INFORMATIVE PRESENCE AND OBSERVATION IN ROUTINELY COLLECTED HEALTH DATA: METHODS TO SUPPORT THE DEVELOPMENT OF CLINICAL PREDICTION MODELS

A thesis submitted to the University of Manchester
for the degree of Doctor of Philosophy
in the Faculty of Biology, Medicine and Health

2022

By
Rose Sisk
School of Health Sciences

[BLANK PAGE]

Contents

Word Count (excluding bibliographies): 30769

List of Tables

List of Figures

Abstract

The development and implementation of clinical prediction models using routinely collected health data is a challenging yet promising avenue of research. When data are collected opportunistically as a result of routine healthcare contacts, information is only collected according to clinical indication, or patient/-clinician concern. This means the patterns of observing the information/data are potentially informative with respect to patient condition: so-called “informative presence” and “informative observation”.

Within this thesis, we aim to assess to what extent informative presence/observation have been considered in the methodological prediction modelling literature and summarise the available methods for doing so to help applied researchers. We then perform simulations and empirical analyses to quantify the impact of allowing clinical prediction models to learn from informative observation processes, and study challenges associated with the use of (informatively) missing data in the development and implementation of clinical prediction models. We provide guidance for applied and methodological researchers on how to approach informative presence and observation, as well as setting out an agenda for further research.

We find that simple ways of harnessing informative measurement patterns (such as including missing indicators, or measures that summarise the observation process, as model predictors) can offer gains in predictive performance, especially within our clinical exemplar where one of the key outcomes can be difficult to predict. The findings and implications of this thesis have the potential to improve the development and implementation of clinical prediction models using routinely collected health data.

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright

Acknowledgements

The completion of this PhD has been aided considerably by the input and support of my friends, family and colleagues. Firstly, I would like to thank my supervisors Dr. Glen Martin, Dr. Matthew Sperrin and Professor Niels Peek for supporting me along this journey. I could not have asked for a more helpful, approachable and supportive supervisory team. I am especially grateful for all that you have taught me, and for the many insightful discussions along the way.

The love and unwavering support of my partner, Amrish, has kept me sane and happy - you have been my rock throughout. Thank you for putting up with my questionable conversational skills during the tough times, and cherishing the highs with me. Your hardworking and determined nature has inspired me to approach my work with a similar mentality.

I would also like to acknowledge the newest member of our household, Cleo the German shepherd, who entered our lives just over a year ago. You have been a constant source of much needed entertainment and affection ever since, and have kept me grounded during the more stressful moments.

Thank you to my mother, Rachel, for your wholehearted belief in me - it means more than you know. Thanks also to my late father, Tim, who taught me to always ask stupid questions and pursue the things that interest me most rather than those that make the most sense.

Last but by no means least, I am grateful to have been in excellent company in the Centre for Health Informatics and broader University of Manchester community. The coffee/lunch/zoom chats with my fellow PhD students and colleagues have undoubtedly enriched the experience.

Chapter 1

Introduction

Due to the existence of modern computer systems, data is constantly being recorded on our day-to-day activities. This is especially true within healthcare, where details of each interaction we have with health services are recorded in huge electronic databases for the primary purpose of supporting delivery of care. From a research perspective, these data sources have an attractive side-effect that they can be used to answer important research questions at scale. The secondary use of routinely collected health data is, however, challenging due to key differences in the way that it is collected compared to more classic research data sources. Specifically, data are only collected upon patient interaction with service, in contrast to other data sources that may have a protocol governing the observation process (e.g. like in an controlled trial or prospective cohort study), meaning that every individual has different types of information recorded at different time points. Although this poses a challenge to the analysis of this type of data, it could also present an opportunity to harness additional information that is "hidden" within potentially informative measurement patterns. This thesis explores exactly this: whether and how informative observation processes can help us to better understand patient condition (present or future).

This introduction chapter of this thesis is laid out as follows: firstly, we introduce Clinical Prediction Models - their uses, development, validation, and methodological challenges. Next, we discuss electronic health record data - what it is, and the associated uses and challenges. We then introduce and define the key topics of this thesis - informative presence and observation, and how they could be considered within the context of prediction modelling and electronic health record data. We propose a clinical exemplar of chronic kidney disease progression in the following section, and finally provide an overview of the specific aims of this thesis.

Note that the entirety of this thesis is written using the plural first-person pronoun “we” - this is to allow consistency in the style of writing across all thesis chapters. The thesis is presented in alternative format and therefore the chapters within the main body are written as journal articles (Chapter 2, 3 and 4), where the “we” voice refers to the full team of co-authors. The introduction and discussion chapters (Chapter 1 and 5) are, however, reflections of the PhD candidate (Rose Sisk) submitting this thesis and should be interpreted as if they were written using the “I” form.

1.1 Clinical Prediction Models

Clinical prediction models (CPMs) are mathematical tools that allow the estimation of an individual’s risk of having or developing a condition of interest [1]. CPMs can be broadly categorised as either diagnostic or prognostic, where diagnostic models establish the probability that an individual currently has a particular condition (the “outcome” or “event”), and prognostic models estimate the risk that they will develop the condition in the future [2]. CPMs estimate risk conditional on a set of predictors that are available at the time a prediction is made. Predictors can consist of patient demographics and lifestyle factors (e.g. age, gender, smoking status, exercise), or key clinical biomarkers such as blood pressure, kidney function or metabolic markers such as serum cholesterol levels.

CPMs have many uses, one of which is to better inform patients and care providers about risk in order to tailor decision-making and interventions to the needs of the individual. Establishing and discussing risk facilitates shared decision-making between the recipients and providers of healthcare services, allowing for preventative action to be taken for those most at risk, and in turn improving patient outcomes [3]. The benefits of early intervention in high risk individuals also extend to health systems since preventing cases of, or complications from, chronic illness through early intervention reduces the burden that these cases would have put on health services.

A notable example of a widely-used CPM that is approved for use in clinical practice in the UK (and recommended by the National Institute for Health and Care Excellence, NICE) is the Kidney Failure Risk Equation (KFRE)[4]. The KFRE estimates the risk of requiring renal replacement therapy within the next 2 or 5 years amongst patients with Chronic Kidney Disease (CKD). Predictions obtained from applying the KFRE are used to guide decisions on when to refer patients for specialist assessment, or in more serious cases, to plan for interventions like kidney dialysis.

Another use of CPMs is in benchmarking and performance monitoring. An example of such a model is APACHE IV [5], used to provide a severity measure to patients admitted to critical care. The severity scores are then used to quantify expected versus observed mortality within critical care units to evaluate their performance.

1.1.1 Developing a CPM

The development of a clinical prediction model is a data-hungry task [6] that, in the case of prognostic modelling, can require long term follow-up of large patient cohorts. The data used to develop CPMs is often referred to as the "development data". Collecting such data prospectively can be both time and cost intensive, so secondary analysis of existing data provides a multitude of advantages such as large available sample sizes and more readily available research data.

Typically, CPMs are developed using a cohort study design, whereby the predictors are measured at a single time point (a cross-sectional snapshot of patient data), and the outcome of interest is either the presence of a condition (for diagnostic models), or when the condition occurs (for prognostic models - patient follow-up is required under the cohort study design to observe this) [7]. The time at which predictors are observed should correspond with a clinically relevant point in the patient's care pathway, at which a prediction would be useful. For example, this may be upon referral to specialist services such as cardiology or renal medicine, or at the beginning of an admission to a critical care unit [8]. The set of candidate predictors considered for possible inclusion in the model should therefore ideally include only items that are routinely available at that particular point in the care pathway.

Guidelines exist that dictate the best practices for model development methodology, covering predictor selection, covariate-outcome association structures, the handling of missing data, model assessment and more [1, 7]. Most commonly, statistical tools such as linear regression, logistic regression, or Cox models underpin CPMs in the case of continuous (numeric), binary and time-to-event outcomes respectively. It is also becoming increasingly common to use machine learning techniques such as neural networks or tree-based methods [9], however the methodological conduct and reporting of the development of such models has been brought into question [10].

In the next sections, we will describe the mathematical details of the most common types of model (logistic regression and Cox models). Details and notation of other models can be found elsewhere [1, 11, 12].

