Impact of sample size on the stability of risk scores from clinical prediction models: a case study in cardiovascular disease

Alexander Pate (alexander.pate@manchester.ac.uk)
The University of Manchester https://orcid.org/0000-0002-0849-3458
Richard Emsley
King's College London
Matthew Sperrin
The University of Manchester
Glen P. Martin
The University of Manchester
Tjeerd van Staa
The University of Manchester

Research

Keywords: risk prediction, sample size, statistical methods, precision, stability

Posted Date: February 28th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-15416/v1

Version of Record: A version of this preprint was published on September 9th, 2020. See the published version at https://doi.org/10.1186/s41512-020-00082-3.

Abstract

Background

Stability of risk estimates from prediction models may be highly dependent on the sample size of the dataset available for model derivation. In this paper, we evaluate the stability of cardiovascular disease risk scores for individual patients when using different sample sizes for model derivation; such sample sizes include those similar to models recommended in national guidelines, and those based on recently published sample size formula for prediction models.

Methods

We mimicked the process of sampling N patients from a population to develop a risk prediction model by sampling patients from the Clinical Practice Research Datalink. A cardiovascular disease risk prediction model was developed on this sample and used to generate risk scores for an independent cohort of patients. This process was repeated 1000 times, giving a distribution of risks for each patient. N = 100 000, 50 000, 10 000 and N min (derived from sample size formula) were considered. The 2.5 – 97.5 percentile range of risks across these models was used to evaluate instability. Patients were grouped by a risk derived from a model developed on the entire population (population derived risk) to summarise results.

Results

For a sample size of 10 000, the median 2.5 – 97.5 percentile range of risks for patients across the 1000 models was approximately 60% of their population derived risk. For example, for patients with a population derived risk of 9 - 10% or 19 - 20%, the median percentile range was 6.25% and 12.59% respectively. Using the formula derived sample size, the range was approximately 170% of their average risk score. Restricting this analysis to models with high discrimination or good calibration reduced the percentile range, but high levels of instability remained.

Conclusions

Widely used cardiovascular disease risk prediction models suffer from high levels of instability induced by sampling variation. Stability of risk estimates should be a criterion when determining the minimum sample size to develop models.

Background

Risk prediction models are used to guide clinical decision-making in a variety of disease areas and settings, ranging from the prevention of cardiovascular disease (CVD) in primary care to intensive care unit based models such as APACHE or SOFA.(1–5) As such, developing risk prediction models appropriately is vital. One aspect of appropriate derivation of prediction models is ensuring sufficient sample size in the development dataset; unfortunately, sample size calculations for models are often not made, or at best are based on the simplistic “10 events per predictor” rule.(6) Risk prediction models that are recommended in treatment guidelines for routine use by clinicians often vary in sample sizes. As an example, QRISK3(7) (recommended by the National Institute for Health and Care Excellence to guide CVD primary prevention in England(8)) was developed on a cohort of 4 019 956 females and 3 869 847 males, whereas the pooled cohort equations (recommended by American

College of Cardiology and American Heart Association to guide CVD prevention in the US(9)) were based on 9098 females and 11 240 males for white ethnicity, and 2641 females and 1647 males for African-American ethnicity.

If the sample size is too small, the most commonly cited issue is that of overfitting, which may cause extreme predictions outside of the development data set. Another potential issue, of which the implications are less clear, is that small sample sizes could lead to instability in the risk scores of individuals depending on which sample of the population is used for model development. By stability, we mean how risk scores for a given individual vary when generated from different prediction models. It is well known that differently defined prediction models may produce different risks for individuals, even if the models perform similarly on the population level.(10–14) However, if a patient's risk score, and therefore treatment decision, is highly unstable due to sample size, this is undesirable. In this scenario, the instability of a patient's risk score is driven by statistical uncertainty around the risk estimate of the subgroup which that patient belongs to, distinguishing this from the reference class problem.(14) Therefore it is important to minimise this instability if wanting to base clinical decisions on risk scores generated from such models.

The aim of this study was to evaluate the stability of CVD risk predictions for individual patients when using different sample sizes in the development of the risk prediction models (including a recommended minimum sample size from work focusing on the issue of overfitting, representing state of the art techniques for sample size calculations in risk prediction models).(15)

Methods

Data source

We defined two cohorts from a Clinical Practice Research Datalink (CPRD)(16) dataset, which comprised primary care data linked with Hospital Episode Statistics(17) (HES), and mortality data provided by the Office for National Statistics(18) (ONS). For the first cohort (referred to as historical cohort) the cohort entry date was the latest of: attaining age 25 years; attaining one year follow up as a permanently registered patient in CPRD; or 1st Jan 1998. The end of follow up was the earliest date of: patient's transfer out of the practice or death; last data collection for practice; or 31st Dec 2015. Patients were excluded if they had a CVD event (identified through CPRD, HES or ONS) or statin prescription prior to their cohort entry date (code lists available in additional file 1). The second cohort comprised patients actively registered on 1st Jan 2016 (referred to as contemporary cohort). This cohort of patients represents a contemporary population, for whom a risk prediction model would subsequently be applied to estimate their CVD risks. To be eligible for this second cohort, a patient had to be aged 25–85 years on 1st Jan 2016, and be actively registered in CPRD with one year prior follow up with no history of CVD or statin treatment.

