



Universiteit
Leiden
The Netherlands

Addressing missingness in predictive models that use electronic health record data

Lin, S.S.; Groenwold, R.H.H.; Mehta, H.B.; Kim, J.S.; Segal, J.B.

Citation

Lin, S. S., Groenwold, R. H. H., Mehta, H. B., Kim, J. S., & Segal, J. B. (2025). Addressing missingness in predictive models that use electronic health record data. *Annals Of Internal Medicine*, 178(10), 1451-1463. doi:10.7326/ANNALS-24-01516

Version: Publisher's Version

License: [Licensed under Article 25fa Copyright Act/Law \(Amendment Taverne\)](#)

Downloaded from: <https://hdl.handle.net/1887/4300255>

Note: To cite this publication please use the final published version (if applicable).

Addressing Missingness in Predictive Models That Use Electronic Health Record Data

Shanshan Lin, ScM; Rolf H.H. Groenwold, MD, PhD; Hemalkumar B. Mehta, MS, PhD; Ji Soo Kim, PhD; and Jodi B. Segal, MD, MPH

Electronic health record (EHR) data are increasingly used to develop prediction models that guide clinical decision making at the point of care. These include algorithms that use high-frequency data, like in sepsis prediction, as well as simpler equations, such as the Pooled Cohort Equations for cardiovascular outcome prediction. Although EHR data used in prediction models are often highly granular and more current than other data, there is systematic and nonsystematic missingness in EHR data as there is with most data. Despite growing use for clinical decisions, algorithms implemented in EHRs are mostly unregulated and are often opaque to the user. Guidelines about the

development, validation, implementation, and reporting on clinical prediction models are sparse in their recommendations regarding missing data. This article characterizes missingness in EHR data, summarizes methods for attending to missing data when developing prediction models, makes recommendations about validation and implementation of models in practice when data are missing, and identifies research needs in this field.

Ann Intern Med. 2025;178:1451-1463. doi:10.7326/ANNALS-24-01516

For author, article, and disclosure information, see end of text.

This article was published at Annals.org on 9 September 2025.

A 62-year-old woman with type 2 diabetes consults her internal medicine physician for guidance. During the visit, they use the Pooled Cohort Equations to estimate her 10-year primary risk for atherosclerotic cardiovascular disease (ASCVD). The risk calculator is programmed into the electronic health record (EHR), and with a simple command, the calculator reports 10-year risk for ASCVD of 6.2%. The calculator had incorrectly assumed that the patient is not being treated for hypertension, as the medication data from the EHR and the text of the clinical notes are not accessible to the risk calculator. The physician then enters the information into a web-based calculator and correctly enters the patient's active treatment for hypertension, and her risk is 8.2%.

The ASCVD prognostic algorithm in this scenario was developed using data from 3 large cohort studies and is now widely implemented in practice (1, 2). Clinicians increasingly use estimates from prognostic models to make decisions with patients, including initiating use of statins or bisphosphonates or preparing for kidney replacement therapy. Astute users of prediction rules will question how the algorithms' creators attended to missing data during the model's development and validation and how the algorithms implemented in their own EHRs respond to missing data elements. In the instance above, the algorithm defaults to "no" treatment for hypertension when the information is unavailable, thus underestimating the risk for ASCVD.

In this article, we address considerations about missing data, specifically in the EHR, in the development, validation, and implementation of prognostic or predictive models intended for use in clinical practice. For brevity, we will use the term *prediction models* throughout this article. We characterize missingness in

EHR data, summarize methods for attending to missing data when developing and validating models, make recommendations about the implementation of prediction models in clinical practice when there are missing data, and identify research needs in this field.

This is a different challenge than using EHR data to make inferences about treatment effects. Incompleteness in EHR data may contribute to biases and health inequities if the missingness varies across populations (3); the predictions will vary in their accuracy across populations, leading to inequitable treatment recommendations. Here, we contextualize missing data in EHRs to raise awareness among clinicians using algorithms deployed in the EHR and to highlight outstanding research needs in the field.

THE PROBLEM OF MISSING DATA WHEN DEVELOPING, VALIDATING, AND IMPLEMENTING PREDICTION MODELS IN THE EHR

Despite growing use for decision making, algorithms implemented in EHRs are mostly unregulated. Currently, few decision support software applications are certified by the U.S. Food and Drug Administration (FDA), although this is evolving. In 2013, a workgroup of the International Medical Device Regulators Forum defined "software as a device" as "software intended to be used for one or more medical purposes ... without being part of a hardware medical device" (4). In 2017, the FDA led this workgroup's efforts in preparing

See also:

Web-Only
Supplement

guidance for industry about the use of software as stand-alone medical devices (5), and in 2021, the FDA released a framework to guide the incorporation of artificial intelligence and machine learning into software as a medical device (6). The Office of the National Coordinator for Health Information Technology has issued a rule to improve transparency in algorithm development (7).

Algorithms deployed in clinical practice are most often models used to inform a decision. These are developed using various sources of data, including administrative claims data (8, 9), data from cohort studies (10, 11), data collected in the setting of clinical trials (12–14), and data from the EHR that are generated during usual care (15, 16). Although still uncommon, EHR data are increasingly used for the development of clinical prediction models to inform point-of-care recommendations (17, 18). These include algorithms that use high-frequency data, like for sepsis prediction (19), and simple equations such as the Padua tool for inpatient venous thromboembolism prediction (20). Any of these data sources will have some missing data that must be considered by the developers. Guidelines about the development, validation, implementation, or reporting on clinical prediction rules are sparse in their recommendations, and none is specific to EHR-based prediction models (**Supplement Table 1**, available at [Annals.org](https://annals.org)). Most of the published recommendations simply advise researchers to report the number of persons with missing predictors or outcomes, compare characteristics of persons with and without missing data, and discuss limitations regarding missing data. This is insufficient in the development and implementation of a tool that will guide clinical care.

The Pooled Cohort Equations model described earlier was developed in 3 cohorts (ARIC [Atherosclerosis Risk in Communities], CARDIA [Coronary Artery Risk Development in Young Adults], and CHS [Cardiovascular Health Study]) supported by the National Heart, Lung, and Blood Institute and was then validated in 2 other large cohorts (21). In 2013, the developers called on EHR vendors “to incorporate these risk algorithms directly into their products to support meaningful use” despite neither development nor validation of these algorithms in EHR data. The developers make no mention of addressing missing data in the development of the algorithms nor any suggestions about how missing data might be attended to when deployed in an EHR. Here, we focus specifically on models that are developed, validated, or implemented using EHR data given the abundant challenges.

Data from EHRs are highly granular and are current, often including data from the visit in which the clinical decision is needed (for example, body mass index or oxygen saturation). The EHR data typically include point-of-care measures; laboratory testing results; diagnoses and procedures coded for submission

for payment; data collected routinely from patient survey, such as social determinants of health or health-related behaviors; and sometimes discrete data recorded from testing, such as left ventricular ejection fraction. The unstructured text recorded in clinical notes is generally less available for modeling, although natural-language processing makes this increasingly viable (22). The primary challenge that we address in this article is that, despite the clinical richness of these data, there is systematic and nonsystematic missingness in EHR data (3). In a systematic review of prediction models that were developed with EHR data, only 54% of the included articles described how the developers addressed missing data, and this was often described very generally (17).

