Informative presence and observation in routine health data: A review of methodology for clinical risk prediction

Rose Sisk, Lijing Lin, Matthew Sperrin, Jessica K. Barrett, Brian Tom, Karla Diaz-Ordaz, Niels Peek, and Glen P. Martin

Division of Informatics, Imaging and Data Sciences, School of Health Sciences, University of Manchester, Manchester, United Kingdom, MRC Biostatistics Unit, University of Cambridge, Cambridge, United Kingdom, Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom, Department of Medical Statistics, London School of Hygiene and Tropical Medicine, London, United Kingdom, NIHR Biomedical Research Centre, Manchester Academic Health Science Centre, University of Manchester, Manchester, United Kingdom, and Alan Turing Institute, University of Manchester, London, United Kingdom

Corresponding Author: Rose Sisk, BSc (Hons), Room G.304, Jean McFarlane Building, University of Manchester M13 9PY, UK; rose.sisk@postgrad.manchester.ac.uk

Received 26 June 2020; Revised 14 September 2020; Editorial decision 15 September 2020; Accepted 17 September 2020

ABSTRACT

Objective: Informative presence (IP) is the phenomenon whereby the presence or absence of patient data is potentially informative with respect to their health condition, with informative observation (IO) being the longitudinal equivalent. These phenomena predominantly exist within routinely collected healthcare data, in which data collection is driven by the clinical requirements of patients and clinicians. The extent to which IP and IO are considered when using such data to develop clinical prediction models (CPMs) is unknown, as is the existing methodology aiming at handling these issues. This review aims to synthesize such existing methodology, thereby helping identify an agenda for future methodological work.

Materials and Methods: A systematic literature search was conducted by 2 independent reviewers using pre-specified keywords.

Results: Thirty-six articles were included. We categorized the methods presented within as derived predictors (including some representation of the measurement process as a predictor in the model), modeling under IP, and latent structures. Including missing indicators or summary measures as predictors is the most commonly presented approach amongst the included studies (24 of 36 articles).

Discussion: This is the first review to collate the literature in this area under a prediction framework. A considerable body relevant of literature exists, and we present ways in which the described methods could be developed further. Guidance is required for specifying the conditions under which each method should be used to enable applied prediction modelers to use these methods.

Conclusions: A growing recognition of IP and IO exists within the literature, and methodology is increasingly becoming available to leverage these phenomena for prediction purposes. IP and IO should be approached differently in a prediction context than when the primary goal is explanation. The work included in this review has demonstrated theoretical and empirical benefits of incorporating IP and IO, and therefore we recommend that applied health researchers consider incorporating these methods in their work.

Key words: clinical prediction model, electronic health records, informative observation, informative presence
INTRODUCTION

Background and significance

Clinical prediction models (CPMs) estimate the risk that a patient currently has (diagnostic), or will develop (prognostic), an outcome of interest based on known clinical and patient measures. Such risk models can guide clinical decision making, among other uses. Widespread adoption of electronic health records (EHRs) facilitates the development of CPMs,1 as detailed clinical and patient information is collected through routine healthcare contacts. Such rich longitudinal information provides long-term patient follow-up without the need to recruit patients and conduct regular follow-up visits. The analysis of routinely collected data is not, however, without challenge. Observation times are not prespecified as they would be in a typical research study (eg, in a prospective cohort study with scheduled follow-up visits). Instead, data are collected opportunistically, in which patient and clinician decisions directly dictate whether we observe clinical biomarkers and patient information.2

For example, general practitioner visits occur more frequently during periods of ill health,3 and only information relevant to the particular consultation will be recorded. Equally, during inpatient care, clinicians will adapt their monitoring frequency to the changing needs and condition of the individual patient (see Figure 1).

We refer to the process by which visits, and hence measurements, occur as the observation process (also known elsewhere as the visiting or monitoring process). We define 2 key properties that an observation process may have, when presence of data is informative:

1. Informative presence (IP) (Figure 1A): The presence or absence of a patient’s data at any given time point carries information about their health status.
2. Informative observation (IO): The timing, frequency, or intensity (rate) of a patient’s longitudinal pattern of observation carries information about their evolving health state. See Figure 1B for an example.

IP is challenging from a statistical perspective as it implies a missing not at random process. IP is, however, conceptually different from missingness, as in the former, there was never any intention of collecting the data at a particular visit. IP has previously been defined elsewhere,4,5 with Phelan et al5 discussing how interactions contained within EHRs are informative with respect to patient health.

IO is the continuous time generalization of IP; a longitudinal visiting (at time t) not at random process, defined as “given data recorded up to time t, visiting at time t is not independent of outcome at time t.” By generalizing the definition of IP above, one can draw value from how frequently a patient is observed over time. This is especially true when no schedule exists dictating when or how often visits should occur; we therefore focus on what an individual’s longitudinal observation process could tell us about their condition.

