Supplementary Appendix

Table of Contents

STATISTICAL ANALYSIS PLAN

Evaluating Mode-Specific FHR Signal Dropout and Perinatal Asphyxia Using Causal Machine Learning

Date of Approval: December 12, 2024

1. Study Objective

To evaluate the causal relationship between mode-specific fetal heart rate (FHR) signal dropout—measured during the final 60 minutes of labor—and the risk of perinatal asphyxia. Specifically, this study investigates whether high dropout in ECG- or ultrasound-based cardiotocography (CTG) recordings is associated with increased asphyxia risk, using doubly robust methods including inverse probability weighting (IPW) and targeted maximum likelihood estimation (TMLE).

2. Primary Hypothesis

Laboring women monitored with CTG have differing risk profiles for perinatal asphyxia based on mode-specific signal quality metrics.

3. Study Design Overview

Multicohort cohort study incorporating causal machine learning methods

Datasets:

- Primary: Mercy Hospital for Women, Melbourne (2010-2021;n=36,792)
- Validation:
  - Czech CTU-UHB dataset (2010-2012;n=552)
  - Werribee Mercy Hospital dataset (2010-2021;n=15,413)

Inclusion: Singleton pregnancies ≥36 weeks gestation, ≥15 minutes of intrapartum CTG(upto 60 mins prior to labour)

Exclusion: Multiple gestations, fetal anomalies, >90% dropout

4. Exposure Definition

Fetal Heart Rate (FHR) dropout is defined as the proportion of time during which the CTG device recorded a fetal heart rate of zero despite active monitoring in a live fetus. Dropout is computed from raw CTG data sampled at 4 Hz over the final 60 minutes of labour. A threshold of >30% dropout is classified as "High Dropout", based on Delphi consensus from clinicians, data scientists, and signal processing experts in the study team, supported by findings from our first signal dropout manuscript.

Participants are assigned into four mutually exclusive exposure groups based on monitoring modality and dropout level:

1. Ultrasound – Low Dropout (reference group)
2. Ultrasound – High Dropout
3. ECG – Low Dropout
4. ECG – High Dropout

Monitoring modality will be determined from raw CTG signals. Patients will be classified as monitored in ECG mode if ≥10 minutes of ECG signal was present; otherwise, they were assigned to ultrasound mode. This definition will reduce misclassification from transient modality switches and allows mode-specific analysis of signal quality.

5. Primary Outcome

Perinatal Asphyxia, defined as a composite of clinical and physiological criteria aligned with international consensus guidelines, including any one of:

1. Umbilical cord pH <7.05 or base excess <–12
2. Apgar score ≤4 at 5 min or ≤6 at 10 min
3. Resuscitation with IPPV at 10 min
4. NICU admission
5. Therapeutic hypothermia
6. Clinical/imaging-confirmed hypoxic–ischemic encephalopathy (HIE)
7. Neonatal seizures due to presumed hypoxia

