicators chosen and real cases about how to compare sites and identify outliers according to data collected from sites [7]. More related papers were then identified from the reference of HQIP report and papers which cite the HQIP report in Google. Papers which involved sites comparison and outlier assessment were included. Similar sites comparison methods used in different papers were synthesised. The second method used was a literature search based on several key words; one paper using information theory to identify outliers was found. It used different methods from HQIP report. One paper including specific details of the information theory method was identified from the reference of the new paper.

Quality of literatures was reviewed based on literatures' data collection, method, setting, research design, funding and interpretation [8]. The quality of these methods was examined into mathematic equation level. Contexts, guidelines, challenges, advantages and disadvantages of sites comparison methods were summarised. Critical appraisal of literatures was based on the principal of the synthesised literature review results and the objective of my PhD. Specifically, reliability, accuracy, computational ability, visualise ability, popularity, advantages, limitations and how much fit to the scope of this PhD of these methods were evaluated.

9.1.3 Results

Table 9.1.1 shows the key statistical challenges in comparing sites. These included the choice of an acceptable threshold, the test for homogeneity, methods for identifying potential outliers and dealing with over-dispersion, choice of performance indicator, approach to dealing with multiple testing and how to use JSD information distance to compare risk factors' distributions.

Table 9.1.1 Key statistical challenges in outlier assessment, possible methods and examples

Statistical Challenges	Statistic technology	Examples	How to use the statistical methods to identify outliers or comparing sites?	Strength	Limitation
Limits or threshold of acceptable performance	Choosing separate thresholds for "safe" and "danger" Safety plots	Safety plots were plotted to identify "safe" and "danger" hospital in terms of the AAA mortality rate [9].	Two safety charts were plotted using different thresholds. One was using the national average and the other one was using the twice of the national average. Hospitals with the significant evidence ($p<0.05$) that the death rate above threshold were outliers of danger, and the controversy (below the threshold) were outliers of safety [9].	Hospitals which were identified as statistical outliers might not be clinical outlier. Using two separate thresholds: one for safety and one for danger, is a suggestion to deal with this scenario [7].	Safety plot was not adjusted for age or sex due to the statistical limitation, so the result might be biased. The results of safety plot would be meaningful on the condition that the two thresholds were well defined. It was suggested not to interpret the result of safety plot alone, but using it with other measures such as other risk and statistical analyses at the same time [9].
Testing homogeneity between sites to decide whether do case-mix (risk) adjustment afterward	overall $\chi 2$ test and individual $\chi 2$ test adjusting for multiple testing (Bonferroni correction) $\chi 2$ statistics and Cramer's V test	Holt et al. [9] tested the homogeneity of AAA mortality rate across hospitals	They first calculated overall $\chi 2$ test by generating the actual counts and expected counts, and deriving $\chi 2$ statistics and degree of freedom, and then comparing to the $\chi 2$ distribution. They also ran individual $\chi 2$ test on the three top outliers comparing to other hospitals multiple times. After adjusting for multiple testing (Bonferroni correction), the p-	$\chi 2$ test was a powerful statistic to test the assumptions of homogeneous and equal variance. As a non-parametric test, $\chi 2$ test can provide reliable results that t-test and ANOVA cannot. It is a convenient way to test homogeneity without need of complex statistic	$\chi 2$ test requires 80% of the cells have expected values of 5, violating this makes p-value unreliable [9 10]. The other similar tests as $\chi 2$ test are fisher's exact test and maximum likelihood ratio Chi-square ($\chi 2$) test. Fisher's exact test is only worked for 2x2 table, and maximum likelihood ratio chi-square

Statistical Challenges	Statistic technology	Examples	How to use the statistical methods to identify outliers or comparing sites?	Strength	Limitation
			value was significant. Overall, they made conclusion that no evidence supported the hypothesis of homogeneity [9].	programming [10]. In the context of true difference provided by treatment, clinical significance in χ2 test can be tested by Cramer's V test [10].	test was used when sample size was small.
Compare sites and identify potential outliers	League table (Visualized by "Caterpillar plot") Caterpillar plot	Spiegelhalter [12] demonstrated using league table and "Caterpillar plot" to compare mortality rate after treatment for fractured hip in 51 hospitals.	League table was generated when people using one measurement (say mortality rate) to rank institutions (hospitals or practices). Considering the confidence interval, "outliers" were identified by whether the CI across the target line or whether the result was far away from more than three standard deviations to the target. Caterpillar plot is one way to visualise the league table [12].	One of the most commonly useful method to identify outliers of institutions after choosing the proper measurement [12].	Misleading readers to focus on ranking institutions by the specific selected measurements, though the measurement cannot measure the quality or performance of institutions [12].
	Funnel plot	Spiegelhalter [12] wanted to identify hospitals and surgeons which have higher mortality than others using data from	Funnel plot was formed by four components, including indicator Y, target, precision parameter p and control limits. The plot showed the observed value against their precision and the target value [12].	1. It is possible to manually add observations to the graph because the axis is interpretable. 2. People's eyes were naturally attracted by outliers and it avoided the problem of ranking institutions.	1. For multiple comparison, ways to adjust are Bonferroni or the False Discovery Rate. It was recommended to compare sites separately to avoid multiple comparison. 2. Irwig et al. [14] pointed out the fact that funnel plots