A recent review of CPMs developed using routinely collected data revealed an apparent lack of understanding of, or proper handling of, IP and IO.1 Moreover, much of the existing methodological literature in this area has focused on IP and IO only in the context of effect estimation (ie, in causal or associational studies),8–14 and has generally viewed it as a “nuisance” (ie, a phenomenon that potentially biases effect estimators and therefore needs to be corrected for in the analysis). However, when developing a CPM, the primary focus is on achieving good predictive performance; predictor effect estimation is less important.

Instead, one could view IP and IO as opportunities to draw information from the EHR that is not explicitly recorded. In this article, we focus on informative measurement patterns in the predictors, and we do not discuss presence or absence of outcome data. Agniel et al15 demonstrated how the timing of a lab test better predicts mortality than the actual result of the test. Others have illustrated how incorporating the presence or absence of a particular test for an individual into a CPM can improve its accuracy.16–18

Objectives

This article aims to review the literature on methodology allowing CPMs to utilize IP or IO, both in overcoming some of the aforementioned challenges, and in harnessing information within informative measurement patterns. In doing so, we also highlight outstanding areas of methodological work that should be prioritized. Finally, we summarize existing software packages capable of implementing the methodology.

MATERIALS AND METHODS

The strategy employed in this review loosely follows a scoping review framework.19 Our protocol has been registered on the Open Science Framework.20

Search strategy

We searched MEDLINE, Embase, and Web of Science for relevant articles using prespecified search terms. Further details of the full search strategy (including search terms and an additional snowballing stage) can be found in the Supplementary Appendix and the published protocol.20

Study selection

We had the following inclusion criteria: any article presenting a method that allows CPMs to incorporate IP or IO. We excluded articles that applied existing methods that had already been published elsewhere, and included those earlier publications instead, nonmedical areas of application, IP and IO in outcome measures, and methods that handle sample selection bias, imputation or censoring only. See the Supplementary Appendix for further justification of these exclusions.

We do not include textbooks within the review; while this could mean we miss some relevant literature, searching within textbooks is not widely feasible. Additionally, we believe that most methodological development in this area will be published in original research articles rather than textbooks.

Two independent reviewers (R.S., L.L.) conducted a 2-stage screening process. Titles and abstracts were screened first, and full texts of remaining articles were reviewed at the second stage. Reviewers met regularly to track agreement. Systematic differences were translated into new inclusion and exclusion criteria, in consultation with a third reviewer (G.P.M.).

Primarily, we extracted information regarding the modeling method employed and any reported advantages and disadvantages. We also extracted information on the form of the observation processes, predictors, and outcome, including any clinical use cases presented.

Widespread adoption of electronic health records (EHRs) facilitates the development of CPMs,1,4 as detailed clinical and patient information is collected through routine healthcare contacts. Such rich longitudinal information provides long-term patient follow-up without the need to recruit patients and conduct regular follow-up visits. The analysis of routinely collected data is not, however, without challenge. Observation times are not prespecified as they would be in a typical research study (eg, in a prospective cohort study with scheduled follow-up visits). Instead, data are collected opportunistically, in which patient and clinician decisions directly dictate whether we observe clinical biomarkers and patient information.2 For example, general practitioner visits occur more frequently during periods of ill health,3 and only information relevant to the particular consultation will be recorded. Equally, during inpatient care, clinicians will adapt their monitoring frequency to the changing needs and condition of the individual patient (see Figure 1).

We refer to the process by which visits, and hence measurements, occur as the observation process (also known elsewhere as the visiting or monitoring process). We define 2 key properties that an observation process may have, when presence of data is informative:

1. Informative presence (IP) (Figure 1A): The presence or absence of a patient’s data at any given time point carries information about their health status.
2. Informative observation (IO): The timing, frequency, or intensity (rate) of a patient’s longitudinal pattern of observation carries information about their evolving health state. See Figure 1B for an example.

IP is challenging from a statistical perspective as it implies a missing not at random process. IP is, however, conceptually different from missingness, as in the former, there was never any intention of collecting the data at a particular visit. IP has previously been defined elsewhere,4,5 with Phelan et al5 discussing how interactions contained within EHRs are informative with respect to patient health.

IO is the continuous time generalization of IP; a longitudinal visiting (at time t) not at random process, defined as “given data recorded up to time t, visiting at time t is not independent of outcome at time t.” By generalizing the definition of IP above, one can draw value from how frequently a patient is observed over time. This is especially true when no schedule exists dictating when or how often visits should occur; we therefore focus on what an individual’s longitudinal observation process could tell us about their condition.

