DOI: 10.6339/22-JDS1037
Philosophy of Data Science

Clinical Prediction Models in Epidemiological Studies: Lessons from the Application of QRISK3 to UK Biobank Data

Ruth E. Parsons[1,*], Glen Wright Colopy[2], David A. Clifton[3,4], and Lei Clifton[1]

[1] *Nuffield Department of Population Health, University of Oxford, United Kingdom*
[2] *Independent Researcher*
[3] *Department of Engineering Science, University of Oxford, United Kingdom*
[4] *Oxford-Suzhou Centre for Advanced Research, Suzhou, China*

Abstract

Statistical models for clinical risk prediction are often derived using data from primary care databases; however, they are frequently used outside of clinical settings. The use of prediction models in epidemiological studies without external validation may lead to inaccurate results. We use the example of applying the QRISK3 model to data from the United Kingdom (UK) Biobank study to illustrate the challenges and provide suggestions for future authors. The QRISK3 model is recommended by the National Institute for Health and Care Excellence (NICE) as a tool to aid cardiovascular risk prediction in English and Welsh primary care patients aged between 40 and 74. QRISK3 has not been externally validated for use in studies where data is collected for more general scientific purposes, including the UK Biobank study. This lack of external validation is important as the QRISK3 scores of participants in UK Biobank have been used and reported in several publications. This paper outlines: (i) how various publications have used QRISK3 on UK Biobank data and (ii) the ways that the lack of external validation may affect the conclusions from these publications. We then propose potential solutions for addressing these challenges; for example, model recalibration and considering alternative models, for the application of traditional statistical models such as QRISK3, in cohorts without external validation.

Keywords *calibration; cardiovascular disease; cohort study; discrimination; external validation; model performance; prospective study; risk prediction; risk scores*

1 Introduction

In clinical practice, risk prediction models, alongside a clinician's judgement, can be used to decide the appropriate treatment program for a patient. A multivariable risk prediction model is a mathematical equation relating multiple predictors (risk factors) for an individual to the risk of future occurrence of a particular disorder or condition (Moons et al., 2015). Prediction models are often derived using data from primary care settings (including general practice, community pharmacy, dental, and optometry services). These models are often used outside of primary care clinical settings in epidemiological cohort studies (for example, to stratify data by the baseline risk of the cohort, as a predictive feature, or as a comparator predictive model). Cohorts that do not rely on primary care data but instead recruit from the general population are likely to have a different case mix. We consider the challenges of using prediction models

[*] Corresponding author. Email: r.parsons21@imperial.ac.uk.

Received January 5, 2022; Accepted January 27, 2022

in populations different from what they were derived on, using the example of applying the QRISK3 cardiovascular disease risk prediction model to the UK Biobank participants. We then recommend solutions for addressing the potentially inaccurate risk predictions.

1.1 The QRISK3 Model

QRISK3 is a Cox proportional hazards model that estimates the 10-year risk of cardiovascular disease (CVD) in women and men (Hippisley-Cox et al., 2017). The output of the model is a risk score, expressed as a percentage, interpreted as a person's risk of developing a heart attack or stroke over the next 10 years. For example, a person with a QRISK3 score of 20% is estimated to have a 2 in 10 chance of developing a CVD within 10 years. The model has been routinely updated since the first version was developed in 2007, with QRISK3 being the most recent version published in 2017 (Hippisley-Cox et al., 2017). QRISK3 includes (i) all risk factors from the previous model version QRISK2 and (ii) additional risk factors deemed important by the National Institute of Health and Care Excellence (NICE) 2014 CVD risk guidelines (National Institute for Health and Care Excellence, 2014) (all risk factors included in QRISK3 are shown in Appendix A).

The QRISK3 model was derived using data from 7.89 million patients aged 25–84 years, across 981 general practices in England (Hippisley-Cox et al., 2017). A separate set of 2.67 million patients of the same age across 328 practices in England were used to internally validate the QRISK3 model (Hippisley-Cox et al., 2017). All of these patients were free of cardiovascular disease and not using prescribed statins at baseline (Hippisley-Cox et al., 2017).

The QRISK3 model plays an important role in the National Health Service (NHS) in the UK. QRISK3 is used as part of the NHS Health Check, a free check-up for adults in England aged 40 to 74 (National Health Service, 2019), to estimate risk of heart disease and stroke (Public Health England, 2021) and aid clinicians in their decision on treatment options for patients. QRISK became the recommended risk assessment model by NICE for use in clinical settings in the UK in 2014 (National Institute for Health and Care Excellence, 2014); the required risk factors are extracted from routinely collected medical records of the individual to calculate the QRISK score for use in primary care (Nuttall and Thompson, 2021).