Statistical Challenges	Statistic technology	Examples	How to use the statistical methods to identify outliers or comparing sites?	Strength	Limitation
		New York State Coronary Artery Bypass Graft (CABG) programme		3. The relationship between list size and outcome can be assessed in formal or informal way. 4. Can deal with over-dispersion. 5. Easy to generate from R package [12].	would be biased when the effects estimated were highly correlated to the standard error. It was highlighted that all asymmetry in funnel plots were due to this bias.
Dealing with over-dispersion in Funnel plots	Multiple options including: Do not use the indicator, improve risk stratification, analysis by clustering, using an interval as a target, estimating an over-dispersion factor and using random effects model	Spiegelhalter [15] demonstrated how to statistically deal with over-dispersion in Funnel plot.	**1. Do not use the indicator.** Over-dispersion suggests there are heterogeneity between sites, people might want to change indicator of sites. **2. Improve risk stratification** Stratify the indicator by factors (such as age or gender) which might be related to the heavy variation. **3. Analysis by clustering** Cluster similar sites together and let them compare to their similar ones. The idea was to use cluster to compare sites in a more homogeneous way. **4. Using an interval as a target** Setting the target to range rather than a precise single value. The	**Do not use the indicator**: Naive and easy to perform once allowed. **Improve risk stratification**: Easy to perform especially know which factors were related to variance. **Analysis by clustering:** Could be considered as a special case of risk stratification, it enables us to stratify risk without knowing what exactly confounders are. In this case, the sites were clustered by five types of NHS trusts. They way to cluster needs to discuss early. **Using an interval as a target:** Naive and easy to perform.	**Do not use the indicator:** Not work for all scenarios especially when there was no alternative outcome variable. **Improve risk stratification:** cannot work successfully in all contexts. **Analysis by clustering:** Not much reduction in over-dispersion. **Using an interval as a target:** It made results arbitrary, most of hospitals were either safe or dangerous. **Estimating an over-dispersion factor:** Require a robust statistical method (such as "Winsorised") to estimate the factor used to

Statistical Challenges	Statistic technology	Examples	How to use the statistical methods to identify outliers or comparing sites?	Strength	Limitation
			range could be determined either from internal (empirical estimate the acceptable range) or external (based on the normal range from previous year). **5. Estimating an over-dispersion factor** "quasi-likelihood" approach to statistic model by adding a fixed factor to each observation to account for more variance. **6. Assuming a random effects model** Adding random effects to estimate the between-sites heterogeneity. The random effects here were considered to follow the normal distribution. Random effects models were preferred as they best mimic the idea that there were unmeasured heterogeneity between sites which caused over-dispersion [15].	**Estimating an over-dispersion factor**: Generally better controlled results than previous methods. **Assuming a random effects model:** Best mimic and quantify the unmeasured between sites heterogeneity [15].	control the effect of outliers, especially when it is impossible to select a group of "in-control" sites. **Assuming a random effects model:** Need a robust statistical method to estimate the standard deviation of random effects distribution [15].
Choice of performance indicator	Risk-standardized rate: (O / E) * Target	Hospital profile means to compare sites' structure,	Calculated average rate of event (death) for a hospital or site (E) and compared to observed number of events (O). Ratio: O/E	One of the most common indicator used in hospital profile [11].	1. Inaccurate when sample size was small. 2. Ignore the sampling variability due to large

Statistical Challenges	Statistic technology	Examples	How to use the statistical methods to identify outliers or comparing sites?	Strength	Limitation
		process of care or outcomes to a target.	multiply by the overall unadjusted mortality for a state or region or country, and the result was the risk-standardised adjusted rate [11].		differences of list size. If we assume a target mortality rate (say 2.19%), it would be expected new hospital has almost 0 mortality. The fact is that patients were not selected randomly to these hospitals, there were heavy heterogeneity in patient population [11]. (Hierarchical models were proposed to overcome this.)
	Alternative Risk-standardized rate: (O – E) / variance (O – E)	In HCFA report, people want to identify hospital outliers which have higher mortality rate than others [11].	Similar to the Risk-standardized rate, after modelling data using logistical model, the difference between expected and actual was calculated as O-E. variance of O-E was calculated using approximated value from Taylor series expansion. The statistic Z was then calculated as Z=(O-E) / Var (O-E) for each practice. Practice with Z over 1.645 (90% significant Z value) would be identified as outlier.	The indicator was used in the first HCFA report and it was still a favour for hospital report cards [11].	Have not done enough risk-adjust for case-mix difference in that time point [11].

Statistical Challenges	Statistic technology	Examples	How to use the statistical methods to identify outliers or comparing sites?	Strength	Limitation
Dealing with multiple testing in comparison of sites	Bonferroni correction	Jones, Ohlssen and Spiegelhalter [13] mentioned how to use Bonferroni correction when comparing multiple health care providers.	Multiple testing increases the possibility to falsely identify too many true positives. Bonferroni proposed a simple method to account for multiple testing based on the number of testing. As Jones, Ohlssen and Spiegelhalter (2008) stated, it simply decided where to draw the "significant line". By doing so, Bonferroni proposed to use the adjusted p value (p*) as unadjusted p value multiply the numbers of testing (m) [13].	It is the common method used to solve multiple testing in hospital profile [13].	It is too strict for outlier assessment, as Jones, Ohlssen and Spiegelhalter [13] argued, they found that there were not many outliers after adjustment. They then criticised this method cost statistical power.
	False discovery rate (FDR)	Jones, Ohlssen and Spiegelhalter [13] demonstrated how to use the false discovery rate when comparing multiple health care providers.	After arguing Bonferroni correction was too strict for multiple comparison in outlier assessment, Jones, Ohlssen and Spiegelhalter [13] wanted to use an alternative way to control the false discovery rate but not too strict. They proposed the false discovery rate (FDR). Instead of adjusting all p-values in Bonferroni way, they proposed to adjust p value based on its significant rank. The most	A good balance of losing statistic power and controlling the false discovery rate [13].	It provides additional allowance for false discovery, so in the scenario people want to make sure their findings are true positive, Bonferroni adjustment (stricter) is more suitable. In the context of "evaluation" of outlier, Jones, Ohlssen and Spiegelhalter [13] argued that the FDR could be used first, and then people could then use hierarchical model to