See supplementary table 1 for harmonizing outcome criteria for Czech dataset

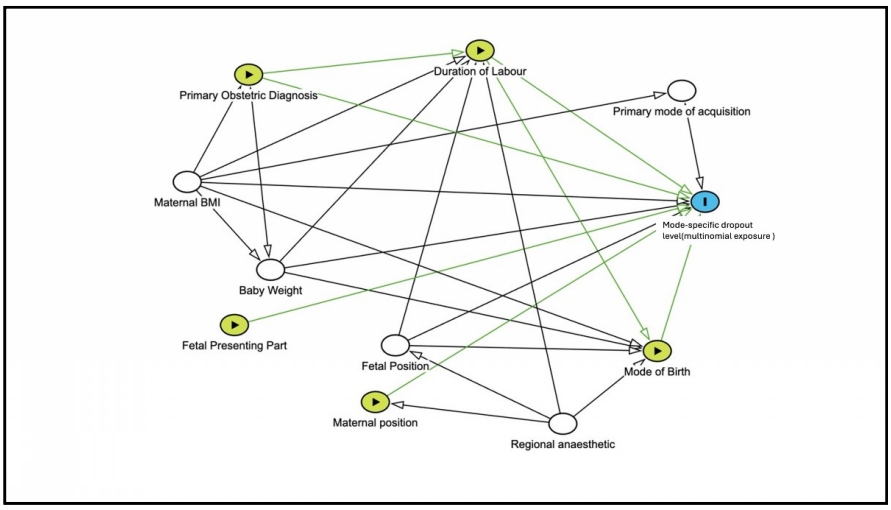
6. Analytic Strategy

6.1. Descriptive Analyses

We will summarize maternal, fetal, and labor characteristics across the four exposure groups (Ultrasound Low Dropout, Ultrasound High Dropout, ECG Low Dropout, ECG High Dropout). Categorical variables will be reported as counts and percentages, while continuous variables will be summarized using means with standard deviations or medians with interquartile ranges, depending on distribution.

6.2. Causal Inference Framework

To estimate the causal effect of mode-specific CTG signal dropout on perinatal asphyxia, we will adopt a doubly robust approach that combines exposure modeling with outcome regression. We will first construct a multinomial exposure model to estimate the probability of each subject being in one of the four exposure groups:

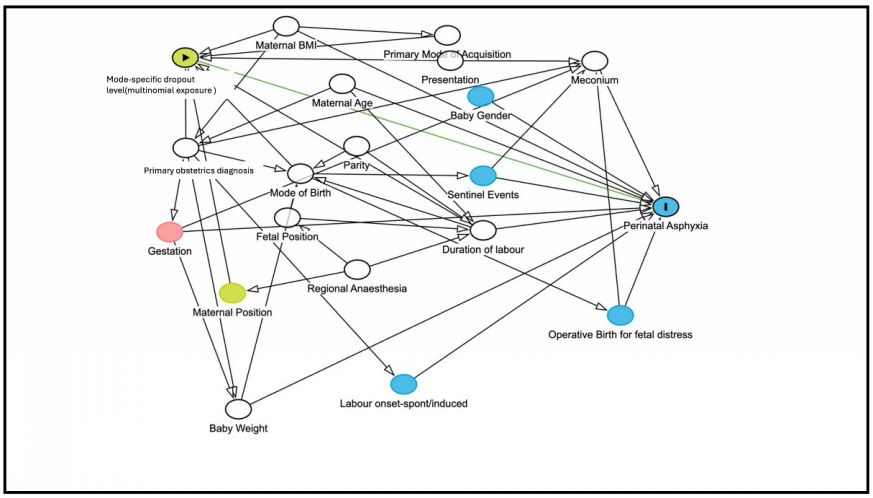


Supplementary Figure 1:CAUSAL DIAGRAM 1:PROPENSITY SCORE MODEL FOR EXPOSURE(created by Daggity)

1. Ultrasound – Low Dropout (reference)
2. Ultrasound – High Dropout
3. ECG – Low Dropout
4. ECG – High Dropout

This model will be built using multinomial logistic regression, with predictor variables selected based on directed acyclic graphs identify minimal ovariates will include key maternal, fetal, and labor characteristics known to influence both signal quality and perinatal outcomes (supplementary table 1).

The propensity scores derived from this exposure model will be used to generate stabilized inverse probability weights (IPWs), clipped between 0.01 and 0.99 to reduce the influence of extreme weights. These weights will be applied to a weighted logistic regression model estimating the association between dropout category and the outcome (perinatal asphyxia). Standard errors will be adjusted using HC3 robust variance estimators. We will report odds ratios (ORs), 95% confidence intervals, and E-values to assess robustness to unmeasured confounding.



Supplementary Figure 2:CAUSAL DIAGRAM 2: OUTCOME MODEL(created using using DAGitty)

6.3. Causal Machine Learning

To refine and validate effect estimates from the IPW model, we will apply Targeted Maximum Likelihood Estimation (TMLE). TMLE integrates machine learning for both the exposure model (treatment assignment) and the outcome model (risk of perinatal asphyxia) the causal effect.

The exposure model will use the same multinomial logistic regression or gradient boosting classifier as in Section 6.2. The outcome model will predict the probability of asphyxia given exposure and confounders, using regularized logistic regression or gradient boosting. Feature selection and variable transformations (e.g., log-scaling of BMI) will follow our DAG-based framework to avoid overadjustment or inappropriate control. The primary contrast of interest will be the mode-specific signal dropout group associated with the highest asphyxia risk, compared to the Ultrasound Low Dropout group (reference group). TMLE will yield doubly robust adjusted odds ratios and confidence intervals based on influence-function-derived standard errors. To enhance interpretability, we will apply SHAP (Shapley Additive Explanations) to both the exposure and outcome model, identifying the most influential variables driving model predictions.