Logistic regression modelling

Logistic regression is a method designed to model a binary outcome - in the context of clinical prediction modelling, this outcome represents the presence or absence of a condition of interest at a particular time point, i.e. the condition that one is trying to predict. A logistic regression model with k predictors developed using data from $i = 1, ..., n$ patients takes the following form:

$$ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki}$$

where $p_i = P(Y_i = 1), Y \in \{0,1\}$, the probability that patient i experiences the outcome Y, ln is the natural logarithm and the $\beta_0, \beta_1, ..., \beta_k$ terms are the model coefficients. β_0 is often referred to as the **intercept**, and represents the (log odds of) the outcome prevalence when all predictor values are set to 0.

The model coefficients (along with observed predictor values) are then used to make risk predictions ($\hat{p}_i$). The above formula can be transformed, and replacing the β's with their estimates ($\hat{\beta}$) we get:

$$\hat{p}_i = \frac{exp(\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + ... + \hat{\beta}_k X_{ki})}{1 + exp(\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + ... + \hat{\beta}_k X_{ki})}$$

The

$$\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + ... + \hat{\beta}_k X_{ki}$$

part of the model formula is sometimes referred to as the **linear predictor** (LP).

Cox proportional hazards regression modelling

Logistic regression models are not appropriate in all cases, for example, when patients within the development data have differing lengths of follow-up, or if predictions are required for multiple time points in the future. The Cox proportional hazards model [13] is commonly used as an alternative in this context.

To fit a Cox model, one must first define the following notation: define the combined information on survival time for patient i (from some pre-defined start time, such as time of diagnosis or onset of admission) as $S_i = (T_i, D_i)$ where T_i is the time of censoring, or occurrence of the outcome of interest, and $D_i \in \{0,1\}$, where $D_i = 0$ if patient i was censored (meaning that they did not experience the outcome by the end of the follow-up period), and $D_i = 1$ if they experienced the outcome.

We must also define the "hazard" of experiencing the outcome at time t, $h(t)$, where the hazard is the instantaneous risk of experiencing the outcome at time t.

This can be expressed as:

$$h_i(t) = h_0(t) exp(\beta_1 X_{1i} + ... + \beta_1 X_{ki})$$

The $h_0(t)$ term acts as a kind of intercept within the Cox model, called the baseline hazard, and represents the instantaneous risk of experiencing the outcome at time t when all predictor values are set to 0.

In order to predict from a fitted Cox model, the above formula is rearranged using the relationship between hazard and survival ($S_i(t)$ - the probability of being event-free by time t) to get the probability of experiencing the outcome by time t:

$$1 - \hat{S}_i(t) = 1 - \hat{S}_0(t)^{exp(\hat{\beta}_1 X_{1i} + ... + \hat{\beta}_k X_{ki})}$$

where $\hat{S}_0(t)$ is the estimated baseline survival probability at time t.

Note that the Cox model-fitting process does not directly fit the baseline hazard or survival, so these must be estimated separately when using this model to predict for new individuals. This can be done by e.g. estimating the baseline hazard within risk-strata from development data, or via methods such as the Breslow estimator [14].

Predictor selection

There are often a multitude of possible predictors that could enter a prediction model, so some form of predictor selection must happen as part of the model fitting process. Data-driven predictor selection techniques such as Lasso and Elastic Net [11] can be used to reduce the risk of overfitting by firstly reducing the set of predictors included in the CPM to only those that are most important, and secondly by applying a penalty term that "shrinks" remaining model coefficients. Such methods can, however, be unstable and it is therefore recommended that wherever possible, predictor selection should be based on expert knowledge over data-driven approaches [12, 15].

1.1.2 Validating a CPM

Following the development of a CPM, the model must be tested to ensure clinical validity and sufficient predictive performance in predicting the target outcome prior to implementation into clinical practice. By this we mean that the model produces predictions that are close to reality, and that the predictions are useful in determining who is/is not at risk. Note that although the underlying models used in the development of CPMs are often similar to those used in descriptive or causal inference work, the primary goal of a CPM is to produce accurate predictions within new patients (different to those used to develop the model), therefore traditional metrics of model fit may not be of primary concern. The process of checking the performance of a CPM is called "**model validation**", and is often performed within the same dataset and cohort within which it was derived (so-called **internal validation**). Models should be further tested in a new dataset or setting (**external validation**), completely independent of the development cohort and data. Ideally both internal and external validation should be conducted, especially if the model will be applied more broadly than in the specific setting within which it was developed.

Internal validation is often performed by randomly splitting the development dataset into two parts; one for developing the model, and a separate one for testing [1]. This approach is, however, statistically inefficient as it does not make full use of the entire sample for model development [16]. Techniques such as cross-validation and bootstrapping have been proposed to overcome this issue whilst allowing correction for optimism [17]. Cross-validation (CV) involves splitting the data into k equal parts, using $k-1$ parts to develop the model and the remaining part to validate it. This is repeated k times for k-fold cross-validation, and performance metrics can be summarised across the k folds. Bootstrapping, on the other hand, takes a sample with replacement from the original development cohort, and uses this to perform the entire model-fitting process, then tests performance on the original development data[12]. This process is repeated any number of times, and (as with CV) measures of predictive performance can be summarised across all bootstrap samples.

The goal of external validation is to test how well the CPM generalises to new settings, populations or time points. The model should therefore be applied within a completely separate cohort of patients to the development data, and model performance evaluated within this new cohort. External validation can take place geographically (from a different region or country), temporally (data collected at a later time point) or clinically (e.g. primary vs secondary care) to the original cohort. Although external validation is essential to ensuring a model is fit-for-purpose, the majority of published CPMs have not been externally validated [18], perhaps due to a lack of availability of suitable datasets.

Model performance can be quantified via the assessment of two key concepts: **discrimination** - how well the model can differentiate individuals with or without the outcome of interest in the sense of estimating higher risks for those that experience the outcome of interest than those that do not, and **calibration** - how "accurate" are predicted risks obtained from a CPM compared with the observed risks, across the full risk range. Discrimination is most commonly assessed via the C-statistic [19], otherwise known as the "concordance statistic", and can be estimated from essentially any underlying model as long as predictions and observed outcomes are available. The C-statistic measures the concordance between all possible (or comparable) pairs of individuals in a given dataset - if one person experiences the outcome of interest, and the other does not then this pair is concordant if the patient with the event has a higher predicted probability according to the model.

Calibration is assessed in many ways [20]. The first of which is sometimes referred to as "**calibration-in-the-large**" (CITL), or "mean calibration" [21]. Mean calibration assesses whether the average event rate matches the average predicted risk from a model, and can be used to identify whether a model over- or under-estimates risks on average. In the case of a logistic-regression based CPM, CITL can be calculated as the intercept from a model fitted to the observed outcomes, and the linear predictor used as an offset [1].

Following the hierarchy of calibration proposed by Van Calster et al. [20], "weak calibration" follows on from "mean calibration". A CPM is said to be weakly calibrated if, on average, it has good mean calibration and additionally its predicted risks are neither too extreme (too close to 0 or 1), nor too modest (too close to prevalence or incidence of outcome of interest). The estimated slope from a model fitted to the observed outcomes and the linear predictor as the sole model predictor can be used to assess weak calibration, often referred to as the "**calibration slope**".

A CPM can be described as "moderately calibrated" if the predicted risks are closely aligned with observed event proportions, i.e. within the subgroup of 100 patients each with a predicted risk of 20%, 20 of them develop the event. Moderate calibration is assessed by plotting the observed vs predicted risks (a calibration plot), ideally smoothed using e.g. a loess smoother [22], but more commonly binned into risk categories based on the distribution of the predicted risks.

Van Calster et al. [20] finally define "strong calibration" - when the observed event rates match up perfectly to the predicted risks for every possible combination of the predictor values. They note, however, that this would mean that the model is perfectly calibrated and is an unrealistic expectation for a real CPM.

Evidence suggests that many CPMs have not been sufficiently validated, as there is often a strong focus on discrimination and little attention paid to calibration [23, 24]. It is, however, essential to properly assess model calibration since it provides insight into how closely risk estimates are to reality - unreliable risk predictions could result in inappropriate decisions regarding the best course of action for a particular patient when CPMs play a key role in the care process.