Statistical Challenges	Statistic technology	Examples	How to use the statistical methods to identify outliers or comparing sites?	Strength	Limitation
			extreme p value (the smallest) would be adjusted as p multiply m, and the second extreme p value would be adjusted as (p * m) / 2, and the k-th extreme p value would be adjusted as (p*m) / k.		identify these identified outliers.
Comparing different data sources by assessing variability of variables' distributions	**Information-theoretic distances** Using Jensen-Shannon distance (JSD) visualised by Information geometric temporal (IGT) plot.	Sáez *et al.* [16] used JSD to measure the distribution of risk factors in different data sources.	Using JSD to calculate the "information-theoretic probabilistic distances" of risk factors (or the combination of risk factors) from one data source to the other [16]. JSD (0 to 1) equals 1 means the two distributions are not joined. They use principal component analysis to calculate probability distribution functions (PDF) for each data source. They only use two dimensions from multidimensional Scaling (MDS) to represent each group of data. MDS is a way to construct coordinates in Euclidean space for points which were known the JSD distance. Principal component analysis	1. Not affected by large sample size [28]. 2. Useful in comparing data when data has continuous variables and categorical variables at the same time[28].	1. Due to a lot of combination of groups of data, a powerful graphic card of computer is required to improve the efficiency [28]. Otherwise, it would take too long to plot the IGT. 2. Instead of PCA, other model might also be suitable to multi-type data [28]. 3. Visualisation of JSD distance can only be maximum in three dimensions, but the JSD is calculated in multiple dimensions [16].

Statistical Challenges	Statistic technology	Examples	How to use the statistical methods to identify outliers or comparing sites?	Strength	Limitation
			would convert and extract information from different risk factors into limited dimensions (components) [29]. To identify outliers, they use simplex and treat each data source as a vertex.		

Table 9.1.2 provides three methods to adjust for case-mix. Methods, such as propensity score and hierarchical (multilevel) models, were listed. Hierarchical models could be divided into hierarchical models using frequentist theory and hierarchical model using Bayesian theory.

Table 9.1.2 Methods for case-mix adjustment in outlier assessment

Statistical methods for case-mix	Examples	Details of methods	Strength	Limitation
Hierarchical model using frequentist theory to deal with patient case-mix	Han et al. [17] derived hierarchical logistic model via SAS Glimmx macro to calculate the within 30 day mortality rate of hospital for patients who have acute myocardial infarction	Pseudo-likelihood estimates were used to estimate coefficients and random effects in multilevel model [30]. The main difference between hierarchical model using frequentist theory and hierarchical model using Bayesian theory is that they use different statistic theory (e.g. frequentists used likelihood, but Bayesian used MCMC chain) to estimate the random effects of hierarchical model. Except this, the definitions of hierarchical model were the same within these two approaches.	When the sample size was large enough, frequentist theory would have similar result to Bayesian [20].	As Normand et all [11] mentioned the estimate of standard deviation of random effects might mask outliers by taking into account too much between-sites variance. There was a statistical way to detect this [31].
Hierarchical model using Bayesian theory to deal with patient case-mix	Normand et all [11] wanted to identify hospital outliers in a more sensible way, they used Bayesian theory to adjust "O" and "E" by estimating or specifying the hyper-parameters	Using Markov chain Monte Carlo methods (MCMC) to estimate hyper-parameters then calculate the risk standardized rate. Or pre-specify an acceptable value for hyper-parameters. Or do cross-	1. Offer a nature way to combine prior information of data to posterior info of data. 2. Inference in small sample was processed similarly in the large sample.	1. Do not know how to choose prior probability. 2. Posterior distribution can be heavily influenced by prior distribution. 3. High computational cost, especially the method MCMC [20].

Statistical methods for case-mix	Examples	Details of methods	Strength	Limitation
	(standard deviation of random effects).	validation (exclude selected sites and use the left sites to estimate the selected site using the way of estimating or specifying hyper-parameters.) Here, "O" is not observed numbers of events but an estimate which shrinks to the average adjusting for case-mix.	3. It is not conflicted to likelihood principle. 4. The results of Bayesian is interpretable. [20].	4. Still, it may mask outliers as stated above [11].
Propensity score (1. matching, 2. regression, 3. propensity score weighting {or inverse probability weighting [IPW]}, 4. stratification) to deal with patient case-mix	Huang *et al.*, [24] compares hierarchical model to propensity score in case-mix adjustment.	Risk factors, such as age and health status, were imbalanced between groups. After adjusting by propensity score, groups were more balanced. In practical, Huang *et al.*, [24] recommended that using general regression-based method to select risk factors first and then apply the propensity score to balance groups.	1. Improve balance of covariates among groups, and remained imbalance would also be expected in randomised trails [24]. 2. When doing case-mix adjustment, linearity assumption between outcome and risk factors were not required [24]. 3. Since risk factors such as age could be more skewed for some groups than others, propensity score would be more robust than regression models in terms of model misspecification [24].	1. The distribution of risk factors across groups might not be perfectly matched due to the random error and shortage of algorithm [24]. 2. Although propensity matching was popular, there were bias from matching algorithms, because the matching results were sensitive to the selected algorithms. 3. Bias was also found in propensity regression. 4. Propensity score weighting has not been used hospital profiling before 2012, but it has been used in economic field [24].

Table 9.1.3 presents possible confounders and its relevant outcome of interest and interventions. For instance, length of hospitalised time of patients was a confounder of new intervention of hospital of interest when comparing hospital mortality rate.

Table 9.1.3 Confounders when comparing sites or identifying outliers

Confounders	Exposure or intervention of interest	Outcome	Descriptions	Solutions
The role of external factors such as healthcare economics and the financial resources within the NHS	Risk factors of hospital	Hospital mortality rate	Natural variation in death rate of surgery may be due to the confounders such as healthcare economics and the financial resources within the NHS [9].	The technology they used had no way to adjust for confounders, so they repeated the investigations twice at two threshold values, and they mentioned confounders can barely explain 15 per cent mortality rate. Therefore, the confounders still had effect, but confounders cannot explain all change.
Length of time patients were hospitalized	Measurement (risk factors) of hospital	Hospital mortality rate	Schreiner, Han and Rapp [25] pointed out that length of hospital-time of patients might bias the measurement of hospital, because hospitals with more patients transferred out early results shorter time of patients staying in hospital which could lower mortality rate.	Measures of hospital have to use a fixed length of follow-up such as 30 days [25].
Provider selection bias	Measurement (risk factors) of hospital	Hospital quality	Unmeasured risk factors from provider selection bias might influence the case-mix adjustment [11].	Normand and Shahian [11] did not focus on this confounder.

