


PROTOCOL

Open Access



# Protocol for development and validation of postpartum cardiovascular disease (CVD) risk prediction model incorporating reproductive and pregnancy-related candidate predictors

Steven Wambua<sup>1\*</sup> , Francesca Crowe<sup>1</sup>, Shakila Thangaratinam<sup>2,3</sup>, Dermot O'Reilly<sup>4</sup>, Colin McCowan<sup>5</sup>, Sinead Brophy<sup>6</sup>, Christopher Yau<sup>7,8,9</sup>, Krishnarajah Nirantharakumar<sup>1</sup>, Richard Riley<sup>1</sup> and on behalf of the MuM-PreDiCT Group

## Abstract

**Background:** Cardiovascular disease (CVD) is a leading cause of death among women. CVD is associated with reduced quality of life, significant treatment and management costs, and lost productivity. Estimating the risk of CVD would help patients at a higher risk of CVD to initiate preventive measures to reduce risk of disease. The Framingham risk score and the QRISK<sup>®</sup> score are two risk prediction models used to evaluate future CVD risk in the UK. Although the algorithms perform well in the general population, they do not take into account pregnancy complications, which are well known risk factors for CVD in women and have been highlighted in a recent umbrella review.

We plan to develop a robust CVD risk prediction model to assess the additional value of pregnancy risk factors in risk prediction of CVD in women postpartum.

**Methods:** Using candidate predictors from QRISK<sup>®</sup>-3, the umbrella review identified from literature and from discussions with clinical experts and patient research partners, we will use time-to-event Cox proportional hazards models to develop and validate a 10-year risk prediction model for CVD postpartum using Clinical Practice Research Datalink (CPRD) primary care database for development and internal validation of the algorithm and the Secure Anonymised Information Linkage (SAIL) databank for external validation. We will then assess the value of additional candidate predictors to the QRISK<sup>®</sup>-3 in our internal and external validations.

**Discussion:** The developed risk prediction model will incorporate pregnancy-related factors which have been shown to be associated with future risk of CVD but have not been taken into account in current risk prediction models. Our study will therefore highlight the importance of incorporating pregnancy-related risk factors into risk prediction modeling for CVD postpartum.

**Keywords:** Prediction modeling, Cardiovascular disease, Pregnant women, Prognosis, Pregnancy complications

## Introduction

CVD is a leading cause of morbidity and mortality globally in both men and women [1, 2]. Estimating the risk of the condition would help patients at a higher risk of CVD to access treatments to reduce the risk of developing CVD. There are several risk prediction models used

\*Correspondence: [ssw107@student.bham.ac.uk](mailto:ssw107@student.bham.ac.uk)

<sup>1</sup> Institute of Applied Health Research, College of Medical and Dental Sciences, University of Birmingham, Edgbaston, Birmingham, UK  
Full list of author information is available at the end of the article



© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

routinely in primary care to predict CVD risk in the general population. These include the Framingham risk score model and the QRISK<sup>®</sup> score [3, 4]. However, studies have shown that they tend to underestimate the risk of CVD in young women [5, 6]. While the most recent QRISK<sup>®</sup> calculator includes several comorbidities (for example diabetes mellitus) and one male-related risk factor (erectile dysfunction), there are no female-specific candidate predictors included in the CVD risk prediction models [5].

During pregnancy, women experience cardiovascular physiological changes such as an increase in cardiac output. A small proportion of pregnant women develop pregnancy-induced hypertension and preeclampsia [7], and a woman's response to such changes could be linked to future cardiovascular health [8]. Several studies have identified a link between certain pregnancy complications (e.g., gestational hypertension, preeclampsia, placental abruption, preterm birth, gestational diabetes mellitus, and stillbirth) and reproductive health factors (e.g., early age at menarche and polycystic ovary syndrome) with risk of CVD [9–11]. More recently, the postpartum period has been identified as a possible window of opportunity to initiate cardiovascular disease preventative measures in women [12, 13]. However, there is lack of guidelines on risk factor management in this population.