Ideally, model validation should not be a one-time exercise, as the performance of a prediction model can be affected by changing clinical practices or policies [25, 26], and strong model performance in a single setting or population does not mean that the model will perform well in others. The very existence of a model could pose a barrier to its own sustained performance [27, 28]. This phenomenon is a result of end-users of CPMs acting upon the resulting predictions (e.g. offering interventions), which in turn changes the risk profile and distribution of the outcome in the CPM's target population. A possible means of mitigating this is in the continuous surveillance and updating of CPMs [29, 30], whereby model performance is continuously assessed. This is, however, a relatively novel area of research and approaches to continuously validating a CPM are underdeveloped [29].

Figure 1.1 illustrates how the model pipeline might look under a continuous validation/updating framework.

1.1.3 Impact assessment

Ideally, prior to implementing a CPM into clinical practice, the effect of the CPM on patient outcomes should be quantified to ensure that, at the very least, the existence of the model is not causing any harm, and that it is impacting the decision-making process in some way [31]. This process is often referred to as "impact assessment", and in practice is rarely undertaken [32, 33]. The gold standard approach to conducting impact assessment would be through the design and conduct of a randomised controlled trial (RCT) [33], which are time and cost intensive to run. The lack of concrete evidence of the clinical efficacy of developed CPMs acts as another potential barrier to their adoption in practice.

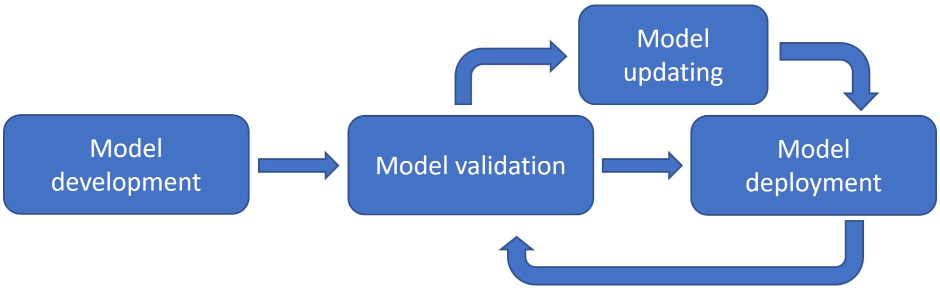


Figure 1.1: An illustration of a CPM pipeline, where the model is repeatedly validated and updated

1.1.4 Applying/deploying a CPM in clinical practice

Another example of a widely adopted CPM in UK clinical practice is the QRISK3 model [34], used to predict 10-year cardiovascular risk in UK primary care. Risk estimates obtained from QRISK3 are used to guide preventative action such as the prescription of statins [35]. CVD risk is an area that has received significant attention within the prediction modelling literature [36], and a range of tools exist to predict cardiovascular risk such as EuroScore [37] and the Framingham Heart

Score [38]. An important step in the CPM model pipeline is model deployment, or implementation into clinical practice. The tool, or the resulting risk predictions, must be available to clinicians at the point-of-care in order to be used to guide decision-making [39]. QRISK2 (and in some cases, QRISK3) has been integrated into existing clinical computer systems for automatic calculation of QRISK scores using data pulled from the patient record. Alternatively, the tool is available to use via a web app [40], where clinicians can manually enter patient information to obtain a risk prediction.

The way in which models are implemented into clinical practice has implications for the methodology that can be used to develop and use CPMs. CPMs based on regression models such as logistic or Cox regression are relatively straightforward to apply, and require minimal computing power at the point of prediction, as it is usually sufficient to simply store the model formula (including the estimated coefficients and intercepts/baseline hazards) to predict for new patients. Other techniques such as joint modelling or neural networks have significantly higher storage and computing power requirements at the development and deployment stages, which could present a barrier to their implementation [41]. A recent review by Sendak et al. [42] of models that have been successfully integrated into EHR systems discusses the various associated challenges and opportunities, and in particular they highlight the potential investment required to achieve success.

1.1.5 Recent methodological developments in CPMs

There has been a recent interest in developing CPMs that consider repeated measures of predictors that vary over time. For example, a model that looks at the entire history of blood pressure measurements when predicting cardiovascular disease [43], or looking at how rapidly kidney function is declining over time in the prediction of kidney failure [44]. The aims of such models are generally twofold: firstly, to develop a model that can update predictions as new information becomes available, and secondly, to allow a model to learn from the entire history of a key measure as opposed to a single instance. This latter goal tends to be more in line with the true clinical decision-making process, where clinicians would rarely consider a single observation in isolation when the entire history is available. A recent review by Bull et al. [45] summarises the methodology that has been proposed and employed to this problem. The process of updating predictions over time (as new information becomes available), conditional on remaining at risk at the new future time point, can be referred to as “dynamic prediction”. Note that the term “dyanmic prediction” has also been used to refer to models that update over time [29], as opposed to updated individual predictions, but within this work we use the term to refer to the prior.

Clinical prediction models are increasingly being developed using machine learning (ML) methods as opposed to classic statistical modelling. ML has its foundations in mathematics and computer science, but many ML methods are also heavily based on statistics, and there is therefore considerable overlap in the methods that are classed as either “statistical modelling” and “machine learning”. For example, regularized versions of logistic regression or Cox regression are often referred to as ML. Alternative modelling strategies such as neural networks, random forests, support vector machines are more typically referred to as ML rather than statistics, and many of these methods are designed to identify and estimate complex, high level interactions among the model predictors. They have proven to be particularly helpful in image [46] and text [47] processing.

Despite the recent increase in the use of ML in the development in CPMs, a recent systematic review observed a high risk of bias in the reporting of the development of ML-based models [48]. Christodoulou et al. [48] advise that comparisons between “classic” statistical models and ML-based models are often flawed, and that there should be a heavier focus on model calibration when performing such comparisons. Typical ML methods also tend to be far more data-hungry than more typical regression-based models: a study by Van Der Ploeg, Austin, and Steyerberg [49] showed that performance metrics for models developed using neural networks, random forests and support vector machines were unstable even when the sample size was very high relative to the number of model predictors. Machine learning models have also been applied to the problem of dynamic prediction [45]. Most commonly, these methods take a “two-stage modelling” approach, whereby any information held within repeatedly measured model predictors is modelled (or aggregated) separately to the primary outcome of the prediction model, and information from the first stage is used as a fixed-time predictor in the main prediction model.

1.2 Missing data

A key challenge in the analysis of health data, and in the development of clinical prediction models, is the handling of missing data. Much of the work on missing data is grounded in the classifications described by Little and Rubin [50]. Within this framework, missing data are categorised into one of three mechanisms, displayed in Table 1.1, that describe the relationship between missing and (un)observed data.

Table 1.1: Description of Rubin's classifications of missing data

Mechanism	Description
Missing Completely at Random (MCAR)	There is no systematic difference between the missing and the observed values. E.g. data are missing due to equipment breakdown.
Missing at Random (MAR)	The difference between the missing and observed values can be explained by differences in another observed measure. For example, blood pressure (BP) may be more likely to be missing in younger patients (and BP is associated with age), but we have collected data on age.
Missing Not at Random (MNAR)	Even after the observed data have been taken into account, there are still systematic differences in missing and observed values. E.g. lower values are less likely to be recorded as there is no cause for concern.

Typically, the choice of missing data handling strategy is driven by the assumed missing data mechanism. Many of the methods developed to handle missing data are developed under the assumption that missing items are MCAR or MAR, and can therefore be imputed by exploiting the relationships between the observed and missing data [51, 52]. MNAR is therefore the most challenging mechanism to handle since we we cannot guarantee that any estimate of the unobserved data based on observed items will be unbiased.

Although one must make an assumption about the assumed missingness mechanism, it is impossible to know whether data are truly MNAR or not, since we know nothing about the data that have not been observed [53], and by definition the missing items are dependent on other unobserved data. It is, however, possible to test the plausibility of MCAR vs MAR mechanisms (using, for example, relatively straightforward procedures discussed by **Enders**, 2010) but this is not widely done in practice [54]. It is helpful to work with clinical experts to better understand the underlying patterns of observation within the EHR to make better informed assumptions about the nature of missing data [55].