6.4. Survival Analysis

We will perform time-to-event analyses to assess the impact of signal dropout and monitoring modality on the timing of perinatal asphyxia. Kaplan–Meier survival curves will be generated for each exposure group and compared using log-rank tests. To control for confounding, we will fit IPW-adjusted Cox proportional hazards models using the same stabilized weights as in the causal inference models. Discrimination will be assessed using Harrell's C-index.

6.5. External Validation

The analytic pipeline will be externally validated using two independent datasets: the Czech CTU-UHB cohort and the Werribee Mercy Hospital cohort. All models, including IPW and TMLE approaches, will be replicated in these cohorts. We will assess consistency of effect estimates (odds ratios and confidence intervals) with the primary training cohort to determine generalizability.

7. Sensitivity Analyses

We will conduct several sensitivity analyses to assess the robustness of our findings. First, we will compare results between complete case analyses and models with multiple imputation to account for missing clinical data, particularly BMI. Next, we will vary the signal dropout threshold to explore the relationship between dropout and asphyxia risk across incremental thresholds (e.g., 10%, 20%, 30%) using inverse probability weighting (IPW) and targeted maximum likelihood estimation (TMLE).

8. Subgroup Analyses

To explore heterogeneity in the association between signal dropout and perinatal asphyxia, we will perform pre-specified subgroup analyses. Subgroups will include stage of labor (first vs second), mode of birth (vaginal vs caesarean), and dropout severity categorized by quintiles. These analyses will help assess whether the strength or direction of the association varies by clinical context or monitoring characteristics.

9. Handling of Missing Data

Missing values in clinical variables will be addressed using multiple imputation via chained equations (MICE). We anticipate notable missingness for maternal BMI (~6.3% in training dataset based on our previous study on signal features). All models will be re-run on both imputed and complete case datasets, and results compared to ensure robustness. Imputation models will include all exposure and outcome variables, as well as auxiliary covariates to improve prediction of missing values.

10. Software & Tools

All analyses will be conducted using Python (including zepid, lifelines, and scikit-learn packages) and Stata 18. Digital CTG data will be extracted using Philips IntelliSpace™ systems and securely stored and processed on the University of Melbourne's Mediaflux™ platform and Spartan High Performance Computing cluster.

11. Ethics and reporting

This study will be conducted in accordance with relevant ethical standards and reporting guidelines. Approval for the study protocol, including data linkage and analysis of de-identified CTG traces and clinical outcomes, has been obtained from the Mercy Health Human Research Ethics Committee (HREC #R2020-077). All data will be de-identified prior to analysis and securely stored on the University of Melbourne's Mediaflux™ platform. Analyses will be conducted on the Spartan High Performance Computing cluster to ensure data integrity and confidentiality.

The study will follow established guidelines for transparent reporting of observational and predictive modeling studies, including:

STROBE (Strengthening the Reporting of Observational Studies in Epidemiology)

TRIPOD-AI (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis – Artificial Intelligence Extension)

Supplementary Table 1. List of Variables Included in the Causal Models (by Dataset)

Variable	Mercy	Werribee	Czech
Log-transformed maternal BMI	✔	✔	
Gestational age (weeks)	✔	✔	✔
Baby weight (kg)	✔	✔	✔
Maternal age group (categorical)	Will be transformed from available data	Will be transformed from available data	Will be transformed from available data
Labour stage duration ratio (quantile bins)	Will be transformed from available data	Will be transformed from available data	Will be transformed from available data
Primary obstetric diagnosis (categorical)	✔	✔	
Fetal presentation (non-cephalic)	✔	✔	✔
Maternal position (categorical)	✔	✔	✔
Parity (categorical or binary)	✔	✔	✔
Labour onset (spontaneous vs induced)	✔	✔	✔
Sentinel events (none, serious, cord accident)	✔	✔	NA
Baby gender (male/female)	✔	✔	✔
Meconium-stained liquor	✔	✔	✔
Log-transformed total labour duration	Will be transformed from available data	Will be transformed from available data	Will be transformed from available data
Log-transformed stage 1 duration	Will be transformed from available data	Will be transformed from available data	Will be transformed from available data
Log-transformed stage 2 duration	Will be transformed from available data	Will be transformed from available data	Will be transformed from available data
Obstetric risk group	✔		Information available on (diabetes, hypertensive disease)
Maternal age (years)	✔	✔	✔
Dropout raw value and transformed quantile (as categorical)	Will be extracted and transformed	Will be extracted and transformed	Will be extracted and transformed