1.2 UK Biobank

The UK Biobank is a large-scale prospective cohort study with linkage to health-related records, containing health information for 500,000 participants (UK Biobank, 2022). The participants are aged between 40 and 69 years and were recruited between 2006 and 2010 (UK Biobank, 2022). The participants provide regular blood, urine, and saliva samples, as well as information about their lifestyle, anthropometrics, and demographics (UK Biobank, 2022).

The UK Biobank participants are more likely to be older, to be white, to be female, and to live in more socioeconomically affluent areas than non-participants (Fry et al., 2017). Compared to the general population, UK Biobank participants are likely to have fewer health conditions and are less likely to be obese, to smoke, and to drink alcohol (Fry et al., 2017). The UK Biobank cohort is not representative of the general population of the UK, with evidence of healthy volunteer selection bias (Fry et al., 2017). Compared with the national death rates among people aged 70 to 74 years, all cause mortality in UK Biobank participants was 46% lower in men and 56% lower in women (Fry et al., 2017). Although a primary care cohort will also not be fully representative of the UK population, the healthy volunteer selection bias in

UK Biobank suggests that the UK Biobank participants have a different baseline risk of adverse health outcomes (including CVD) than a more general primary care cohort.

1.3 Model Validation

A prediction model should not enter clinical practice unless acceptable performance has been demonstrated in validation studies. Levels of validation vary, but for a model to be externally validated (and some would say for it to be useful (Altman et al., 2009)) it should perform well in a dataset collected independently from the dataset used to derive the original model (Royston and Altman, 2013).

There are two main aspects of validation: discrimination and calibration:

Discrimination is the extent to which risk estimates from a model characterise different patient prognoses. The model should discriminate between individuals of varying risk, such that those with greater risk should have estimates higher than those with lower risk.

Calibration is the accuracy of the estimates provided by the model such that a well-calibrated model will assign approximately correct event probabilities (i.e., estimates of absolute risk) at all levels of predicted risk.

Poor discrimination is arguably worse than poor calibration, as a model can be re-calibrated (Royston and Altman, 2013) without being retrained, whereas poor discrimination can be improved only by retraining.

The QRISK3 model was externally validated using the primary care data from the Clinical Practice Research Datalink (CPRD) in 2021 and was found to perform well at the overall population level (Livingstone et al., 2021). CPRD contains primary care data with a similar case mix to the derivation cohort used for QRISK3, and therefore the external validation using this data measures the reproducibility of the model's performance. The transportability of QRISK3 (i.e., how well it performs in a population with different characteristics to the derivation cohort) must be explored in each independent population that differs considerably in setting to the derivation cohort (Ramspek et al., 2021) before the model can be used reliably in that independent population.

2 The Application of QRISK3 to UK Biobank Data: Literature Review

A literature search was performed in the MEDLINE database using the Ovid interface for articles up to 7 December 2021 that included 'QRISK3' AND 'UK Biobank' anywhere in the text. We excluded results in languages other than English. The initial search resulted in the identification of 27 studies, of which 10 met our inclusion criteria of having used the QRISK3 scores of UK Biobank participants in their analysis.

In six of the 10 included studies, the authors compared the performance of a model they had developed using UK Biobank data, or the QRISK3 model with additional risk factors, to the original QRISK3 model applied to UK Biobank data. In four of the studies the authors compared the discriminatory performance of the QRISK3 model to: (i) a polygenic risk score (Elliott et al., 2020), (ii) the addition of lipoprotein(a) to the QRISK3 model as well as a genetic risk score using lipoprotein(a) (Trinder et al., 2021), (iii) a machine learning model with selected predictors (Agrawal et al., 2021), and (iv) the "DiCAVA" machine learning risk model (Dolezalova et al., 2021). In one study the authors used the risk factors and the outcome definition from the original QRISK3 model to build their own Cox model using UK Biobank data; the aim of this study was to explore whether the discrimination of their Cox model was improved by the addition of the risk factor HbA1c (Welsh et al., 2020) (glycated haemoglobin (A1c)); higher levels indicate greater risk of diabetes-related complications (Diabetes.co.uk, 2019)). In the last study authors measured the correlation between the QRISK3 scores of participants in UK Biobank and the novel coronary artery disease (CAD) polygenic risk scores that the authors had developed (Riveros-Mckay et al., 2021). The aim of this polygenic risk prediction study was to prove the utility of polygenic risk factors for CAD, independent of those included in QRISK3 (Riveros-Mckay et al., 2021).

Of the six studies that applied QRISK3 to UK Biobank data as a comparator for other risk prediction tools, the authors of only one study mentioned the imperfect mapping of risk factors between UK Biobank and QRISK3 (supplementary material of Riveros-Mckay et al. (2021)). All six studies had a complete case analysis (using only dataset cases for which there are no missing values for any of the risk factors). In one study the authors recalibrated QRISK3 before comparing the performance of their developed model to it (Agrawal et al., 2021).