Table 9.1.4 shows results of other data quality issues and heterogeneities. Heterogeneity of data included coding accuracy, interpretation of observations, data completeness, accuracy of critical data and case ascertainment.

Table 9.1.4 Heterogeneity in data and data quality issues

Heterogeneity of data	Type	Effect	Possible solutions
Coding accuracy	Coding	Coding accuracy had significant impact on risk-adjusted outcome results	Choosing trustworthy database
Interpretation of observations might be different [16].	Definition and interpretation	Contribute to unexpected variability of datasets	Interpret observations considering its context.
Data completeness (missing value) especially in performance indicator and variables used in case-mix adjustment [7].	Missing value	Contribute to unexpected variability of datasets	Missing value imputation
Accuracy of critical data [7].	Data accuracy	Inaccurate data might have wrong data.	Internally check range of variables and consistency, or externally compare to other data[7].
Case ascertainment [7].	Representative of sample	Case ascertainment is a percentage calculated by number of included cases over number of eligible cases. This relates to statistic power [7].	Need related to external data sources [7].

Choosing acceptable threshold is a challenge, because sometimes statistical significant is not necessarily clinical significant [7]. This implies hospitals which were identified as outliers might not necessarily be in danger or in safe, safety plots with two separate thresholds were proposed to solve this [9]. Two thresholds were chosen in safety plots to identify safe or danger "hospitals" separately to deal with this threshold challenge. In order to identify hospitals which have higher mortality rate of after abdominal aortic aneurysm (AAA) repair than others, Holt *et al.* [9] developed "Safety charts" to show outliers of hospitals. Two safety charts were built using two different thresholds to identify outliers of safety and danger separately, one was the national average mortality rate and the other one was the twice of the average mortality rate. Hospitals with significant evidence that the mortality rate were below the threshold would be identified as outliers of safety, and the contrary ones (above the threshold) would be identified as outliers of danger. RR of each hospital was calculated by comparing the death rate of one hospital to the overall average (or twice the average) death rate of other hospitals exclude the hospital. They applied bivariate binomial analysis between mortality rate of one hospital and average mortality rate of other left hospitals to plot the safety charts. The main drawback is that the safety plots were not adjusted for risk factors such as age and sex, which means outliers might be because they have more high BMI patients or smokers. The strength is using separate thresholds to identify outliers of safety and danger, because sometimes significance in statistic was not equal to significance in clinical [7].

Testing homogeneity between sites is to see whether there is heterogeneity between sites, and it is often recommended as a first step before adjusting for case-mix [9]. If sites are homogeneous, then the differences between sites would be all due to chance. An overall $\chi 2$ test was performed to test the homogeneity of death rates of after AAA repair in hospitals [9]. $\chi 2$ test provides p value by comparing observed events to expected events. Specifically, p value was generated from the $\chi 2$ distribution using $\chi 2$ statistics and degree of freedom. However, they found that 55% of hospitals had cell counts below 5, which means the significant p value is not reliable. They also performed individual $\chi 2$ test by comparing the three outliers to all the other hospitals. P values were adjusted using Bonferroni correction due to multiple testing. The significant p values suggested no evidence for the hypothesis of homogeneity, but most of hospitals had cell counts less than 5. The strength of the test is it is a non-

parametric test, and provide reliable results that t-test and ANOVA cannot [10]. The main drawback is that the cell counts less than 5 would make the p value unreliable [9].

Identifying potential outliers is a challenge to visualise and statistically prove part of sites are significantly different from others [11]. “League table” was a common method to identify outliers when people using one measurement (say mortality rate) to rank institutions (hospitals or practices). Considering the confidence interval, "outliers" were identified by whether the CI across the target line or whether the value of hospital was far away from more than three standard deviations to the target [12]. Caterpillar plot is one way to visualise the league table. The strength of league table is easy to generate and used to identify outliers, and it is the most common way to identify outliers [13]. The main drawback is it may mislead people to rank institutions using an incorrect measurement [12]. Usually, the league table would be standardised by age and sex, but there were other confounders. When there were too many institutions, the caterpillar plot would be unreadable, but outliers could still be identified by filtering in league table.

“Funnel plots” is an alternative way of “League table” to identify outliers. Using data of recorded 3-year moving totals of all operations plus outcomes of hospitals and surgeons from New York State Coronary Artery Bypass Graft (CABG) program, Spiegelhalter [12] used “Funnel plots” to demonstrate how to identify outliers of hospitals and surgeons. By fitting a logistic model using the event (such as death) as the outcome variable, he calculated the expected number of events (E) in each hospitals or surgeons, and then calculated the actual number of events (O). He then derived risk-adjusted death rate by multiplying standardised mortality ratio (O/E) by the overall state-wide mortality rate (2.2% in 1997 -1999). A funnel plot was plotted using the overall mortality rate (2.2%) as the expectation rate (target), risk-adjusted mortality rate of each hospital or surgeon as Y-axis and volume (list size) of hospital or surgeon as X-axis. The plot clearly identified one hospital with high-mortality but with low-volume and one hospital with low-mortality but with high-volume. The drawback is that the funnel plot would be biased (plot would be asymmetry) when estimates and their standard error are correlated [14]. Also, when there is heavy heterogeneity between hospitals, a large amount of hospitals might not lie within the funnel (Over-dispersion). Statistical adjustments, such as “Winsorised” estimate and random effects, were proposed to adjust for over-dispersion, but the advice of

investing on the reason of heterogeneity were highly recommended. False discovery rate (FDR) adjusted p-value were proposed to re-calculate the confidence interval line in funnel plot, because comparing hospitals multiple times would increase false discovery rate [13]. The strength of funnel plot is that it does not mislead readers to rank hospitals or surgeons by inappropriate quantities like “League table”. Also, it enables people to investigate formally or informally on the relationship between list size and the outcome [12]. The funnel plot was recommended as an initial analysis to identify outliers.

