



# Methodological Challenges for Risk Prediction in Perinatal Epidemiology

Sonia M. Grandi<sup>1,2</sup> · Jennifer A. Hutcheon<sup>3</sup> · Kristian B. Filion<sup>1,2,4</sup> · Robert W. Platt<sup>1,2,5,6</sup>

Published online: 22 September 2018  
© Springer Nature Switzerland AG 2018

## Abstract

**Purpose of Review** To discuss the methodological challenges in developing risk prediction models in perinatal epidemiology and barriers to their implementation in clinical practice.

**Recent Findings** In perinatal epidemiology, risk prediction models have been created to examine the risk of adverse health outcomes in pregnancy, delivery, and post-partum periods. However, only a limited number of prediction models are being used to guide clinical decisions.

**Summary** The accuracy and utility of prediction models for clinical decision making are contingent on the use of robust methods to develop risk prediction models and appropriate metrics to assess their performance and clinical impact. In order to increase the transportability (i.e., generalizability) of prediction models, careful consideration of the patient populations represented in the data used to develop and externally validate prediction models and the mechanism for data collection are needed. The era of big data provides researchers the opportunity to leverage existing databases, such as birth and pregnancy registries, through linkage to electronic health records, disease registries, and census data in order to enrich the breadth of clinical and sociodemographic information available for prediction modeling. However, these data sources introduce new challenges that require thorough assessment to evaluate their impact on the accuracy of resulting prediction models and their transportability to the general population.

**Keywords** Prediction model · Perinatal epidemiology · Model development · Model validation · Clinical impact

## Introduction

Since the introduction of the first risk prediction tool for cardiovascular disease in 1976 [1], there has been a steady rise in the number of prediction models in various fields of clinical epi-

demiology including cardiology, oncology, and pediatrics. In the field of perinatal epidemiology, the motivation for prediction modeling has been to identify women at highest risk of a negative health outcome to guide prevention strategies for the mother and infant. More specifically, prediction modeling has enabled physicians to provide individualized care to women and their infants through evidence-based decision making [2].

Prediction models have been broadly used in various fields of perinatal epidemiology to predict treatment success for women undergoing fertility treatments [3], predict complications of pregnancy (e.g., preeclampsia [4] and fetal growth restriction [5]), predict outcomes at delivery (e.g., vaginal birth after a cesarean section [6]) and in the post-partum period (e.g., post-partum hemorrhage [7] and neonatal mortality in preterm infants [8]), and to rule out women at risk of an adverse outcome [9]. More recent literature has examined the association between pregnancy-related exposures and long-term outcomes in mothers and children [10, 11]. Despite the increasing number of risk prediction models being developed, few models are of sufficiently high quality or easily implemented in routine clinical practice [2, 12, 13]. This discrepancy can be attributed to a number of factors including (1)

---

This article is part of the Topical Collection on *Reproductive and Perinatal Epidemiology*

---

✉ Robert W. Platt  
robert.platt@mcgill.ca

<sup>1</sup> Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, QC, Canada

<sup>2</sup> Centre for Clinical Epidemiology, Lady Davis Institute, Jewish General Hospital, Montreal, QC, Canada

<sup>3</sup> Department of Obstetrics and Gynaecology, Faculty of Medicine, University of British Columbia, Vancouver, BC, Canada

<sup>4</sup> Department of Medicine, McGill University, Montreal, QC, Canada

<sup>5</sup> McGill University Health Center Research Institute, Montreal, QC, Canada

<sup>6</sup> Department of Pediatrics, McGill University, Montreal, QC, Canada

inappropriate methods for model development and validation; (2) the choice of data sources and populations for model development and validation; (3) absence or imprecision in the measurement of important predictors; and (4) lack of external validation. The implications of these factors for the implementation of prediction models in routine practice will be discussed in further detail in this review.

## From Development to Use in Clinical Practice

Prior to implementation of prediction models in clinical practice, researchers need to (1) develop and internally validate the model; (2) perform external validation; and (3) assess the clinical impact of the model (Fig. 1). All three components are needed to provide clinicians an objective measure for risk stratification above clinical judgement [14].