In three studies the authors used QRISK3 to compare predicted risk between groups of UK Biobank participants; the authors of one study examined the association between the QRISK3 scores of participants and their educational attainment and statin use (Carter et al., 2021). QRISK3 scores were used in the analysis of another study to estimate the contribution of risk factors included in QRISK3 to the higher risk of CAD in the Scottish compared to the English populations in the UK Biobank (Yang et al., 2021). Authors of the last study calculated the QRISK3 score to estimate the risk scores of South Asian and European UK Biobank participants. Comparing the medians of these two populations yielded a hazard ratio, which was subsequently compared to the observed hazard ratio of these two populations (Patel et al., 2021).

In the final study, the QRISK3 scores of participants in UK Biobank were used to adjust the prevalence of CVD for the participants by their predicted risk (Berry et al., 2021). The authors subsequently used this adjusted ratio to estimate the prevalence of CVD events among people with schizophrenia who had experience of sleep disturbance, sedentary behaviour, or muscular weakness (Berry et al., 2021).

3 Commentary on the Literature Review Findings

A prediction model may show excellent performance when applied to individuals from the population on which it was developed but perform poorly for individuals from an independent cohort (Siontis et al., 2015). Prediction models may be unintentionally fitted to the idiosyncrasies of the developmental dataset, and it is therefore important to test the reproducibility of the model's performance in new, but similar, individuals. The reproducibility of the QRISK3's performance was assessed in the CPRD external validation study (Livingstone et al., 2021). However, the transportability of QRISK3 to a population with a different case mix has not been assessed. Transportability should be examined in each population for which the use of the model is desired, if the population differs in setting, baseline characteristics, or outcome incidence to the population that the model was derived on (Ramspek et al., 2021).

All studies that have applied the QRISK3 model to data from UK Biobank, have applied a model which was developed and validated on primary care data to individuals from a cohort made up of volunteers. The UK Biobank participants are more likely to be older, to be female, and to live in more socioeconomically affluent areas than non-participants (Fry et al., 2017).

Compared to the general population, UK Biobank participants are likely to have fewer health conditions and are less likely to be obese, to smoke, and to drink alcohol (Fry et al., 2017). The UK Biobank cohort is not representative of the general population of the UK, with evidence of healthy volunteer selection bias (Fry et al., 2017).

The extent to which the identified studies address the discrimination and the calibration of QRISK3 applied to UK Biobank data varies. Some studies do not address discrimination or calibration at all (Welsh et al., 2020; Carter et al., 2021; Yang et al., 2021; Patel et al., 2021; Berry et al., 2021) and some only address discrimination (Elliott et al., 2020; Trinder et al., 2021; Riveros-Mckay et al., 2021) (with one quoting a measure of discrimination calculated in another study (Dolezalova et al., 2021)). One study discusses both discrimination and calibration, finding that the mean 10-year QRISK3 predicted CVD risk was markedly greater than the observed 10-year event rate of CAD (Agrawal et al., 2021). This study recalibrated QRISK3 to the incidence of CAD in the developmental cohort before comparing the performance of a machine learning model to this recalibrated model (Agrawal et al., 2021).

Without considering the discrimination and calibration of the QRISK3 model applied to UK Biobank data, the accuracy of the CVD risk scores is unknown, and conclusions from this application may be misleading.

4 Recommendations for Future Research

Prediction models derived using primary care data have been used outside of clinical settings, as evidenced in the literature review. For a model's generalisability to be ascertained, the model's performance must be evaluated on data other than that which was used in the model's development (Moons et al., 2015). There are strategies for minimising the model's inaccuracy and to improve model's performance when applying a risk prediction model to an external dataset that is independent to the derivation dataset. In this section, we recommend potential solutions for addressing such limitations, using the example of the application of QRISK3 to UK Biobank participants.

4.1 Model Calibration

Calibration is the agreement between the predicted and observed number of events such that a well-calibrated model would show that for every 100 individuals given a risk score of x%, close to x individuals would have the event (van Calster and Vickers, 2015) within the specified time horizon; for example, the observed frequency of CVD over 10-years should be approximately 20 out of 100 for individuals with a QRISK3 score of 20%. Calibration is poor if these predictions do not correspond with the realised outcome. An example of poor calibration is risk estimates being systematically too high for all individuals, whether they had an event or not (van Calster et al., 2019).

Calibration has been labelled as the 'Achilles heel' of predictive analytics; researchers often overlook that estimated risks may be unreliable even when prediction models have good discrimination (van Calster et al., 2019). Authors of systematic reviews have reported that calibration is assessed far less than discrimination in studies using prediction models (van Calster et al., 2019), as is the case for QRISK3 applied to UK Biobank participants.