A recent review of CPMs developed using routinely collected data revealed an apparent lack of understanding of, or proper handling of, IP and IO.1 Moreover, much of the existing methodological literature in this area has focused on IP and IO only in the context of effect estimation (ie, in causal or associational studies),8–14 and has generally viewed it as a “nuisance” (ie, a phenomenon that potentially biases effect estimators and therefore needs to be corrected for in the analysis). However, when developing a CPM, the primary focus is on achieving good predictive performance; predictor effect estimation is less important.

Instead, one could view IP and IO as opportunities to draw information from the EHR that is not explicitly recorded. In this article, we focus on informative measurement patterns in the predictors, and we do not discuss presence or absence of outcome data. Agniel et al15 demonstrated how the timing of a lab test better predicts mortality than the actual result of the test. Others have illustrated how incorporating the presence or absence of a particular test for an individual into a CPM can improve its accuracy.16–18

Objectives

This article aims to review the literature on methodology allowing CPMs to utilize IP or IO, both in overcoming some of the aforementioned challenges, and in harnessing information within informative measurement patterns. In doing so, we also highlight outstanding areas of methodological work that should be prioritized. Finally, we summarize existing software packages capable of implementing the methodology.

MATERIALS AND METHODS

The strategy employed in this review loosely follows a scoping review framework.19 Our protocol has been registered on the Open Science Framework.20

Search strategy

We searched MEDLINE, Embase, and Web of Science for relevant articles using prespecified search terms. Further details of the full search strategy (including search terms and an additional snowballing stage) can be found in the Supplementary Appendix and the published protocol.20

Study selection

We had the following inclusion criteria: any article presenting a method that allows CPMs to incorporate IP or IO. We excluded articles that applied existing methods that had already been published elsewhere, and included those earlier publications instead, nonmedical areas of application, IP and IO in outcome measures, and methods that handle sample selection bias, imputation or censoring only. See the Supplementary Appendix for further justification of these exclusions.

We do not include textbooks within the review; while this could mean we miss some relevant literature, searching within textbooks is not widely feasible. Additionally, we believe that most methodological development in this area will be published in original research articles rather than textbooks.

Two independent reviewers (R.S., L.L.) conducted a 2-stage screening process. Titles and abstracts were screened first, and full texts of remaining articles were reviewed at the second stage. Reviewers met regularly to track agreement. Systematic differences were translated into new inclusion and exclusion criteria, in consultation with a third reviewer (G.P.M.).

Primarily, we extracted information regarding the modeling method employed and any reported advantages and disadvantages. We also extracted information on the form of the observation processes, predictors, and outcome, including any clinical use cases presented.
Figure 1. (A) An illustration of informative presence and how this could impact the information available at prediction time. We see the longitudinal pattern of blood pressure for 2 patients, with both their observed and unobserved values shown. Patient 1 has 1 single observed value of systolic blood pressure (BP), and this happens when their BP was at its highest. Patient 2 has no observed values, but their BP remains in the normal range—either the patient or clinician saw no clinical need to take a blood pressure measurement at any time. (B) An illustration of informative observation, taken from the MIMIC (Medical Information Mart for Intensive Care) dataset. Patient 1 has many more in-hospital measurements of blood glucose than patient 2 throughout their intensive care unit admission, likely due to the fact that their blood glucose is much higher and much more variable than patient 2. A more severe condition often means more intense monitoring.
RESULTS

Our database searches identified 6127 studies, of which 111 were retained for full text screening. Eleven of these were deemed eligible for inclusion. We identified a further 25 articles through forward and backward citation searching, giving a final set of 36 included articles (Figure 2).

Throughout this section, we will illustrate each method with the following notation. Consider a binary outcome $Y_t$ (or $Y$ if only observed once) for patients $i = 1, \ldots, n$, at time $t$, where $Y = 1$ denotes that the event occurred, with marginal probability $P(Y = 1)$. Define a potentially time-varying continuous covariate process $X(t)$, with potential realizations $x_{ij}$ for $i = 1, \ldots, n$ and $j = 1, \ldots, m_i$, or simply $x_i$ if $X$ is not time-varying. The timing of the $j$th realization of $X(t)$ is $t_{ij} \in \mathbb{R}^+$. Denote $R = 1$ if $X(t)$ is ever observed, and $R = 0$ if not. Define $r_{ij} = 1$ if the covariate process is observed at time $t_{ij}$. We assume that $Z$ is a completely observed time-invariant covariate.

The methods described in this section address IP or IO by deriving some representation of the observation process and including this as a separate predictor in the model to exploit the informativeness for predictive value. These approaches tend to be straightforward and have been proposed to handle both IP and IO. However, attention must be paid to the intended use of the final model, particularly where the model will be applied in clinical settings different to the one in which it was developed. Where measurement protocols change across different settings, these models may lack generalizability when transported to a new setting.57–59 This should not be a concern where the development and application settings remain the same.