Overview

We mimicked the process of sampling an overarching target population for the development of a risk prediction model by randomly sampling N patients from the historical cohort. A risk prediction model was developed on this sample and used to generate risk scores for the contemporary cohort. This process was repeated 1000 times, giving 1000 risk scores for each patient, for each sample size. The sample sizes considered were N = 10 000, 50 000, 100 000 and N_{min} (minimum sample size required to meet criteria outlined by Riley et al.(15)). We

chose 10 000 as it is similar to the number of females and males used to develop ASSIGN(19) (6540 and 6757), Framingham(20) (3969 and 4522) and Pooled Cohort Equations(9) (9098 and 11 240). The upper limit of 100 000 was chosen to match the SCORE(21) equations, which were developed on 117 098 and 88 080 females and males respectively. Derivation of N_{min} = 1434 (female) and 1405 (male) is described in additional file 2.

Generation of risk scores

The historical cohort and contemporary cohort were both split into female and male cohorts and imputed using one stochastic imputation using the mice package.(22) All variables included in QRISK3(7), including the Nelson Aalen estimate of the baseline cumulative hazard at the event time and the outcome indicator, were included in the imputation process. The following process was then carried out separately for females and males: 100 000 individuals were chosen at random from the historical cohort to form an internal validation cohort, the remaining individuals formed the development cohort. The development cohort was then viewed as the population. For each value of N, we sampled N patients from this population without replacement, 1000 times.

The following process was repeated within each sample. A Cox model was fit to the sampled data, where the outcome was defined as the time until the first CVD event. Predictor variables included in the model were continuous variables, and categorical variables with > 1% prevalence in all categories calculated from the entire development cohort (age, body mass index, cholesterol/high density lipoprotein ratio, family history of CVD, treated hypertension, smoking status, systolic blood pressure, Townsend deprivation index and type 2 diabetes). This set of variables reflects the smaller number of variables used in models with lower sample sizes in practice. (9,19,20) The developed model was used to generate 10 year risk scores for the contemporary cohort. Harrell's C(23) statistic for this model, and the calibration-in-the-large (mean predicted risk – observed/Kaplan Meier risk) were calculated in the validation cohort. A graphical representation of this process is given in Fig. 1.

Finally, we calculated a 10 year risk for each patient in the contemporary cohort using a model developed on the entire development cohort, called the population derived risk, and also calculated the Harrell's C and calibration-in-the-large of this model in the validation cohort.

Analysis of stability of risk scores

For each sample size, four different analyses were carried out to summarise the stability of risks across the 1000 models. First, the 2.5–97.5 percentile range of risks was calculated for each patient across the 1000 models. The distribution of these ranges was then plotted in box plots stratified by the population derived risk. Second, we split the models into three groups of equal size that had the lowest, medium or highest C statistics. We then calculated the 2.5–97.5 percentile range of risks within these subsets of models, and presented in box plots stratified by population derived risk. This allowed us to explore whether models with high C statistics had more stability than those with lower C statistics. Third, we split the models into groups defined by their calibration-in-the-large, and presented boxplots of the 2.5–97.5 percentile range of risks within these subsets of models. Here, the groups were defined as models with calibration-in-the-large deviating from the population derived model by less than 0.1%, 0.25%, 0.5%, and then all models. This allowed us to explore how much of the instability of the risk scores was driven by variation in overall calibration. Finally, we grouped patients into risk groups of width 1% as defined by their population derived risk. The proportion of the 1000 models that classified a patient above/below the 10% risk threshold (threshold for statin eligibility according to the recommended guidelines in the UK(8)) was then calculated.

Results

The baseline characteristics for the female development cohort, validation cohort, and the contemporary cohort are provided in Table 1. See additional file 3 for the equivalent table for the male cohort.

Table 1
Baseline characteristics of each female cohort

		Development (n = 1 865 079)	Validation (n = 100 000)	Contemporary (n = 387 557)
Outcome	CVD events	82 065	4482	NA
	Follow up (years)	13 098 449	703 471	NA
Age		43.07 (15.94)	43.14 (15.96)	48.38 (14.43)
Systolic blood pressure		123.91 (18.28)	124 (18.22)	123.97 (15.17)
Body mass index		25.6 (5.60)	25.56 (5.56)	27.1 (6.31)
Cholesterol/high density lipoprotein ratio		3.72 (1.20)	3.72 (1.21)	3.46 (1.04)
Smoking status	Never	56.04	56.15	46.05
	Ex	16.97	16.98	31.66
	Current	27.00	26.87	22.29
Townsend	1 (least deprived)	21.96%	21.96%	24.95%
	2	21.99%	21.81%	22.35%
	3	21.17%	21.46%	21.56%
	4	20.46%	20.36%	18.70%
	5 (most deprived)	14.42%	14.41%	12.44%
Treated hypertension		6.18%	6.19%	8.45%
Family history of CVD		15.08%	15.13%	20.86%
Type 2 diabetes		1.16%	1.19%	1.15%