The consequences of poorly calibrated risk prediction models have been explored more in clinical settings than in epidemiological studies, and in the latter, they may lead to spurious findings. For example, studies that aim to compare risk estimates between groups in a cohort

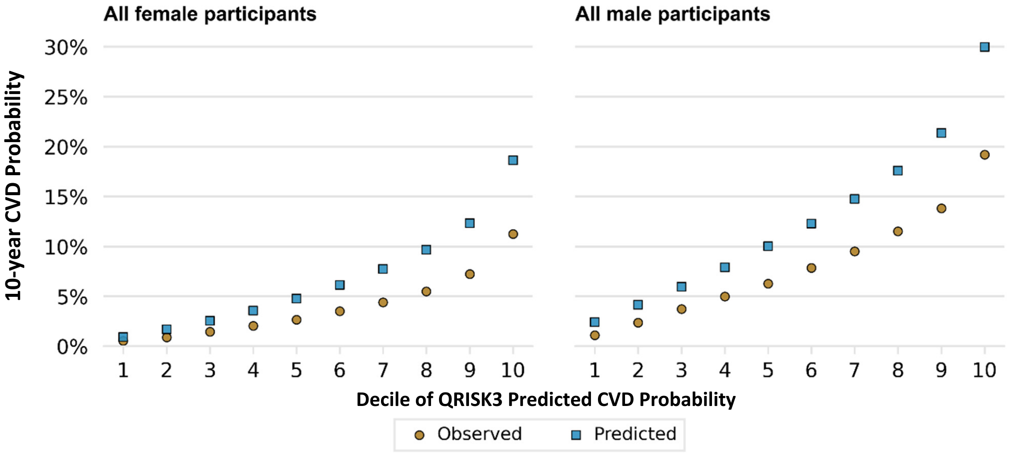

Figure 1: Calibration plot of the QRISK3 scores at ten years for UK Biobank participants by sex. The QRISK3 predicted probability of CVD consistently exceeds the observed CVD probability for both male and female UK Biobank participants. The magnitude of this over-prediction is greater at higher probability deciles.

using a model with poor calibration are unlikely to be comparing true risks, and hence should be aware of this when reporting their results.

Applying a prediction model to independent external data that is not from the derivation set is to assume that the influence of each covariate is correctly specified for the independent data. For models that provide risk estimates at a single time point, a graphical assessment of calibration can be used.

A calibration plot can be used to display the agreement between a prognostic model's predictions and the true outcomes. As an example, to produce Figure 1, the QRISK3 scores (the probability of a CVD event within 10 years) were calculated for all participants in UK Biobank that met several inclusion criteria. The participants were then grouped according to the decile of their respective QRISK3 score, such that the 10% of patients with the lowest QRISK3 score were binned into the first decile and so forth. The predicted probability within each decile group was calculated as the average QRISK3 score within that group (plotted as the blue squares in Figure 1). The observed 10-year CVD probability within each group (plotted as the orange circles in Figure 1) was calculated using the Kaplan-Meier method (to account for right censoring). The observed and predicted 10-year CVD probabilities can then be compared within each decile in Figure 1. The plot suggests that QRISK3 systematically over-predicts the probability of CVD for UK Biobank participants, with the magnitude of overprediction increasing at higher risk deciles. Royston provides advice on assessment methods for the calibration of models where estimates vary over time (Royston and Altman, 2013).

If model calibration is poor for individuals in an independent dataset, updating the model can improve its predictive accuracy for these individuals (Steyerberg, 2009). van Houwelingen 2000 suggests a solution for poorly calibrated survival models (which Agrawal et al. 2021 used to recalibrate QRISK3 for application on UK Biobank data), this approach involves a conservative procedure of recalibration where a coefficient is only changed if necessary. This is achieved by a stepwise forward procedure (van Houwelingen, 2000) where extra covariates are added to the prognostic index of the original model only if they are significant. Through this process, this model re-estimates the baseline hazard of Cox risk prediction models. While recalibrating a Cox model the proportional hazards assumption of the Cox model should be additionally checked and corrected (van Houwelingen, 2000).

4.2 Model Discrimination

Discrimination is the ability of the model to separate individuals who have the event of interest from those who do not (Pencina and D'Agostino Sr, 2015). For time-to-event models, discrimination is the ability of the model to predict who will develop an event earlier and who will develop on later or not at all (Pencina and D'Agostino Sr, 2015). Inadequate model discrimination in a population independent to the derivation population may be the result of heterogeneity in the magnitude of the risk factor between populations.

The discrimination of a model with no time element can be measured using the C-statistic; the C-statistic is the probability that, given two individuals (one who has an outcome, and one that does not or has the outcome later), the model will output a higher risk for the first individual than for the second (Pencina and D'Agostino Sr, 2015). For models with a time-to-event element, Harrell's C-index can be used; this is the probability that individuals with shorter time-to-disease have higher risk estimates. The C-statistic and the C-index range from 0 to 1, with values near 0.5 indicating that the model predictions are no better than chance alone, and values near 1 indicating that the risk estimates are good at separating individuals.