Also worrisome is the scant attention to the deployment of prediction tools in a clinical setting where data elements are incompletely available because of the usual care seeking and care delivery behaviors. Despite this expected missingness, most EHRs do not have analytic capabilities to fill in missing information; therefore, most prediction tools are deployed as simple calculators. An outcome is only generated if the patient has information for each required variable or the calculator is programmed to default to a value when data are missing. Imputation—filling in missing values using a model—requires that data be transferred from the EHR system to a separate computing environment, with results then returned to the EHR. Epic Systems has a computing environment, called Nebula, for that purpose, but it is largely used for models that assess health system performance retrospectively and not in real time to guide individual patient care (23). Even the carefully designed Severe COVID-19 Adaptive Risk Predictor (SCARP) model for prediction of death due to COVID-19 (24) has no automated provisions for addressing missing data upon deployment; it requires the user to enter the data when prompted by the algorithm.

In this article, we assume that data have been prepared by the clinical informaticians within the institution following best practices, meaning that the processes for extracting, transforming, and harmonizing the data elements are appropriate and complete and the data are ready for use for model development. Broad considerations about the use of clinical data for research, regulatory concerns, and human subjects protection are outside the scope of this discussion (25).

KEY MISSING DATA CONCEPTS

The key concern is that missingness is often informative, meaning that if a measurement or observation is missing in the EHR, that missingness may convey information about the patient that is predictive of the future (26–28). Because missing data cannot be ignored in inferential studies testing hypotheses, it cannot be ignored in the development, validation, and deployment of prediction models (17, 18).

Defining Missingness and Its Causes

In any data set, when data are missing, there are unobserved values that would be meaningfully included in analyses or models if they were observed (29). Missingness can occur at the level of the individual, the encounter, or values from that encounter. To conceptualize the reasons for which data are missing, we have created a few new definitions (Glossary). We define *absent data* as data that are not generated because there was no clinical encounter—the patient did not connect with care anywhere. This may be because they did not need care or because they were unable to receive it. We consider *ghost data* to be data that are invisible within the health system’s data because care was delivered outside the health system—the patient visits a clinic that does not transmit data to their primary health system. We define *vanished data* as data that are expected to be present but are not—for example, there was no note from the encounter or the patient’s weight was not recorded.

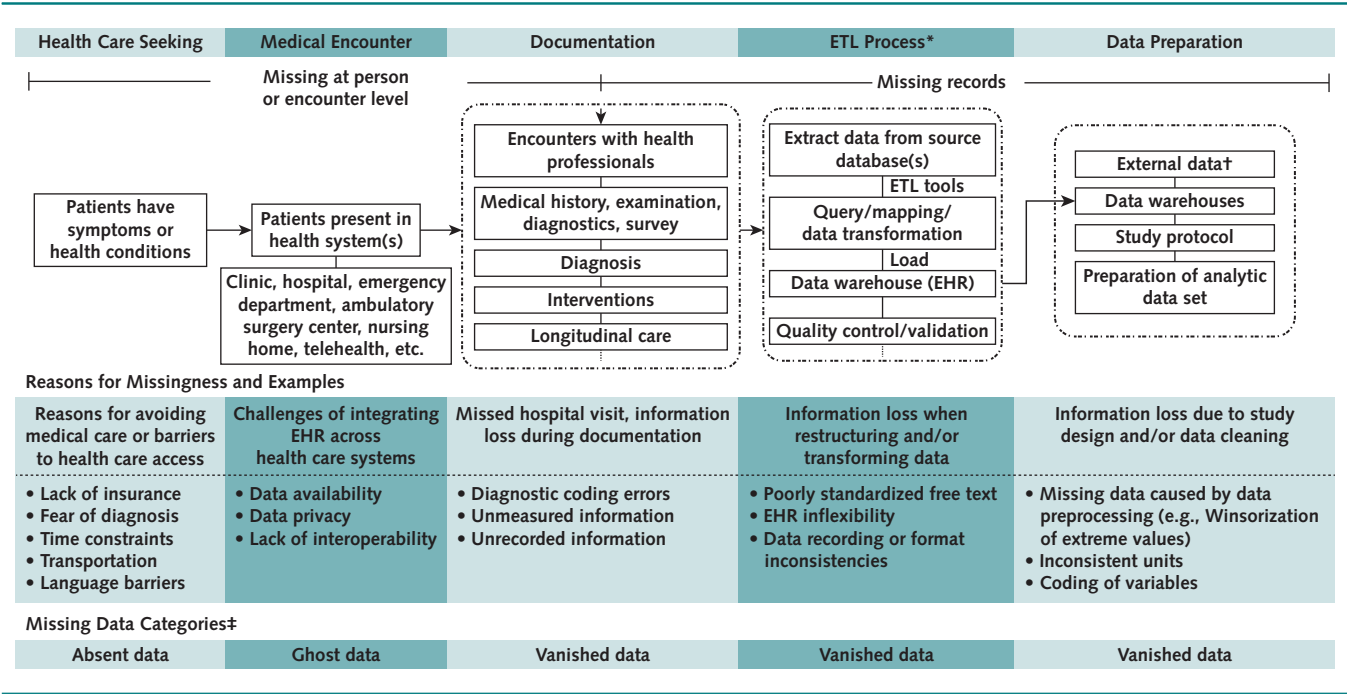
This framing necessarily precedes the consideration of the missing data mechanisms that are commonly assumed when making inferences with data; data are missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR) (29). Consideration of the reasons for the missingness in the data (absent, ghost, or vanished) suggests the most likely missing data mechanisms (Figure 1). These considerations are similar to what others have termed

“submechanisms,” which lead to missingness in exposures, confounders, and outcomes (26). Although the reasons for missingness in EHR data are informative and are expected to affect a prediction model’s validity, there are few empirical data to support this. Assuming a single reason for missingness of a variable and then a single missing data mechanism is an oversimplification, but this is typical in practice (26).

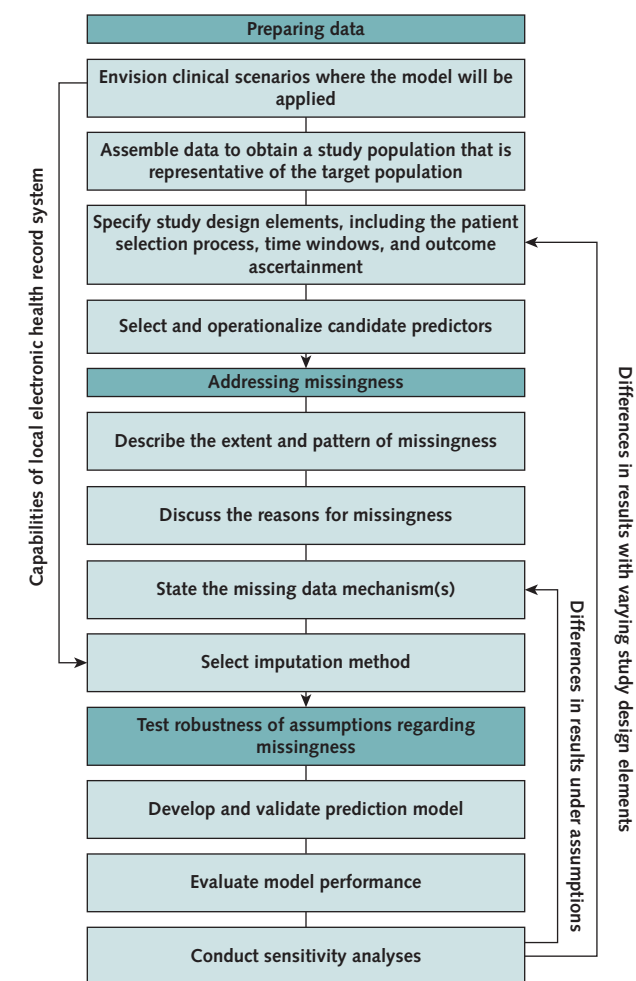
We illustrate here that the reason for missing data suggests the plausible mechanism that may inform the modeling. To illustrate, if a clinic’s automated sphygmomanometer malfunctions with no backup and there is no record of the blood pressure measurement, this vanished blood pressure can be considered to be MCAR because the malfunction is a seemingly random event that led to missing data. In contrast, persons with severe mental illness have less regular measurement of hemoglobin A_{1c} because they miss scheduled visits due to impairment. Here, the extent of absent hemoglobin A_{1c} values will differ in persons with and without severe mental illness, and missing values are best considered to be MNAR (28).