There have been recent efforts to quantify the predictive value of pregnancy-related candidate predictors to established CVD risk prediction models [5, 14–16]. A study by Markovitz et al. [14] showed that adding pregnancy complications history to the NORRISK 2 risk model improved the c-index by 0.004, while another study by Marziah et al. [15] established that the Framingham risk score was enhanced (c-statistic of 0.0053) after adding these factors. The National Institute for Health and Care Excellence (NICE) recommends using QRISK<sup>®</sup> assessment tool to calculate a person's 10-year risk of CVD in the UK, but there have been no attempts to evaluate the added value of pregnancy factors in the development of the risk prediction model in women.

We plan to develop a robust CVD risk prediction model postpartum to assess whether adding reproductive health and pregnancy-related candidate predictors to the QRISK<sup>®</sup>-3 risk prediction model improves the performance of the individual risk prediction of CVD in women.

### Objectives

The main aim of this study is to update the QRISK<sup>®</sup>-3 tool to include candidate predictors related to women's health to help predict the risk of CVD postpartum in women without a history of CVD. This tool will be

important to help healthcare professionals in their decision making about the need for targeted care. The specific objectives of the study are as follows:

- i. To externally validate the QRISK<sup>®</sup>-3 score in the postpartum period using a large, representative study population of women from UK primary care
- ii. To develop a clinical prediction model for 10-year risk of CVD postpartum (15 months after conception as index date) and internally validate its performance (overall model fit, calibration, and discrimination) using the study population in objective (i)
- iii. To externally validate the risk prediction model developed in objective (ii), by examining its performance and clinical utility in a separate large, representative study population of women from UK primary care, both overall and within relevant subgroups

### Research design and methods

#### Data sources

Two databases of anonymized Electronic Health Records will be used for this study. They are as follows:

1. Clinical Practice Research Datalink (CPRD) [17], which has over 19 million patient records in the UK from over 940 participating general practices, with a mean follow-up of 13 years as of February 2021.

The CPRD pregnancy register is used to capture information from maternity, antenatal, and delivery records to identify pregnancies within CPRD GOLD [18]. According to recent data, the CPRD register captured 5.8 million pregnancies among 2.4 million women in the period January 1987–February 2018 [18]. We will use the register to extract pregnancy data from CPRD.

2. Secure Anonymised Information Linkage (SAIL) [19], which has data from over 4 million patient records in Wales and covers 80% of Welsh general practices [20]. Follow-up is longer than CPRD databases as SAIL tracks patient journeys even when they transfer practice within Wales.

The National Community Child Health Database will be used to identify pregnancies and will be linked to the Welsh Longitudinal General Practice database (for diagnosis and medications data) and Welsh Demographic Service database (for demographics data) within the SAIL databank. Using these databases, 27,783 pregnant women were identified in SAIL in 2018 in a study conducted within the MuM-PreDiCT consortium [21]. We expect to have more pregnant women within a follow-up period of 10 years.

Both databases contain data from GP practices captured primarily using Vision software. CPRD will be used to develop the risk prediction model and for the internal validation process while the SAIL database will be used for independent external validation of the risk prediction model. We will exclude data on patients from Wales in CPRD to ensure no overlap with patients in the SAIL database.

### **Target population**

The target population is women between the ages of 15 and 49 years who have a history of pregnancy and registered with participating GPs between 1 of January 2000 and 31 of December 2021. Women with pre-existing CVD before study entry will be excluded as the risk prediction model is for those who have not been diagnosed with CVD.

Each woman can contribute to the cohort after a minimum registration period with their GP for at least 12 months to ensure sufficient quality data at baseline. The index date will be 15 months after date of conception of the last pregnancy (estimated to be 6 months postpartum). The index date has been chosen to be around 6 months postpartum because this allows for normal physiological changes of pregnancy to resolve and time lag for postpartum information to be recorded in the GP database [22, 23].

Participants will be followed from the index date until the earliest of outcome date, transfer date (CPRD GOLD), last date of data collection, death date, or study end date. Participants will be censored 10 years after the index date.

Flow chart of participants from baseline (6 months postpartum) through completion of the study (study end, 31 December 2021) will be presented in the final report.