There exists a breadth of methodological literature covering scenario 2 (S2) (without accounting for IP and IO), which has recently been synthesized by Bull et al.22 We therefore focus on modeling strategies that have specifically been proposed or extended to accommodate IP or IO.

Identified approaches to handle IP and IO

We identified 3 broad categories of method based on the included articles: (1) methods that incorporate IP or IO through derived predictors; (2) methods for modeling under IP; and (3) methods that incorporate IP or IO using latent structures. Within these 3 categories, we identified 8 modeling strategies. A summary of the methods can be found in Table 2. Table 3 summarizes the advantages, disadvantages, software, and assumptions for each method—here, the reported advantages and disadvantages were inferred by the research team because they are not consistently mentioned in the included literature. A summary table at article level can be found in Supplementary Appendix 3.

Category 1: Derived predictors

The methods described in this section address IP or IO by deriving some representation of the observation process and including this as a separate predictor in the model to exploit the informativeness for predictive value. These approaches tend to be straightforward and have been proposed to handle both IP and IO. However, attention must be paid to the intended use of the final model, particularly where the model will be applied in clinical settings different to the one in which it was developed. Where measurement protocols change across different settings, these models may lack generalizability when transported to a new setting.57–59 This should not be a concern where the development and application settings remain the same.

Missing indicators or separate class

The missing indicator approach16,23–30 handles IP in a straightforward manner, by deriving a binary variable that indicates whether a predictor has been observed at a specific time (IP) or over a defined window of time. The indicators enter the prediction model as a separate predictor alongside other patient and clinical information.
Table 1. A description of different prediction scenarios, covering cross-sectional vs longitudinal predictors and outcomes

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Scenario name</th>
<th>Description</th>
<th>Example (SOFA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>Cross-sectional prediction</td>
<td>Interest lies in obtaining a single prognostic estimate (prediction) using a single value for each predictor.</td>
<td>Use values of bilirubin and BP obtained upon ICU admission to predict in-hospital survival (binary).</td>
</tr>
<tr>
<td>S2</td>
<td>Cross-sectional prediction with longitudinal predictor measurements</td>
<td>Interest lies in obtaining a single prognostic estimate but using the longitudinal history of predictor values.</td>
<td>Use all repeated lab tests obtained throughout inpatient admission for bilirubin and BP to predict in-hospital survival.</td>
</tr>
<tr>
<td>S3</td>
<td>Longitudinal prediction with longitudinal predictors and outcomes</td>
<td>Interested in prognostic estimates at multiple time points, potentially using the longitudinal history of predictor values.</td>
<td>Use all repeated measures of BP and bilirubin obtained throughout inpatient and ICU admission to predict survival at multiple future time points.</td>
</tr>
</tbody>
</table>

BP: blood pressure; ICU: intensive care unit; S: scenario; SOFA: Sequential Organ Failure Assessment;

For example, if a prediction model requires an entry for bilirubin but this test has not been conducted, then a missing indicator would be included as a predictor with value 1 (or 0 when observed). For categorical variables, a separate “missing” category could instead be created. Because most prediction models require a value for every predictor, the missing indicator approach is usually combined with imputation at both model development and prediction time (not necessary for categorical predictors with a separate class). The missing indicator approach results in a model of the form:

\[
g(P|Y=1; X; Z) = \beta_0 + \beta_1 X + \beta_2 Z + \gamma R
\]  
for continuous predictors within cross-sectional prediction (S1).

Similarly, for a categorical predictor \( x \) with \( k \) categories, then the missing indicator approach would set \( x_i \in \{ \text{Cat}_1, \ldots, \text{Cat}_k \} \). Missing and our model would be

\[
g(P|Y=1; X; Z) = \beta_0 + \beta_1 X + \beta_2 Z
\]  

The previous 2 equations could be combined to consider prediction models with both continuous and categorical predictors. Alternatively, missing indicators and separate classes have been well developed in tree-based prediction algorithms.

Including a missing indicator or separate class is straightforward and has demonstrated improved predictive performance over models omitting them. However, their inclusion could double the number of candidate predictors for a model. The approach also fails to capture complex representations of the measurement process.

Summary measures

An extension to missing indicators, capable of incorporating both IP and IO, is to derive a summary of the measurement process and include this as a predictor. Examples include a count of the number of measurements (eg, throughout a critical care admission), weighted counts, combined missing indicators, missingness rates over time, time intervals between measurements, embedding vectors that represent missing values, or information relating to hospital processes.

In some cases, combined missing indicators and time intervals also alter the relationship between a predictor and outcome. The method stipulates that the longer a measure has been missing, the less influence it should have on an individual’s prediction; therefore, the last observed measurement is decayed toward a mean value.