Over-dispersion means the measurement of most of sites are far away from the mean, which would then identify most of sites as outliers [15]. Multiple suggestions were made to deal with over-dispersion and random effects model were preferred. After drawing “Funnel plot” for emergency re-admission rates of 140 NHS acute trusts in 2002-2003, Spiegelhalter [15] found that the most of NHS acute trusts were identified as outliers (Over-dispersing). In order to deal with over-dispersion, methods such as change the indicator, “improve risk stratification, analysis by clustering”, change the single value of target to an interval, “estimate an over-dispersion factor” and adding random effects were proposed [15]. Because the over-dispersion might be due to the between-sites heterogeneity, change the indicator would be the first thought. However, it cannot work for all scenarios. Risk stratification suggests stratifying on factors which might contribute to “over-control”, but it cannot work successfully for all context and sometimes it is unclear whether it is worthy to do so especially when indicators were already stratified due to other reasons. Analysis by clustering suggests to group sites in a more homogeneous way and compare sites to those similar ones, but the effect of clustering on over-dispersion might be low. Using an interval means replacing target to an interval range, but often it makes the results arbitrary (Whether safe or danger). Adding an over-dispersion factor or random effects means to consider the source of over-dispersion as a factor or random effects, but both require robust statistical methods. Random effect models were preferred because they best mimic the idea that the over-dispersion was due to unmeasured between-sites heterogeneity [12].

Performance indicators measure one or two aspects of sites, choosing inappropriate performance indicator might cause the over-dispersion [11]. Risk-standardised rate was a common performance indicator used in hospital profile.

Hospital profile means to compare sites' structure, process of care or outcomes to a reasonable target [11]. The probability of event (death or CVD) was usually calculated by a statistic model (logistic model or Cox model) adjusting for possible confounders. The average rate of event in a site were used to calculate the expected incidence (E) of events in a site. With the aid of the actual events (O) in sties, an event ratio could be calculated by O/E. The risk-standardised adjusted rate was then calculated by multiplying the event ratio (O/E) by the overall unadjusted event rate of a state or region or country. The risk-standardised adjusted rate could be visualized to compare sites and identify outliers. The drawback of the risk-standardised adjusted rate is inaccurate when sample size is small, patients are not independent to sites and multiple comparison increases false discovery rates. The advantage is that it enables people easily to identify the correlation between outcome and risk factors, and it is the most common indicator in hospital profiling [11].

Alternative risk-standardised rate as a performance indicator would be using minus between O and E (O-E) instead of division (O / E). Similar to the process above, after deriving O and E for each site, then alternative risk-standardized rate was calculated as Z = (O – E) / var (O – E), and var (O - E) was estimated using Taylor series expansion. Practices with Z over 1.645 (90% significant Z value) would be identified as outlier. For example, in HCFA report people used the indicator to identify hospitals which had higher mortality rate than others [11]. The advantage of this indicator is that it was the first indicator used in HCFA report and it is still a favour in hospital profile. The disadvantage is that the indicator was not adjusted for case-mix at that time point [11].

Multiple testing would increase the false positive discovery rate when comparing sites repeatedly [13]. Bonferroni correction and False discovery rate (FDR) were methods to deal with multiple testing in sites outlier assessment. Bonferroni controls the familywise error rate – P[at least one incorrect H0 rejection], while FDR controls P[incorrect H0 rejection | H0 rejected]. For example, if people want to confirm whether their finding are completely positive, they need to use Bonferroni correction (stricter). If people only want to control the percentage of false finding in low level, then FDR could be used. Bonferroni correction is a simple method to account for multiple testing based on number of tests, and it often uses adjusted significant level

(e.g. normal significant level 0.05 / number of tests) [13]. The advantage of this correction is that it is simple to apply, and it is the most common method used in hospital profile. Disadvantage of the method is that it is too strict so there would be not many outliers identified (cost of statistic power) [13].

Based on the specific requirement, false discovery rate (FDR) could also be considered. Rather than adjust significant level using number of tests for every test, FDR adjusts significant level according to the test's significant rank [13]. The way to adjust for multiple testing is either increase p-value or decrease significant level. In terms of increasing p-value, FDR would adjust the most extreme p value as p *m (m is the number of tests), and this is the same adjusting way as the Bonferroni correction. If k represents the k-th ($1 \leq k \leq m$) test where ranked by p-value, then the k-th p value would be adjusted as (p*m) / k. In Bonferroni correction, every p value would be adjusted as p*m. One can see that (p*m) / k ≤ p*m, which means that every adjusted p value by FDR would be smaller than p value adjusted by Bonferroni correction. Therefore, FDR adjust is more relaxed than Bonferroni correction. The strength of FDR is that it offers a good balance between losing statistic power and controlling the false discover rate. The weakness of FDR is that it allows more finding due to false discovery rate [13]. In the scenario that people want to be definitely sure their finding is true; Bonferroni correction would be preferred. In context of outlier assessment, Jones, Ohlssen and Spiegelhalter (23) advised that the FDR could be used first.