The distribution of the C statistic and the calibration-in-the-large of the 1000 models are given in Table 2. The 97.5th percentile of C statistics was similar for each sample size, but as the sample size decreased, the 2.5th percentile got smaller (0.802 vs 0.868 female and 0.805 vs 0.843 male). All C statistics in the 2.5–97.5 percentile range were > 0.8. The variation in the calibration-in-the-large decreased as the sample size increased. The 2.5–

97.5 percentile ranges of the calibration-in-the-large values was 2.61% (female) and 3.12% (male) for N = N_{min}, decreasing to 0.32% (female) and 0.36% (male) for N = 100 000.

Table 2
Quantiles of C statistics and calibration-in-the-large of the 1000 models, for each sample size

		Quantiles of C statistic					Quantiles of calibration-in-the-large (as a %)				
	Sample size	2.5%	25%	50%	75%	97.5%	2.5%	25%	50%	75%	97.5%
Female	N_{min}	0.802	0.852	0.857	0.861	0.864	-2.22	-1.43	-0.95	-0.47	0.39
	10000	0.865	0.866	0.867	0.867	0.868	-1.45	-1.13	-0.95	-0.78	-0.44
	50000	0.867	0.868	0.868	0.868	0.868	-1.18	-1.03	-0.95	-0.87	-0.73
	100000	0.868	0.868	0.868	0.868	0.868	-1.11	-1.01	-0.96	-0.90	-0.79
Male	N_{min}	0.805	0.827	0.831	0.835	0.839	-2.56	-1.49	-1.01	-0.45	0.56
	10000	0.840	0.841	0.842	0.843	0.843	-1.61	-1.20	-1.01	-0.80	-0.39
	50000	0.843	0.843	0.843	0.843	0.844	-1.28	-1.11	-1.02	-0.93	-0.77
	100000	0.843	0.843	0.843	0.844	0.844	-1.21	-1.08	-1.02	-0.95	-0.85

*C statistics of population models in the validation dataset are 0.868 (female) and 0.844 (male). Calibration-in-the-large of the population models in the validation dataset are − 0.95% (female) and − 1.02% (male).

Figure 3 plots the 2.5 – 97.5 percentile range in risks for patients across models stratified by the C statistic of the models (female cohort, N = 10 000). The median 2.5 - 97.5 percentile range for models with high C statistics was 2.42%, 5.02%, 7.60% and 10.20% for patients in the respective risk groups. This equates to an 18 – 20% reduction in the median percentile range when using well discriminating models compared to all models (2.98%, 6.25%, 9.46% and 12.59%). Results for other sample sizes presented in additional file 3.

Figure 4 plots the 2.5 – 97.5 percentile range in risks for patients across models stratified by the calibration-in-the-large of the models (female cohort, N = 10 000). The median 2.5 - 97.5 percentile range across models with the best calibration-in-the-large was 2.72%, 5.70%, 8.72% and 11.69%, for the respective risk groups. This equates to a 7-9% reduction in the median percentile range compared to when using all models (2.98%, 6.25%, 9.46% and 12.59%). Results for other sample sizes presented in additional file 3.

Table 3 shows the probability that a patient from a given risk group (according to population derived model) may be classified on the opposite side of the 10% threshold by a randomly chosen model. For example when using a sample size of N_{min}, 26.91% of patients with a population derived risk between 14–15% would be classified as having a risk below 10%, whereas this is only 2.50% for N = 10 000, 0.01% for 50 000 and < 0.01% for 100 000.

Table 3

probability of being classified above/below the treatment threshold, stratified by population derived risk

		Population derived risk									
	Sample size	5–6%	6–7%	7–8%	8–9%	9–10%	10–11%	11–12%	12–13%	13–14%	14–15%
Female	N_{min}	6.46	12.55	20.49	29.63	38.69	52.48	44.46	37.72	31.95	26.91
	10,000	0.08	0.74	4.25	15.24	35.07	41.17	22.55	11.40	5.50	2.50
	50,000	0.00	0.00	0.08	2.29	24.49	27.67	4.44	0.46	0.06	0.01
	100,000	0.00	0.00	0.00	0.50	18.56	21.50	1.09	0.04	0.00	0.00
Male	N_{min}	4.32	9.98	18.18	28.54	39.14	50.87	41.97	34.37	28.13	22.97
	10,000	0.03	0.33	2.51	12.40	34.43	38.89	18.84	8.13	3.33	1.34
	50,000	0.00	0.00	0.02	1.28	21.80	26.51	3.07	0.26	0.03	0.00
	100,000	0.00	0.00	0.00	0.23	16.02	19.79	0.63	0.01	0.00	0.00

*For patients with a population derived risk < 10%, the probabilities represent the chance of being classified above the threshold, for patients with a population derived risk > 10%, the probabilities represent the chance of being classified below the threshold.