### **Study outcome**

The outcome will be the first recorded diagnosis of cardiovascular disease (coronary heart disease, stroke, myocardial infarction, or transient ischemic attack). The outcome will be ascertained using Read codes, a clinical terminology system used for record-keeping in general practice in the National Health Service (NHS) [24]. For comparability, Read codes for the outcome of CVD have been obtained from the article on the development and validation of the QRISK<sup>®</sup>-3 and are presented in Additional file 1.

### **Clinical predictor variables**

#### **Determining candidate predictors for model development**

Candidate predictors are features that will be investigated for their potential predictive value towards

risk prediction of CVD postpartum. The features will include any information that precedes cardiovascular disease and are available at the start-point (moment of intended prediction) and are linked to an increased risk of CVD. Examples will include pregnancy-related risk factors for example gestational diabetes, pre-eclampsia, and gestational hypertension.

We will use two approaches to identify candidate predictors: (1) clinical and patient expertise and (2) evidence from previous studies [25]. For the clinical and patient expertise approach, candidate predictors will be selected through discussions with clinicians and patient research partners while for the evidence from previous studies approach, and risk factors will be identified through literature review [26]. Potential candidate predictors for CVD postpartum have been chosen based on the umbrella review identified from the literature and clinical significance and through discussions with clinicians and patient research partners. We plan to assess the data quality of the potential candidate predictors chosen, including by evaluating missing data and any outliers, and the timing and method of their measurement. We will then perform variable selection using the least absolute shrinkage and selection operator (LASSO) to determine predictors that will be included in the final model [27, 28].

### **Proposed candidate predictors**

Table 1 shows the list of the proposed candidate predictors from QRISK<sup>®</sup>-3, an umbrella review of reproductive health factors associated with CVD in young women, and from discussions with clinicians and patient research partners [4, 26].

The list of potential candidate predictors can be expanded to include more potential factors based on new information that emerges from literature or through discussions with clinicians and patients as the project progresses.

### **Statistical analysis**

#### **Steps for development and validation of the updated risk prediction model**

- i) Externally validate the QRISK<sup>®</sup>-3 using CPRD-GOLD data
- ii) Use the QRISK<sup>®</sup>-3 model coefficients in (i) above as a single predictor and add additional candidate predictors to develop and internally validate an updated risk prediction model (Model 1).
- iii) Develop and internally validate a risk prediction model using all predictors, i.e., QRISK<sup>®</sup>-3 predic-

**Table 1** Proposed candidate predictors

Candidate predictors identified from QRISK-3 relevant to women [4]	Candidate predictors identified from Umbrella Review [26]	Candidate predictors identified from discussions within the research team
<p><b>Patient characteristics</b></p> <ul style="list-style-type: none"> <li>• Age</li> <li>• Ethnicity</li> <li>• Deprivation</li> <li>• Systolic blood pressure (SBP)</li> <li>• Body mass index (BMI)</li> <li>• Total cholesterol: high-density lipoprotein cholesterol ratio</li> <li>• Smoking status</li> </ul> <p><b>Medical history predictors</b></p> <ul style="list-style-type: none"> <li>• Family history of coronary heart disease</li> <li>• Diabetes mellitus</li> <li>• Treated hypertension</li> <li>• Rheumatoid arthritis</li> <li>• Atrial fibrillation</li> <li>• Chronic kidney disease (stages 4 or 5)</li> <li>• 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)</li> <li>• The standard deviation of SBP (variability of repeated SBP measures)</li> <li>• Diagnosis of migraine</li> <li>• Corticosteroid use</li> <li>• Systemic lupus erythematosus</li> <li>• Second generation "atypical" antipsychotic use</li> <li>• Diagnosis of severe mental illness</li> </ul>	<p><b>Reproductive health factors</b></p> <ul style="list-style-type: none"> <li>• Early age at menarche (&lt;12 years old)</li> <li>• Use of hormonal contraceptive agents</li> <li>• Use of oral contraceptives</li> <li>• Polycystic ovary syndrome</li> <li>• Parity/gravidity</li> </ul> <p><b>Adverse pregnancy outcomes</b></p> <ul style="list-style-type: none"> <li>• Hypertensive disorders of pregnancy</li> <li>• Gestational diabetes mellitus</li> <li>• Placental abruption</li> <li>• Pregnancy loss</li> <li>• Preterm birth</li> <li>• Low birth weight and small for gestational age</li> </ul>	<ul style="list-style-type: none"> <li>• Idiopathic intracranial hypertension</li> <li>• Postnatal depression</li> <li>• Hypothyroidism</li> <li>• Hyperthyroidism</li> <li>• Endometriosis</li> <li>• Menstrual irregularity</li> <li>• Antiphospholipid syndrome</li> </ul>