Perinatal asphyxia in the Czech dataset will be defined using a modified, harmonised outcome: the presence of any one of the following — pH < 7.05, Apgar score at 5 minutes ≤ 4, or base excess < –12 mmol/L.

Supplementary findings

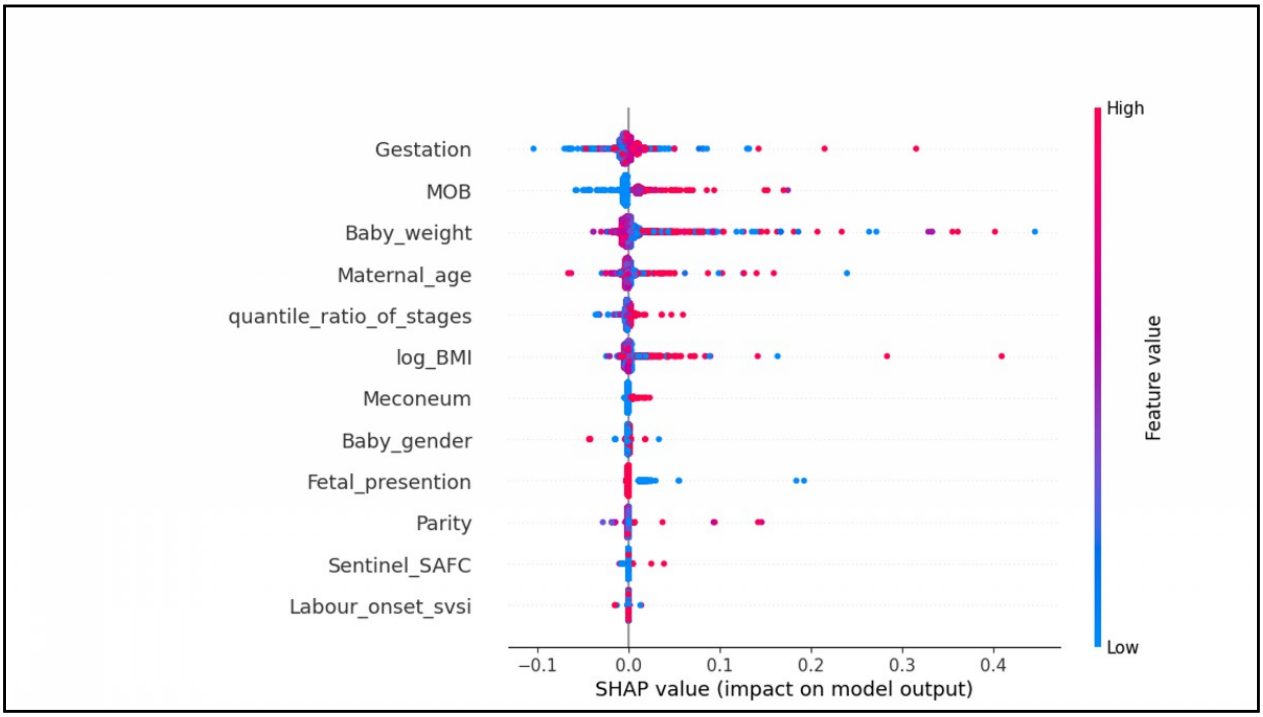
Supplementary Table 2: Summary Statistics of Raw FHR Dropout by Mode of Acquisition Across Datasets

Dataset	Acquisition Mode	Mean (%)	SD (%)	Median (%)	IQR (%)
Mercy	US	15.9	14.5	11.9	6.7–20.1
	ECG	7.6	9.0	4.7	1.0–10.9
Werribee	US	11.7	11.6	8.7	4.7–15.1
	ECG	10.8	13.9	6.3	1.1–15.3
Czech	US	18.0	20.5	9.0	1.7–30.7
	ECG	9.0	17.2	0.5	0.0–8.2

Note: All values are expressed as percentages of signal loss; SD=Standard deviation; IQR = interquartile range (25th–75th percentile).