In addition, the reasons for missing data may differ across clinical scenarios, settings, or populations. The clinical data collected in a subspecialty visit differ from those collected in a primary care visit, and the completeness of coding differs for patients in a capitated versus a noncapitated insurance plan (33). We note also that data in the EHR are generated at irregular intervals, unlike the data collected in clinical trials or in

Figure 1. Possible reasons for missing data at different stages of the data-generation process.



EHR = electronic health record; ETL = extract, transform, load.
* ETL is defined as the process of combining data from multiple sources into a large central repository known as a data warehouse. It is sometimes used interchangeably with extract, load, transform (ELT).
† External data may include information from sources such as REDCap, imaging, and monitors.
‡ We identified the most likely missing data category in these scenarios, but it may vary depending on the context. (Informed by references 30, 31, and 32.)

Figure 2. Process of addressing missing data in prediction models.

prospective cohort studies with data collection protocols. These reasons for missingness in the data should motivate choices during model development that are informed by an understanding of the population in whom the tools will be deployed.

PREPARING A PREDICTION MODEL WITH ATTENTION TO MISSINGNESS

The developers of prediction models have options at all stages of development, validation, and implementation of a prediction model in an EHR, beginning with consideration of how the model will be used in practice (Figure 2). We expect that the developers of the Pooled Cohort Equations anticipated its implementation in EHRs, but it is unclear how this informed their model development.

Specify the Clinical Scenarios in Which the Models Will Be Applied

The first step is to specify the clinical scenarios in which the clinical prediction models will be applied,

including the target population for whom the model will be used. The developer should start by consulting with informatics professionals about the capabilities of deploying candidate models in the local settings, including local computing opportunities for addressing missing data. Given the limitations of most EHRs (that is, the inability to do the real-time model fitting needed with imputation), the developer must think about the feasibility of using these methods while the model is being developed. Developers should consider how the model will be used in practice. Will the clinician be able to query or measure the patient to fill in missing values at the point of care ("Did your mother have a hip fracture?"), or is this a model that will generate a predicted risk without additional input from the user? This suggests whether the missingness in some variables will be remediable at the point of care. In our setting, to use the Pooled Cohort Equations, the clinician needs to go outside the EHR to use a calculator at the point of care to update the calculator with information learned during the visit.

Prediction models ideally are transportable—that is, the model can be validly applied in later years or across disparate hospitals and clinics. Therefore, the developer must seek data for model development from a population that is representative of populations in which the model will be used. The targeted population is expected to have some degree of missing data; thus, the developer generally should not include only the patients having complete data for model development, as this will not be generalizable to most patient populations. Sometimes, however, this is a reasonable choice. The developers of a tool to predict relapse in patients with multiple sclerosis opted to include only the small set of variables that they confidently expected would not be missing in a clinical setting (34). Therefore, they required complete data in the cohort in which they developed the model.

Minimize Missing Encounters in the Data

The developer should aim to reduce missing encounters in the development data. With growing use of health information exchanges since the passage of the Health Information Technology for Economic and Clinical Health (HITECH) Act in 2009, statewide or regional data are more likely to be available to a given health system (35). For example, the Chesapeake Regional Information System for our Patients (CRISP) facilitates the sharing of information across 57 hospitals in Maryland and the District of Columbia. Within an EHR platform, such as Epic's Care Everywhere, there may be sharing of patient information across institutions (36). This shared information may be particularly valuable if the model is predicting utilization outcomes, such as emergency department visits or hospitalizations.

Select Potential Predictors for the Model, With Consideration of Completeness

The developer will choose how the candidate predictors and outcomes are operationalized during the development of the model and might select variables that are expected to have less missingness than others. For example, some predictor variables may require in-person contact to be observed (for example, a laboratory measure), whereas others do not (for example, a prescription for an antihypertensive medicine). A laboratory measurement may have greater missingness in settings where patients are more challenged in accessing health care and may limit transportability of the model (36). Some measures may be predictive of outcomes, yet if they are infrequently measured in practice, they are not valuable in a deployed prediction model; an example is apolipoprotein B for ASCVD prediction.

Apply Sound Epidemiologic Methods for Defining the Cohort in Whom the Model Is Developed

The developmental cohort is the population in whom the model is developed. As in any cohort study, there will be inclusion and exclusion criteria that should reflect the population to whom the prediction model will be applied in practice. The developers of the Pooled Cohort Equations explicitly sought existing cohorts that were biracial, community-based populations so that the models would be more applicable to the U.S. population (21). Also, the relevant time periods will be specified in development of the model. There is a baseline period during which data accrue for inclusion in the model. This span of time depends on the use of the model: Sepsis prediction models use very granular data from a short interval (hours), whereas ASCVD prediction models may use information on high-density lipoprotein cholesterol from 2 years previously given the stability of that measure. There is a period after time zero (the end of the baseline data collection) during which there will be observation for the outcome's occurrence (for example, 1 day, 30 days after discharge, 2 years). Similar inclusion and exclusion criteria and interval definitions will be applied when the model is validated in a new population. Importantly, these intervals need to be clinically relevant for the deployment of the prediction tool in practice. If we anticipate that there will be scant data from previous encounters, such as in an emergency department setting, a model that uses predictors from the past 6 months is inappropriate.

Describe the Missingness Pattern of the Predictors

The developer should use descriptive analytics about the frequency of missingness and its distribution in the population in whom the model is being developed. Missing data patterns are the distribution of missing values in the rows and columns of data that are derived from EHR data (29). The user can describe the extent to which values are missing across the population (for

example, persons with severe mental illness have infrequently recorded screening colonoscopy) and whether there are predictable patterns to the missingness (for example, a missing sodium measure is highly correlated with a missing potassium measure). There is software to assist with this process; for example, software developed for missingness when answering inferential questions may be equally applicable here (37). Evaluation of missing patterns will help with articulation of the missing data reasons and mechanisms. In addition, incorporating information about the patterns of missingness into the prediction model may improve its performance (38). For example, the missing indicator approach, though not widely recommended, is valuable in some situations (39). In this approach, the prediction model includes an additional variable that indicates whether the predictor variable was present or absent.

State the Assumption of the Missing Data Mechanism for Key Variables

The developer will consider why data are missing to inform assumptions about the mechanisms and inform the selection of methods to reduce its effect. The most likely missing data mechanisms should be proposed for the variables being considered for inclusion in the model as well as for the outcome of interest. The most problematic missingness is predictor variables that are MNAR—there is a reason that these data are missing, and it is related to the outcome that is being predicted (40, 41) (**Supplement Table 2**, available at [Annals.org](https://www.annals.org)). Having a single missing data assumption for all variables is generally an oversimplification; even individual variables may be missing due to a mixture of mechanisms. The developer may turn to external resources or published evidence to support the missing data assumption for a given variable. It may be most conservative to consider all the missing variables to be MNAR and choose appropriate models to accommodate this assumption, yet these models come with additional assumptions.