tors plus additional candidate predictors (Model 2), allowing for variable selection via the LASSO.

- iv) Compare predictive performance measures (calibration and discrimination) of QRISK<sup>®</sup>-3, Model 1 and Model 2, using internal validation techniques.
- v) Externally validate the best risk prediction model (based on predictive performance measures obtained after internal validation) between Model 1 and Model 2, using the SAIL database, and again compare to QRISK<sup>®</sup>-3
- vi) Compare predictive performance measures (calibration, discrimination, and net benefit analysis) in (i) and in (v) above.

#### **Development, internal and external validation of models**

**Missing data** Missing data in each candidate predictor will be investigated before analysis. Missing data will be classified based on whether the values are expected to be missing at deployment. Missing data will then be handled using three approaches. Firstly, missing entry for a condition (e.g., diabetes) will be taken to indicate the absence of the comorbidity (no history of diabetes). Secondly, the missing indicator method will be used for variables where we expect informative missingness at deployment. For example, if a biomarker test (e.g., blood pressure measurements, blood cholesterol, HbA1c, etc.) has been carried out, then the perceived need for the test of the biomarker might be informative of the patient's health [29]. Thirdly, multiple imputations with chained equations will be applied for candidate predictors where we do not expect missing data at deployment. The three approaches will be used to ensure missing data methods match at both the development and deployment stages of the risk prediction model as recommended in recent studies [29].

**Externally validating QRISK<sup>®</sup>-3 equation using CPRD-GOLD dataset** The QRISK<sup>®</sup>-3 risk prediction model was developed using the QResearch primary care database [30]. The first step will be to externally validate the QRISK<sup>®</sup>-3 risk prediction model using the CPRD-GOLD dataset and assess its performance for women with a history of pregnancy. This will form a benchmark for risk prediction models incorporating additional candidate predictors.

We will calculate 10-year risk of CVD (predicted risk) for women with a history of pregnancy using the QRISK<sup>®</sup>-3 algorithm. The observed 10-year risk (observed risk) of CVD will be estimated using the method of Kaplan-Meier.

Missing data in each predictor will be handled in a similar way as during the development of the QRISK<sup>®</sup>-3 [4]. We will examine the predictive performance of QRISK<sup>®</sup>-3 in the population using calibration (plots, curves, and slope) to see how closely the predicted risk agrees with the observed risk and discrimination (the model's ability to distinguish between those who develop post-partum CVD and those who do not, summarized as time-dependent C-statistics and Royston's D statistic). These measures will be obtained overall and for sub-groups of women defined by ethnicity, socio-economic status, and age. We have chosen to validate the risk prediction model within these subgroups for a start because algorithmic biases in the risk prediction models used in healthcare occur in various subgroups defined by ethnicity, socio-economic status, and age [31]. Additionally, previous studies have considered these type of subgroups in the validation of risk prediction models for CVD [4, 32].

**Primary model development** Using all the candidate predictors identified previously, we will develop our models using a Cox proportional hazards regression (combined with a non-parametric estimate of the baseline survival) following practical approaches for clinical prediction models [33–35]. If competing risks are prevalent, for example due to the risk of dying from causes other than CVD, then this will be accounted for using sub-distribution (Fine Gray) approaches, with the Aalen-Johansen estimator used to obtain the baseline survival [36]. Model parameter coefficients will be pooled across imputations using Rubin Rules to produce the model.