Supplementary table 3: Statistically significant predictors of perinatal asphyxia in the IPW-Adjusted Cox Proportional Hazards Model in the training dataset

Variable	Hazard Ratio (HR)	95% CI	p-value
Baby weight (kg)	0.81	0.73-0.90	0.0001
Labor stage ratio (quantiles)	0.85	0.77-0.94	0.0012
Fetal presentation (non-cephalic)	1.10	1.03-1.17	0.0037
Parity = 1	1.59	1.45-1.75	<0.0001
Parity = 2	1.55	1.41-1.71	<0.0001
Parity ≥ 3	1.40	1.28-1.54	<0.0001
Spontaneous labor onset	0.69	0.60-0.79	<0.0001
Male infant	1.10	1.02-1.19	0.021
Meconium-stained liquor	1.09	1.02-1.18	0.019
US High Dropout	1.90	1.39-2.60	<0.0001
ECG Low Dropout	1.42	1.20-1.69	<0.0001
ECG High Dropout	4.76	2.98-7.62	<0.0001



Supplementary Figure 3. SHAP Summary Plot of the Outcome Model from TMLE Analysis in the Training Cohort

SHAP summary plot showing feature contributions to predicted asphyxia risk in the TMLE outcome model using the training cohort. Features are ranked by importance; color indicates feature value (red = high, blue = low), and SHAP value reflects impact on model output.

STROBE Statement—Checklist of items that should be included in reports of *cohort studies*

	Item No	Recommendation	Page No
Title and abstract	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract	1 (Title and abstract: "Multicohort Study" and "causal machine learning")
		(*b*) Provide in the abstract an informative and balanced summary of what was done and what was found	1–6
Introduction			
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported	7 (Introduction)
Objectives	3	State specific objectives, including any prespecified hypotheses	8 (Introduction)
Methods			
Study design	4	Present key elements of study design early in the paper	8 (Methods: "Study Design & Data Sources")
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection	8 (Cohorts from 2010–2021; hospitals named)
Participants	6	(*a*) Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up	9-0 (Inclusion: ≥36 weeks, exclusions listed)
		(*b*) For matched studies, give matching criteria and number of exposed and unexposed	Propensity score
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable	9 (Exposure classification + Outcome measures),Supplementary table 1
Data sources/ measurement	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group	12 (Philips Intellispace™, BOS)
Bias	9	Describe any efforts to address potential sources of bias	10,12,13 (Use of DAGs, IPW, E-values)
Study size	10	Explain how the study size was arrived at	12
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why	10-11 (BMI log-transformed; dropout as %; standardization)
Statistical methods	12	(*a*) Describe all statistical methods, including those used to control for confounding	10-13(Multinomial PS, IPW, TMLE, Cox)
		(*b*) Describe any methods used to examine subgroups and interactions	9-10 (Subgroup by mode, labor stage, etc.)
		(*c*) Explain how missing data were addressed	12(Multiple imputation, complete case analysis)
		(*d*) If applicable, explain how loss to follow-up was addressed	NA (no long-term follow-up; outcome at delivery)
		(*e*) Describe any sensitivity	10,12,13 (Thresholds, imputation, DAGs)

		analyses	
Results			
Participants	13*	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed (b) Give reasons for non-participation at each stage (c) Consider use of a flow diagram	figure 1
Descriptive data	14*	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders (b) Indicate number of participants with missing data for each variable of interest (c) Summarise follow-up time (eg, average and total amount)	13; Table 1,supplementary tables 1,2 13 (BMI = 6.32%) NA (not longitudinal)
Outcome data	15*	Report numbers of outcome events or summary measures over time	13

272

Main results	16	(*a*) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included	12-17, table 2
		(*b*) Report category boundaries when continuous variables were categorized	9,11 (e.g., Dropout >30%, BMI transformations)
		(*c*) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period	15 (e.g., 1 in 44 women)
Other analyses	17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses	14 (TMLE thresholds, reclassification)
Discussion			
Key results	18	Summarise key results with reference to study objectives	17
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias	19
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence	19-20
Generalisability	21	Discuss the generalisability (external validity) of the study results	20
Other information			
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based	20