Methods for Addressing Missing Data in Prediction Models

The choice of the analytic plan to address missing data in prediction models is guided by the assumptions about missing data mechanisms. In the rare situations where the data are clearly MCAR, straightforward methods are available, including use of the last observation carried forward (if a variable is time-invariant or a baseline covariate that does not change over time) or simply complete-case analysis if the population with complete data is large enough (42). In the more common situation, in which data are MAR, there are many analytic options but limited information from simulations in the context of prediction to inform selection of the most appropriate method (41) (**Table 1**). The developer will weigh the tradeoffs of the methods, recognizing that some may be infeasible at the point of care

Table 1. Commonly Used Methods When Missing Data Are MAR or MCAR*

Classification†	General Description	Examples	Pros	Cons
Complete-case analysis (43)	Uses only observations with complete information on prognostic factors	–	Easy to implement	Loss of information Unrealistic assumptions in EHR
Missing indicator approach	Assigns an indicator to missing values and retains all observations in the outcome model	–	Shown to improve model performance (41)	Impractical to use in clinical settings (39)
Weighting methods (44)	Estimate values based on the observed values while accounting for the probability of missing values	Propensity-weighted method	Easy to apply when there is monotone missingness due to dropout (44)	Model that generates the weight needs to be correctly specified (44) May require large sample (43)
Likelihood-based methods (44, 45)	Impute missing values using conditional means based on the observed values and parameters estimated from the previous iterations, thus maximizing the likelihood of complete data	E-M algorithm (ECM, ECME, AECM, and PX-EM algorithms)	Valuable method when data are MAR	Model specification must be correct (44)
Multiple imputation (46, 47)	Iteratively estimates multiple values for the missing data based on the distribution of the observed data	Multiple imputation, multilevel multiple imputation	Valuable when data are MAR or MCAR (44) Leverages auxiliary variables	Computationally demanding at the stage of implementation (28)
Bayesian approach (48-50)	Chooses reasonable priors and computes a posterior distribution for the missing values	Gibbs sampling, Bayesian multilevel modeling approach, Bayesian additive regression trees	Useful when the sample is small Accounts for uncertainty in the parameters	Bayesian inference requires a high degree of model specification (full specification of a likelihood distribution and priors)
Machine-learning method (51, 52)	Uses algorithms with built-in capabilities for handling missing values or imputing missing data	Random forest, decision tree, kernel-based methods, extreme gradient boosting (XGBoost), deep learning/neural network, support vector machine	Valuable for high-dimensional data Allows interactions and nonlinearity Does not need model specification	May not actually address the missingness May be computationally intensive and unintuitive

AECM = alternating expectation conditional maximization; ECM = expectation/conditional maximization; ECME = expectation/conditional maximization either; EHR = electronic health record; E-M = expectation-maximization; MAR = missing at random; MCAR = missing completely at random; PX-EM = parameter-expanded expectation-maximization.

* The current framework was adopted from references 53 and 54. The assumption of the mechanism for missingness in the table is the underlying assumption of the method to obtain an unbiased estimate, mostly discussed in the context of inferential studies.

† Items in the classification are not mutually exclusive. For example, the likelihood-based method could be regarded as an imputation technique, while multiple imputation (using the Markov-chain Monte Carlo algorithm) is also a Bayesian approach.

due to computational demands. Given that individual predictor variables may vary in their missing mechanisms, the developer may make the conservative choice of assuming that missing data among the predictors are MNAR, meaning that the missingness is informative. Generally, the methods for addressing this informative missingness can be classified into 2 broad categories: selection models, and pattern-mixture models (55, 56), both of which are also options when data are MAR (Supplement Table 2). These are statistical approaches for handling informative missing data. Selection models assume that the probability of missingness depends on both observed and unobserved data through a joint model of the outcome and missingness process, and pattern-mixture models account for missing data by modeling the outcome distribution separately for each missing data pattern, often requiring assumptions about the relationship between patterns to identify parameters (56, 57). These categories are differentiated by how they

address the missing data before the data are used in the prediction models.

Single-imputation models are simple regressions where the missing predictor value is estimated from a model based on the other predictor data. A variant of this is stochastic imputation, which allows for some uncertainty around the predicted value when the missing data are MAR. Multiple (or multivariate) imputation fills in the missing predictors with a distribution of values from a series of models to more correctly reflect the uncertainty around the true value. In the context of a clinical prediction tool for sepsis prediction, Sisk and colleagues did a set of simulations that began with their creating missing variables that were MCAR, MAR, and MNAR (41). They then compared the performance of single- and multiple-imputation models while also including missing indicators in the prediction model. In many scenarios, Sisk and colleagues observed comparable predictive performance under multiple

imputation and regression (single) imputation. They note that the strategy for handling missing data at development depends on whether missing data will be present at deployment; if missing data are expected at both the development and deployment of the prediction model, which is often the case, single regression might be a more practical alternative to multiple imputation. They discourage inclusion of the outcome variable in the imputation model during development. They encourage inclusion of missing indicators in the prediction model only if the data are most likely MNAR. In their simulations, they found that inclusion of missing indicators in the model was inconsistently beneficial and required the assumption that the missing mechanism was the same in the development data and the deployment setting, which is hard to confirm. Groenwold cautions that the use of missing indicators in a clinical setting may be problematic; the predictive value of a missing indicator may change over time as clinicians change their information-gathering behaviors as prompted by clinical decision support based on the model. This feedback loop can render the missing data mechanism nonreproducible in practice, ultimately reducing the model's transportability (40).

There have been other simulations to compare imputation methods for clinical prediction. Deforth and colleagues simulated 36 data sets to predict risk for long COVID. The data sets varied in the proportion of noncomplete cases and the mechanisms for the missingness in the 9 predictor variables (58). They applied 3 existing imputation algorithms that are based on multivariate imputation by chained equations (MICE) and are available as R packages (*mice*, *aregImpute*, and *missForest*). MICE is also implemented in other statistical software, including Stata (MI command) and SAS (PROC MI and MIANALYZE), both of which provide thorough documentation. They compared the performance of the resultant models and found that the differences in model performance with the 3 imputation methods were generally small, although *missForest* had the lowest predictive accuracy. They found that the choice of algorithm affected discrimination more than it affected model calibration. Similarly, Hoogland and colleagues compared methods using both simulated data and data from a project in which they predicted mortality among recipients of implantable cardioverter defibrillators (59). They stressed that the methods used in development and validation should be those that are appropriate for making predictions in a "new" patient with missing information. They compared 6 modeling approaches and concluded that the differences in performance of the methods were small when the amount of missing data was low (<5%). As Sisk and colleagues concluded, they also recommend avoidance of inclusion of the outcome in the test set when validating a model for use in practice. These simulations provide valuable information to support that only missing predictors should be imputed when prediction models

are developed, which had been raised as an unresolved question by Little and Rubin (60).

As mentioned earlier, a further complication when using EHR data is that the predictors are values that come at irregular intervals given that patients seek care irregularly. There are methods to interpolate missing values that come at irregular intervals. These include models that capture subject-specific trends as random effects as well as the population mean trend as a fixed effect (61–65). In addition, Kalman filtering (66, 67), which is often used to describe trajectories, can also be used for recursive updating of the missing values while incorporating time dependence in a state transition model. State transition models are used when there are observed or latent time-dependent measures. A current state is modeled using past states and other predictors that define how longitudinal measures evolve over time (68, 69). We have not seen examples of simulations that have tested these methods for development or validation of clinical prediction rules.

Evaluation of the Developed Model

Considerable research has focused on the effect of missing data in inferential studies where the goal is to understand a relationship between an exposure and an outcome. When making inferences, the missing data management choices affect bias and precision in the estimates of interest. When doing prediction modeling, the missing data management choices will affect the model's validity, which is often reported as the model's discrimination and calibration (70) (Glossary). Best calibration may be particularly important as these models are intended to guide decision making based on predicted risk, which should mirror the actual risk (71). It is increasingly recommended to report metrics beyond traditional measures such as calibration and discrimination to assess the model's potential effect on decision making, patient outcomes, and equity (72). We summarize the metrics that have been used and reported in **Supplement Table 3** (available at [Annals.org](https://annals.org)). There are many metrics for model evaluation, although these have not been specifically assessed for their appropriateness across missing data management options. Similarly, there is little in the literature to suggest which evaluation metrics are better in small populations or for evaluating model performance in subgroups.