Perhaps the most straightforward method of handling missing data is to simply remove patients with missing items, so-called “complete case analysis”. This method was found to be the most commonly applied strategy in the development of CPMs in a recent review [56]. Within this thesis, we consider the presence of missing data to be informative and therefore do not focus on the complete case analysis method since there is no way to draw information from potentially informative missing data if this data is excluded from the analysis. It is therefore not recommended as a means of handling informatively missing data.

Multiple imputation (MI) is often considered the gold standard approach to handling missing data in the context of health data analysis [1, 3]. MI is performed by defining a model for the missing predictor based on other available information in our dataset, and repeatedly sampling from the posterior predictive distribution of the missing variable to create a set of complete datasets [57]. By creating and analysing more than one imputed dataset, the process accommodates and carries through the uncertainty associated with the “filling in” of missing data - something that simpler single imputation models cannot achieve. An illustration of how MI works is shown in Figure 1.2. In the following section, the secondary

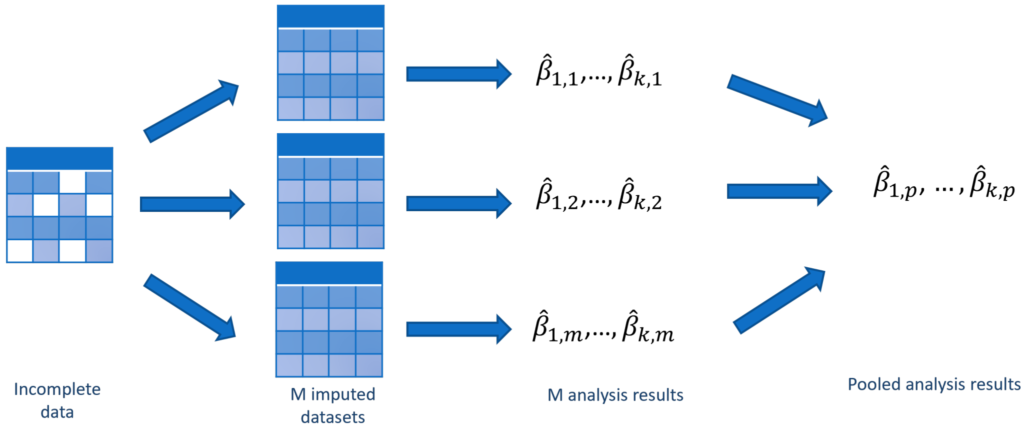

$\hat{\beta}_{1,1},\ldots,\hat{\beta}_{k,1}$
$\hat{\beta}_{1,2},\ldots,\hat{\beta}_{k,2}$
$\hat{\beta}_{1,m},\ldots,\hat{\beta}_{k,m}$
$\hat{\beta}_{1,p}, \ldots, \hat{\beta}_{k,p}$
Incomplete data
M imputed datasets
M analysis results
Pooled analysis results

Figure 1.2: An illustration of multiple imputation, where model coefficients are pooled across imputed datasets

use and analysis of patient medical records will be discussed. Analysing such data presents opportunities for these methodological advances to be applied in clinical practice.

1.3 Electronic Health Records

The recent widespread adoption of Electronic health records (EHRs) [58] presents a multitude of opportunities for health research [59, 60], and particularly in the development and evaluation of CPMs [61–63], since they provide a relatively inexpensive means of obtaining data on a large number of individuals. Patient interactions with health services are captured within the EHR, resulting in databases that contain rich longitudinal clinical information that facilitates long-term follow-up of large patient cohorts. Prospectively collecting such data is a time and cost intensive task, so it is not surprising that research studies based on the secondary use of electronic health record data have proliferated in recent years [62, 64, 65].

The record can capture the entirety of a patient's care pathway through the healthcare system, especially in the UK where the majority of patients are seen within a single healthcare provider, and linking data across care providers is feasible [66]. It is, however, often difficult or impossible to obtain the entirety of every patient's record for research purposes, and EHR-based analysis datasets therefore represent a small subsection of this journey. Examples of information that is found in the EHR are: details of GP visits, procedures and diagnoses recorded during inpatient or outpatient hospital visits, and prescribed medications. Not every comorbidity is recorded at every visit, however, and point-of-care testing procedures that do not require involvement or reporting from analytical labs are likely to be omitted from the record.

Recent widespread uses of EHR data in clinical research include observational studies exploring prevalence and incidence of disease, comparative effectiveness research, epidemiological studies, feasibility studies and patient recruitment in clinical trials [64, 67, 68]. Additionally, the development of clinical prediction models has received considerable attention from both academic and industry researchers in recent years [62].

Traditionally, health research studies have followed more classical research designs that dictate the way in which patient data is observed over time. Under prospective study designs, data are all collected according to the requirements of the specific study at hand - for example, randomised controlled trials are the gold standard in establishing the effectiveness of a drug or intervention. Under such a design, patient cohorts are defined and recruited according to strict inclusion/exclusion criteria, and a study protocol will govern the types and timings of all patient observations. Retrospective study designs, on the other hand, instead make use of data that already exists at the beginning of the study, and therefore was not collected with the goal of answering the specific research questions at hand. Moreover, EHR data is not even collected with research in mind. Clearly, when data are collected without a research agenda in mind, no data collection protocol exists. Instead, patient visits and observations are driven by clinical indication, the patients' and clinicians' judgements, and the nature of the EHR system and how it is interacted with. This opportunistic nature of data collection in EHR systems can introduce methodological challenges into the development, validation and deployment of CPMs [69, 70]. In order to frame these challenges, it is useful consider the “who”, “what” and “why” of EHR-derived datasets. By this we mean:

- **Who** is included in the cohort?
- **What** data have we observed on these patients?
- **Why** have these particular data items been observed?

Each of the above questions can shed light on analytical and methodological considerations that are relevant to the analysis of routinely collected health data.

1.3.1 Who?

As part of the conceptualisation of a CPM, the intended target population within which the model will be used should be pre-defined. Since EHR data can (theoretically) cover entire populations, establishing well-defined patient cohorts seems like a relatively simple task, however careful attention should be paid to any conditions for inclusion that are based on the level of completeness of observed data. By this, we mean the sufficiency of the observed data items to run our intended analysis. Weiskopf, Rusanov, and Weng [71] illustrate that there is an association between EHR completeness and health status: they observe that patients with more complete data tend to be of poorer health. This finding could have important implications for CPMs that are intended for use within specific patient populations, or even in the general population, since our development cohort may be, as a whole, sicker than the target population, and have a different risk profile as a result. If the cohort used to develop a CPM is not representative of the population that it will be applied to, then the underlying model could be based on incorrect specifications of prognosis or predictor-outcome relationships. This has further implications for clinical-decision making - an ill-fitting model is likely to lead to ill-informed decisions and potentially improper patient care.

To illustrate this, consider a critical care setting, where we wish to use the first day of data to predict key outcomes such as mortality. We could impose a “completeness” inclusion criteria on our patient cohort, whereby only patients with a full 24 hours of data will be included in our development cohort. Many patients will therefore be omitted if they do not meet this criteria. Such patients could either have been discharged safely out of critical care as their condition stabilised, or at the other extreme, they may have died within the first 24 hours. Both of these scenarios have implications for the overall cohort’s risk profile, and could in turn result in the estimation of a poorly calibrated CPM.

Throughout this thesis, it is assumed that the derived cohort does represent the intended use population of our exemplar CPMs and we do not consider the issue of selection bias in any chapter of this thesis. The consideration of selection bias in EHR studies is a related (and important) yet separate issue to the methodological challenges studied herein, since we focus on the observation patterns of CPM predictors, and not who has or has not formed part of the patient cohort used to develop, validate or apply the model.

1.3.2 What?

An important consideration in the secondary analysis of EHR data is what data has been observed on our intended patient cohort, and, more importantly in the context of CPMs, whether the relevant candidate predictors are all present within the record. In the case of a typical cross-sectional model, we only require a single measurement of each predictor. Predictor measurements should occur within a suitable time frame relative to the prediction time e.g. within the 6 months prior to a referral to secondary care services, or within 24 hours of an inpatient admission. Observations falling outside of this time-frame should be classified as missing, but most importantly, the model should never be developed using data that would not be available at the time of prediction.