Piecewise-constant intensity models have also been proposed to handle informatively observed predictors. Piecewise-constant intensity models use decision trees to assign an intensity rate to the observation process, conditional on its history (timings, values, and events).

Define a summary measure of the observation process \( Q \), eg, a count of the number of times \( X(t) \) (whether continuous or categorical) has been observed: \( Q = m_t \). For cross-sectional prediction with a time-varying covariate, we then have:

\[
g(P|Y=1; X; Z) = \beta_0 + \beta_1 X + \beta_2 Z + \gamma Q
\]  
where \( X \) is a summary of \( X(t) \) deemed to have predictive value (eg, the mean, most recent, or most extreme value). If \( X(t) \) has never been observed, this should be imputed. Like missing indicators, summary measures are easily derived and implemented in any prediction model using standard software (since they are included as standard predictors). Combining missing indicators into one summary, or implementing a dimension-reduction technique such as Lasso, also overcomes the issue of including multiple missing indicators. However, selecting the most appropriate summary measure for a model requires careful consideration, and will depend on the clinical application. No current guidance exists on how best to choose the most appropriate summary measure. The association between a chosen summary measure and the outcome might lack generalizability where measurement processes vary across locations.

Simple summary measures such as counts may also fail to capture the complex relationship between the observation process and outcome.

Category 2: Modeling under IP

While the methods in the other categories can be used to handle both IP and IO, this category comprises methods that have specifically been proposed to handle IP.

Pattern-specific models

The pattern-specific approach derives separate models for each missingness pattern, generalizing the missing indicator approach. The model corresponding to the observed pattern in a new individual is then used for prediction. For example, in a model with a single partially-observed time-invariant continuous predictor, \( X \) we would derive the following submodels:

\[
g(P|Y=1; X; Z) = \beta_{01} + \beta_{11} X + \beta_{21} Z
\]
<table>
<thead>
<tr>
<th>Modeling approach</th>
<th>Broad category</th>
<th>Refs</th>
<th>Scenario(s)</th>
<th>IP or IO</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Missing indicators &amp; Separate class</td>
<td>Derived predictors</td>
<td>16,23–30</td>
<td>S1</td>
<td>IP</td>
<td>Creating a binary indicator, representing presence/absence of a predictor at a given time point or in a given window.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Create a binary indicator taking 0 when bilirubin is observed, and 1 if missing. Enter this as an additional predictor alongside observed bilirubin and BP.</td>
</tr>
<tr>
<td>Summary measures</td>
<td>Derived predictors</td>
<td>15,24,31–44</td>
<td>S2</td>
<td>IO</td>
<td>Summarizing the observation process into a single variable, e.g., counting visits, rates of visits over a window, weighted counts.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Count the number of times bilirubin has been measured over the first 24 hours of each ICU admission. Enter this count as an additional predictor in the model.</td>
</tr>
<tr>
<td>Pattern-specific models</td>
<td>Modeling under informed presence</td>
<td>45,46</td>
<td>S1</td>
<td>IP</td>
<td>Derive separate models for each missingness pattern.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Develop models for: bilirubin and BP observed, and only BP observed. Bilirubin is missing not at random.</td>
</tr>
<tr>
<td>Likelihood-based methods</td>
<td>Modeling under informed presence</td>
<td>47,48</td>
<td>S1</td>
<td>IP</td>
<td>Incorporating missingness mechanism into maximum-likelihood estimation of parameter estimates.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Estimate model parameters using method-of-weights and EM algorithm.</td>
</tr>
<tr>
<td>Similarity measures</td>
<td>Derived predictors</td>
<td>49</td>
<td>S2</td>
<td>IO</td>
<td>Calculate similarity between target patient and all others, based on predictor values and measurement timings. Develop models separately for “similar” groups of patients.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Develop separate models amongst cohorts of patients with similar bilirubin, BP and timings of those measures.</td>
</tr>
<tr>
<td>Latent variable</td>
<td>Latent structures</td>
<td>50,51</td>
<td>S1, S3</td>
<td>IP</td>
<td>Outcome can be partially latent, and the observation process infers the latent state.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>The occurrence of a bilirubin measurement is used to infer patient state in a hierarchical model.</td>
</tr>
<tr>
<td>HMMs</td>
<td>Latent structures</td>
<td>52,53</td>
<td>S3</td>
<td>IO</td>
<td>Outcome is a partially latent process, and the observation process infers the state at any time.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>The intensity of bilirubin measurements over the course of a patient’s admission infers their severity at any time point.</td>
</tr>
<tr>
<td>Joint modeling/shared random effects</td>
<td>Latent structures</td>
<td>54–56</td>
<td>S2, S3</td>
<td>IP and IO</td>
<td>Model the outcome, predictor and observation processes separately, but join them through random effects shared across the models.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Model the number of times bilirubin is measured throughout the admission as a point process, the repeated measures of bilirubin using a linear mixed model, and the binary outcome using a logistic regression. Link these via at least 1 shared random effect across the models.</td>
</tr>
</tbody>
</table>

BP: blood pressure; HMM: hidden Markov model; ICU: intensive care unit; IO: informative observation; IP: informative presence; S: scenario; SOFA: Sequential Organ Failure Assessment.
Where $Z$ is completely observed. Note that formulas 4 and 5 can also be combined by including interaction terms with the missing indicator, illustrating how this approach extends the missing indicator method.