External Validation Considerations

A prediction model needs to be externally validated—that is, its performance must be assessed in data other than those in which it was developed. This may be in a population that is external to the health system or within the health system but later in time. The developer must confront the possibility that the extent or pattern of missing data or the mechanisms of missingness differ in the new data relative to the development data. If the missingness metrics (extent, patterns, mechanism) are markedly different, that may deleteriously affect model performance. The developer may need to reconsider

Table 2. Proposed Recommendation About Methods for Addressing Missing Data in Prediction Studies, Extended From the TRIPOD Checklist*

Section/Topic	Item	Development or Validation?	Checklist Item	Suggested Missingness Considerations
Methods				
Source of data	4a	D;V	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data) separately for the development and validation data sets, if applicable	Mitigate the amount of missing data through data integration or data linkage. [Section 3. Minimize missing encounters]
Predictors	7a	D;V	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured	Report any missingness caused by data preparation choices [Section 3. Minimize missing encounters; Select potential predictors for the model with consideration of completeness] [Delete row and discuss elsewhere]
Missing data	9	D;V	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation), with details of any imputation method	
Statistical analysis methods	10b	D	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation	Provide possible reasons for missing predictors and articulate plausible missing data mechanisms [Section 3. State the assumption of missing data mechanism for key variables] Specify analytic choices based on missing mechanism assumptions [Section 3. Methods for addressing missing data in prediction models] Describe sensitivity analyses assessing robustness of missing data assumptions and analytic choices [Section 3. Evaluation of the developed model] Describe sensitivity analyses assessing impact of study design choices on missingness [Section 3. Evaluation of the developed model]
Development vs. validation	12	V	For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors	Inspect missingness and missingness patterns, including separately in data sets used for development and validation [Section 3. External validation considerations]
Results				
Participants	13a	D;V	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.	Describe missingness throughout the patient selection process [Section 3. Apply sound epidemiologic methods for defining the cohort in whom the model is developed]
	13b	D;V	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome	Compare the characteristics of individuals with and without missing data. [Section 3. Describe the missingness pattern of the predictors] Describe the association of missing predictors with observed outcome (assumption for predictors) [Section 3. State the assumption of missing data mechanism for key variables]
Model specification	15b	D	Explain how to use the prediction model	Specify clinical scenarios in which the clinical prediction models will be applied, and discuss the availability of predictors and possible changes in missing mechanism assumptions if the implementation setting differs substantially from the development setting [Section 3. Specify the clinical scenarios in which the clinical prediction models will be applied, including the target population for whom the model will be used]

Continued on following page

Table 2–Continued

Section/Topic	Item	Development or Validation?	Checklist Item	Suggested Missingness Considerations
Model performance	16	D;V	Report performance measures (with CIs) for the prediction model	Include performance metrics from sensitivity analyses to assess the effect of missingness [Section 3. Evaluation of the developed model]

D = development; TRIPOD = Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis; V = validation.
* Only relevant rows are included. Boldface indicates the original checklist items that contain information about missing data for comparison.

the variables in the model and select variables that are more consistently measured.

IMPLEMENTATION

Despite their importance in development models, imputation methods may not be feasible in models implemented at the point of care because the computing environment may not accommodate imputation. When clinical prediction rules are deployed in practice, they often use nothing more complex than a calculator. There may be a simple point system that is assigned given the patient information, or perhaps a weight where the presence of a covariate is multiplied by a coefficient from a model to generate a score. There are more recent examples where the EHR data are directed to a cloud-based application that allows more complex analytics, with the “answer” returned to the EHR to inform clinical care (73, 74). Different health care systems have different tolerances for allowing their clinical data to leave the protected environment.

Some methods are more feasibly implemented within the environment of a health system’s EHR than other methods. For example, pattern-mixture models are considered more computationally feasible (56). Tree-based models that use sorting functions can be implemented within EHRs more easily than models that require regression or other real-time model fitting. Notably, BARTm (75), a tree-based Bayesian pattern-mixture model that incorporates missingness as part of the splitting criterion in the decision tree, may be a valuable model for use in practice. SCARP is a tree-based prognostic tool that accommodates missing data and has been used to predict risk for progression or death among adults hospitalized with COVID-19 (24). The more similar the patients and their missingness patterns are to those in the development and validation cohorts, the more likely it is that the prediction model will have similar performance in the setting of its intended use. When selecting the missing data approach for the implemented model, the planned approach at implementation should generally match the methods used at validation (76, 77).

Transparent reporting of decisions by prediction model developers will inform use of the model. It is important to report sufficient information about the missing data so the user can evaluate the appropriateness of assumptions and the selection of statistical methods. Our review of existing guidelines found that

missing data recommendations are insufficiently detailed (78) (Supplement Table 1). We suggest extensions to the TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) statement to promote more complete reporting about missing data considerations (79) (Table 2). Furthermore, as prediction models are often developed with data from multiple health systems or hospitals, we urge investigators to review the TRIPOD-Cluster checklist, although its recommendations for missing data do not differ from those in the original TRIPOD statement and are also incomplete (80). In addition, we highlight the publication of TRIPOD + AI (81), which explicitly updates and replaces the original TRIPOD statement for studies developing or validating prediction models, regardless of whether regression or machine-learning methods are used.

MOVING FORWARD

Additional methodological work is needed to establish best practices for development and validation of clinical prediction rules using EHR data when there is missingness. A prominent need is best practices when the observations are clustered, which is more likely than not to occur. There may be missingness at any level of the multilevel clustered data, such as missing clinics, missing encounters, or missing measures within an encounter. In 2006, Reiter and colleagues prepared simulations and a real-world example using data from the National Health and Nutrition Examination Survey (not EHR data) (82). These investigators suggested that there are benefits to using hierarchical models for imputation of missing data relative to using fixed-effects models, although they may be difficult to fit. They suggested that it may be possible to fit sequential hierarchical models (83).

More recent work has used Markov-chain Monte Carlo (MCMC) methods, assuming that the missingness was MAR, to impute data for an upper level using weighted data from the lower levels (84). These investigators also demonstrated the use of auxiliary variables, which are included in the model to predict missing values that are not included in the outcome prediction model. Another recent analysis used New York State Vital Records Registry data rather than EHR data and used a joint imputation model where the variance-covariance structure was modified to reflect the correlated data units (85). As in the previous example, the

GLOSSARY

Absent data: Data not generated because there was no clinical encounter.

Calibration: The absolute risk estimates and the correspondence between predicted and observed risk.

Clinical prediction tool: Algorithm deployed in clinical practice based on a prognostic or prediction model.

Discrimination: The extent to which the model can distinguish between those who will develop the outcome of interest and those who will not. In the case of a binary outcome, discrimination is often quantified with the c-statistic; for time-to-event (survival) models, the c-index is reported.

Ghost data: Data that are invisible to the user because care was delivered outside the health system of record.

Missing at random: The missingness depends on the values of the data only through the observed components.

Missing completely at random: The missingness is independent of the values of the data.

Missing not at random: The missingness depends on the missing components.

Prediction model: A mathematical model that generates patients' expected outcomes conditional on characteristics measured at baseline and treatment.

Prognostic model: A mathematical model that generates patients' expected outcomes conditional on characteristics measured at baseline.

Vanished data: Data where information is expected but is not present.

researchers used MCMC sampling techniques. They also described an alternative approach, a variable-by-variable imputation method (similar to sequential regression imputation), although they stated that this would be uncommonly used in multilevel settings.