Discussion

This study found that at sample sizes typically used for developing risk models (e.g. in the CVD domain, the pooled cohort equations(9) and ASSIGN(19) were based on approximately 10 000 individuals or less), there is substantial instability in risk estimates attributable to sampling error. Furthermore, when restricting the analysis to models with high discrimination or good calibration, high levels of instability remained.

This variability in individual risk is especially relevant if using the model to make clinical decisions based on whether a risk score is above or below a fixed threshold (a common use for risk prediction models). From an individual's and clinician's perspective, it is unsatisfactory that a different treatment decision may be made depending on the model used. However this is also an issue at the population level. Consider statin therapy in the UK. Initiating statins in patients who have a 10-year risk of CVD > 10% has been shown to be cost effective. (24) This intervention becomes more cost effective the better the performance (calibration and discrimination) of the model used to calculate the risk scores. Sample size is strongly correlated with model performance, and a small sample size will likely lead to a poorly performing model, and less events prevented. However, it is difficult to assess when increasing sample size will improve model performance, given that model performance is affected by many other factors (prevalence of outcome, inclusion of important predictors, strength of association between predictors and outcome). Sample size affects model performance through the precision of coefficients, and imprecise estimates will cause the risk of fixed subgroups in the population to be miss-calculated (the central theme of this paper). Therefore, if the coefficients are precise, and risk estimates are stable, one will unlikely be able to improve model performance by increasing the sample size unless doing so allows for incorporating more predictors. The stability of risk scores (and ultimately precision of coefficients) could therefore be used as a proxy to determine whether increasing sample size will improve model

performance. When N = 10 000 we see levels of instability that indicate the performance of the model could be improved by increasing sample size, resulting in fewer CVD events.

At the sample size suggested by Riley et al.,(15) the instability in risk is even higher and the issues are heightened. However, there are no CVD risk prediction models used in practice with such small sample sizes, so the implications are more general. There is often ample data to produce CVD risk prediction models; however this may not be the case for other disease areas, where the outcomes are not well recorded in routinely collected datasets. In this scenario one may have to actively recruit patients into a cohort and the work by Riley et al.,(15) could be used in order to derive a sample size. We propose that if risk scores from a model are going to be used to drive clinical decision making above or below a fixed threshold, Sect. 6 of Riley et al.,(15) "Potential additional criterion: precise estimates of predictor effects" should be properly considered. It is imprecise estimates of the predictor effects that leads to instability of risk scores. If this criterion is not met, as is the case for $N = N_{min}$ in this paper, risks scores have high levels of instability and models poorer performance. The number of patients required to ensure stable risk scores will depend on the prevalence of the outcome, the number of predictors and the strength of the association between outcomes and predictors among other things, and therefore will vary for each model.

In practice, to ascertain whether a given development cohort has a sufficient sample size, the process undertaken in this manuscript could be replicated using bootstrap resampling methods. Instead of sampling the population without replacement (not possible in practice), sampling the development cohort with replacement (i.e. bootstrapping) can replicate this process and one could obtain a similar range of risks for each patient. The stability of the risk scores could then be assessed, and a decision made on whether more patients should be recruited. One proposal on how to use this information to determine a sufficient sample size could be to ensure the bootstrapped 2.5–97.5 percentile range for all patients must be smaller than x% of their estimated risk. Another proposal may be to ensure that for patients whose estimates are a certain distance away from a treatment threshold, that there is a less than an x% chance of deriving a risk on the other side of the treatment threshold if one resampled.

There are some limitations that warrant discussion. The first is that the calibration-in-the-large of the population derived model was poor. We don't believe this is a problem as a similar miss calibration-in-the-large is found in QRISK3,(7) despite the model being well calibrated within risk deciles. It is likely caused by incompatible assumptions under how the observed risks (Kaplan Meier assumes unconditional independent censoring) and predicted risks (Cox model assumes independent censoring only after conditioning on the covariates) are estimated. When looking within risk deciles, the difference in assumptions is not as large and good calibration was found. Centring these measurements thus allowed the evaluation of whether the instability in risk was being driven by over and under predicting models. A second limitation was that one may argue that variation in predicted risk was observed because the proper process for deriving risk prediction models wasn't followed. We didn't do this as it would have resulted in different variables and non-linear terms being selected across the models, and we believe this would have increased the variation in risks across the models, rather than reduce it. Finally, this study concerned the outcome CVD and used a specific set of variables for prediction. However the results are likely to be generalizable to other disease areas as the study evaluated the effects of random variability in sampling.

Conclusions

In conclusion, CVD risk prediction models developed on randomly sampled cohorts of size 10 000 or less suffer from high levels of instability in individual risk predictions. There are multiple models used in practice that are developed on sample sizes this small. To avoid this, models should be developed on larger cohorts such as the QRISK3(7) and SCORE(21) models. More generally, if developing a risk prediction model to guide treatment for patients above a fixed threshold, consideration should be given to the stability of risks scores and precision of effect estimates when choosing a sample size.

Additional Files

File name: Additional file 1.txt

Title: Predictor variable information and code lists

Description: More detailed information on how variables were extracted from the electronic health record to be used for analysis, including code lists.