## Development and Internal Validation

The first step in model development is the identification of potentially relevant predictors based on substantive knowledge and the existing literature. Considerations for selection of candidate predictors are discussed below. Once a list of candidate predictors has been created, data reduction is performed to remove predictors with narrow distributions (limited ability to explain variation in outcome) or a large degree of missingness to increase model validity and parsimony [15]. Collinearity between predictors should be assessed and minimized either by choosing predictors based on objective criteria, which may include clinical relevance, availability, reliability, or cost of measurement. A full model is then estimated using variables not previously eliminated (i.e., the strength and direction of association between each predictor and outcome is estimated). Ideally, continuous variables are modeled using restricted cubic splines or other smoothing

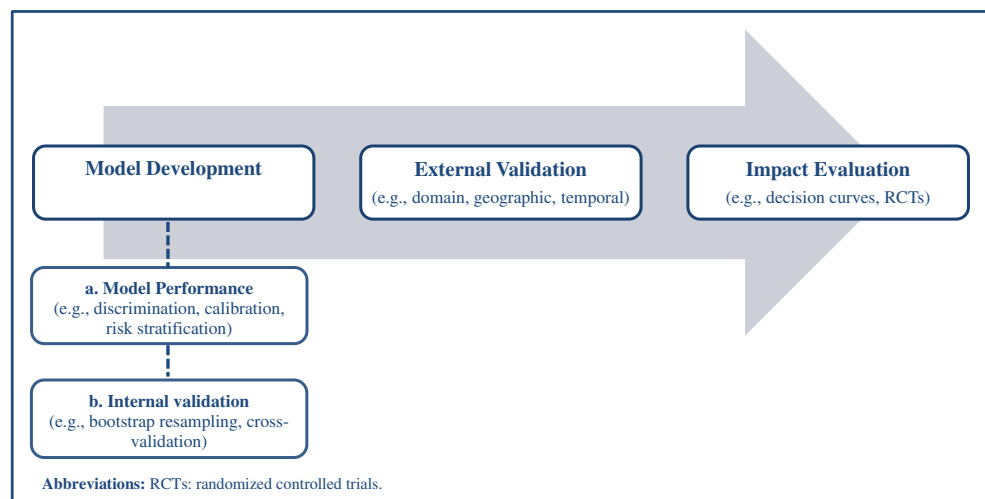
functions such as fractional polynomials, and categorical variables are modeled using indicator variables. When fitting the model, shrinkage methods should be considered when dealing with small sample sizes to reduce the potential for model overfit [16, 17]. The final step of model development involves further data reduction for which various methods have been proposed [18]. A well-established approach to data reduction is the stepdown approach of Harrell et al. [15, 17]. The benefit of this approach is that it is done independently of the outcome, which reduces systematic bias and avoids using  $p$  values for variable selection, which tend to result in model overfit and poor model performance [15, 17].

## Assessing Model Performance

Once the final model is established, its predictive performance is examined using measures of accuracy and validity. Performance can be grouped into three main categories, (1) discrimination, (2) calibration, and (3) risk stratification. Although risk stratification is not commonly used to examine model performance, its addition provides a comprehensive assessment since it evaluates a model's capacity to appropriately stratify patients. The utility of all three metrics to assess the performance of prediction models is illustrated by the stillbirth calculator to identify the risk of stillbirth in women [19] and the fullPIERS model to identify the risk of adverse maternal outcomes in women with preeclampsia [20]. Although both models were found to perform well based on standard metrics of discrimination and calibration, risk stratification allowed investigators to identify optimal thresholds (based on the rate of false positive and true positive predictions) to assist clinicians in their choice of treatment options for these women.

Discrimination refers to how well the model discriminates between individuals with and without the outcome [18]. A commonly used method to assess discrimination is the

**Fig. 1** Steps to building a risk prediction model



estimation of the area under the receiver operating curve (AUC) or the c statistic [21], in which a value of 1 refers to perfect discrimination, and 0.5 is equivalent to random chance. If the prediction model involves time-to-event data, standard metrics pose problems due to unobserved event times as a result of right censoring [22]. Moreover, the presence of censoring during follow-up warrants additional consideration, since the ordering of events becomes difficult to decipher; Harrell’s concordance (c) index [18, 23], Royston and Sauerbrei’s D statistic [24], and the weighted Brier score [18, 23, 25] have been proposed to address this limitation. Harrell’s c index in the context of time-to-event data is a rank order statistic measuring the ability of a model to discriminate between individuals with different event times [18, 23]. It is a measure of the probability of concordance between observed and predicted survival probabilities given that pairs are useable ( $\geq 1$  individual experiences the event of interest) [26]. Therefore, a model with good discriminative properties will assign a higher predicted probability to an individual with the event compared to an individual without the event at the same time point [15]. Royston and Sauerbrei’s D statistic is an absolute measure of separation of survival curves that measures discrimination between strata of risk groups and the baseline hazard [24]. The Brier score is a quadratic scoring rule that estimates the squared distance between the observed and predicted outcomes [18, 23]. As a measure of explained variation, it can be used to assess calibration, as well as, goodness-of-fit of the model. A Brier score can take values from 0 to 1, with a value of 0 suggesting perfect prediction. Brier scores are generated based on the prediction times from the models calculated at fixed time points (e.g., 6- or 12-month intervals) to generate time-dependent curves [23]. If censoring is found to be substantial and informative, a weighted Brier score using inverse probability of censoring weights should be used [25, 27].