TRIPOD AI Checklist

TRIPOD+AI

Version: 11-January-2024

Section/Topic	Item	Development / evaluation[1]	Checklist item	Reported on page
TITLE				
Title	1	D;E	Identify the study as developing or evaluating the performance of a multivariable prediction model, the target population, and the outcome to be predicted	1
ABSTRACT				
Abstract	2	D;E	See TRIPOD+AI for Abstracts checklist	4-6
INTRODUCTION				
Background	3a	D;E	Explain the healthcare context (including whether diagnostic or prognostic) and rationale for developing or evaluating the prediction model, including references to existing models	7-8
	3b	D;E	Describe the target population and the intended purpose of the prediction model in the context of the care pathway, including its intended users (e.g., healthcare professionals, patients, public)	7-9
	3c	D;E	Describe any known health inequalities between sociodemographic groups	7-9
Objectives	4	D;E	Specify the study objectives, including whether the study describes the development or validation of a prediction model (or both)	8
METHODS				
Data	5a	D;E	Describe the sources of data separately for the development and evaluation datasets (e.g., randomised trial, cohort, routine care or registry data), the rationale for using these data, and representativeness of the data	8
	5b	D;E	Specify the dates of the collected participant data, including start and end of participant accrual; and, if applicable, end of follow-up	8-9
Participants	6a	D;E	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including the number and location of centres	8-9
	6b	D;E	Describe the eligibility criteria for study participants	8-9,fig 1
	6c	D;E	Give details of any treatments received, and how they were handled during model development or evaluation, if relevant	9
Data preparation	7	D;E	Describe any data pre-processing and quality checking, including whether this was similar across relevant sociodemographic groups	8-13
Outcome	8a	D;E	Clearly define the outcome that is being predicted and the time horizon, including how and when assessed, the rationale for choosing this outcome, and whether the method of outcome assessment is consistent across sociodemographic groups	8-13
	8b	D;E	If outcome assessment requires subjective interpretation, describe the qualifications and demographic characteristics of the outcome assessors	
	8c	D;E	Report any actions to blind assessment of the outcome to be predicted	NA
Predictors	9a	D	Describe the choice of initial predictors (e.g., literature, previous models, all available predictors) and any pre-selection of predictors before model building	8,7 Supplementary – Predictors defined from literature and DAGs; include maternal, fetal, and labor variables.
	9b	D;E	Clearly define all predictors, including how and when they were measured (and any actions to blind assessment of predictors for the outcome and other predictors)	
	9c	D;E	If predictor measurement requires subjective interpretation, describe the qualifications and demographic characteristics of the predictor assessors	
Sample size	10	D;E	Explain how the study size was arrived at (separately for development and evaluation), and justify that the study size was sufficient to answer the research question. Include details of any sample size calculation	12
Missing data	11	D;E	Describe how missing data were handled. Provide reasons for omitting any data	12
Analytical methods	12a	D	Describe how the data were used (e.g., for development and evaluation of model performance) in the analysis, including whether the data were partitioned, considering any sample size requirements	8,11,16,17
	12b	D	Depending on the type of model, describe how predictors were handled in the analyses (functional form, rescaling, transformation, or any standardisation).	IPW, TMLE
	12c	D	Specify the type of model, rationale[2], all model-building steps, including any hyperparameter tuning, and method for internal validation	Ridge, GBM,
	12d	D;E	Describe if and how any heterogeneity in estimates of model parameter values and model performance was handled and quantified across clusters (e.g., hospitals, countries). See TRIPOD-Cluster for additional considerations[3]	SHAP,
	12e	D;E	Specify all measures and plots used (and their rationale) to evaluate model performance (e.g., discrimination, calibration, clinical utility) and, if relevant, to compare multiple models	survival analyisis
	12f	E	Describe any model updating (e.g., recalibration) arising from the model evaluation, either overall or for particular sociodemographic groups or settings	E VALUES
	12g	E	For model evaluation, describe how the model predictions were calculated (e.g., formula, code, object, application programming interface)	includes tuning ar standardisation
Class imbalance	13	D;E	If class imbalance methods were used, state why and how this was done, and any subsequent methods to recalibrate the model or the model predictions	IPW
Fairness	14	D;E	Describe any approaches that were used to address model fairness and their rationale	Mode specific risk
Model output	15	D	Specify the output of the prediction model (e.g., probabilities, classification). Provide details and rationale for any classification and how the thresholds were identified	SHAP 8-10,13-17,sup fig3

[1] D=items relevant only to the development of a prediction model; E=items relating solely to the evaluation of a prediction model; D;E=items applicable to both the development and evaluation of a prediction model

[2] Separately for all model building approaches.