File name: Additional file 2.txt

Title: Calculation of Nmin

Description: Calculation of the minimum required sample size according to published sample size formula references in the manuscript. Separate calculations for male and female cohorts.

File name: Additional file 3.txt

Title: Supplementary tables and figures

Description: Baseline demographics of male cohorts and results from simulations that could not be included in the main manuscript for space reasons.

Declarations

Ethical approval and consent to participate

The study was approved by the independent scientific advisory committee for Clinical Practice Research Datalink research (protocol no. 17_125RMn2.). The interpretation and conclusions contained in this study are those of the authors alone.

Consent for publication

Not applicable

Availability of data and materials

The datasets generated and/or analysed during the current study are not publicly available as this would be a breach of the contract with CPRD. However it can be obtained by a separate application to CPRD after getting

approval from Independent Scientific Advisory Committee (ISAC). To apply for data follow the instructions here: https://www.cprd.com/research-applications .

Competing interests

All authors state they have nothing to disclose.

Funding

This project was funded by the MRC, grant code: MR/N013751/1. The funder played no other role in the study.

Acknowledgements

This study is based in part on data from the Clinical Practice Research Datalink obtained under licence from the UK Medicines and Healthcare products Regulatory Agency. The data were 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) were re-used with the permission of The Health & Social Care Information Centre. All rights reserved.

Authorship statement

AP lead conception and design of the study, acquired the data, ran all analyses, lead interpretation of results and drafted the article

RE was involved in conception and design of the study, acquiring the data, interpretation of results and made significant revisions to the article, and gave final approval for submission

MS was involved in conception and design of the study, interpretation of results and made significant revisions to the article, and gave final approval for submission

GM was involved in conception and design of the study, interpretation of results and made significant revisions to the article, and gave final approval for submission

TVS was involved in conception and design of the study, acquiring the data, interpretation of results and made significant revisions to the article, and gave final approval for submission

FUNDING

This work was supported by the Medical Research Council [MR/N013751/1]

References

1. 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 [Internet]. 2016;353(1 Pt 2):i2416. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27184143%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4868251

2. Tangri N, Kitsios GD, Inker LA, Griffith J, Naimark DM. Risk Prediction Models for Patients With Chronic Kidney Disease: A systematic Review. Ann Intern Med. 2014;158(8):596–603.
3. Abbasi A, Peelen LM, Corpeleijn E, Van Der Schouw YT, Stolk RP, Spijkerman AMW, et al. Prediction models for risk of developing type 2 diabetes: Systematic literature search and independent external validation study. BMJ. 2012;345(7875):1–16.
4. Jentzer JC, Bennett C, Wiley BM, Murphree DH, Keegan MT, Gajic O, et al. Predictive value of the Sequential Organ Failure Assessment score for mortality in a contemporary cardiac intensive care unit population. J Am Heart Assoc. 2018;7(6).
5. Knaus WA, Wagner DP, Draper EA, Zimmerman JE, Bergner M, Bastos PG, et al. The APACHE III prognostic system: Risk prediction of hospital mortality for critically III hospitalized adults. Chest [Internet]. The American College of Chest Physicians; 1991;100(6):1619–36. Available from: http://dx.doi.org/10.1378/chest.100.6.1619
6. Van Smeden M, De Groot JAH, Moons KGM, Collins GS, Altman DG, Eijkemans MJC, et al. No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. BMC Med Res Methodol [Internet]. BMC Medical Research Methodology; 2016;16(1):1–12. Available from: http://dx.doi.org/10.1186/s12874-016-0267-3
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 [Internet]. 2017;357(3):j2099. Available from: http://dx.doi.org/10.1136/bmj.j2099
8. NICE. Cardiovascular disease: risk assessment and reduction, including lipid modification [Internet]. 2016 [cited 2018 May 3]. Available from: https://www.nice.org.uk/guidance/cg181/chapter/1-recommendations
9. Goff DC, Lloyd-jones DM, Bennett G, Coady S, Agostino RBD, Gibbons R, et al. 2013 ACC/AHA guideline on the assessment of cardiovascular risk. Circulation. 2014.
10. Stern RH. Individual Risk. J Clin Hypertens. 2012;14(4):261–4.
11. Kent DM, Box G. Risk Models and Patient-Centered Evidence: Should Physicians Expect One Right Answer? Health Policy (New York). 2012;307(15):1585–6.
12. Steyerberg EW, Eijkemans MJC, Boersma E, Habbema JDF. Equally valid models gave divergent predictions for mortality in acute myocardial infarction patients in a comparison of logical regression models. J Clin Epidemiol. 2005;58(4):383–90.
13. Lemeshow S, Klar J TD. Outcome prediction for individual intensive care patients: useful, misused, or abused? Intensive Care Med. 1995;21:770–6.
14. Hájek A. The reference class problem is your problem too. Synthese. 2007;156(3):563–85.
15. Riley RD, Snell KIE, Ensor J, Burke DL, Harrell FE, Moons KGM, 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.
16. Herrett E, Gallagher AM, Bhaskaran K, Forbes H, Mathur R, Staa T van, et al. Data Resource Profile: Clinical Practice Research Datalink (CPRD). Int J Epidemiol. 2015;44(3):827–36.
17. Digital N. Hospital Episode Statistics [Internet]. [cited 2018 May 3]. Available from: https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics
18. Office for National Statistics [Internet]. [cited 2020 Jan 17]. Available from: https://www.ons.gov.uk/