The discriminative ability of QRISK3 to separate UK Biobank participants as described above has been measured in previous studies that applied the model. The C-indices reported in these studies range from 0.64 (Trinder et al., 2021) to 0.79 (Elliott et al., 2020). The magnitude of this range is driven by varying study methodologies. For example, different outcome definitions, methods for handle missing values, and exclusion criterion are used when applying QRISK3 to data from UK Biobank; this leads to the observed heterogeneity between studies and difficulty in comparing statistics, such as the C-index, between them. As such, the C-indices from the studies applying QRISK3 to data from UK Biobank contrast with the C-index of 0.88 for women and 0.86 for men when QRISK3 was applied to individuals from the internal validation cohort (Hippisley-Cox et al., 2017).

How much consideration should be given to the discriminative abilities of a model (applied to an independent dataset) depends on the research objective of the specific study. For example, when using the risk factors and outcome definition of QRISK3 and recalculating the baseline hazard such as in Welsh et al. (2020), the discrimination of QRISK3 matters little. Researchers that apply QRISK3 to UK Biobank data without validation are likely to see poor model performance, as predictive performance of a model is commonly poorer for independent individuals than for the individuals on whom the model was derived (Moons et al., 2015). This means that applying QRISK3 to UK Biobank data is likely to yield spurious performance metrics, and if a researcher aims to compare performance of a model to QRISK3 in this context then they may not be able to accurately conclude on which model performs better. Therefore, in the case that the ability of a model to separate individuals by outcome is necessary for a study, and the chosen model has poor discrimination, researchers may consider using an alternative model.

4.3 Alternative Models

One solution to avoid the pitfalls of using a model with poor discrimination and/or calibration is to use a different model. Researchers have compared QRISK3 to other models such as polygenic risk scores (Elliott et al., 2020; Riveros-Mckay et al., 2021; Yang et al., 2021), genetic risk scores (Trinder et al., 2021), machine learning methods (Agrawal et al., 2021; Dolezalova et al., 2021), or just using the covariates included in QRISK3 to create their own risk prediction models (Welsh et al., 2020). In some of these studies (Elliott et al., 2020; Trinder et al., 2021; Riveros-Mckay et al., 2021) the authors report that these models have markedly better performance in UK Biobank than QRISK3 applied crudely to this population. Where researchers aim to predict risk at a specific time point, they may benefit from using a logistic regression model instead of thresholding the risk prediction from the Cox model at one time point.

UK Biobank contains a wealth of information on genetic risk factors and other variables beyond what is available in primary care datasets. These variables can be used for more advanced risk prediction models. Researchers looking to predict CVD risk in UK Biobank may consider using machine learning or polygenic risk prediction techniques over traditional primary care prediction models such as QRISK3. Polygenic risk prediction models are developed on data containing information on genomic variants that are associated with a disease to estimate a person's risk of developing that disease. Sun et al. 2021 developed a polygenic risk score using genomic and traditional risk factor data from UK Biobank; they tested this model using 2.1 million individuals from CPRD and found that the addition of these polygenic risk scores to traditional CVD risk factors could help prevent 7% more CVD events compared to using traditional risk factors alone if this technique was used at scale.

Alternatively, researchers comparing models to one another using discrimination measures may consider using models that have been developed in UK Biobank participants. Examples of the latter include machine learning models such as the elastic net-based Cox model by Agrawal et al. 2021 and the DeepSurv model by Dolezaloza et al. 2021, or the polygenic tools by Riveros-McKay et al. 2021 and Elliott et al. 2020.

Researchers may consider that the successful use of machine learning algorithms over traditional statistical approaches like QRISK3 depends on the transparency and standardisation of reporting, the testing and training of these algorithms on large scale databases, and formal independent external validations (Allan et al., 2021). The diversity of available machine learning algorithms also gives researchers a wide range of options, from which they may select a model that's well-aligned with the underlying dynamics of the data. (For example, the linearity assumptions of traditional models are unrealistic for many biological phenomena, so nonlinear machine learning models are better at approximating the underlying relationship for improved predictive performance.)

Other advanced techniques could include transfer learning (Li et al., 2020) to use data from the larger primary care data set to enhance the modelling of the UK Biobank participants. Successful application of transfer learning techniques to clinical prediction is non-trivial and beyond the scope of this discussion.

4.4 Collecting the Most Appropriate Risk Factors

A challenge in the application of QRISK3 to UK Biobank data is the lack of exact matches for some of the risk factors required in the QRISK3 model (Table 1). Without accurate matching of risk factors, the effect size of the covariates set by the model may be applied to the wrong

Table 1: Examples of some risk factors required in the QRISK3 model for which exact matches (Field Identifier (FID)) are not available in the UK Biobank study.