[3] TRIPOD-Cluster is a checklist of reporting recommendations for studies developing or validating models that explicitly account for clustering or explore heterogeneity in model performance (eg, at different hospitals or centres). Debray et al, BMJ 2023; 380: e071018 [DOI: 10.1136/bmj-2022-071018]

Training versus evaluation	16	D;E	Identify any differences between the development and evaluation data in healthcare setting, eligibility criteria, outcome, and predictors	8-10
Ethical approval	17	D;E	Name the institutional research board or ethics committee that approved the study and describe the participant-informed consent or the ethics committee waiver of informed consent	13
OPEN SCIENCE				
Funding	18a	D;E	Give the source of funding and the role of the funders for the present study	20
Conflicts of interest	18b	D;E	Declare any conflicts of interest and financial disclosures for all authors	20
Protocol	18c	D;E	Indicate where the study protocol can be accessed or state that a protocol was not prepared	12,SAP in supplementary
Registration	18d	D;E	Provide registration information for the study, including register name and registration number, or state that the study was not registered	NA
Data sharing	18e	D;E	Provide details of the availability of the study data	21
Code sharing	18f	D;E	Provide details of the availability of the analytical code[4]	21
PATIENT & PUBLIC INVOLVEMENT				
Patient & Public Involvement	19	D;E	Provide details of any patient and public involvement during the design, conduct, reporting, interpretation, or dissemination of the study or state no involvement.	Not applicable
RESULTS				
Participants	20a	D;E	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	Page 8,
	20b	D;E	Report the characteristics overall and, where applicable, for each data source or setting, including the key dates, key predictors (including demographics), treatments received, sample size, number of outcome events, follow-up time, and amount of missing data. A table may be helpful. Report any differences across key demographic groups.	
	20c	E	For model evaluation, show a comparison with the development data of the distribution of important predictors (demographics, predictors, and outcome).	Figure 1
Model development	21	D;E	Specify the number of participants and outcome events in each analysis (e.g., for model development, hyperparameter tuning, model evaluation)	Pages 8–11
Model specification	22	D	Provide details of the full prediction model (e.g., formula, code, object, application programming interface) to allow predictions in new individuals and to enable third-party evaluation and implementation, including any restrictions to access or re-use (e.g., freely available, proprietary)[5]	ORs, CIs, reclassification, SHAP importance, Harrell's C-index
Model performance	23a	D;E	Report model performance estimates with confidence intervals, including for any key subgroups (e.g., sociodemographic). Consider plots to aid presentation.	
	23b	D;E	If examined, report results of any heterogeneity in model performance across clusters. See TRIPOD Cluster for additional details[3].	
Model updating	24	E	Report the results from any model updating, including the updated model and subsequent performance	
DISCUSSION				
Interpretation	25	D;E	Give an overall interpretation of the main results, including issues of fairness in the context of the objectives and previous studies	13-17
Limitations	26	D;E	Discuss any limitations of the study (such as a non-representative sample, sample size, overfitting, missing data) and their effects on any biases, statistical uncertainty, and generalizability	13-17
Usability of the model in the context of current care	27a	D	Describe how poor quality or unavailable input data (e.g., predictor values) should be assessed and handled when implementing the prediction model	13-17
	27b	D	Specify whether users will be required to interact in the handling of the input data or use of the model, and what level of expertise is required of users	13-17
	27c	D;E	Discuss any next steps for future research, with a specific view to applicability and generalizability of the model	13-17

From: Collins GS, Moons KGM, Dhiman P, et al. *BMJ* 2024;385:e078378. doi:10.1136/bmj-2023-078378

[4] This relates to the analysis code, for example, any data cleaning, feature engineering, model building, evaluation.
[5] This relates to the code to implement the model to get estimates of risk for a new individual.