Based on information-theoretic distance, Sáez et al. [16] proposed using Jensen-Shannon distance (JSD) to compare distributions of different data sources. Information geometric temporal (IGT) plot was the visualisation to identify data sources which were not clustered with other data sources. This is a new method to identify outliers from sites or data sources, because it uses information distance and distribution to identify outliers rather than statistic models and thresholds. Basically, they calculated JSD distances according to risk factors from data source, and if there were multi-dimensions, principal component analysis (PCA) was used to decrease dimensions. Information geometric temporal (IGT) plot was constructed by multidimensional Scaling (MDS) method, the MDS is a way to construct coordinates in Euclidean space for data sources which are known of the JSD distances [16]. Specifically, each data source represents one point and their JSD distances of each

other were calculated, then the IGT plot was produced by projecting the Euclidean space from MDS to two or three dimensions. The advantage of information-theoretic distance is that 1. It is generally not affected by the sample size, unless the sample size is too small and accurate estimate of distribution cannot be acquired 2. It is useful when data has categorical variables and continuous variables at the same time. The drawback including 1. Since the coordinates of points in IGT plot is constructed by MDS method, a powerful graphic card would be needed to efficiently visualise the result. 2. PCA is not the only way to decrease dimensions, other methods might be more suitable in different scenarios. 3. Visualisation of JSD distances could be only visualised maximum in three dimensions, but JSD distance may be calculated in more than three dimension [16].

Case-mix was often adjusted when sites had huge heterogeneity (i.e. it was "unfair" to compare the indicator without adjustment) [17]. Hierarchical model using frequentist theory, hierarchical model using Bayesian theory and propensity score were methods to adjust for case-mix in sites comparison. Case-mix adjustment represents a process to adjust for differences from patients or sites, so sites can be compared "fairly" [18]. Hierarchical model (multilevel model) is a kind of model which involves random effects of between-sites [19]. Frequentist and Bayesian are two different theories in statistic, and they have their own definitions. The key difference is that Bayesians consider everything unobserved to be random – while Frequentists only allow randomness to enter through hypothetical repeated sampling. Both approaches use likelihood. When the sample size was large enough, frequentist theory would have similar result to Bayesian [20]. Propensity matching is a way to balance differences between sites, and it uses propensity score - a conditional probability of assignment to a measurement\treatment given other covariates in sites [21].

Hierarchical logistic model - frequentist was used to calculate the within 30-day mortality rate of hospital for patients who have acute myocardial infraction, and they used Pseudo-likelihood to estimate coefficients and random effects in multilevel model [17]. One drawback of hierarchical model in sites comparison is that it may mask outliers by taking into account too much between-sites variance, but there is a statistical approach, which uses replication to measure the differences between model and data, to detect this problem [22].

Normand et al. [23] used Bayesian theory to adjust O and E by estimating or specifying the standard deviation of random effects from hierarchical logistic model – Bayesian. Specifically, they used Markov chain Monte Carlo methods (MCMC) to estimate hyper-parameters and then calculated the standardised risk rate. The advantages of hierarchical logistic model – Bayesian include 1. It offers a natural way to combine prior information of data to posterior information, 2. The inference in small sample was dealt similarly to the large sample, 3. It is not conflicted to likelihood principle (frequentist) and 4. The result is interpretable [20]. The disadvantages are 1. It is hard to choose prior probability, 2. Posterior distribution can be heavily influenced by prior distribution, 3. High computational cost especially in the method of MCMC [20].

Propensity score matching algorithms include matching, regression, propensity score weighting (or inverse probability weighting) and stratification [24]. Huang et al., [24] compared hierarchical model to propensity score in case-mix adjustment, and it was found that age and health status were more balanced after propensity score matching. Advantages of propensity score include 1. Improving balance of covariates among groups, 2. Do not need assumption of linearity between outcome and risk factors as required in linear model and 3. Propensity score is more robust than regression models for those skewed variables [24]. Disadvantages include 1. The match cannot be perfect due to the shortage of algorithm and there is still unbalance between groups, 2. The match results were sensitive to the selected algorithms, 3. Bias was found in one of algorithms – propensity regression, and 4. One of algorithms – propensity score weighting has not been publicly used in hospital profiling before 2012, thought it was used in economic field beforehand [24].

Confounders are defined as factors which have influence on both exposure or intervention and the outcome of interest. In terms of sites comparison and sites outlier assessment, confounders are factors which influence risk factors of sites of interest and performance indicator at the same time. It was noticed that no matter what case-mix adjustment or model adjustment used, there were still unknown or unmeasured confounders existed [7]. Several confounders were mentioned in literature. Holt *et al.* [9] pointed out healthcare economics and the financial resources within the NHS might be a confounder of natural variation in death rate of surgery, but they mentioned the confounder effect cannot explain all the variation. Schreiner, Han and Rapp [25] pointed

out that length of time of patients might bias the measurement of hospital, because hospitals with more patients transfer out early end up with shorter time of patients stay in hospital which lower the mortality rate. Normand and Shahian [11] mentioned unmeasured risk factors from provider selection bias might confound the case-mix adjustment.

Data quality issues are important, because the outliers identified or sites comparison would not be reliable with poor data [7]. Other heterogeneity between data of sites would also influence the sites comparison. Heterogeneity of data recording including coding accuracy, interpretation of observation, data completeness, accuracy and case ascertainment were found from literatures. Coding accuracy directly reflects the data quality of data, and it has significant impact on risk-adjusted outcome results [26]. Sáez, Robles and García-Gómez [16] pointed out that different interpretation of observations of datasets might result more unexpected variability of datasets. Data completeness (missing value) especially in performance indicator and variables used in case-mix adjustment influences much on case-mix adjustment [7]. Accuracy of data means whether the value of variables is correct and sensible, this can be checked "internally" and "externally" [7]. Case ascertainment is a percentage calculated by number of included cases over number of eligible cases [7]. One can deduce that if different sites have different case ascertainment, the outliers identified might be just because data of that site is not representative.