Likelihood-based methods

A different approach assumes that missingness in the predictors is nonignorable, and incorporates this into parameter estimates via likelihood-based methods. The model formulation would take, for example, the same form as equation 2, with parameter estimates obtained according to estimation procedures detailed in the following examples. Escarela et al\(^57\) assumed a bivariate copula-based probability function for the missing covariates and the missingness mechanism. Kirkham\(^48\) instead applied the “method of weights,” which assumes a parametric model for the missingness mechanism and incorporates this into the maximum likelihood estimation of parameter estimates.

### Table 3. Summary of (subjective assessments of) advantages, disadvantages, software, and assumptions for each method described in this review

<table>
<thead>
<tr>
<th>Modeling approach</th>
<th>Advantages</th>
<th>Disadvantages</th>
<th>Software</th>
<th>Assumptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Missing indicators &amp; Separate class</td>
<td>Straightforward</td>
<td>Potentially doubles no. of predictors</td>
<td>Easily applied in common statistical software</td>
<td>Assumes absence is a proxy for some unmeasured patient feature</td>
</tr>
<tr>
<td></td>
<td>Flexible</td>
<td>Too simplistic for complex relationships between missingness and outcome</td>
<td></td>
<td>Linear relationship with outcome</td>
</tr>
<tr>
<td></td>
<td>Low computational cost</td>
<td>Assumes discrete time intervals</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Easy to communicate</td>
<td>May fail to capture complex relationships between observation process and outcome</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Summary measures</td>
<td>Straightforward</td>
<td>Generalizability of models across centers may be questioned</td>
<td>Easily applied in common statistical software</td>
<td>Assumes observation process is a proxy for some unmeasured patient feature</td>
</tr>
<tr>
<td></td>
<td>Flexible</td>
<td>May fail to capture complex relationships between observation process and outcome</td>
<td></td>
<td>Largely assumes linear relationship with outcome</td>
</tr>
<tr>
<td></td>
<td>Low computational cost</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Easy to communicate</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pattern-specific models</td>
<td>Straightforward</td>
<td>Number of models becomes large as no. of predictors increases</td>
<td>Easily applied in common statistical software</td>
<td>No assumptions placed on how missingness relates to observed or unobserved variables</td>
</tr>
<tr>
<td>Similarity measures</td>
<td>Also allows for imputation</td>
<td>Computationally intensive</td>
<td>None provided</td>
<td>Assumes absence is related to the unobserved value</td>
</tr>
<tr>
<td>Latent variable</td>
<td>Flexible</td>
<td>Computationally intensive</td>
<td>None provided</td>
<td>None provided</td>
</tr>
<tr>
<td></td>
<td>Improved performance over methods not incorporating informative presence</td>
<td>Computationally intensive</td>
<td>R code provided by Coley and Hubbard</td>
<td>Assumes same functional form for all pattern-specific models</td>
</tr>
<tr>
<td>HMMs</td>
<td>Using a Hawkes process for intensity allows for time-varying intensity</td>
<td>Complex and computationally intensive</td>
<td>None provided</td>
<td>Assumes longitudinal predictors are normally distributed</td>
</tr>
<tr>
<td>Joint modeling/shared random effects</td>
<td>Flexible to different forms of outcome and observation process</td>
<td>Complex</td>
<td>Frailtypack in R, WinBUGS, merlin in STATA for flexible user-defined models</td>
<td>Assumes processes (outcome, observation) are independent conditional on random effects</td>
</tr>
<tr>
<td></td>
<td>Complex Computationally intensive</td>
<td>Often requires independence assumption between processes given random effects</td>
<td></td>
<td>Existing methods assume constant intensity of observation</td>
</tr>
</tbody>
</table>

HMM: hidden Markov model.
Escarela et al. described how their missing not at random model can also be used to impute missing values. However, this does not remove the need to make untenable assumptions on the missing data mechanism.

**Category 3: Latent structures**

**Similarity measures**

Patient similarity measures apply a sequencing algorithm to establish the alignment of 2 sequences of patient data (eg, longitudinal EHR data). Sha et al. presented a novel similarity measure, which recognizes that the type of tests ordered and the time between tests can be indicative of patient condition. Their metric is therefore based on a distance measure incorporating the type, timings and results of tests and they assume that more intense monitoring indicates a more severe condition.