19. Woodward M, Brindle P, Tunsfall-Pedoe H. Adding social deprivation and family history to cardiovascular risk assessment: The ASSIGN score from the Scottish Heart Health Extended Cohort (SHHEC). Heart. 2007;93(2):172–6.
20. D'Agostino RB, Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, et al. General cardiovascular risk profile for use in primary care: The Framingham heart study. Circulation. 2008;117(6):743–53.
21. Conroy RM, Pyörälä K, Fitzgerald AP, Sans S, Menotti A, De Backer G, et al. Estimation of ten-year risk of fatal cardiovascular disease in Europe: The SCORE project. Eur Heart J. 2003;24(11):987–1003.
22. Groothuis-oudshoorn K. mice : Multivariate Imputation by Chained. 2011;45(3).
23. Harrell FEH, Lee KL, Mark DB. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 1996;15:361–87.
24. NICE. CG181 Lipid modification Appendicies - Cardiovascular risk assessment and the modification of blood lipids for the primary and secondary prevention of cardiovascular disease. 2014; Available from: https://www.nice.org.uk/guidance/cg181/evidence/lipid-modification-update-appendices-pdf-243786638

Abbreviations

Cardiovascular disease, CVD; Clinical Practice Research Datalink, CPRD; Hospital Episode Statistics, HES; Office for National Statistics, ONS