Clearly, not all required data will be available for each individual in EHR-derived cohorts, since we (as analysts) are not in control of the data-generating mechanism. Consequently, the handling of missing data is a commonly encountered methodological challenge in the development, validation and use of CPMs [62] in EHR data. There exists a broad range of methods to handling missing data. Perhaps the simplest approach is complete case analysis (CCA) that simply omits any patient records containing missing items from the analysis. Alternatively, to avoid discarding data or reducing our sample size, imputation methods such as MI, introduced previously, exist that attempt to fill in missing items with some replacement value or set of possible values. Other possibilities include imputing missing items with 0 (risk factor absent approach), the mean in the observed data, or a conditional mean based on one or more other available data items.

In addition to considering the type of information recorded in the patient record, the timing and frequency of predictor measurements will also vary between individuals, resulting in irregularly observed longitudinal data [72]. This is most relevant to dynamic CPMs, introduced in section 1.1.5, that consider repeated predictor measurements. A key consequence of the routine nature of data collection in the EHR is that observation times will differ between patients, as well as the intensity or frequency of monitoring. Predictor measurement schedules are adapted according to the (potentially changing) clinical requirements of the patient and the judgements of care providers. Resulting EHR-derived datasets will contain irregularly observed longitudinal data, which poses further methodological challenges to the development of CPMs.

1.3.3 Why?

Once we have established what data are available for our specific patient cohort, we must then consider **why** we have observed (or failed to observe) this particular set of information. Data collection in the EHR takes place as a result of patient contact with healthcare services. As such, patient condition and clinical requirements will likely determine both the data items that we observe in the record as well as the frequency or intensity of observation [73]. For example, upon visiting a General Practitioner (GP), only the tests relevant to that particular consultation will be ordered, and we tend to visit our GP more frequently during periods of ill health.

We can draw parallels between the presence/absence of data in the EHR and the existing literature on missing data. Given the unplanned nature of data collection in the EHR, it seems reasonable to assume that data are not “missing completely at random”, as there is often some clinical indication that results in data being observed for a particular patient [69]. This could be the development of a new symptom that prompts the patient to visit their GP, or a change in the perceived condition of a patient admitted to critical care that triggers a new set of lab tests to be ordered. It is also highly likely that data are missing not at random within this context, since the decision to order a particular test will be driven by the anticipated outcome i.e. only testing kidney function when it is expected to be poor. Previous work has described how it is possible to derive information about patient condition from the presence or absence of data alone [74].

We can extend the notion of the presence or absence (and underlying reasons) of specific data items to the monitoring frequency of various clinical parameters. As described in the previous section, each patient will be monitored and observed at varying rates, determined by their clinical requirements. Guidelines on the management of chronic conditions such as diabetes dictate that diabetic patients should have a set of key markers measured at least once annually [75]. We would also expect that, in general, those with overall poorer health tend to engage more frequently with health services than those in good health, resulting in varying levels of completeness and richness of the patient record across individuals. There may also exist differences in access to health services across deprivation levels and ethnic groups [76, 77].

The classifications of the missing data mechanisms described in Table 1.1 have been extended to the case of irregularly observed longitudinal data to describe visiting completely at random (VCAR), visiting at random (VAR) and visiting not at random (VNAR) [72]. These classifications are based on the relationship between visit times and outcomes, e.g. visiting completely at random indicates that visit times and outcomes are independent, whereas visiting at random requires that visiting at time t is independent of outcome at time t given data recorded upto time t. As with MNAR, VNAR is generally the most challenging mechanism to account for in a statistically sound way.

Others have further discussed how, especially within routinely collected health data, it is useful to distinguish between scheduled and unscheduled patient contacts [78]. Unscheduled contacts are more irregular and occur more unpredictably than scheduled visits, but likely tell us far more about the overall condition of a patient. In practice, however, it can be difficult to differentiate between scheduled and unscheduled visits by simply looking at the data, so this distinction is often impossible. The exception to this would be to distinguish between unplanned admissions and regular community-based follow-up visits.

We can further distinguish between patient vs. clinician-initiated data, using definitions proposed by Gruger, Kay, and Schumacher [79] (and later revisited by Gasparini et al. [80]), four possible observation schedules can be defined:

1. Examination at regular intervals - measurement times are fixed and defined for all patients. This is typical in prospective study designs.
2. Random sampling - measurement times are irregular across patients, but not informative with respect to patient condition, for example in screening or epidemiological studies.
3. "Doctor's care" - measurement times are specified by the care provider, but adapted according to the varying needs of the patient. The times only depend on the current (observed) status of the patient
4. Patient self-selection - measurement times are decided by the patient according to their perceived need to engage with health services. For example, a patient may choose to visit their GP when unwell. Generally the reasons for instigating or skipping scheduled visits are unknown under this process.

Schedules 1 and 2 above could be classed as "visiting completely at random", since measurement times are considered to be independent of health status. Under this definition, Schedule 3 could be viewed as "visiting at random", since the timings of (and between) measurements only depend on information available at that visit time. Schedule 4 is a form of "visiting not at random", since the drivers of measurement times are generally unobserved in the patient record.

At this stage I introduce two further definitions used throughout this thesis:

1. Informative presence (IP): The presence or absence of a patient's data at any given time point carries information about their health status that is not available elsewhere in the observed data.

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 that is not available elsewhere in the observed data.

Others have proposed a definition of informative presence that captures both of these phenomena (IP and IO) [73, 81], however (for the purpose of this work) it is necessary to separate out the cross-sectional and longitudinal scenarios, as the methodology suited to each will vary.

IP refers mainly to the cross-sectional prediction setting, i.e. where we wish to predict an outcome using a single predictor measurement, or considering only a single fixed window of time (and whether the predictors have been measured at all during that window - presence vs absence). IO extends this to the setting where one is interested in learning from the entire trajectory of repeated predictor measurements, and acknowledges that there is potentially information held within the intensity of measurement, not only the presence or absence of predictor measurements.

Methodological challenges associated with IP and IO will be discussed in the next section, particularly in relation to how they affect the development and implementation of CPMs.

1.4 Informative Observation processes and CPMs - Methodological implications

The previous sections described the nature of the data found in the EHR, and key differences with traditional research databases and designs. Many of these differences have implications for the choice of methodology employed.

1.4.1 Missing data and Informative Presence

More specifically, we first consider the issue of informative presence as a missing data problem. There exists a breadth of literature on the handling of missing data in health research studies [82–85]. We have discussed how data within the EHR is likely to be MNAR, and that the presence or absence of data itself can be informative with respect to patient condition.

MI, as with many other missing data handling strategies, was developed with the primary goal of recovering unbiased parameter estimates in the context of descriptive or causal analysis work, and is shown to work well under the assumption that data are missing at random (MAR) [84, 86, 87]. MI performs less well and is more likely to introduce bias into parameter estimates under MNAR mechanisms [87, 88]. This is a situation in which MI may fail to recover the underlying parameters that form the basis of a CPM, resulting in less accurate predictions from the derived model. It should be noted, however, that recovering unbiased parameter estimates is not the primary focus of prediction research, and in fact a model that is based on biased estimates can still perform well [89].

The creation and inclusion of missing indicators as model predictors is a promising yet straightforward method of potentially overcoming these issues related to informative presence. Firstly, under some missingness mechanisms, missing indicators combined with multiple imputation have been shown to reduce the bias in both parameter estimates [90] and model predictions [91] under MNAR when compared against multiple imputation alone, as they allow us to directly model the missingness mechanism. Missing indicators further provide a simple means of incorporating information about the observation process into the prediction model, allowing it to harness this additional information for gains in predictive performance. They should, however, be used with caution as they can introduce further bias into parameter estimates under some uninformative missingness mechanisms [92, 93]. Note that studies illustrating this effect generally only apply a simple imputation method such as zero or mean imputation to the missing item, so by using a more sophisticated imputation process it may be possible to mitigate this problem.

Furthermore, predictive modelling constitutes more than simply fitting a model and estimating the relevant model coefficients. Since missing data can occur at any stage of the model pipeline, we must first distinguish between handling missing data at model development, validation and deployment [56]. At the development stage, existing guidelines suggest that MI generally results in models with the best predictive performance [1, 94]. Some existing models also allow missing data to be present at the point of prediction (e.g. QRISK3), whereas others insist on obtaining a complete set of predictors to predict for new patients [56]. In the development of QRISK3, multiple imputation was used to account for missing items, but at the point of prediction missing items are imputed using conditional imputation based on age, sex and ethnicity (or zero imputation for the standard deviation of systolic blood pressure).