The sequencing algorithm produces a similarity matrix, defining the similarity between each pair of patients. We do not present the model formulation for this method since there are various approaches to using this matrix in prediction (described by Shara-foddini et al.). One such method defines cohorts of “similar” patients within which to develop separate models. This approach can be viewed as an extension of the pattern submodel approach with longitudinally and irregularly measured predictors, in which the patterns are defined by similar longitudinal sequences.

The benefit of this method is that, as with others, it can be applied to any form of prediction framework. Drawbacks include the computational burden of redefining multiple models, and requiring access to the training data at prediction time to train a model using similar patients.

**Latent variable**

A simple way of representing a latent clinical condition is to use a single (partially) latent binary variable, representing 1 of 2 states. This approach was used by Coley et al. and Hubbard et al. in which IP and IO are incorporated by allowing the measurement process to infer a latent patient condition under a hierarchical structure.

Define the partially latent binary outcome $Y^l \sim Bern(\eta)$ representing 1 of 2 patient states, where only 1 state is entirely observed. In Coley et al., example, “true” cancer state (aggressive vs indolent) is the outcome, but is only observed for a subset of patients who underwent surgery. We then assume that the value of the outcome can influence the presence of $x_t$ within the hierarchical model.

$$ R \ Y^l, \ Z \sim Bern(P|R = 1 \ Y^l, \ Z; \ b) \quad (6) $$

We have not provided the outcome model formulation since predictions are obtained by sampling from the posterior of the full hierarchical model.

Both studies note improved predictive performance in which the measurement process influences predictions compared with a model that ignores IP and IO. These models can, however, be computationally intensive to fit.

**Hidden Markov models**

Hidden Markov models extend the latent variable approach by allowing a time-varying latent process. Zheng et al. and Alaa et al. used hidden Markov models to capture IO, but the way they incorporated the observation process differs. Hidden Markov model-based prediction models incorporate IO by allowing the measurement frequency or rate to infer the clinical state at any given time.

Alaa et al. proposed a latent semi-Markov process to capture a patient’s evolving clinical state. The “state” variable, $Y^l(t) \in \{1, \ldots, 4\}$, ranges from clinical stability to clinical deterioration, where stability (state 1) and deterioration (state 4) are observed states, but intermittent states are latent. Here the model aims to predict eventual clinical deterioration, that is, $P(Y(\infty) = 4)$. The observation process (ie, timings) of $X(t)$ is used to infer this clinical state, in which it is assumed that increased monitoring indicates a less stable condition. A marked point process model (in this case a Hawkes process) is adopted to model the rate of patient monitoring, with the marks corresponding to the observed value. IO is captured through state-specific intensity functions for the monitoring frequency as follows:

$$ \lambda(t | Y^l(t) = 1) = \lambda_1 + \sum_{t < t_m} e^{-\beta_1(t - t_m)} \quad (7) $$

$$ \ldots $$

$$ \lambda(t | Y^l(t) = 4) = \lambda_4 + \sum_{t < t_m} e^{-\beta_4(t - t_m)} \quad (8) $$

$\lambda_1, \ldots, \lambda_4, \beta_1, \ldots, \beta_4$ are state-specific parameters to be estimated. $t_m$ is the time of the last measure of $X(t)$, $t$ is the time of the most recent change in $Y^l(t)$, which is only observed if the state is absorbing. Details of the learning and prediction algorithm are presented in more detail in their article.

A key advantage is that the Hawkes process allows for a time-varying intensity in the observation process. Model fitting and interpretation are, however, complex because there are multiple components to be estimated simultaneously.

**Joint modeling**

Joint modeling has been developed extensively within the prediction context, particularly for dynamic prediction, that is, incorporating time-updated variables ($S_2$ in Table 1). Joint modeling can be extended to handling IP and IO, by linking the outcome to the observation process via a shared random effect, which can be seen as an alternative approach to modeling latent variables. Separate models are defined for the outcome occurrence and the observation process, each of them containing an individual-level random effect representing individual “frailty.” By sharing these random effects across the 2 models, the outcome and observation processes are linked. Liang et al. and Choi et al. both allow for irregularly observed visits, and therefore specify a hazard or intensity function that defines how often visits occur. The random effect, or frailty term, controls how an individual’s visit rate differs from average. As this effect also appears in the model for the outcome, the visit rate indirectly affects the prediction for the outcome.

The method outlined in Zhang et al. only allows for scheduled, regular observations. Therefore, rather than specifying a model for the intensity/hazard of visiting, the “observation process” model is a repeated measures logistic regression model, in which the outcome indicates whether an individual provided data at a specific time point.