Perhaps more important, additional work with simulation is needed to evaluate whether more complex methods have important effects on predictions at deployment. Sisk and colleagues' work suggests that simple regression is often adequate (41); Hoogland and Deforth and their respective colleagues suggest that there are only small differences in model fit between imputation approaches, particularly when the extent of missing data is small (58, 59). Indeed, simple may be better if computing environments and patient data protection practices do not allow advanced imputation in real time. Earlier, we mentioned Epic's cloud-

based platform, Nebula. This is computing space outside the EHR where data can be sent for calculations, including imputation, with results then reimported. With HL7 FHIR (Fast Health Interoperability Resources), this can be accomplished in real time, although more typically the data are sent once daily. Currently, Nebula is mostly used for operational decision making and not for individual-level predictions. A preprint manuscript describes the Amazon Web Services cloud computing platform for individual-level prediction of diabetes (86). In this example, investigators integrated data from an EHR, biochemical indicators, and lifestyle data to predict diabetes. They describe using the "Imputer" to fill missing values for important features, such as blood glucose, blood pressure, body mass index, and age. The authors do not describe the results with and without use of the Imputer, but this manuscript nicely describes the real-time use of cloud computing and imputation for missing values, although without much detail about that process.

CONCLUSION

The Pooled Cohort Equations, which were rigorously developed and validated using cohort data, illustrate some challenges that arise with a prediction model that is intended for direct clinical care. Implementation details, such as how missing data are addressed in the embedded software of users' EHRs, are needed to understand the model's usability in practice; this was not considered at the time of development and validation of this ASCVD prediction tool.

As EHR data are increasingly used for clinical prediction and to guide care decisions, there is an urgent need to think carefully about how missing data are addressed, particularly as there may be less transparency in these models than in models developed for inference. This is becoming increasingly important as machine-learning methods are used in prediction models, which may further reduce transparency. Simulations are still needed for missing data approaches, where there are assumptions that the missing data mechanisms are representative of those encountered in clinical practice, and for evaluation of the technical feasibility of applying these methods at the point of care.

From Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland (S.L.); Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, the Netherlands (R.H.H.G.); Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland (H.B.M.); Division of Rheumatology, Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland (J.S.K.); and Department of Medicine, Johns Hopkins University School of Medicine, and Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland (J.B.S.).

Note: Dr. Jodi B. Segal, Associate Editor for *Annals*, had no role in the editorial review of or the decision to publish this article.

Grant Support: Drs. Lin and Segal were supported by grant K24AG049036 from the National Institute on Aging. Dr. Mehta was supported by grant K01AG070329 from the National Institute on Aging. The National Institute on Aging had no role in the content of this manuscript.

Disclosures: Disclosure forms are available with the article online.

Corresponding Author: Jodi B. Segal, MD, MPH, Division of General Internal Medicine, Johns Hopkins School of Medicine, 624 North Broadway, Room 644, Baltimore, MD 21205; e-mail, jsegal@jhmi.edu.

Author contributions are available at Annals.org.

References

1. Stone NJ, Robinson JG, Lichtenstein AH, et al; American College of Cardiology/American Heart Association Task Force on Practice Guidelines. 2013 ACC/AHA guideline on the treatment of blood cholesterol to reduce atherosclerotic cardiovascular risk in adults: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *J Am Coll Cardiol*. 2014;63:2889-2934. [PMID: 24239923] doi:10.1016/j.jacc.2013.11.002
2. Muntner P, Colantonio LD, Cushman M, et al. Validation of the atherosclerotic cardiovascular disease Pooled Cohort risk equations. *JAMA*. 2014;311:1406-1415. [PMID: 24682252] doi:10.1001/jama.2014.2630
3. Getzen E, Ungar L, Mowery D, et al. Mining for equitable health: assessing the impact of missing data in electronic health records. *J Biomed Inform*. 2023;139:104269. [PMID: 36621750] doi:10.1016/j.jbi.2022.104269
4. IMDRF Software as a Medical Device (SaMD) Working Group. "Software as a Medical Device": Possible Framework for Risk Categorization and Corresponding Considerations. 18 September 2014. Accessed at <https://www.imdrf.org/sites/default/files/docs/imdrf/final/technical/imdrf-tech-140918-samd-framework-risk-categorization-141013.pdf> on 10 May 2025.
5. Software as a Medical Device Working Group. Software as a Medical Device (SaMD): Clinical Evaluation. 22 June 2017. Accessed at <http://fda.gov/media/100714/download> on 10 May 2025.
6. U.S. Food and Drug Administration. Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan. January 2021. Accessed at <https://www.fda.gov/media/145022/download> on 10 May 2025.
7. Office of the National Coordinator for Health Information Technology. Health Data, Technology, and Interoperability: Certification Program Updates, Algorithm Transparency, and Information Sharing. Accessed at <https://www.healthit.gov/sites/default/files/page/2023-12/hti-1-final-rule.pdf> on 10 May 2025.
8. Held U, Forzy T, Signorelli A, et al. Development and internal validation of a prediction model for long-term opioid use—an analysis of insurance claims data. *Pain*. 2024;165:44-53. [PMID: 37782553] doi:10.1097/j.pain.0000000000003023
9. Vimont A, Béliard S, Valéro R, et al. Prognostic models for short-term annual risk of severe complications and mortality in patients living with type 2 diabetes using a national medical claim database. *Diabetol Metab Syndr*. 2023;15:128. [PMID: 37322499] doi:10.1186/s13098-023-01105-x
10. Donahue PT, Xue QL, Carlson MC. Peak expiratory flow predicts incident dementia in a representative sample of U.S. older adults: the National Health and Aging Trends Study (NHATS). *J Gerontol A Biol Sci Med Sci*. 2023;78:1427-1435. [PMID: 36524396] doi:10.1093/gerona/glac235
11. Zelnick LR, Shlipak MG, Soliman EZ, et al; CRIC Study Investigators. Prediction of incident heart failure in CKD: the CRIC study. *Kidney Int Rep*. 2022;7:708-719. [PMID: 35497796] doi:10.1016/j.ekir.2022.01.1067
12. Alssaad A, Melhus H, Hammarlund-Udenaes M, et al. A tool for prediction of risk of rehospitalisation and mortality in the hospitalised elderly: secondary analysis of clinical trial data. *BMJ Open*. 2015;5:e007259. [PMID: 25694461] doi:10.1136/bmjopen-2014-007259
13. Luo W, Deng X, Xu X, et al. Development of a prognostic model for predicting multiple sclerosis after optic neuritis: a secondary analysis of data from the Optic Neuritis Treatment Trial. *J Neuroophthalmol*. 2022;42:88-96. [PMID: 34860745] doi:10.1097/WNO.0000000000001424
14. Tammemägi MC, Ten Haaf K, Toumazis I, et al. Development and validation of a multivariable lung cancer risk prediction model that includes low-dose computed tomography screening results: a secondary analysis of data from the National Lung Screening Trial. *JAMA Netw Open*. 2019;2:e190204. [PMID: 30821827] doi:10.1001/jamanetworkopen.2019.0204
15. Goldstein BA, Xu C, Wilson J, et al. Designing an implementable clinical prediction model for near-term mortality and long-term survival in patients on maintenance hemodialysis. *Am J Kidney Dis*. 2024;84:73-82. [PMID: 38493378] doi:10.1053/j.ajkd.2023.12.013
16. Ming DY, Zhao C, Tang X, et al. Predictive modeling to identify children with complex health needs at risk for hospitalization. *Hosp Pediatr*. 2023;13:357-369. [PMID: 37092278] doi:10.1542/hpeds.2022-006861
17. Goldstein BA, Navar AM, Pencina MJ, et al. 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. [PMID: 27189013] doi:10.1093/jamia/ocw042
18. Yang C, Kors JA, Ioannou S, et al. Trends in the conduct and reporting of clinical prediction model development and validation: a systematic review. *J Am Med Inform Assoc*. 2022;29:983-989. [PMID: 35045179] doi:10.1093/jamia/ocac002
19. Henry KE, Hager DN, Pronovost PJ, et al. A targeted real-time early warning score (TREWScore) for septic shock. *Sci Transl Med*. 2015;7:299ra122. [PMID: 26246167] doi:10.1126/scitranslmed.aab3719
20. Barbar S, Noventa F, Rossetto V, et al. A risk assessment model for the identification of hospitalized medical patients at risk for venous thromboembolism: the Padua Prediction Score. *J Thromb Haemost*. 2010;8:2450-2457. [PMID: 20738765] doi:10.1111/j.1538-7836.2010.04044.x
21. Goff DC Jr, Lloyd-Jones DM, Bennett G, et al; American College of Cardiology/American Heart Association Task Force on Practice Guidelines. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Circulation*. 2014;129:S49-S73. [PMID: 24222018] doi:10.1161/01.cir.0000437741.48606.98
22. de Hond A, van Buchem M, Fanconi C, et al. Predicting depression risk in patients with cancer using multimodal data: algorithm development study. *JMIR Med Inform*. 2024;12:e51925. [PMID: 38236635] doi:10.2196/51925
23. Nikitin M. Epic's Nebula: Next Milestone in Healthcare Innovation. Itirra; 10 September 2024. Accessed at <https://itirra.com/blog/epics-nebula-next-milestone-in-healthcare-innovation> on 10 May 2025.
24. Wongvibulsin S, Garibaldi BT, Antar AAR, et al. Development of Severe COVID-19 Adaptive Risk Predictor (SCARP), a calculator to predict severe disease or death in hospitalized patients with COVID-19. *Ann Intern Med*. 2021;174:777-785. [PMID: 33646849] doi:10.7326/M20-6754
25. Ahalt SC, Chute CG, Fecho K, et al; Biomedical Data Translator Consortium. Clinical data: sources and types, regulatory constraints,