Although the handling of missing data at the model development stage has received considerable attention in the literature, the opposite is true for the latter stages of validation and model application. Guidelines for the application of multiple imputation become challenging to follow once a model is deployed in practice for two primary reasons: firstly, it is recommended that the outcome is used in the imputation model for MI [52], however this is, by definition, missing for new patients. Secondly, there may be a lack of computational power available at the point of prediction, which is required to apply MI to new patient data. Recent work by Hoogland et al. [95], Fletcher Mercaldo and Blume [91] and Nijman et al. [96], however, has begun to consider the application of MI within the context of applying a CPM to new patients with missing predictor values. We may also find that the very existence of a CPM has implications for the way in which data are observed, i.e. key model predictors are observed more consistently across patients when a model exists as part of the clinical care process [97].

Fletcher Mercaldo and Blume [91] explore the issue of imputing with or without the outcome in the imputation model at both model development and implementation. Both Fletcher Mercaldo and Blume [91] and Hoogland et al. [95] perform multiple imputation in such a way that uses the outcome to impute missing predictors in the development data, then imputes the outcome as part of the MI process for new individuals once the model is applied in practice. Fletcher Mercaldo and Blume [91] also explore omitting the outcome entirely from the imputation process at model development (and implementation), and find that this is the preferred approach under all missingness mechanisms in terms of reducing overall prediction error, despite the fact that the “MI without Y” approach results in biased parameter estimates. This is a particularly notable finding, that reinforces the fact that the underlying parameters of a model do not have to be unbiased for it to predict well. The work of Fletcher Mercaldo and Blume [91] does not consider the impact of the missing data handling strategy at development, when complete data must be obtained at deployment.

Hoogland et al. [95] and Nijman et al. [96] both propose methods of multiply imputing at the point of prediction based on information derived from the development data e.g. means and variance-covariance matrices of the observed predictors, or imputation models. This is a promising avenue of research that overcomes the issue of requiring access to the full development dataset at the time of prediction, which is rarely available due to privacy constraints and difficulties in sharing and securely storing patient-level data. Their work shows promise of the “conditional mean imputation” approach (what we refer to here as "regression imputation"), however it does not consider the impact of potentially informative missingness mechanisms, or how each method performs when used at development, but complete data are present at deployment. These additional considerations are explored in this thesis.

The process of applying multiple imputation can be computationally intensive. In the clinic, the necessary computing power to run MI may not be available, so the suitability of MI for handling missing data at the time of prediction is questionable. Since CPMs are deployed into a clinical setting and designed to be used within the clinic, we must consider whether the proposed missing data handling strategy can be feasibly applied in such a setting, and whether it can be applied consistently across all stages of the model pipeline.

The practical and methodological limitations of MI might explain the inconsistency in the missing data handling strategy between the development, validation and implementation stages for established models such as QRISK3. A recent review by Tsvetanova et al. [56] further found that missing data was handled inconsistently across the different stages of a CPM, and suggested that there could be consequences for model performance resulting from such inconsistency.

It is therefore evident that missing data handling strategies should be studied and adapted according to the differing requirements of prediction research [89]. The existing literature fails to consider that missing data can occur at any stage of the model pipeline, and therefore methods should be adapted to acknowledge this.

Although thus far, we have predominantly provided an introduction to the methodological challenges posed by informative presence to the analysis of EHR data, we could conversely perceive it as an opportunity to draw additional information from the patient record. Previous work has shown how the very presence of lab tests within the EHR can be informative with respect to patient condition [74, 98], and can be exploited for predictive benefit. Moreover, we have distinguished between the aims of prediction research, and descriptive/inferential work, and discussed how traditional missing data methods have generally been developed with the goals of the latter in mind. The analytic methods best suited to handling informative presence in the context of prediction research may therefore be different.

1.4.2 Informative Observation and Dynamic Prediction

As introduced earlier, there is a growing interest in "dynamic prediction" using models that can harness repeated measures of key clinical biomarkers, and produce updated individual-level predcitions as new information becomes available. Two methods dominate the (methodological) literature in this area: joint modelling and landmarking. Joint modelling, in short, separately models the primary outcome of the CPM via a time-to-event model, and a separate model is fitted to the repeated measurements of the longitudinally observed biomarker. Since the two processes are correlated, they share random effects in common, and the models are estimated simultaneously [99, 100]. The fitting of joint models can more accurately capture the dependence structure between the longitudinal evolution of a biomarker and the occurrence of an event of interest compared to e.g. time-varying covariates in Cox models [101], as well as accommodating informative censoring of the biomarker process due to occurrence of the time-to-event outcome [99, 102, 103].

Landmarking is a separate approach to dynamic prediction that identifies a series of "landmark times", at which a prediction of the time-to-event outcome is desired. The basic premise is that at each of these landmark times, the most recently observed set of information for each patient is used to predict the risk of the event occurring[104, 105]. Landmark models are fit within patient risk sets: patients still at risk (i.e. already entered the study, but not yet experienced the event/been censored) at the landmark time. The key difference between landmarking and joint modelling is that landmarking does not adopt a model for the biomarker process, and instead updates predictions based on the last observed value.

Informative observation in the context of inferential studies has recently received considerable attention within the academic literature. Various methods have been proposed, including extensions to the joint modelling framework [106, 107], to adjust for potentially informative observation processes, resulting in unbiased estimates of association parameters within and between the longitudinal biomarker process and the time-to-event outcome. An explanation of these methods has been conducted by Pullenayegum and Lim [72].

Given the recent interest in developing CPMs that learn from the entire trajectory of a longitudinally measured biomarker, this could be extended to further extract information with predictive value from the monitoring frequency of the predictors. Alaa, Hu, and Schaar [108] illustrate that the intensity of monitoring within an inpatient admission can be used to improve predictions of mortality. They note that by doing so, they are essentially allowing the model to learn from judgments and perceptions of the care providers that are not explicitly recorded in the patient record. Further work by Agniel, Kohane, and Weber [109] illustrates how the time of day, day of the week and time gaps between lab test monitoring are all predictors of mortality within inpatient admissions. There is therefore a growing body of evidence to suggest that allowing CPMs to harness IO for predictive benefit may provides gains in model performance.

Much of the existing literature around informative observation has focused on the need to estimate association parameters without bias (similarly to informative presence), as is the case in descriptive, inferential and causal inference work. As previously discussed, however, this is not the primary focus of prediction research, and therefore the extent to which it has been considered in the context of prediction should be explored.

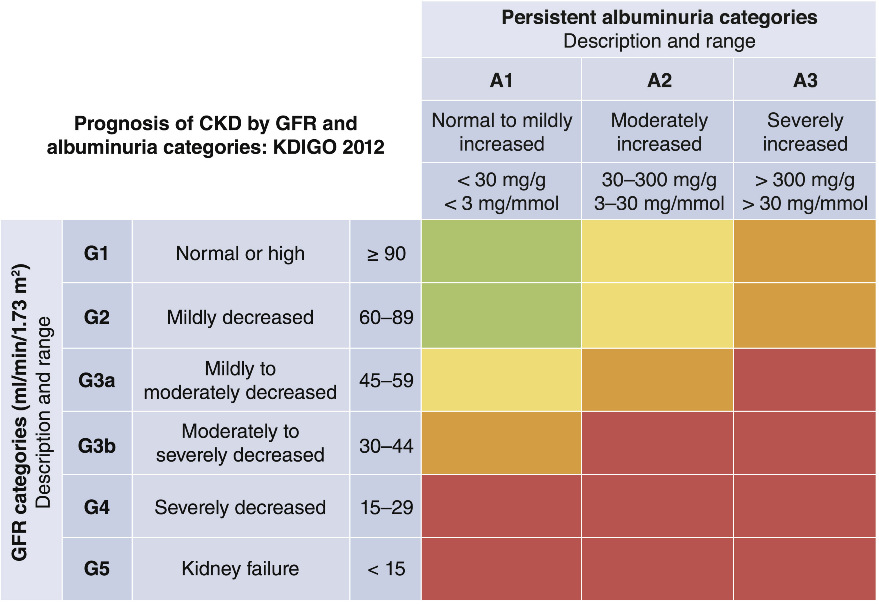
1.5 Clinical Exemplar: Chronic Kidney Disease & End-Stage Renal Disease

We now discuss a clinical exemplar that will be used as an illustrative example of how clinical prediction models could harness informative observation for gains in predictive performance.