9.1.4 Discussion

They key finding of this literature review was the methods found to deal with the key statistical challenges in sites comparison. Specifically, acceptable threshold could be chosen separately according to "safe" and "danger". $\chi 2$ test is one option to test the homogeneity between different sites. Identifying potential outliers could be done via methods of "League table", "Funnel plot' or JSD information distance. Over-dispersion could be dealt with "Funnel plot" using hierarchical model (the highly recommended one) or other six methods (table 9.1.1). Risk-standardised rate ((O / E) * target) and the alternative one ((O-E) / var(O-E)) are the most two common indicators used as performance indicators in hospital mortality rate comparison [11]. Multiple testing could be handled by Bonferroni correction or False discovery rate (FDR). The methods to adjust for case-mix (confounders) include propensity score

and Hierarchical models using statistical theory of frequentist or Bayesian. Other findings include potential confounders and possible data quality issues identified from literatures, though they (often hospital mortality rate) might not directly connect to our specific research scenario (CVD prediction).

Overall, the literature review result shows there are multiple methods to identify outliers in sites comparison, and the challenges from sites comparison could be dealt with statistical methods. With identified statistical methods, we can investigate on the performance and accuracy of QRISK3 in different sites and then evaluate its generalisability to other population (e.g. Chinese datasets).

9.1.5 References

1. Steyerberg EW. *Clinical Prediction Models : A Practical Approach to Development, Validation, and Updating*. Springer; 2009.

2. Sitbon O, Benza RL, Badesch DB, et al. Validation of two predictive models for survival in pulmonary arterial hypertension. *Eur Respir J*. 2015;46(1):152-164. doi:10.1183/09031936.00004414

3. Diverse Populations Collaborative Group DPC. Prediction of mortality from coronary heart disease among diverse populations: is there a common predictive function? *Heart*. 2002;88(3):222-228. http://www.ncbi.nlm.nih.gov/pubmed/12181209. Accessed October 2, 2017.

4. Riley RD, Ensor J, Snell KIE, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. *BMJ*. 2016;353:i3140. doi:10.1136/BMJ.I3140

5. Lorenc T, Felix L, Petticrew M, et al. Meta-analysis, complexity, and heterogeneity: a qualitative interview study of researchers' methodological values and practices. *Syst Rev*. 2016;5(1):192. doi:10.1186/s13643-016-0366-6

6. Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. *Bmj*. 2017;2099(May):j2099. doi:10.1136/bmj.j2099

7. Clinical N, Advisory A, England NHS, Improvement NHS. Detection and management of outliers for National Clinical Audits. 2017;(May).

8. Fink A. *Conducting Research Literature Reviews : From the Internet to Paper*. Sage Publications; 2005.

9. Holt PJE, Poloniecki JD, Loftus IM, Thompson MM. Demonstrating safety through in-hospital mortality analysis following elective abdominal aortic aneurysm repair in England. 2008:64-71. doi:10.1002/bjs.5990

10. Mchugh ML. The Chi-square test of independence Lessons in biostatistics. *Biochem Medica*. 2013;23(2):143-149. doi:10.11613/BM.2013.018

11. Normand S-LT, Shahian DM. Statistical and Clinical Aspects of Hospital Outcomes Profiling. *Stat Sci*. 2007;22(2):206-226. doi:10.1214/088342307000000096

12. Spiegelhalter DJ. Funnel plots for comparing institutional performance. 2005;(May 2004):1185-1202. doi:10.1002/sim.1970

13. Jones HE, Ohlssen DI, Spiegelhalter DJ. Use of the false discovery rate when comparing multiple health care providers. *J Clin Epidemiol*. 2008;61(3). doi:10.1016/j.jclinepi.2007.04.017

14. Stuck AE, Rubenstein LZ, Wieland D, et al. Bias in meta-analysis detected by a simple, graphical. *Bmj*. 1998;316(7129):469-469. doi:10.1136/bmj.316.7129.469

15. Spiegelhalter DJ. Handling over-dispersion of performance indicators. *Qual Saf*

Heal Care. 2005;14(5):347-351. doi:10.1136/qshc.2005.013755

16. Sáez C, Robles M, García-Gómez JM. Stability metrics for multi-source biomedical data based on simplicial projections from probability distribution distances. *Stat Methods Med Res*. 2017;26(1):312-336. doi:10.1177/0962280214545122

17. Han LF, Rapp MT, Mattera J a, Medicine C, Wood R, Clinical J. NIH Public Access. 2012:1-21. doi:10.1161/CIRCOUTCOMES.110.957498.An

18. Jencks SF, Dobson A. Refining Case-Mix Adjustment. *N Engl J Med*. 1987;317(11):679-686. doi:10.1056/NEJM198709103171106

19. Leckie G. Module 5: Introduction to Multilevel Modelling. 2010:1-5.

20. Introduction to Bayesian Analysis Procedures: Bayesian Analysis: Advantages and Disadvantages :: SAS/STAT(R) 9.2 User's Guide, Second Edition. https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_introbayes_sect006.htm. Accessed December 5, 2017.

21. ROSENBAUM PR, RUBIN DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70(1):41-55. doi:10.1093/biomet/70.1.41

22. Joseph L. 4. Bayesian data analysis (2nd edn). Andrew Gelman, John B. Carlin, Hal S. Stern and Donald B. Rubin (eds), Chapman & Hall/CRC, Boca Raton, 2003. No. of pages: xxv + 668. Price: $59.95. ISBN 1-58488-388-X. *Stat Med*. 2004;23(21):3401-3403. doi:10.1002/sim.1856

23. Normand SL, Glickman M, Gatsonis C. Statistical Methods for Profiling Providers of Medical Care: Issues and Applications. *J Am Stat Assoc*. 1997;92(439):803-814. doi:10.1080/01621459.1997.10474036

24. Huang IC, Frangakis C, Dominici F, Diette GB, Wu AW. Application of a propensity score approach for risk adjustment in profiling multiple physician groups on asthma care. *Health Serv Res*. 2005;40(1):253-278. doi:10.1111/j.1475-6773.2005.00352.x

25. Schreiner GC, Han L, Rapp M. NIH Public Access. 2013;156:19-26. doi:10.1059/0003-4819-156-1-201201030-00004.Comparison

26. Rangachari P. Coding for quality measurement: the relationship between hospital structural characteristics and coding accuracy from the perspective of quality measurement. *Perspect Heal Inf Manag*. 2007;4:3. http://www.ncbi.nlm.nih.gov/pubmed/18066353. Accessed December 22, 2017.