applications. *Clin Transl Sci*. 2019;12:329-333. [PMID: 31074176] doi:10.1111/cts.12638

26. Haneuse S, Arterburn D, Daniels MJ. Assessing missing data assumptions in EHR-based studies: a complex and underappreciated task. *JAMA Netw Open*. 2021;4:e210184. [PMID: 33635321] doi:10.1001/jamanetworkopen.2021.0184

27. Haneuse S, Bogart A, Jazic I, et al. Learning about missing data mechanisms in electronic health records-based research: a survey-based approach. *Epidemiology*. 2016;27:82-90. [PMID: 26484425] doi:10.1097/EDE.0000000000000393

28. Sterne JAC, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*. 2009;338:b2393. [PMID: 19564179] doi:10.1136/bmj.b2393

29. Little RJA, Rubin DB. Introduction. In: Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. J Wiley; 2020:3-28.

30. Taber JM, Leyva B, Persoskie A. Why do people avoid medical care? A qualitative study using national data. *J Gen Intern Med*. 2015;30:290-297. [PMID: 25387439] doi:10.1007/s11606-014-3089-1

31. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. *J Clin Epidemiol*. 2005;58:323-337. [PMID: 15862718] doi:10.1016/j.jclinepi.2004.10.012

32. Holmes JH, Beinlich J, Boland MR, et al. Why is the electronic health record so challenging for research and clinical care? *Methods Inf Med*. 2021;60:32-48. [PMID: 34282602] doi:10.1055/s-0041-1731784

33. Ghoshal-Datta N, Chernew ME, McWilliams JM. Lack of persistent coding in traditional Medicare may widen the risk-score gap with Medicare Advantage. *Health Aff (Millwood)*. 2024;43:1638-1646. [PMID: 39626136] doi:10.1377/hlthaff.2024.00169

34. Efthimiou O, Seo M, Chalkou K, et al. Developing clinical prediction models: a step-by-step guide. *BMJ*. 2024;386:e078276. [PMID: 39227063] doi:10.1136/bmj-2023-078276

35. Washington V, DeSalvo K, Mostashari F, et al. The HITECH era and the path forward. *N Engl J Med*. 2017;377:904-906. [PMID: 28877013] doi:10.1056/NEJMp1703370

36. Ross MK, Sanz J, Tep B, et al. Accuracy of an electronic health record patient linkage module evaluated between neighboring academic health care centers. *Appl Clin Inform*. 2020;11:725-732. [PMID: 33147645] doi:10.1055/s-0040-1718374

37. Weberpals J, Raman SR, Shaw PA, et al. smdi: an R package to perform structural missing data investigations on partially observed confounders in real-world evidence studies. *JAMIA Open*. 2024;7:00ae008. [PMID: 38304248] doi:10.1093/jamiaopen/ooae008

38. Tan ALM, Getzen EJ, Hutch MR, et al; COVID-19 by EHR (4CE). Informative missingness: what can we learn from patterns in missing laboratory data in the electronic health record? *J Biomed Inform*. 2023;139:104306. [PMID: 36738870] doi:10.1016/j.jbi.2023.104306

39. van Smeden M, Groenwold RHH, Moons KG. A cautionary note on the use of the missing indicator method for handling missing data in prediction research. *J Clin Epidemiol*. 2020;125:188-190. [PMID: 32565213] doi:10.1016/j.jclinepi.2020.06.007

40. Groenwold RHH. Informative missingness in electronic health record systems: the curse of knowing. *Diagn Progn Res*. 2020;4:8. [PMID: 32699824] doi:10.1186/s41512-020-00077-0

41. Sisk R, Sperrin M, Peek N, et al. Imputation and missing indicators for handling missing data in the development and deployment of clinical prediction models: a simulation study. *Stat Methods Med Res*. 2023;32:1461-1477. [PMID: 37105540] doi:10.1177/09622802231165001

42. Hughes RA, Heron J, Sterne JAC, et al. Accounting for missing data in statistical analyses: multiple imputation is not always the answer. *Int J Epidemiol*. 2019;48:1294-1304. [PMID: 30879056] doi:10.1093/ije/dyz032

43. Little RJA, Rubin DB. Complete-case and available-case analysis, including weighting methods. In: Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. J Wiley; 2020:47-66.

44. Molenberghs G, Fitzmaurice G. Incomplete data: Introduction and overview. In: Fitzmaurice G, Davidian M, Verbeke G, et al, eds. *Longitudinal Data Analysis*. Chapman & Hall/CRC; 2009:395-408.

45. Little RJA, Rubin DB. Maximum likelihood for general patterns of missing data: introduction and theory with ignorable nonresponse. In: Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. J Wiley; 2020:185-212.

46. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. J Wiley; 1987.

47. van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. *Stat Methods Med Res*. 2007;16:219-242. [PMID: 17621469] doi:10.1177/0962280206074463

48. Sparapani RA, Logan BR, McCulloch RE, et al. Nonparametric survival analysis using Bayesian additive regression trees (BART). *Stat Med*. 2016;35:2741-2753. [PMID: 26854022] doi:10.1002/sim.6893

49. Chipman HA, George EI, McCulloch RE. BART: Bayesian additive regression trees. *Ann Appl Stat*. 2010;4:266-298. doi:10.1214/09-AOAS285

50. Ji J, Hu L, Liu B, et al. Identifying and assessing the impact of key neighborhood-level determinants on geographic variation in stroke: a machine learning and multilevel modeling approach. *BMC Public Health*. 2020;20:1666. [PMID: 33160324] doi:10.1186/s12889-020-09766-3

51. Nijman S, Leeuwenberg AM, Beekers I, et al. Missing data is poorly handled and reported in prediction model studies using machine learning: a literature review. *J Clin Epidemiol*. 2022;142:218-229. [PMID: 34798287] doi:10.1016/j.jclinepi.2021.11.023

52. Emmanuel T, Maupong T, Mpoeleng D, et al. A survey on missing data in machine learning. *J Big Data*. 2021;8:140. [PMID: 34722113] doi:10.1186/s40537-021-00516-9

53. Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. J Wiley; 2020.