Risk factor required for QRISK3	Mapping in UK Biobank
Family history of coronary heart disease (CHD) in a first degree relative aged less than 60 years	UK Biobank fields include illnesses in father (FID 20107), illnesses in mother (FID 20110), and illnesses of siblings (FID 20111). Applying these fields to the QRISK3 model is with the caveat that the level 'heart disease' of these illnesses is assumed to be CHD and that the relative is assumed to be aged less than 60 years at diagnosis.
Smoking status (non-smoker, former smoker, light smoker (1–9/day), moderate smoker (10–19/day), or heavy smoker ($\geqslant$20/day))	The availability of UK Biobank fields mean that this variable is derived such that individuals reporting their current smoking as 'only occasionally' at baseline (FID 1239) are be classed as light smokers, individuals reporting their smoking status as 'never' (FID 20116) are non-smokers, and individuals reporting their smoking status as 'previous' (FID 20116) are former smokers. For individuals reporting that they currently smoked in FID 1239 or 20116, the number of cigarettes smoked daily (FID 3456) can be used to derive their level of smoking.
Measure of systolic blood pressure (SBP) variability (standard deviation of repeated measures)	There is no field in the UK Biobank for variability in SBP, this variable can be derived as the standard deviation between two automated or manual SBP readings at baseline (FID 4080 and 93). These two readings are not available for everyone which leads to missing values.

covariates, leading to inaccurate predictions from the model, and limiting the use of QRISK3 in UK Biobank.

One way to mitigate against including imprecise risk factors in risk prediction is to plan during the study design stage, and collect the correct covariates required for traditional statistical models, such as QRISK3, in large scale cohort studies. As suggested by Wolford et al. 2021, biobanks should collect high quality family history risk factors for use in future prediction models. During our preliminary work on an independent external validation of QRISK3 using UK Biobank data, we found that UK Biobank lacks information on the age at which family members were diagnosed with coronary heart disease (a requirement for the QRISK3 model), as shown in Table 1.

Additionally, we found it difficult to derive the smoking status of UK Biobank participants because of multiple records of smoking risk factors (some shown in Table 1). UK Biobank could provide a single derived variable that concatenates the existing overlapping smoking variables. This would allow researchers to access more information with ease.

It is sometimes not feasible to collect the variables required for traditional statistical models in epidemiological cohorts and this should be considered when choosing the appropriate model for a study. For example, repeated measurements of blood pressure are collected routinely in primary care which allows the calculation of systolic blood pressure (SBP) variability. These repeated measurements are usually not available in cohort studies, owing to the cost and time involved in collecting these variables. For example, in the UK Biobank study, automated and manual SBP were measured at baseline for almost all participants, however only 20,286 and 220 participants have first repeat assessment readings for automated and manual SBP, respectively.

5 Conclusion

Caution should be taken when applying risk prediction models derived from primary care data to independent datasets without first validating model performance on the new data set. We recommend that researchers explore the limitations of applying these models in the context of their research aims by considering: (i) the lack of exact matches for some risk factors; (ii) the differences in case mix between the model derivation population and the data they are using; (iii) potential model miscalibration; and (iv) any issues with discrimination. Potential solutions for inaccurate conclusions when applying a model to independent data include: (i) model recalibration; (ii) using an alternative model; (iii) collecting the correct covariates in studies; and (iv) considering discrimination in context.

A Appendix

Box 1

Risk factors included in the QRISK3 model (Hippisley-Cox et al., 2017)

Risk factors in the QRISK3 model that exist in the QRISK2-2017 model (Hippisley-Cox et al., 2017):

- Age at study entry (baseline)
- Ethnic origin (nine categories)
- Deprivation (as measured by the Townsend score, where higher values indicate higher levels of material deprivation)
- Systolic blood pressure
- Body mass index
- Total cholesterol: high density lipoprotein cholesterol ratio
- Smoking status (non-smoker, former smoker, light smoker (1–9/day), moderate smoker (10–19/day), or heavy smoker (⩾20/day))
- Family history of coronary heart disease in a first degree relative aged less than 60 years
- Diabetes (type 1, type 2, or no diabetes)
- Treated hypertension (diagnosis of hypertension and treatment with at least one antihypertensive drug)
- Rheumatoid arthritis (diagnosis of rheumatoid arthritis, Felty's syndrome, Caplan's syndrome, adult onset Still's disease, or inflammatory polyarthropathy not otherwise specified)
- Atrial fibrillation (including atrial fibrillation, atrial flutter, and paroxysmal atrial fibrillation)
- Chronic kidney disease (stage 4 or 5) and major chronic renal disease (including nephrotic syndrome, chronic glomerulonephritis, chronic pyelonephritis, renal dialysis, and renal transplant)