**Model performance** The main follow-up time-point will be 10 years, but earlier time points (e.g., at 5 years) will also be considered. The choice of the 10-year time point is because NICE guidelines recommend clinicians to offer a statin based on the risk of CVD within 10 years. However, we are also considering shorter-time scale (5 years) as sensitivity analysis and to enable early interventions to reduce the risk of CVD. The initial model will include all candidate predictors (no variable selection). This model will then be compared with a model employing least absolute shrinkage and selection operator (LASSO) for variable selection. Continuous variables will be analyzed on their continuous scale, with non-linear trends modeled using fractional polynomials. Overfitting and optimism are expected to be minimal (due to the large sample size) but will be evaluated using bootstrapping (incorporating all model development steps) and heuristic shrinkage estimates and adjusted for using a uniform shrinkage factor if necessary, to produce the final model.

To evaluate the validity of our data, we will compare the representativeness of our datasets to published CVD

populations by summarizing clinical features. Each model's apparent performance will be evaluated at specific time points of 5 and 10 years post-partum, and, if necessary, recalibration by time will be applied by refitting the models linear predictor to the pseudo-observations at each time-point using a generalized linear model [37], to produce a separate prediction model for each time-point of interest. The ability of the model to correctly classify disease status will be evaluated by calculating the models' discrimination in terms of time-dependent C-statistics and Royston's D statistic; the models' calibration by plotting the observed probability of the outcome against predicted probability (smoothed calibration curves), at particular time-points using the pseudo-value approach [38], alongside summary measures of calibration, and clinical utility across a range of risk thresholds deemed clinically relevant by our user groups.

**The sample size for development** The CPRD pregnancy register captured 5.8 million pregnancies among 2.4 million women in the period January 1987–February 2018 [18]. We will use the register to extract pregnancy data from CPRD for the period 1 of January 2000 and 31 of December 2021. Once the data are obtained, further assessment will be done to ensure the sample size (and outcome event proportion) available meets the minimum sample size that ensures accurate estimation of regression coefficients and reduces overfitting during model development [39, 40]. Given we anticipate a large sample size, it is highly likely to do this. If not, we will reduce the number of candidate predictors accordingly.

#### **External validation of the models**

The SAIL database will be used for external validation to evaluate the predictive performance and clinical utility of the newly developed risk prediction model. The performance will be assessed at various time points as a whole and within important subgroups (e.g., age groups, socio-economic status, and ethnicity) using the following predictive performance measures; calibration, discrimination, and clinical utility (using net benefit analysis and decision curves).

The clinical utility of incorporating the risk prediction model into clinical practice will be assessed using decision curves [41]. The net benefit, which is the fraction of true positives gained by making decisions based on risk predictions over a range of possible risk thresholds will be evaluated [42]. We will define the threshold probability as the population risk of CVD postpartum. The net benefit of the risk prediction model will be compared using a decision curve analysis assuming all are at high risk (“treat all [offer a statin to all people who have

10% or greater risk of developing cardiovascular disease within the next 10 years according to NICE guidelines]”) and assume all are at low risk (“treat none”). We will also compare the net benefit of the risk prediction model with current practice guidelines on postpartum CVD.

Missing data will be handled in a similar way as during the development stage of the risk prediction model.

**The sample size for validation** Using the QRISK<sup>®</sup>-3 risk prediction model for women, we estimated the distribution of the linear predictor and calculated the minimum sample size needed for external validation of QRISK<sup>®</sup>-3 using a recommended simulation-based approach for calculating a risk prediction model with a time-to-event outcome [43]. We established that a minimum sample size of about 24,000 patients and 264 CVD events would result in precise estimates of prediction model performance, for example with a calibration slope CI width of 0.3 (i.e., CI width of 0.85–1.15 assuming the true value is 1), with an assumed 20% censoring rate by 10 years. The validation datasets will be evaluated to confirm the number of CVD events postpartum exceeds 264, but this is expected.

#### **Statistical software**

The computer software programs R version 4.2.1 and Stata (StataCorp. 2021. *Stata Statistical Software: Release 17*. College Station, TX: StataCorp LLC.) will be used for all analyses.