Figures

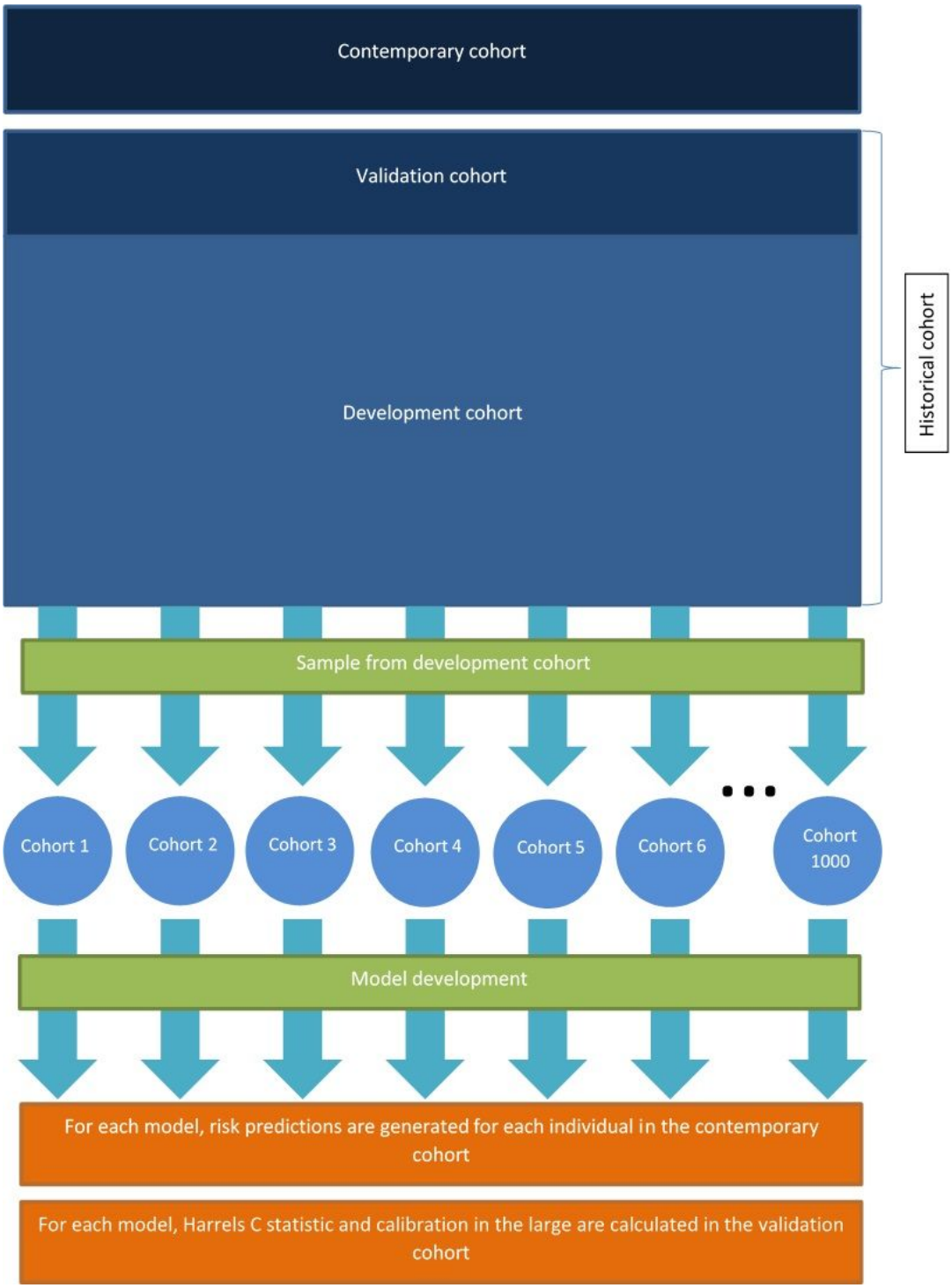

Figure 1

A graphical representation of the sampling process

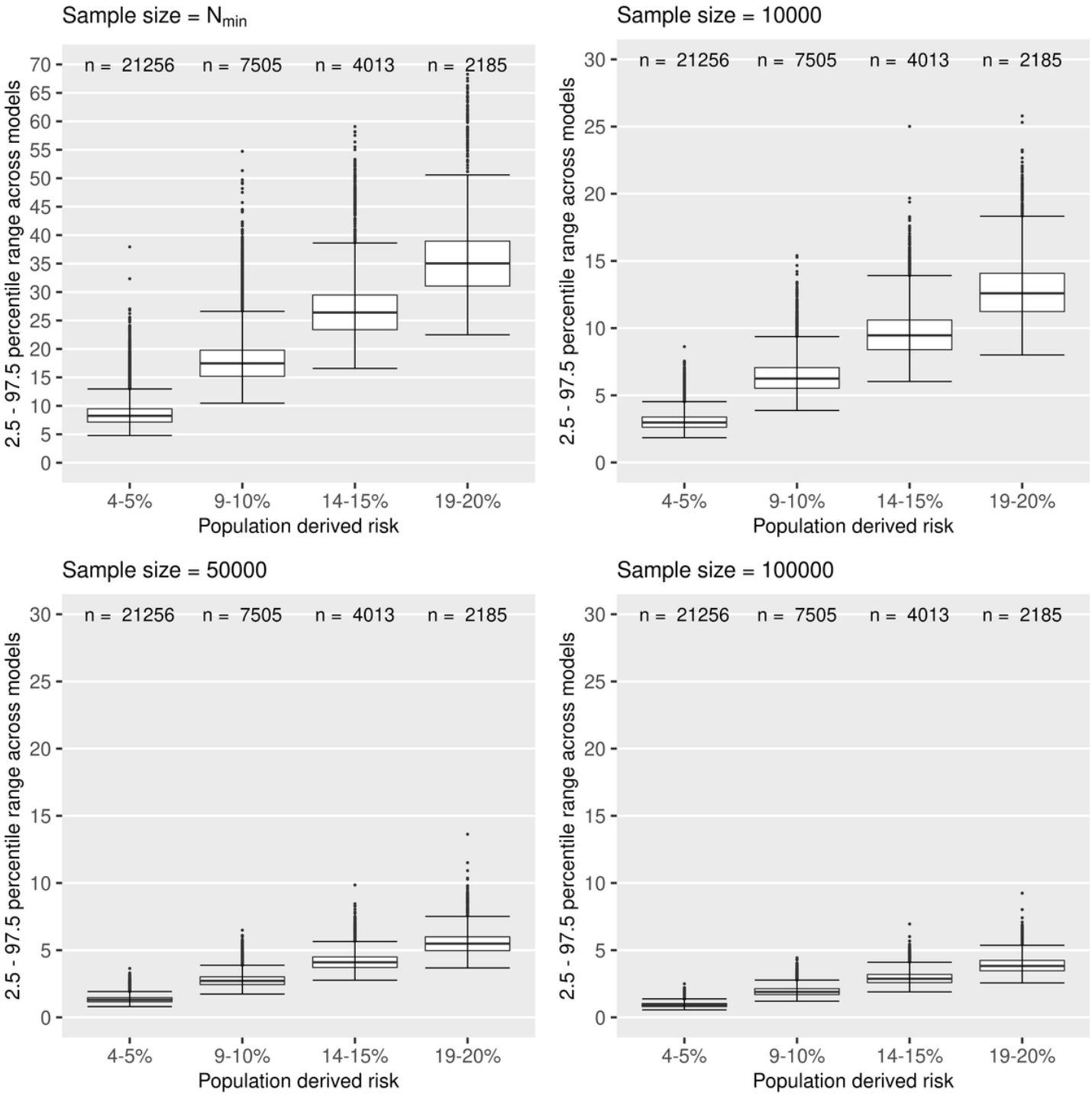

Figure 2

Boxplots of the percentile ranges of risk for individuals across the 1000 models (female cohort). Each data point represents the 2.5-97.5 percentile range in risk for an individual across the 1000 models