Chronic Kidney Disease (CKD) is defined by a gradual deterioration in the function of the kidneys over time [110]. The kidneys' main function is to filter out waste and excess fluid from the blood, which are then passed through urination. Since CKD is a progressive condition, it generally worsens over time. In serious cases, CKD can progress to kidney failure, also referred to as End Stage Renal Disease (ESRD; when the kidneys can no longer sufficiently function), or cardiovascular disease [110], both of which can be fatal. Once a patient progresses to ESRD, they will require either renal replacement therapy (RRT) or a kidney transplant to survive.

CKD can be classified into one of five stages (shown in Figure 1.3), each describing differing extents of kidney damage [110]. The classification system is based on two key biomarkers that are used to diagnose, and assess the severity and risk of progression of the condition. These are: glomerular filtration rate

(GFR) - a measure of how well the kidneys are filtering, and urine albumin-to-creatinine ratio (UACR) which measures the level of albumin protein in the urine, an early indicator of kidney damage. GFR can be complicated to measure, therefore GFR is generally estimated (eGFR) based on a measure of creatinine in the blood. Based on these two values (eGFR and UACR), CKD stage can be classified from G1-G5 based on eGFR, and A1-A3 based on UACR. Note that for a diagnosis of **chronic** kidney disease to be made, there should be evidence of a sustained drop in either eGFR or UACR for a total of at least three months, captured by at least three repeated measurements of eGFR or UACR. Chronic Kidney Disease (CKD) is a major public health issue worldwide [112], with increasing prevalence and incidence [113] due to an increase in the prevalence of key risk factors such as hypertension and diabetes. Managing the condition and in turn reducing the risk of CKD progression is therefore essential to improving patient outcomes and reducing the burden placed on healthcare systems globally as a result of CKD [114]. Multiple tools have been developed that enable patients and care-providers to estimate the risk that an individual patient will progress to ESRD. Applying prognostic tools within the care pathway for patients diagnosed with CKD can facilitate timely referral to a nephrologist, and allow for the planning of renal replacement therapy. The most commonly used and well-validated



Prognosis of CKD by GFR and albuminuria categories: KDIGO 2012				Persistent albuminuria categories Description and range		
				A1	A2	A3
				Normal to mildly increased	Moderately increased	Severely increased
				< 30 mg/g < 3 mg/mmol	30–300 mg/g 3–30 mg/mmol	> 300 mg/g > 30 mg/mmol
GFR categories (ml/min/1.73 m^2) Description and range	G1	Normal or high	≥ 90	🟩	🟨	🟧
	G2	Mildly decreased	60–89	🟩	🟨	🟧
	G3a	Mildly to moderately decreased	45–59	🟨	🟧	🟥
	G3b	Moderately to severely decreased	30–44	🟧	🟥	🟥
	G4	Severely decreased	15–29	🟥	🟥	🟥
	G5	Kidney failure	< 15	🟥	🟥	🟥

Figure 1.3: KDIGO classifications of Chronic Kidney Disease: green - no kidney damage; Yellow: moderately increased risk; Orange: high risk; Red: very high risk [111]

CPM for the prediction of ESRD is the Kidney Failure Risk Equation (KFRE) [4], which has two versions - a long 8-variable version and a shorter 4-variable one. The short KFRE contains requires only age, sex, estimated glomerular filtration rate (eGFR) and albuminuria as inputs, and the longer version additionally includes serum calcium, phosphate, albumin and bicarbonate. The KFRE has been validated in multiple international cohorts [115–118], and performance has generally been good in each of these studies.

The original KFRE is a classic cross-sectional CPM that, during each stage of the model pipeline, uses only a single measurement of each input predictor. This does not, however, reflect the way that clinicians would assess risk within the clinic. Instead of considering only e.g. the most recent eGFR reading, the entire longitudinal history of eGFR, and more specifically its rate of decline, is commonly used to identify high risk patients. The original authors of the KFRE have developed a separate CPM that attempts to mimic this process via the inclusion of time-varying values of eGFR at the development stage of the model [119]. Their model does not, however, directly model the relationship between change in eGFR and outcome (ESRD).

Further studies have explored the idea of considering the entire longitudinal trajectory of eGFR to enable more accurate prognostic predictions. Brand et al. [120] consider and landmark-type (Cox) models with joint models for this purpose, and explicitly use the estimated eGFR slope in the Cox model for prediction of ESRD. They find that a simpler model that uses the most recently observed value of eGFR, in addition to the slope, is the best performing model, but all options had similar discriminative ability in the prediction of ESRD.

1.5.1 Informative Observation in CKD progression

As well considering how eGFR has progressed over time, monitoring frequencies of key kidney health markers are likely to be observed informatively over time, which becomes relevant when prognostic models for ESRD are developed using routinely collected health data. Patient monitoring is adapted according to the evolving needs of the individual, and episodes of poor kidney health may result in more intense periods of observation or admission to inpatient care. There is the potential to learn from the patients' perceived needs, as reflected by how closely care providers decide to observe markers such as eGFR. A key hypothesis to test is whether information contained within the monitoring frequency can improve predictive performance beyond using the longitudinally observed values of eGFR (and other key predictors) alone.

1.6 Thesis aims, objectives and outline

Previous sections have demonstrated that the handling of informative presence and informative observation in the context of prediction modelling research is under-studied, but could offer gains in predictive performance through careful incorporation into CPMs. This thesis studies this idea further by establishing to what extent IP and IO have been considered in the existing CPM literature (via a scoping review in Chapter 2), and additionally through methodological and applied research to demonstrate some of the ways in which CPMs may be able to learn from IP and IO.

Given existing gaps in the literature around the occurrence of missing data at multiple stages of the model pipeline, we look at the effect of including and excluding the outcome from the imputation model on a range of predictive performance metrics, and further extend this to the scenario where the complete data must be collected in order to apply the model in practice. We explore whether regression/conditional imputation could offer a more practical alternative to MI in the clinic as it does not require access to the original dataset or significant computational power. This is all studied in the context of informative presence, whereby we incorporate missing indicators (in combination with imputation techniques) as a means of mitigating potential bias (and affording potential gains in predictive performance) resulting from informatively missing data.

In Chapter 4 of this thesis, we use CKD progression as a clinical exemplar (described in the previous section) to illustrate existing and novel methods of incorporating informative observation in the development of clinical prediction models. Our analysis is based on a cohort of lab data collected in the Grampian region of Scotland, UK. The dataset consists of all biochemistry results collected in Grampian during the period 2009-2014, and has been linked to further NHS Scotland sources to collect information on hospital admissions, comorbidities and death.

The exact nature of the methods to be applied in this chapter will depend on the findings in Chapter 2 - a scoping review to establish the current state of methodological development in this area. We hypothesise that the inclusion of some representation of the observation process will result in gains in predictive performance of established models for the prediction of ESRD in CKD patients, and we aim to illustrate a range of potential representations of this process that could be adapted for use in other clinical areas.

More specifically, the aims of this thesis are threefold:
1) Identify the extent to which IP and IO have been considered in the context of methodological prediction modelling research, and review the key methods for doing this. 2) Evaluate whether the use of missing indicators in combination with single and multiple imputation techniques to develop and apply CPMs under informative presence improves performance. 3) Assess the added value of incorporating informative observation in a clinical prediction model, within the clinical exemplar of CKD progression.

This thesis is submitted in "journal format", and therefore each of the above aims is addressed in the form of a manuscript that has been, or will be, submitted to a journal for publication. The rationale for submitting an alternative format thesis is that this work embodies three separate (yet related) bodies of work, each of which will be (or have been) sumitted for publication in different journals.