Joint models take many different forms and provide the most general framework. We present an example of a trivariate joint model, with submodels for the repeatedly and informatively measured covariate, the binary outcome, and the observation process of the covariate $x_4$. Assuming that measurement times are regular ($t_i = t_j \neq i, j$),

$$ X = x_0 + x_1 Z + x_2 t + U \quad (9) $$
DISCUSSION

This study has identified 3 broad categories of approaches to incorporate IP or IO into clinical prediction models: derived predictors, modeling under informed presence, and latent structures. This is a growing area of research, and much of the included literature illustrates that IP and IO can be incorporated into clinical prediction models in a meaningful way. Where missing data and nonrandom visit processes have been seen as a nuisance in effect estimation, a tractable way is to incorporate IP or IO into clinical prediction models: derived predictors, modeling under informed presence, and latent structures. This is an attractive but also a concern. Simple summaries of the entire process do not capture important changes in the observation process over time, such as a sudden increase in monitoring frequency which indicates worsening state. Latent structure approaches (eg, modeling measurement times via a nonhomogeneous point process) may be better suited to capturing longitudinal variability but are computationally intensive. Developing a more sophisticated representation of the observation process to use as a predictor is a promising avenue of further research, offering a potential trade-off between the simplicity of summary measures and the sophistication of joint modeling. These more complex measures should be compared with both joint modeling techniques and simple summary measures to assess their added benefit in terms of predictive performance and computational efficiency. We plan to perform such comparisons in a separate full empirical study.

There already exists a vast body of literature on joint modeling for prediction, particularly covering 32 (incorporating longitudinal predictors). Such methods have also recently been extended to functional data, allowing them to accommodate complex structures in longitudinal predictors. Joint models have also been proposed to handle IO under an inferential framework, so it follows that there is scope to extend joint models further to exploit IO for predictive benefit, as this review revealed that the method remains underdeveloped for this particular purpose.

There are broader challenges associated with exploiting IP and IO for prediction. First, because the association between the observation process and outcome is unlikely to be causal, this relationship
may not generalize well to different settings. For example, clinicians’ monitoring behaviors are likely to vary across units or clinical guidelines could recommend changes in the way patients are observed. This is particularly true following the introduction of a CPM into clinical practice; once this happens, the predictor variables in the model are far more likely to be observed. The predictive utility of any model incorporating the observation process should therefore be regularly validated and potentially updated.

A second challenge described by Alaa et al. 53 concerns models that use the observation process to inform predictions, but also update predictions as new information becomes available. An issue arises when clinicians change their monitoring behavior based on predictions produced by the model; any changes in the way they monitor patients will be fed back into future predictions via the observation process. This should be accounted for to avoid the feedback loop, potentially by developing causal models to account for the possible time-varying confounding, 72 or by explicitly modeling the effects of previous predicted values.

Despite these challenges, we view IP and IO as opportunities to improve the performance of predictive models, as opposed to a nuisance. The literature is divided on this point; much of the work in this review proposes methods that “overcome” the challenges of IP and IO, whereas others illustrate the added benefit of incorporating informative measurement patterns. Missing data have typically been seen as a threat to the estimation of parameters, but because this is not the key focus of prediction research, it may be useful to move away from terms such as missingness and instead focus on what the presence of an observation can tell us.

CONCLUSION

We have demonstrated that there is a growing recognition of both IP and IO within prediction research. Although parallels exist with missing data, IP should not be considered the same way, especially within the context of prediction and routinely collected data in which there is no prespecified observation process. By synthesizing the available methods and software that could be applied to incorporate IP and IO into CPMs, this article can assist applied researchers in adopting suitable methods. Future research should investigate the challenges presented herein, which will require the development of formal guidelines and making existing methodology more accessible.

FUNDING

This work was supported by Medical Research Council grants MC UU_00002/5 (JKB), MC UU_00002/2 (BT), and MR/N013751/11 (RS) and the Alan Turing Institute under the “Predictive Healthcare” project (Health and Medical Sciences Programme).

AUTHOR CONTRIBUTIONS

RS designed the study, conducted screening, and wrote the manuscript. LL conducted screening and provided critical revisions to the final manuscript. GPM, MS, and NP provided substantial contributions to the conception, design and conduct of the work, and provided critical revisions to the final manuscript. JKB, BT, and KD-O contributed to discussions on study design and conduct, and provided critical revisions to the final manuscript.

SUPPLEMENTARY MATERIAL

Supplementary material is available at Journal of the American Medical Informatics Association online.

ACKNOWLEDGMENTS

We thank 2 anonymous reviewers for their thoughtful comments on our manuscript, which have undoubtedly strengthened the final version.

CONFLICT OF INTEREST STATEMENT

The authors have no competing interests to declare.

REFERENCES