Risk factors new in the QRISK3 model, compared with the QRISK2-2017 model (Hippisley-Cox et al., 2017):

- Expanded definition of chronic kidney disease (to include general practitioner recorded diagnosis of chronic kidney disease stage 3 in addition to stages 4 and 5 as well as major chronic renal disease)
- Measure of systolic blood pressure variability (standard deviation of repeated measures)
- Diagnosis of migraine (including classic migraine, atypical migraine, abdominal migraine, cluster headaches, basilar migraine, hemiplegic migraine, and migraine with or without aura)
- Corticosteroid use (British National Formulary (BNF) chapter 6.3.2 including oral or parenteral prednisolone, betamethasone, cortisone, depo-medrone, dexamethasone, deflazacort, efcortesol, hydrocortisone, methylprednisolone, or triamcinolone)
- Systemic lupus erythematosus (including diagnosis of SLE, disseminated lupus erythematosus, or Libman-Sacks disease)
- Second generation "atypical" antipsychotic use (including amisulpride, aripiprazole, clozapine, lurasidone, olanzapine, paliperidone, quetiapine, risperidone, sertindole, or zotepine)
- Diagnosis of severe mental illness (including psychosis, schizophrenia, or bipolar affective disease)
- Diagnosis of erectile dysfunction or treatment for erectile dysfunction (BNF chapter 7.4.5 including alprostadil, phosphodiesterase type 5 inhibitors, papaverine, or phentolamine)

Conflicts of Interest

Glen Wright Colopy is employed by LifeBell AI. David A Clifton declares grants from GlaxoSmithKline and personal fees from Oxford University Innovation, Biobeats, and Sensyne Health, in areas unrelated to this work.

Funding

The research was supported by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC). The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

References

Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. *Annals of Internal Medicine*, 162(1): W1–W73. https://doi.org/10.7326/M14-0698.

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

National Institute for Health and Care Excellence (2014). *Clinical Guideline 181: Lipid Modification: Cardiovascular Risk Assessment and the Modification of Blood Lipids for the Primary and Secondary Prevention of Cardiovascular Disease*. National Institute for Health, London.

National Health Service (2019). NHS Health Check. Available: https://www.nhs.uk/conditions/nhs-health-check/. [Accessed November 2021].

Public Health England (2021). *NHS Health Checks: QRISK3 Explained*. Public Health England, London.

Nuttall M, Thompson K (2021). *NHS Health Checks: QRISK3 Explained*. Public Health England, London.

UK Biobank (2022). About Us. [Online]. Available: https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/about-us. [Accessed January 2022].

Fry A, Littlejohns TJ, Sudlow C, Doherty N, Adamska L, Sprosen T, et al. (2017). Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. *American Journal of Epidemiology*, 186(9): 1026–1034. https://doi.org/10.1093/aje/kwx246.

Altman DG, Vergouwe Y, Royston P, Moons KGM (2009). Prognosis and prognostic research: validating a prognostic model. *BMJ (Online)*, 338(7708): 1432–1435. https://doi.org/10.1136/bmj.b605.

Royston P, Altman DG (2013). External validation of a Cox prognostic model: principles and methods. *BMC Medical Research Methodology*, 13: 33. http://www.biomedcentral.com/1471-2288/13/33.

Livingstone S, Morales DR, Donnan PT, Payne K, Thompson AJ, Youn JH, et al. (2021). Effect of competing mortality risks on predictive performance of the QRISK3 cardiovascular risk prediction tool in older people and those with comorbidity: external validation population cohort study. *The Lancet Healthy Longevity*, 2(6): e352–e361. https://doi.org/10.1016/S2666-7568(21)00088-X.

Ramspek CL, Jager KJ, Dekker FW, Zoccali C, van Diepen M (2021). External validation of prognostic models: what, why, how, when and where? *Clinical Kidney Journal*, 14(1): 49–58. https://doi.org/10.1093/ckj/sfaa188.

Elliott J, Bodinier B, Bond TA, Chadeau-Hyam M, Evangelou E, Moons KGM, et al. (2020). Predictive accuracy of a polygenic risk score-enhanced prediction model vs a clinical risk score for coronary artery disease. *JAMA. Journal of the American Medical Association*, 323(7): 636–645. https://doi.org/10.1001/jama.2019.22241.

Trinder M, Uddin MM, Finneran P, Aragam KG, Natarajan P (2021). Clinical utility of lipoprotein(a) and LPA genetic risk score in risk prediction of incident atherosclerotic cardiovascular disease. *JAMA Cardiology*, 6(3): 287–295. https://doi.org/10.1001/jamacardio.2020.5398.

Agrawal S, Klarqvist MDR, Emdin C, Patel AP, Paranjpe MD, Ellinor PT, et al. (2021). Selection of 51 predictors from 13,782 candidate multimodal features using machine learning improves coronary artery disease prediction. *Patterns*, 2(12): 100364. https://doi.org/10.1016/j.patter.2021.100364.