Calibration refers to the agreement between the predicted and observed outcomes [18]. For prognostic estimates, calibration is important since it provides a measure of model reliability [18]. Calibration plots are constructed as a function of the predictions from the model on the x-axis and the observed outcomes on the y-axis with perfect predictions falling on the 45° line [18]. Plots can also be generated by grouping individuals based on their predicted probability of the outcome with a larger division between groups indicating improved discrimination. In the context of time-to-event data, calibration curves are created from predicted probabilities obtained from the final models and compared to observed probabilities obtained from Kaplan-

Meier estimates at fixed time intervals [15]. Calibration is then measured as the difference between the observed and expected survival estimates at specified time intervals. This difference can then be used to correct the performance measures for the degree of optimism or overfit of the model. A second measure of calibration is goodness-of-fit, which is commonly measured using the Hosmer-Lemeshow goodness-of-fit test for binary outcomes. The number of expected and observed outcomes is compared within groups of individuals using a  $\chi^2$  statistic. Goodness-of-fit for survival models is typically assessed using calibration curves and the Brier score or by comparing the Cox-Snell residuals and the cumulative hazard function within risk categories [18].

Risk stratification is an important measure of performance since it assesses the capacity of a prediction model to stratify individuals into clinically relevant risk groups [28]. Stratification entails dichotomizing or categorizing predicted risks based on meaningful cut-offs and assessing the capacity of the model to classify patients into the defined risk categories [29]. Risk stratification can also be used to compare the incremental value of predictors to existing models using reclassification tables [28]. However, reclassification tables do not account for improvements in risk stratification. The net reclassification improvement (NRI) index was developed to quantify improvements resulting from appropriate risk reclassification by assigning scores based on upward reclassifications in individuals with the disease and downward reclassifications for individuals without the disease (Table 1) [30]. An extension of the NRI, the integrated discrimination improvement (IDI) index, assesses the NRI over all possible cut-offs (Table 1) [30]. Risk stratification is important for implementation of prediction models in clinical practice since it facilitates the identification of high-risk patients and clinically relevant thresholds for targeting prevention strategies. An example of the utility of risk stratification to guide decision making in clinical practice is the stratification of patients into low- and high-risk of perinatal death compared to gestational age alone using the miniPIERS model [31]. Based on the performance of the model across various thresholds of predicted risk, the investigators were able to determine the incremental value of the miniPIERS model above the current standard.

### Internal Validation

Once performance measures have been established, internal validation is needed to determine the degree of overfit of the

**Table 1** Measures of Incremental Value and Clinical Utility

Net reclassification improvement	$NRI = P(\text{up} \text{event}) - P(\text{down} \text{event}) + P(\text{down} \text{non-event}) - P(\text{up} \text{non-event})$
Integrated discrimination improvement	$\text{Slope} = E(\text{events}) - E(\text{non-events})$
Net benefit	$\text{Net benefit} = \frac{TP - wFP}{N}$ where $w = \text{odds of cut-off} \left[ \frac{p_c}{1-p_c} \right]$

*E*, expectation; *FP*, false positives; *P*, probability; *p<sub>c</sub>*, predicted probability of event; *NRI*, net reclassification improvement; *TP*, true positives

model. Commonly used techniques such as bootstrap resampling and cross-validation allow investigators to report optimism-corrected performance measures [32]. Bootstrap resampling is preferable in particular when dealing with small sample sizes to provide more precise estimates of the variability associated with modeling [15].

## External Validation

Once the final model has been found to perform adequately, external validation should be completed to improve the generalizability of the model. External validation is performed in a study population with a different data collection strategy from that used for model development. Several methods for external validation have been proposed including domain, geographic, and temporal validation [18, 33]. All three forms of external validation attempt to capture the potential for differences in model performance based on temporal and geographical trends or heterogeneity in patient populations. A recent study externally validating the fullPIERS model for prediction of adverse outcomes in women with preeclampsia provides an example of the various methods used for external validation [34]. Using three cohorts including women from different geographic locations, with varying periods of follow-up time, and with a broader range of disease (hypertensive disorders in versus of pregnancy versus preeclampsia), the investigators were able to assess the transportability of the model across time and clinical settings.