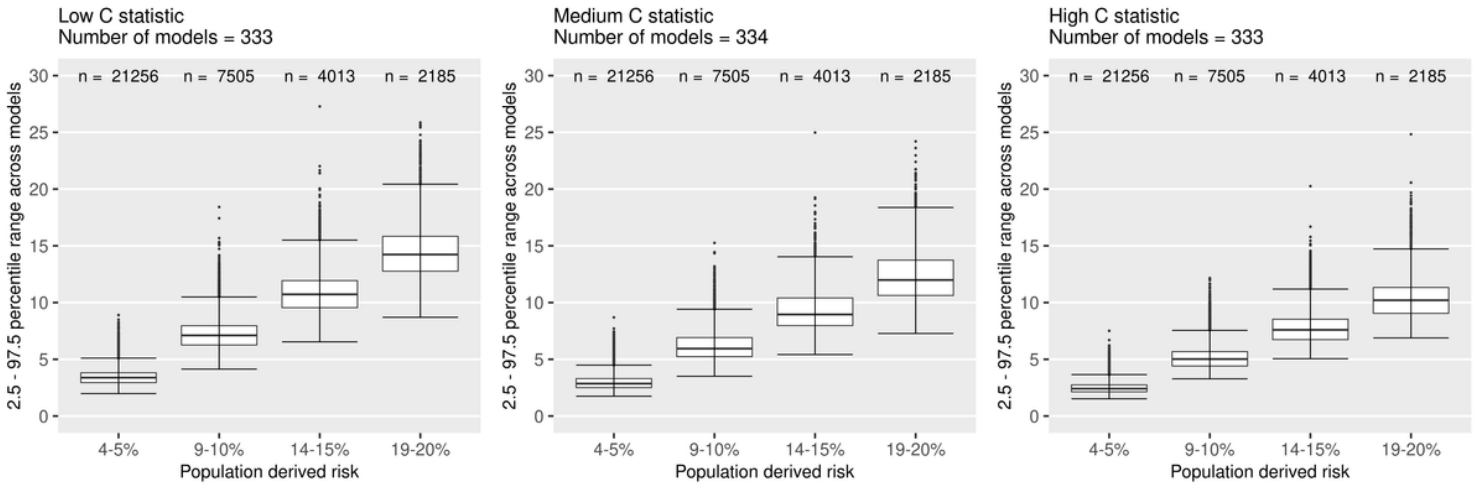

Figure 3

Percentile ranges of risk for individuals, stratified by C-statistic of the models (female cohort, N=10000). Each data point represents the 2.5-97.5 percentile range in risk for an individual across a group models defined by their C-statistics.

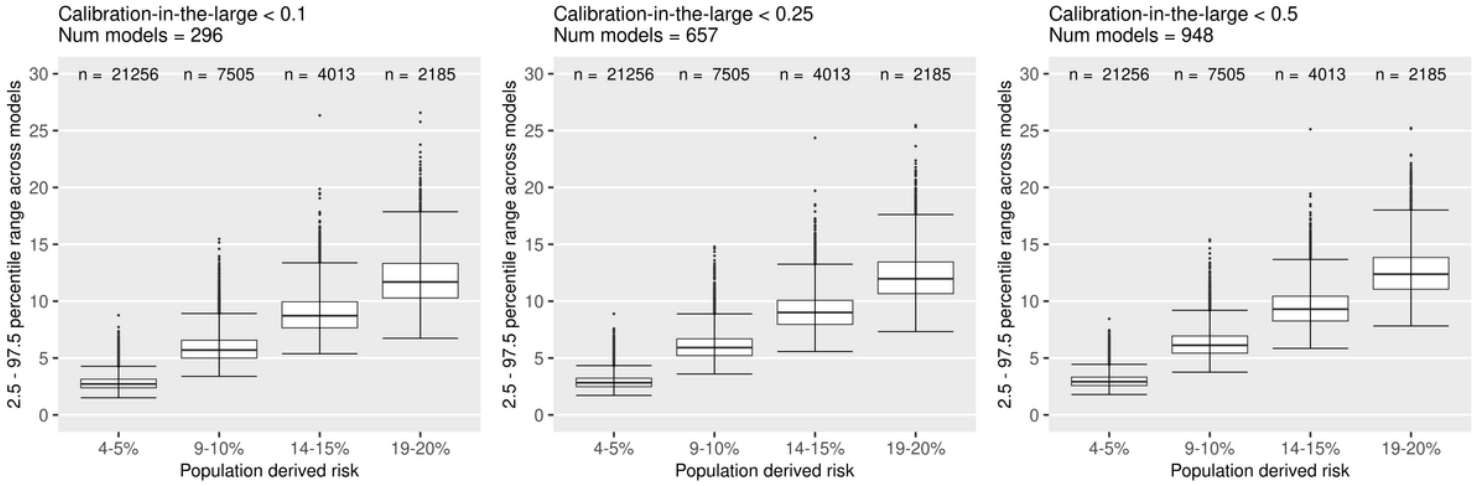

Figure 4

Percentile ranges of risk for individuals, stratified by calibration-in-the-large of the models (female cohort, N=10000). Each data point represents the 2.5-97.5 percentile range in risk for an individual across a group models defined by their calibration-in-the-large.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile3.docx
- Additionalfile2.docx
- Additionalfile1.docx