27. Fearnhead P. Computational methods for complex stochastic systems: a review of some alternatives to MCMC. *Stat Comput*. 2008;18(2):151-171. doi:10.1007/s11222-007-9045-8

28. Sáez C, Zurriaga O, Pérez-Panadés J, Melchor I, Robles M, García-Gómez JM. Applying probabilistic temporal and multisite data quality control methods to a public health mortality registry in Spain: A systematic approach to quality control of repositories. *J Am Med Informatics Assoc*. 2016;23(6):1085-1095. doi:10.1093/jamia/ocw010

29. Abdi H, Williams LJ. Principal component analysis. *Wiley Interdiscip Rev Comput Stat*. 2010;2(4):433-459. doi:10.1002/wics.101

30. Wolfinger R, O'connell M. Generalized linear mixed models a pseudo-likelihood approach. *J Stat Comput Simul*. 1993;48(3-4):233-243. doi:10.1080/00949659308811554

31. Gelman A, El-Shaarawi AH, Piegorsch WW. Prior distribution. 3:1634-1637. http://www.stat.columbia.edu/~gelman/research/published/p039-_o.pdf. Accessed January 15, 2018.

Blank page

9.2 Comparison of incidence rate between CPRD and Qresearch

As part of quality control, we compared the incidence rate of CPRD data to Qresearch database by replicating the procedure which were done by QRISK3 developers in table 9.2.1. The table below shows that generally there are no big differences between CPRD data and Qresearch database in terms of CVD incidence rate. Although CPRD only contains about 40% practices of Qresearch, it contains enough information to represent UK population. Also, the average observed risk calculated by life table was about 7.8 (std: 2.6), which was closer to the average QRISK3 predict risk score 6.9 (std: 1.7). This supports the correctness of our localised QRISK3 program and the correctness of extrapolation in life table calculation.

Table 9.2.1 Comparison of incidence rate between CPRD and Qresearch

	Women						Men					
	Incident cases		Person years		Rate per 1000 person years (95%CI)		Incident cases		Person years		Rate per 1000 person years (95%CI)	
Age group (years)	CPRD	QResearch*	CPRD	QResearch	CPRD	QResearch	CPRD	QResearch	CPRD	QResearch	CPRD	QResearch
25-29	135	832	626840.89	3455662	0.22 (0.18, 0.25)	0.24 (0.22, 0.26)	186	1351	654744.46	3379716	0.28 (0.24, 0.32)	0.40 (0.38, 0.42)
30-34	245	1878	593983.49	3802577	0.41 (0.36, 0.46)	0.49 (0.47, 0.52)	419	3823	595991.6	3880890	0.70 (0.64, 0.77)	0.99 (0.95, 1.02)
35-39	511	3636	648320.64	3551460	0.79 (0.72, 0.86)	1.02 (0.99, 1.06)	1113	7963	656688.65	3748285	1.69 (1.60, 1.79)	2.12 (2.08, 2.17)
40-44	1004	5651	641431.12	2971995	1.57 (1.47, 1.66)	1.90 (1.85, 1.95)	2214	12750	654637.28	3192048	3.38 (3.24, 3.52)	3.99 (3.92, 4.06)
45-49	1477	8272	573915.85	2581104	2.57 (2.44, 2.70)	3.20 (3.14, 3.27)	3295	17763	571825.61	2672642	5.76 (5.57, 5.96)	6.65 (6.55, 6.74)
50-54	2023	12022	515582.21	2490263	3.92 (3.75, 4.09)	4.83 (4.74, 4.91)	4470	24040	499274.29	2437106	8.95 (8.69, 9.22)	9.86 (9.74, 9.99)
55-59	2717	14524	455592.8	1944140	5.96 (5.74, 6.19)	7.47 (7.35, 7.59)	5616	25464	423954.18	1796342	13.25 (12.90, 13.59)	14.18 (14.00, 14.35)
60-64	3349	18471	368655.4	1625795	9.08 (8.78, 9.39)	11.36 (11.20, 11.53)	5695	27021	315462.19	1372104	18.05 (17.58, 18.52)	19.69 (19.46, 19.93)
65-69	4240	22510	285178.29	1314303	14.87 (14.42, 15.32)	17.13 (16.90, 17.35)	5809	26903	224919.51	1013291	25.83 (25.16, 26.49)	26.55 (26.23, 26.87)
70-74	5589	25462	231372.62	1015263	24.16 (23.52, 24.79)	25.08 (24.77, 25.39)	5911	24549	161544.92	691866	36.59 (35.66, 37.52)	35.48 (35.04, 35.93)
75-79	7340	26883	210338.58	765681	34.90 (34.10, 35.69)	35.11 (34.69, 35.53)	6090	19820	127356.06	438861	47.82 (46.62, 49.02)	45.16 (44.53, 45.79)
80-84	10365	20408	227658.19	424994	45.53 (44.65, 46.41)	48.02 (47.36, 48.68)	6288	11569	112501.96	198481	55.89 (54.51, 57.27)	58.29 (57.23, 59.35)
Total (25-84)	47272	160549	6609359.6	25943236	7.15 (7.09, 7.22)	6.19 (6.16, 6.22)	58125	203016	6149098.9	24821632	9.45 (9.38, 9.53)	8.18 (8.14, 8.21)

*** QResearch statistical data was from Hippisley-Cox et al. [3]**
Average 10 years observed CVD risk in practices calculated by life table, mean (SD): 7.8 (2.6)
Average 10 years CVD risk in practices predicted by QRISK3, mean (SD): 6.9 (1.7)