## Assessing Clinical Impact

The final step in prediction modeling involves assessing the clinical impact of the model. The presentation of absolute risks of an outcome without clearly defined decision thresholds is unlikely to modify a clinician's decision for patient management. Decision curve analyses were developed as a means of quantifying the harms and benefits of treatment over a range of decision thresholds [35, 36]. A decision curve is based on a measure of net benefit (NB) defined as the proportion of true positives penalized for false positives (Table 1) [37]. This measure is weighted by the ratio of over diagnosis (false positives) versus appropriate diagnosis (true positives), which is directly related to the decision threshold. The clinical utility of the final models can be assessed by plotting the range of threshold probabilities using the final models against a “treat all” and “treat none” scenario [36, 38]. For the purpose of establishing clinical utility, discrimination should be prioritized relative to calibration since it facilitates decision making. However, discrimination in isolation cannot determine the impact of the model for use in clinical settings since miscalibrated models can result in increased harm and reductions in the NB of prediction models [18].

The final step prior to implementation of prediction models is to perform impact evaluation studies for clinically relevant outcomes [39]. Impact evaluation studies can be assessed using randomized trials; however, due to time and cost constraints associated with the conduct of such trials, observational or quasi-experimental designs (e.g., pre- and post-designs, regression discontinuity, or differences-in-differences comparing outcomes in populations in which risk prediction models are used to standard of care) can provide a more efficient means of evaluating the impact of prediction models.

## Considerations in the Choice of Study Populations for Development and Validation

The big data era has seen an upsurge of prediction models developed using new data sources, including electronic health records (EHR) and administrative health and insurance claims databases. In perinatal epidemiology, these databases have been used to develop prediction models for the risk of early onset gestational diabetes [40], neonatal encephalopathy [41], neonatal sepsis [42], and adverse pregnancy outcomes [43]. EHRs are digital versions of a patient's medical chart containing medical and treatment history including laboratory and diagnostic test results, prescriptions, and hospital admissions. The breadth of clinical data in EHRs facilitates the sharing of clinical information across healthcare providers to improve continuity of care. Unlike EHRs, administrative health and insurance claims databases include data collected for administrative or billing purposes (e.g., Medicaid, Medicare, and Kaiser Permanente). The advantages of these databases are that they include a large number of patients followed longitudinally over time. Since these data are not collected for research purposes, their use for development and validation of risk prediction models is limited by the absence of detailed clinical information, inconsistencies in reporting, and discontinuous coverage resulting from changes in insurance providers or eligibility status. However, administrative and claims databases can be leveraged for research purposes through linkage to EHRs, disease registries, or census data.

The availability of more data and larger data sets affords an opportunity to identify novel predictors not previously considered or to include a larger set of predictors. However, the availability of new data sources and machine learning methods may also contribute to the surplus of unvalidated and poorly performing models. For example, there are approximately 1000 prognostic models developed to assess the risk of cardiovascular disease. However, only a limited number of these models have been externally validated, and even fewer are used for decision making in clinical practice [44].

Although EHRs and administrative and insurance databases allow for the inclusion of a larger set of candidate predictors, the data for predictors and outcomes may be less detailed, are subject to measurement error or inconsistencies in

reporting due to between-center or between-healthcare provider heterogeneity, and may have a large degree of missingness [45]. Although multiple imputation methods are able to circumvent issues of missing data (with an acceptable degree of missingness), they do not account for the inconsistencies in reporting and data collection. For example, the reporting of spontaneous abortions may vary by time (primarily in earlier databases), by institution, and by healthcare provider in administrative databases largely due to the passive nature of data collection. Moreover, private insurers often provide incentives to improve documentation of clinical and sociodemographic characteristics by healthcare providers compared to administrative databases where documentation is left to the discretion of the provider. EHRs or administrative databases may also lack important predictors that are not routinely collected or recorded. A recent study by Dalton and colleagues showed that a summary measure for neighborhood deprivation outperformed traditional risk factors in the pooled cohort equations risk model for prediction of cardiovascular risk [46]. As articulated by Galea and Keyes, the study by Dalton highlights the uncertainty of the accuracy of individual risk predictions based on a small set of clinical and demographic characteristics [47].