#### **Model presentation**

The whole study will be reported following Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) guidelines [44].

#### **Discussion**

In the proposed study, we aim to evaluate the added value of reproductive and pregnancy related risk factors, which have been shown to be associated with future risk of CVD but have not been taken into account in current risk algorithm for CVD, QRISK<sup>®</sup>-3. Our study will therefore highlight the importance of incorporating reproductive and pregnancy-related risk factors into risk prediction modeling post-partum.

Our study will develop and internally validate the risk prediction model developed using a large cohort of primary care data from CPRD. This implies large sample sizes to enable stability of the parameters estimated. We will also use a separate dataset (SAIL) for external validation of the developed risk prediction model and hence quantify the generalizability of the algorithm in the UK population and within relevant subgroups (e.g., age

groups, socio-economic status, and ethnicity). This will enable us to detect any algorithmic biases and therefore develop frameworks to reduce them.

We foresee some limitations to this study. While the CPRD database is representative of the UK population, recording rates for primary care record codes may vary significantly between practices. We plan to assess the data quality before analysis. We also expect missing data in some of the candidate predictors, and we have outlined a framework to address this challenge during both the development and validation phases. Finally, although data such as genetics could be potential predictors in the risk prediction modelling of CVD, we will not consider them in our study as there is still little information in primary care datasets.

## Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s41512-022-00137-7>.

**Additional file 1.** Read codes to be used to identify patients with CVD from GP records (obtained from QRISK<sup>®</sup>-3 for comparability of models).

## Acknowledgements

Patient representatives and MuM-PreDiCT team.

## Authors' contributions

SW was responsible for the study conceptualization and design and drafted the initial manuscript. FC, ST, DO, CM, SB, CY, KN, and RR were responsible for the study conceptualization and design and revised the manuscript critically for important intellectual content. The authors have approved the final submitted version and are accountable for all aspects of the work.

## Funding

This work is funded by the Strategic Priority Fund "Tackling multimorbidity at scale" programme (grant number MR/W014432/1) delivered by the Medical Research Council and the National Institute for Health Research in partnership with the Economic and Social Research Council and in collaboration with the Engineering and Physical Sciences Research Council. SW PhD studentship is funded by the British Heart Foundation (BHF) Data Science Centre (BHF grant number SP/19/3/34678, awarded to Health Data Research (HDR)). His PhD is also supported through the HDR-UK-Turing Wellcome PhD Programme.

The funders will have no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

## Availability of data and materials

The datasets that will be used for this study are available from CPRD and SAIL and can be accessed upon reasonable request.

## Declarations

### Ethics approval and consent to participate

CPRD has ethics approval from the Health Research Authority to support research using anonymized patient data. This study protocol has been submitted to CPRD for approval by Independent Scientific Advisory Committee for CPRD. Protocol number: 22\_001909.

### Consent for publication

As the study data are de-identified, consent for publication is not required.

### Competing interests

The authors declare that they have no competing interests.

## Author details

<sup>1</sup>Institute of Applied Health Research, College of Medical and Dental Sciences, University of Birmingham, Edgbaston, Birmingham, UK. <sup>2</sup>WHO Collaborating Centre for Global Women's Health, Institute of Metabolism and Systems Research, University of Birmingham, Birmingham, UK. <sup>3</sup>Department of Obstetrics and Gynaecology, Birmingham Women's and Children's NHS Foundation Trust, Birmingham, UK. <sup>4</sup>Centre for Public Health, Queen's University Belfast, Belfast, UK. <sup>5</sup>School of Medicine, University of St Andrews, St Andrews, UK. <sup>6</sup>Data Science, Medical School, Swansea University, Swansea, UK. <sup>7</sup>Big Data Institute, University of Oxford, Li Ka Shing Centre for Health Information and Discovery, Old Road Campus, Oxford OX3 7LF, UK. <sup>8</sup>Nuffield Department of Women's & Reproductive Health, University of Oxford, Level 3 Women's Centre, John Radcliffe Hospital, Oxford OX3 9DU, UK. <sup>9</sup>Health Data Research, London, UK.