Dolezalova N, Reed AB, Despotovic A, Obika BD, Morelli D, Aral M, et al. (2021). Development of an accessible 10-year Digital CArdioVAscular (DiCAVA) risk assessment: a UK Biobank study. *European Heart Journal – Digital Health*, 2(3): 528–538. https://doi.org/10.1093/ehjdh/ztab057.

Welsh C, Welsh P, Celis-Morales CA, Mark PB, Mackay D, Ghouri N, et al. (2020). Glycated hemoglobin, prediabetes, and the links to cardiovascular disease: data from UK Biobank. *Diabetes Care*, 43(2): 440–445. https://doi.org/10.2337/dc19-1683.

Diabetes.co.uk (2019). Guide to HBA1c. [Online]. Available: https://www.diabetes.co.uk/

what-is-hba1c.html. [Accessed January 2022].

Riveros-Mckay F, Weale M, Moore R, Selzam S, Krapohl E, Sivley R, et al. (2021). An integrated polygenic and clinical risk tool enhances coronary artery disease prediction. *Circ Genom Precis Med.*, 14(2).

Carter AR, Gill D, Davey Smith G, Taylor AE, Davies NM, Howe LD (2021). Cross-sectional analysis of educational inequalities in primary prevention statin use in UK Biobank. *Heart*. https://doi.org/10.1136/heartjnl-2021-319238.

Yang C, Starnecker F, Pang S, Chen Z, Güldener U, Li L, et al. (2021). Polygenic risk for coronary artery disease in the Scottish and English population. *BMC Cardiovascular Disorders*, 21(1): 586. https://doi.org/10.1186/s12872-021-02398-4.

Patel AP, Wang M, Kartoun U, Ng K, Khera Av (2021). Quantifying and understanding the higher risk of atherosclerotic cardiovascular disease among South Asian individuals: results from the UK Biobank prospective cohort study. *Circulation*, 144(6): 410–422. https://doi.org/10.1161/CIRCULATIONAHA.120.052430.

Berry A, Yung AR, Carr MJ, Webb RT, Ashcroft DM, Firth J, et al. (2021). Prevalence of major cardiovascular disease events among people diagnosed with schizophrenia who have sleep disturbance, sedentary behavior, or muscular weakness. *Schizophrenia Bulletin Open*, 2(1): sgaa069. https://doi.org/10.1093/schizbullopen/sgaa069.

Siontis GCM, Tzoulaki I, Castaldi PJ, Ioannidis JPA (2015). External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *Journal of Clinical Epidemiology*, 68(1): 25–34. https://doi.org/10.1016/j.jclinepi.2014.09.007.

van Calster B, Vickers AJ (2015). Calibration of risk prediction models: impact on decision-analytic performance. *Medical Decision Making*, 35(2): 162–169. https://doi.org/10.1177/0272989X14547233.

van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW, Bossuyt P, et al. (2019). Calibration: the Achilles heel of predictive analytics. *BMC Medicine*, 17(1): 230. https://doi.org/10.1186/s12916-019-1466-7.

Steyerberg E (2009). *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Springer, New York, NY. https://doi.org/10.1007/978-0-387-77244-8.

van Houwelingen HC (2000). Validation, calibration, revision and combination of prognostic survival models. *Statistics in Medicine*, 19: 3401–3415. https://doi.org/10.1002/1097-0258(20001230)19:24<3401::AID-SIM554>3.0.CO;2-2.

Pencina MJ, D'Agostino Sr RB (2015). Evaluating discrimination of risk prediction models: the C statistic. *JAMA*, 314(10): 1063–1064. https://doi.org/10.1001/jama.2015.11082.

Sun L, Pennells L, Kaptoge S, Nelson CP, Ritchie SC, Abraham G, et al. (2021). Polygenic risk scores in cardiovascular risk prediction: a cohort study and modelling analyses. *PLoS Medicine*, 18(1): e1003498. https://doi.org/10.1371/JOURNAL.PMED.1003498.

Allan S, Olaiya R, Burhan R (2021). Reviewing the use and quality of machine learning in developing clinical prediction models for cardiovascular disease. *Postgraduate Medical Journal*. BMJ Publishing Group. https://doi.org/10.1136/postgradmedj-2020-139352.

Li S, Cai TT, Li H (2020). Transfer learning for high-dimensional linear regression: prediction, estimation, and minimax optimality. arXiv preprint: http://arxiv.org/abs/2006.10593.

Wolford BN, Surakka I, Graham SE, Nielsen JB, Zhou W, Gabrielsen ME, et al. (2021). Utility of family history in disease prediction in the era of polygenic scores. medRxiv preprint: https://doi.org/10.1101/2021.06.25.21259158.