An additional consideration is the transportability of prediction models to different healthcare settings (e.g., socialized versus private healthcare and insurance claims versus population-based). For example, if we are interested in developing a risk prediction model to predict the occurrence of preeclampsia in low-resource settings, using an EHR (from a tertiary care setting) for the development of this model may not reflect the distribution of predictors or outcomes in the target population and impact its generalizability. An additional concern regarding the use of these databases is the potential for selection bias due to informative censoring. Differences in the case-mix in EHRs compared to the general population could result in substantial selection bias as a result of competing events or admissions to different hospitals. In perinatal epidemiology, however, losses to follow-up may be less of a concern since women tend to be younger and have fewer chronic illnesses. Although EHR and administrative and claims databases have become increasingly available, researchers need to consider the limitations of these data and the implications for the accuracy of individual-level predictions and the potential harm to patients based on miscalibrated models [13•].

### Considerations for Selection of Predictors

The performance of prediction models is determined by the strength of the predictors included in the final model. The strength of a predictor is a function of both the magnitude of its association with the outcome and its distribution in the population [18]. However, additional considerations are needed to optimize selection of predictors. First, a predictor can

only have a small degree of missingness to be considered. If there is an acceptable degree of missingness (30–50%) [48], multiple imputation is preferred to minimize potential selection bias that may occur when using complete cases only. Moreover, investigators need to ensure that predictors not routinely collected or readily available at the time of risk assessment are included in the model as this will reduce the generalizability of the model. For example, the usefulness of the gold standard for assessment of proteinuria, 24-h urine protein, versus a rapid dip-stick for management of women with gestational hypertension at > 37 weeks gestation is debatable due to the lag time associated with laboratory testing. Second, predictors need to be clearly defined using standardized and clinically relevant definitions [33]. Using arbitrary cut-offs or categories for predictors will reduce its transportability to clinical settings. For example, if gestational diabetes is included as a candidate predictor and the threshold for diagnosis used for developing the model is different from the threshold used in clinical practice, it will impact the predictive performance of the model and its transportability into practice. In addition, researchers should be cautious of data-driven categorization of continuous predictors since they may be fitting the idiosyncrasies of the data rather than true associations [18]. Third, the approach to data collection or capture needs to be considered as this may impact the distribution of predictors in the population used for model development or the accuracy of the model for external validation. For example, developing a prediction model to predict adverse obstetrical outcomes using a general practitioner's database may not capture women at higher risk of experiencing the outcome since these women are typically seen by obstetricians, thus, impacting the generalizability of the model to all pregnant women. Fourth, the temporality of predictors is essential for predictor selection. Prediction models should only include patient characteristics available to clinicians at the time of risk assessment and not those that occur after the outcome. For example, although infant birth weight is a strong predictor of success of vaginal birth after a cesarean section, it should not be included as a predictor since it is not available prior to delivery. Fifth, predictors do not need to be causally related to the outcome. Candidate predictors should be chosen based on substantive and clinical knowledge and not based on their causal relationship with the outcome. For example, there is a lack of evidence to support a causal link between demographic and certain clinical characteristics and stillbirth. However, previous research suggests that socioeconomic status and smoking are strong predictors of stillbirth [49, 50•]. These risk factors should therefore be considered as candidate predictors when developing a prognostic model for stillbirth. Sixth, the predictive value of predictors should not be assessed using measures of association (e.g., odds ratios (OR), risk ratios, and risk differences) [51]. As demonstrated in simulations, predictors would need to have associations of magnitude of OR > 25 to

be deemed strong predictors [51]. Researchers should therefore avoid using univariate analyses for selection of predictors and rely on more relevant measures such as the discriminatory ability of models. Finally, investigators tend to measure more predictors than can reasonably be included in the model. For prediction modeling, the number of predictors should be determined by the number of outcomes. To minimize the risk of overfitting or overly optimistic models (higher than expected false positives), the convention is to use the 10:1 rule (ratio of events to predictor) to improve model accuracy [39]. However, more recent work suggests that the 10:1 rule may be too conservative and that the number of predictors should be based on the prevalence of the outcome in the population, the total sample size, and the number of events in the population used for model development [52].

### Generalizability of Prediction Models

External validation is essential for the implementation of prediction models in clinical practice. However, it is rarely performed as a result of the limited availability of suitable data, [2, 13•] and when performed, prediction models rarely perform well in external validation, mainly due to study-level or population-level differences with the development dataset [12, 18]. These differences can occur in part due to differences in study design, which can lead to differences in the incidence of outcomes as a consequence of the sampling strategy (e.g., case-control versus cohort) or the mechanism of data collection (e.g., self-report versus physician diagnoses). Differences in the incidence of the outcome between the development and the validation set can reduce the transportability of the model largely due to poor discrimination [13•]. The distribution of predictors may also differ as a result of variations in the case-mix. For example, using a disease registry may result in more-severe patients compared to a primary care or population-based cohort. The accuracy of prediction models in validation sets may also decrease as a function of temporal trends in patient characteristics and outcome distributions. To accommodate such temporal changes and to avoid inappropriately rejecting a potentially useful prediction model, investigators can recalibrate or update models based on population-level differences in the validation set [18, 33]. Discrepancies may also result from differences in standard of care across jurisdictions or availability of resources (e.g., tertiary versus primary care settings or rural versus urban settings). Variations in the strength of predictors can result from overfitting of models or from variations in the definition of predictors and outcomes. This can be minimized through the use of standardized definitions for predictors and outcomes and transparent reporting as described in the Transparent Reporting of a multivariate prediction model for Individual Prognosis Or Diagnosis (TRIPOD) guidelines [53].