Received: 2 August 2022 Accepted: 7 December 2022

Published online: 19 December 2022

## References

- Appelman Y, et al. Sex differences in cardiovascular risk factors and disease prevention. *Atherosclerosis*. 2015;241(1):211–8.
- Roth GA, et al. Global, regional, and national burden of cardiovascular diseases for 10 causes, 1990 to 2015. *J Am College Cardiol*. 2017;70(1):1–25.
- Wilson PW, et al. Prediction of coronary heart disease using risk factor categories. *Circulation*. 1998;97(18):1837–47.
- 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:j2099. <https://doi.org/10.1136/bmj.j2099>.
- Baart SJ, et al. Cardiovascular risk prediction models for women in the general population: a systematic review. *PLoS one*. 2019;14(1):e0210329.
- Damen JAAG, Hooft L, Schuit E, Debray TPA, Collins GS, Tzoulaki I, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. *BMJ*. 2016;353:i2416. <https://doi.org/10.1136/bmj.i2416>.
- Umesawa M, Kobashi G. Epidemiology of hypertensive disorders in pregnancy: prevalence, risk factors, predictors and prognosis. *Hypertens Res*. 2017;40(3):213–20.
- Williams D. Pregnancy: a stress test for life. *Curr Opin Obstetr Gynecol*. 2003;15(6):465–71.
- Bellamy L, et al. Pre-eclampsia and risk of cardiovascular disease and cancer in later life: systematic review and meta-analysis. *Bmj*. 2007;335(7627):974.
- Grandi SM, et al. Cardiovascular disease-related morbidity and mortality in women with a history of pregnancy complications: systematic review and meta-analysis. *Circulation*. 2019;139(8):1069–79.
- O'Kelly AC, et al. Pregnancy and reproductive risk factors for cardiovascular disease in women. *Circ Res*. 2022;130(4):652–72.
- Haas DM, et al. Pregnancy as a window to future cardiovascular health: design and implementation of the nuMoM2b Heart Health Study. *Am J Epidemiol*. 2016;183(6):519–30.
- Rich-Edwards JW, et al. Pregnancy characteristics and women's future cardiovascular health: an underused opportunity to improve women's health? *Epidemiol Rev*. 2014;36(1):57–70.
- Markovitz AR, et al. Does pregnancy complication history improve cardiovascular disease risk prediction? Findings from the HUNT study in Norway. *Eur Heart J*. 2019;40(14):1113–20.
- Saei Ghare Naz M, Sheidaei A, Aflatounian A, Azizi F, Ramezani Tehrani F. Does Adding Adverse Pregnancy Outcomes Improve the Framingham Cardiovascular Risk Score in Women? Data from the Tehran Lipid and Glucose Study. *J Am Heart Assoc*. 2022;11(2):e022349. <https://doi.org/10.1161/JAHA.121.022349>.
- Timpka S, et al. The value of pregnancy complication history for 10-year cardiovascular disease risk prediction in middle-aged women. *Eur J Epidemiol*. 2018;33(10):1003–10.
- Clinical Practice Research Datalink. CPRD GOLD February 2021 (Version 2021.02.001) [Data set]. Clinical Practice Research Datalink. 2021. <https://doi.org/10.48329/S0M3-8M14>.