54. Fitzmaurice G, Davidian M, Verbeke G, et al, eds. *Longitudinal Data Analysis*. Chapman & Hall/CRC; 2009.

55. Little RJA, Rubin DB. Missing not at random models. In: Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. J Wiley; 2020:351-404.

56. Little R. Selection and pattern-mixture models. In: Fitzmaurice G, Davidian M, Verbeke G, et al, eds. *Longitudinal Data Analysis*. Chapman & Hall/CRC; 2009:409-432.

57. Little RJA. Pattern-mixture models for multivariate incomplete data. *J Am Stat Assoc*. 1993;88:125-134. doi:10.2307/2290705

58. Deforth M, Heinze G, Held U. The performance of prognostic models depended on the choice of missing value imputation algorithm: a simulation study. *J Clin Epidemiol*. 2024;176:111539. [PMID: 39326470] doi:10.1016/j.jclinepi.2024.111539

59. Hoogland J, van Barneveld M, Debray TPA, et al. Handling missing predictor values when validating and applying a prediction model to new patients. *Stat Med*. 2020;39:3591-3607. [PMID: 32687233] doi:10.1002/sim.8682

60. Little RJA, Rubin DB. Multivariate normal examples, ignoring the missingness mechanism. In: Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. J Wiley; 2020:249-284.

61. Brown H, Prescott R. *Applied Mixed Models in Medicine*. Third Edition. J Wiley; 2015.

62. Diggle PJ, Heagerty P, Liang KY, et al, eds. *Analysis of Longitudinal Data*. Second Edition. Oxford Univ Pr; 2013.

63. Laird NM, Ware JH. Random-effects models for longitudinal data. *Biometrics*. 1982;38:963-974. [PMID: 7168798]

64. Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. *J Am Stat Assoc*. 1993;88:9-25. doi:10.2307/2290687

65. Fitzmaurice GM, Laird NM, Ware JH. *Applied Longitudinal Analysis*. J Wiley; 2011.

66. Kalman RE. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*. 1960;82:35-45. doi:10.1115/1.3662552
67. Niako N, Melgarejo JD, Maestre GE, et al. Effects of missing data imputation methods on univariate blood pressure time series data analysis and forecasting with ARIMA and LSTM. *BMC Med Res Methodol*. 2024;24:320. [PMID: 39725886] doi:10.1186/s12874-024-02448-3
68. Siebert U, Alagoz O, Bayoumi AM, et al; ISPOR-SMDM Modeling Good Research Practices Task Force. State-transition modeling: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force-3. *Value Health*. 2012;15:812-820. [PMID: 22999130] doi:10.1016/j.jval.2012.06.014
69. Anderson WJ. *Continuous-Time Markov Chains: An Applications-Oriented Approach*. Springer; 2012.
70. Collins GS, Dhiman P, Ma J, et al. Evaluation of clinical prediction models (part 1): from development to external validation. *BMJ*. 2024;384:e074819. [PMID: 38191193] doi:10.1136/bmj-2023-074819
71. Van Calster B, McLernon DJ, van Smeden M, et al; Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative. Calibration: the Achilles heel of predictive analytics. *BMC Med*. 2019;17:230. [PMID: 31842878] doi:10.1186/s12916-019-1466-7
72. Localio AR, Goodman S. Beyond the usual prediction accuracy metrics: reporting results for clinical decision making. *Ann Intern Med*. 2012;157:294-295. [PMID: 22910942] doi:10.7326/0003-4819-157-4-201208210-00014
73. Solomon J, Dauber-Decker K, Richardson S, et al. Integrating clinical decision support into electronic health record systems using a novel platform (EvidencePoint): developmental study. *JMIR Form Res*. 2023;7:e44065. [PMID: 37856193] doi:10.2196/44065
74. Sendak MP, Gao M, Brajer N, et al. Presenting machine learning model information to clinical end users with model facts labels. *NPJ Digit Med*. 2020;3:41. [PMID: 32219182] doi:10.1038/s41746-020-0253-3
75. Kapelner A, Bleich J. Prediction with missing data via Bayesian additive regression trees. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*. 2015;43:224-239. doi:10.1002/cjs.11248
76. Tsvetanova A, Sperrin M, Jenkins D, et al. Compatibility in missing data handling across the prediction model pipeline: a simulation study. *Stud Health Technol Inform*. 2024;310:1476-1477. [PMID: 38269704] doi:10.3233/SHTI231252
77. Sperrin M, Martin GP, Sisk R, et al. Missing data should be handled differently for prediction than for description or causal explanation. *J Clin Epidemiol*. 2020;125:183-187. [PMID: 32540389] doi:10.1016/j.jclinepi.2020.03.028
78. Tsvetanova A, Sperrin M, Peek N, et al. Missing data was handled inconsistently in UK prediction models: a review of method used. *J Clin Epidemiol*. 2021;140:149-158. [PMID: 34520847] doi:10.1016/j.jclinepi.2021.09.008
79. Moons KGM, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med*. 2015;162:W1-W73. [PMID: 25560730] doi:10.7326/M14-0698
80. Debray TPA, Collins GS, Riley RD, et al. Transparent reporting of multivariable prediction models developed or validated using clustered data: TRIPOD-Cluster checklist. *BMJ*. 2023;380:e071018. [PMID: 36750242] doi:10.1136/bmj-2022-071018
81. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e078378. [PMID: 38626948] doi:10.1136/bmj-2023-078378
82. Reiter JP, Raghunathan TE, Kinney SK. The importance of modeling the sampling design in multiple imputation for missing data. *Survey Methodology*. 2006;32:143.
83. Raghunathan TE, Lepkowski JM, Van Hoewyk J, et al. A multi-variate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*. 2001;27:85-96.
84. McCoy DC, Seiden J, Cuartas J, et al. Estimates of a multidimensional index of nurturing care in the next 1000 days of life for children in low-income and middle-income countries: a modelling study. *Lancet Child Adolesc Health*. 2022;6:324-334. [PMID: 35429451] doi:10.1016/S2352-4642(22)00076-1
85. Akkaya Hocagil T, Yucel RM. A computationally efficient sequential regression imputation algorithm for multilevel data. *J Appl Stat*. 2024;51:2258-2278. [PMID: 39157267] doi:10.1080/02664763.2023.2277669
86. Zhang Y, Wang F, Huang X, et al. Optimization and application of cloud-based deep learning architecture for multi-source data prediction. *arXiv*. Preprint posted online 16 October 2024. doi:10.48550/arXiv.2410.12642

Author Contributions: Conception and design: S. Lin, R.H.H. Groenwold, H.B. Mehta, J.B. Segal.
Drafting of the article: S. Lin, J.B. Segal.
Critical revision for important intellectual content: S. Lin, R.H.H. Groenwold, H.B. Mehta, J.B. Segal.
Final approval of the article: S. Lin, R.H.H. Groenwold, H.B. Mehta, J.S. Kim, J.B. Segal.
Statistical expertise: R.H.H. Groenwold, H.B. Mehta, J.S. Kim, J.B. Segal.
Obtaining of funding: J.B. Segal.
Administrative, technical, or logistic support: S. Lin, J.B. Segal.