### Conclusions

Risk prediction modeling provides clinicians with an objective measure of an individual's absolute risk to guide treatment and prevention strategies. The increasing availability of prediction models developed to predict outcomes during pregnancy and delivery and in the post-partum period highlights the importance of targeting high-risk patients for prevention strategies [54]. However, the utility of risk prediction models in perinatal epidemiology is contingent on the use of appropriate modeling strategies for model development and validation, transparency in reporting of results, and assessment of clinical impact. Additionally, data linkage and data quality need to be optimized in order to facilitate the use of EHRs and administrative and claims databases for development and validation of prediction models and to improve the transportability of models across clinical settings and geographic locations. Population-based pregnancy registries linked to various databases including information from obstetrical visits (including genetic screening, ultrasound, and diagnostic tests), the antepartum and delivery period (including maternal and infant outcomes), neonatal outcomes, past clinical history, and vital statistics should be prioritized for the development and validation of prediction models in perinatal epidemiology. Birth and perinatal registries, similar to those available in Denmark (Danish Medical Birth Register), Norway (Medical Birth Registry of Norway), Finland (Medical Birth Registry), Canada (British Columbia Perinatal Data Registry), and the UK (Clinical Practice Research Datalink Pregnancy Registry) are a few databases that could be exploited for risk prediction in perinatal epidemiology due to the large number of individuals included in these databases, the longitudinal follow-up, and their representativeness of the general population. However, the quality of the data and linkage to external databases (as previously described) needs to be optimized in order to reduce the potential for measurement error and missing data and to improve the accuracy and generalizability of prediction models.

Future research in risk prediction modeling in perinatal epidemiology should focus on updating existing models and adjusting or recalibrating them to the local circumstances or settings rather than developing new models. This way, prediction models may strengthen evidence-based, individualized decision making and can contribute to a rational use of scarce resources. When new prediction models are needed, considerations regarding the clinical setting and the outcomes of greatest importance should be prioritized to increase their transportability to the target population. Despite the challenges of implementing prediction models in clinical practice, they are useful in improving our understanding of how risk factors contribute to the burden of disease and for identifying women and infants who would benefit from available treatments.

## Compliance with Ethical Standards

**Conflict of Interest** Sonia Grandi is supported by a Doctoral award from the Fonds de recherche du Québec – Santé (FRQS) and reports grants from the Canadian Institutes of Health Research (CIHR), outside the submitted work. Kristian Filion holds a Junior 2 award from the FRQS and a William Dawson Scholar award from McGill University; he reports grants from CIHR and the Quebec Foundation for Health Research, outside the submitted work. Robert Platt and Jennifer Hutcheon each declares no potential conflicts of interest.

**Human and Animal Rights and Informed Consent** This article does not contain any studies with human or animal subjects performed by any of the authors.

## References

Papers of particular interest, published recently, have been highlighted as:

- Of importance

1. Kannel WB, McGee D, Gordon T. A general cardiovascular risk profile: the Framingham study. *Am J Cardiol.* 1976;38:46–51.
2. Steyerberg EW, Moons KG, van der Windt DA, et al. Prognosis research strategy (PROGRESS) 3: prognostic model research. *PLoS Med.* 2013;10:e1001381.
3. Coppus SF, van der Veen F, Opmeer BC, Mol BW, Bossuyt PM. Evaluating prediction models in reproductive medicine. *Hum Reprod.* 2009;24:1774–8.
4. Ukah UV, De Silva DA, Payne B, et al. Prediction of adverse maternal outcomes from pre-eclampsia and other hypertensive disorders of pregnancy: a systematic review. *Pregnancy Hypertens.* 2018;11:115–23.
5. Bakalis S, Peeva G, Gonzalez R, Poon LC, Nicolaides KH. Prediction of small-for-gestational-age neonates: screening by biophysical and biochemical markers at 30–34 weeks. *Ultrasound Obstet Gynecol.* 2015;46:446–51.
6. Grobman WA, Lai Y, Landon MB, Spong CY, Leveno KJ, Rouse DJ, et al. Development of a nomogram for prediction of vaginal birth after cesarean delivery. *Obstet Gynecol.* 2007;109:806–12.
7. Koopmans CM, van der Tuuk K, Groen H, Doornbos JPR, de Graaf IM, van der Salm PCM, et al. Prediction of postpartum hemorrhage in women with gestational hypertension or mild preeclampsia at term. *Acta Obstet Gynecol Scand.* 2014;93:399–407.
8. Ge WJ, Mirea L, Yang J, Bassil KL, Lee SK, Shah PS, et al. Prediction of neonatal outcomes in extremely preterm neonates. *Pediatrics.* 2013;132:e876–85.
9. Garite TJ, Casal D, Garcia-Alonso A, Kreaden U, Jimenez G, Ayala JA, et al. Fetal fibronectin: a new tool for the prediction of successful induction of labor. *Am J Obstet Gynecol.* 1996;175:1516–21.
10. Heida KY, Bots ML, de Groot CJ, et al. Cardiovascular risk management after reproductive and pregnancy-related disorders: a Dutch multidisciplinary evidence-based guideline. *Eur J Prev Cardiol.* 2016;23:1863–79.
11. Maher GM, O’Keefe GW, Kearney PM, Kenny LC, Dinan TG, Mattsson M, et al. Association of hypertensive disorders of pregnancy with risk of neurodevelopmental disorders in offspring: a systematic review and meta-analysis. *JAMA Psychiat.* 2018;75: 809–19.
12. Shah ND, Steyerberg EW, Kent DM. Big data and predictive analytics: recalibrating expectations. *JAMA.* 2018;320:27–8.
13. Riley RD, Ensor J, Snell KI, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. *BMJ.* 2016;353:i3140 **This article highlights the challenges and opportunities for predictive modeling with big data. The authors discuss opportunities for external validation and for the assessment of the incremental value of new predictors. However, the authors also discuss the methodological challenges relating to the quality of data and inconsistencies in reporting of data with the use of large datasets.**
14. Leushuis E, Tromp M, Ravelli AC, van Huis A, Mol BW, Visser GHA, et al. Indicators for intervention during the expulsive second-stage arrest of labour. *BJOG.* 2009;116:1773–81.
15. Harrell FEJ. Regression modeling strategies - with applications to linear models, logistic and ordinal regression, and survival analysis. New York: Springer-Verlag; 2001.
16. Ambler G, Brady AR, Royston P. Simplifying a prognostic model: a simulation study based on clinical data. *Stat Med.* 2002;21:3803–22.
17. Harrell FE Jr, 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.
18. Steyerberg EW. Clinical prediction models. a practical approach to development, validation, and updating. New York: Springer; 2009.
19. Trudell AS, Tuuli MG, Colditz GA, Macones GA, Odibo AO. A stillbirth calculator: development and internal validation of a clinical prediction model to quantify stillbirth risk. *PLoS One.* 2017;12: e0173461.
20. von Dadelszen P, Payne B, Li J, Ansermino JM, Pipkin FB, Côté AM, et al. Prediction of adverse maternal outcomes in pre-eclampsia: development and validation of the fullPIERS model. *Lancet.* 2011;377:219–27.
21. Tosteson AN, Begg CB. A general regression methodology for ROC curve estimation. *Med Decis Mak.* 1988;8:204–15.
22. Royston P, Altman DG. External validation of a Cox prognostic model: principles and methods. *BMC Med Res Methodol.* 2013;13:33.
23. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology.* 2010;21:128–38.
24. Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. *Stat Med.* 2004;23:723–48.
25. Robins JM, Finkelstein DM. Correcting for noncompliance and dependent censoring in an AIDS Clinical Trial with inverse probability of censoring weighted (IPCW) log-rank tests. *Biometrics.* 2000;56:779–88.
26. Pencina MJ, D’Agostino RB. Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Stat Med.* 2004;23:2109–23.
27. Gerds TA, Schumacher M. Consistent estimation of the expected Brier score in general survival models with right-censored event times. *Biom J.* 2006;48:1029–40.
28. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation.* 2007;115:928–35.
29. Janes H, Pepe MS, Gu W. Assessing the value of risk predictions by using risk stratification tables. *Ann Intern Med.* 2008;149:751–60.
30. Pencina MJ, D’Agostino RB Sr, D’Agostino RB Jr, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med.* 2008;27:157–72 discussion 207–12.
31. Payne BA, Groen H, Ukah UV, Ansermino JM, Bhutta Z, Grobman W, et al. Development and internal validation of a multivariable model to predict perinatal death in pregnancy hypertension. *Pregnancy Hypertens.* 2015;5:315–21.