18. Minassian C, et al. Methods to generate and validate a pregnancy register in the UK clinical practice research Datalink primary care database. *Pharmacoepidemiol Drug Safe*. 2019;28(7):923–33.
19. Ford DV, et al. The SAIL Databank: building a national architecture for e-health research and evaluation. *BMC Health Serv Res*. 2009;9(1):1–12.
20. Jones KH, Ford DV, Thompson S, Lyons RA. A Profile of the SAIL Databank on the UK Secure Research Platform. *Int J Popul Data Sci*. 2019;4(2):1134. <https://doi.org/10.23889/ijpds.v4i2.1134>.
21. Lee SI, et al. Epidemiology of pre-existing multimorbidity in pregnant women in the UK in 2018: a population-based cross-sectional study. *BMC Preg Childbirth*. 2022;22(1):1–15.
22. Smith GN, Louis JM, Saade GR. Pregnancy and the postpartum period as an opportunity for cardiovascular risk identification and management. *Obstet Gynecol*. 2019;134(4):851–62.
23. Brodribb WE, Mitchell BL, Van Driel ML. Continuity of care in the post partum period: general practitioner experiences with communication. *Aust Health Rev*. 2015;40(5):484–9.
24. NHS, N.B. What are the read codes? *Health Lib Rev*. 1994;11(3):177–82.
25. Shipe ME, et al. Developing prediction models for clinical use using logistic regression: an overview. *J Thorac Dis*. 2019;11(Suppl 4):S574.
26. Okoth K, Chandan JS, Marshall T, Thangaratinam S, Thomas GN, Nirantharakumar K, Adderley NJ. Association between the reproductive health of young women and cardiovascular disease in later life: umbrella review. *BMJ*. 2020;371:m3502. <https://doi.org/10.1136/bmj.m3502>. Erratum in: *BMJ*. 2020;371:m3963.
27. Pavlou M, et al. Review and evaluation of penalised regression methods for risk prediction in low-dimensional data with few events. *Stat Med*. 2016;35(7):1159–77.
28. Riley RD, et al. Penalization and shrinkage methods produced unreliable clinical prediction models especially when sample size was small. *J Clin Epidemiol*. 2021;132:88–96.
29. Sperrin M, et al. Missing data should be handled differently for prediction than for description or causal explanation. *J Clin Epidemiol*. 2020;125:183–7.
30. Hippisley-Cox J, Stables D, Pringle M. QRESEARCH: a new general practice database for research. *Inform Prim Care*. 2004;12(1):49–50.
31. Chen IY, et al. Ethical machine learning in healthcare. *Ann Rev Biomed Data Sci*. 2021;4:123–44.
32. Tillin T, et al. Ethnicity and prediction of cardiovascular disease: performance of QRISK2 and Framingham scores in a UK tri-ethnic prospective cohort study (SABRE—Southall And Brent REvisited). *Heart*. 2014;100(1):60–7.
33. Royston P, Moons KG, Altman DG, Vergouwe Y. Prognosis and prognostic research: Developing a prognostic model. *BMJ*. 2009;338:b604. <https://doi.org/10.1136/bmj.b604>.
34. Steyerberg EW. *Clinical prediction models*: Springer; 2019.
35. Riley RD, et al. *Prognosis research in healthcare: concepts, methods, and impact*: Oxford University Press; 2019.
36. Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. *Stat Med*. 2007;26(11):2389–430.
37. Royston P. Tools for checking calibration of a cox model in external validation: prediction of population-averaged survival curves based on risk groups. *Stat J*. 2015;15(1):275–91.
38. Andersen PK, Pohar Perme M. Pseudo-observations in survival analysis. *Stat Methods Med Res*. 2010;19(1):71–99.
39. Riley RD, Ensor J, Snell KIE, Harrell FE Jr, Martin GP, Reitsma JB, Moons KGM, Collins G, van Smeden M. Calculating the sample size required for developing a clinical prediction model. *BMJ*. 2020;368:m441. <https://doi.org/10.1136/bmj.m441>.
40. Riley RD, et al. Minimum sample size for developing a multivariable prediction model: PART II-binary and time-to-event outcomes. *Stat Med*. 2019;38(7):1276–96.
41. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Dec Making*. 2006;26(6):565–74.
42. Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ*. 2016;352:i6. <https://doi.org/10.1136/bmj.i6>.
43. Riley RD, Collins GS, Ensor J, Archer L, Booth S, Mozumder SI, Rutherford MJ, van Smeden M, Lambert PC, Snell KIE. Minimum sample size calculations for external validation of a clinical prediction model with a time-to-event outcome. *Stat Med*. 2022;41(7):1280–95. <https://doi.org/10.1002/sim.9275>.
44. Collins GS, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *J Bri Surg*. 2015;102(3):148–58.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