32. Efron B, Tibshirani R. Improvements on cross-validation: the .632+ bootstrap method. *JASA*. 1997;92:548–60.
33. Moons KG, Kengne AP, Grobbee DE, et al. Risk prediction models: II. External validation, model updating, and impact assessment. *Heart*. 2012;98:691–8.
34. Ukah UV, Payne B, Hutcheon JA, et al. Assessment of the fullPIERS risk prediction model in women with early-onset pre-eclampsia. *Hypertension*. 2018;71:659–65.
35. Pauker SG, Kassirer JP. The threshold approach to clinical decision making. *N Engl J Med*. 1980;302:1109–17.
36. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Mak*. 2006;26:565–74.
37. Steyerberg EW, Pencina MJ, Lingsma HF, Kattan MW, Vickers AJ, Van Calster B. Assessing the incremental value of diagnostic and prognostic markers: a review and illustration. *Eur J Clin Investig*. 2012;42:216–28.
38. 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.
39. Moons KG, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? *BMJ*. 2009;338:b375.
40. Qiu H, Yu HY, Wang LY, Yao Q, Wu SN, Yin C, et al. Electronic health record driven prediction for gestational diabetes mellitus in early pregnancy. *Sci Rep*. 2017;7:16417.
41. Li T, Gao C, Yan C, Osmundson S, Malin BA, Chen Y. Predicting neonatal encephalopathy from maternal data in electronic medical records. *AMIA Jt Summits Transl Sci Proc*. 2017;2018:359–68.
42. Kuzniewicz MW, Puopolo KM, Fischer A, Walsh EM, Li S, Newman TB, et al. A quantitative, risk-based approach to the management of neonatal early-onset sepsis. *JAMA Pediatr*. 2017;171:365–71.
43. Goyal NK, Hall ES, Greenberg JM, Kelly EA. Risk prediction for adverse pregnancy outcomes in a medicaid population. *J Women's Health (Larchmt)*. 2015;24:681–8 **This article identifies the potential limitations of using electronic health records for prediction modeling including missing data, losses-to-follow-up, and the selective availability of clinical information as a result of patient contact with the healthcare system. The article also highlights the limitations of the most commonly used evaluation metrics (e.g., c-statistic) to assess the performance of prediction models and their inability to accurately assess the clinical utility of models for use in clinical practice.**
44. Wessler BS, Paulus J, Lundquist CM, Ajlan M, Natto Z, Janes WA, Jethmalani N, et al. Tufts PACE Clinical Predictive Model Registry: update 1990 through 2015. *Diagn Progn Res*. 2017;1:20–7.
45. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JP. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *J Am Med Inform Assoc*. 2017;24:198–208.
46. Dalton JE, Perzynski AT, Zidar DA, Rothberg MB, Coulton CJ, Milinovich AT, et al. Accuracy of cardiovascular risk prediction varies by neighborhood socioeconomic position: a retrospective cohort study. *Ann Intern Med*. 2017;167:456–64.
47. Galea S, Keyes KM. Population health science and the challenges of prediction. *Ann Intern Med*. 2017;167:511–2.
48. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. *Stat Med*. 2011;30:377–99.
49. Flenady V, Koopmans L, Middleton P, Frøen JF, Smith GC, Gibbons K, et al. Major risk factors for stillbirth in high-income countries: a systematic review and meta-analysis. *Lancet*. 2011;377:1331–40.
50. Varner MW, Silver RM, Rowland Hogue CJ, Willinger M, Parker CB, Thorsten VR, et al. Association between stillbirth and illicit drug use and smoking during pregnancy. *Obstet Gynecol*. 2014;123:113–25 **This article discusses the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement which underscores the importance of transparency in reporting of prediction models. The article also emphasizes the need for external validation and assessment of the clinical impact of previously developed prediction models in order to facilitate their use in clinical decision making.**
51. Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *Am J Epidemiol*. 2004;159:882–90.
52. van Smeden M, Moons KG, de Groot JA, et al. Sample size for binary logistic prediction models: beyond events per variable criteria. *Stat Methods Med Res*. 2018;962280218784726. <https://doi.org/10.1177/0962280218784726>
53. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD). *Ann Intern Med*. 2015;162:735–6.
54. Zulman DM, Vijan S, Omenn GS, Hayward RA. The relative merits of population-based and targeted prevention strategies. *Milbank Q*. 2008;86:557–80.