1.7 Author Contributions

- Chapter 2: Informative presence and observation in routine health data: A review of methodology for clinical risk prediction. Journal of the American Medical Informatics Association, 2021, 28(1): 10.1093/jamia/ocaa242
  - RS, GPM, and NP conceptualised the study. RS and GPM designed the study. RS and LL conducted the screening. RS collated and reviewed included literature. RS wrote the initial draft of the manuscript. All other authors critically reviewed the content and writing of the manuscript.
- Chapter 3: Imputation and Missing Indicators for handling missing data in clinical prediction models: a simulation study. Under review with Statistical Methods in Medical Research.
  - RS, GPM, MS and NP conceptualised and designed the study. RS wrote the code for the simulation, conducted all analysis and wrote

the initial draft of the manuscript. All other authors interpreted the results and provided critical review and revisions of the manuscript.

- Chapter 4: Harnessing informative patterns of eGFR measurement for improved prediction of key outcomes in chronic kidney disease. In preparation for submission.
    - RS, GPM, MS and SS designed the study. RS conducted the analysis and interpreted findings with GPM, MS, NP and SS. RS wrote the initial draft of the manuscript, with important review and revisions by all other authors.

References

[1] E. Steyerberg. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Second. Springer, 2019. DOI: 10.1007/978-0-387-77244-8.

[2] M. van Smeden et al. "Clinical prediction models: diagnosis versus prognosis". In: *Journal of Clinical Epidemiology* 132 (Apr. 2021), pp. 142–145. DOI: 10.1016/J.JCLINEPI.2021.01.009.

[3] E. W. Steyerberg et al. "Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research". In: *PLOS Medicine* 10.2 (2013), e1001381. DOI: 10.1371/JOURNAL.PMED.1001381.

[4] N. Tangri et al. "A predictive model for progression of chronic kidney disease to kidney failure". In: *JAMA - Journal of the American Medical Association* 305.15 (Apr. 2011), pp. 1553–1559. DOI: 10.1001/jama.2011.451.

[5] J. E. Zimmerman et al. "Acute Physiology and Chronic Health Evaluation (APACHE) IV: Hospital mortality assessment for today's critically ill patients". In: *Critical Care Medicine* 34.5 (2006), pp. 1297–1310. DOI: 10.1097/01.CCM.0000215112.84523.F0.

[6] R. D. Riley et al. "Calculating the sample size required for developing a clinical prediction model". In: *BMJ* 368 (Mar. 2020). DOI: 10.1136/BMJ.M441.

[7] E. W. Steyerberg and Y. Vergouwe. “Towards better clinical prediction models: seven steps for development and an ABCD for validation.” In: *European heart journal* 35.29 (Aug. 2014), pp. 1925–31. DOI: 10.1093/eurheartj/ehu207.

[8] J. L. Vincent et al. “The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure”. In: *Intensive Care Medicine* 22.7 (1996), pp. 707–710. DOI: 10.1007/BF01709751.

[9] W.-H. Weng and W.-H. Weng. “Machine Learning for Clinical Predictive Analytics”. In: *Leveraging Data Science for Global Health* (2020), pp. 199–217. DOI: 10.1007/978-3-030-47994-7_12. arXiv: 1909.09246.

[10] P. Dhiman et al. “Reporting of prognostic clinical prediction models based on machine learning methods in oncology needs to be improved”. In: *Journal of Clinical Epidemiology* 138 (Oct. 2021), pp. 60–72. DOI: 10.1016/J.JCLINEPI.2021.06.024/ATTACHMENT/FB81B338-94B8-4599-8EAB-B43C2D237B35/MMC1.DOCX.

[11] J. H. Friedman, R. Tibshirani, and T. Hastie. *The Elements of Statistical Learning.* New York: Springer, 2001.

[12] F. E. J. Harrell. *Regression modeling strategies.* Cham, Switzerland: Springer International Publishing, 2016.

[13] D. R. Cox and D. Oakes. *Analysis of survival data.* London: Chapman and Hall, 1984.

[14] N. Breslow. “Covariance Analysis of Censored Survival Data”. In: *Biometrics* 30.1 (Mar. 1974), p. 89. DOI: 10.2307/2529620.

[15] G. Heinze and D. Dunkler. “Five myths about variable selection”. In: *Transplant international : official journal of the European Society for Organ Transplantation* 30.1 (Jan. 2017), pp. 6–10. DOI: 10.1111/TRI.12895.

[16] E. W. Steyerberg et al. “Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis”. In: *Journal of Clinical Epidemiology* 54.8 (Aug. 2001), pp. 774–781. DOI: 10.1016/S0895-4356(01)00341-9.

[17] F. E. Harrell, K. L. Lee, and D. B. Mark. “Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors”. In: *Statistics in Medicine* 15 (1996), pp. 361–387. DOI: 10.1002/(SICI)1097-0258(19960229)15:4.

[18] K. G. M. Moons et al. “Risk prediction models: II. External validation, model updating, and impact assessment”. In: *Heart* 98.9 (May 2012), pp. 691–698. DOI: 10.1136/heartjnl-2011-301247.

[19] F. E. Harrell et al. “Evaluating the Yield of Medical Tests”. In: *JAMA* 247.18 (May 1982), pp. 2543–2546. DOI: 10.1001/JAMA.1982.03320430047030.

[20] B. Van Calster et al. “A calibration hierarchy for risk models was defined: from utopia to empirical data”. In: *Journal of clinical epidemiology* 74 (June 2016), pp. 167–176. DOI: 10.1016/J.JCLINEPI.2015.12.005.

[21] B. Van Calster et al. “Calibration: The Achilles heel of predictive analytics”. In: *BMC Medicine* 17.1 (Dec. 2019), p. 230. DOI: 10.1186/s12916-019-1466-7.

[22] P. C. Austin and E. W. Steyerberg. “Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers”. In: *Statistics in Medicine* 33.3 (Feb. 2014), p. 517. DOI: 10.1002/SIM.5941.

[23] W. Bouwmeester et al. “Reporting and methods in clinical prediction research: A systematic review”. In: *PLoS Medicine* 9.5 (May 2012). DOI: 10.1371/JOURNAL.PMED.1001221.

[24] G. S. Collins et al. “External validation of multivariable prediction models: a systematic review of methodological conduct and reporting”. In: *BMC medical research methodology* 14.1 (Mar. 2014). DOI: 10.1186/1471-2288-14-40.

[25] G. L. Hickey et al. “Dynamic trends in cardiac surgery: why the logistic EuroSCORE is no longer suitable for contemporary cardiac surgery and implications for future risk models”. In: *European journal of cardio-thoracic surgery* 43.6 (June 2013), pp. 1146–1152. DOI: 10.1093/EJCTS/EZS584.

[26] S. E. Davis et al. "Calibration drift in regression and machine learning models for acute kidney injury". In: *Journal of the American Medical Informatics Association* 24.6 (Nov. 2017), pp. 1052–1061. DOI: 10.1093/JAMIA/OCX030.

[27] M. C. Lenert, M. E. Matheny, and C. G. Walsh. "Prognostic models will be victims of their own success, unless…" In: *Journal of the American Medical Informatics Association* 26.12 (Dec. 2019), pp. 1645–1650. DOI: 10.1093/JAMIA/OCZ145.

[28] M. Sperrin et al. "Explicit causal reasoning is needed to prevent prognostic models being victims of their own success". In: *Journal of the American Medical Informatics Association : JAMIA* 26.12 (Nov. 2019), p. 1675. DOI: 10.1093/JAMIA/OCZ197.

[29] D. A. Jenkins et al. "Dynamic models to predict health outcomes: current status and methodological challenges". In: *Diagnostic and Prognostic Research 2018 2:1* 2.1 (Dec. 2018), pp. 1–9. DOI: 10.1186/S41512-018-0045-2.

[30] D. A. Jenkins et al. "Continual updating and monitoring of clinical prediction models: time for dynamic prediction systems?" In: *Diagnostic and Prognostic Research 2021 5:1* 5.1 (Jan. 2021), pp. 1–7. DOI: 10.1186/S41512-020-00090-3.

[31] K. G. M. Moons et al. "Prognosis and prognostic research: application and impact of prognostic models in clinical practice." In: *BMJ (Clinical research ed.)* 338 (June 2009), b606. DOI: 10.1136/bmj.b606.

[32] E. W. Steyerberg et al. "Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research". In: *PLoS Medicine* 10.2 (Feb. 2013), e1001381. DOI: 10.1371/journal.pmed.1001381.

[33] B. M. Reilly and A. T. Evans. "Translating Clinical Research into Clinical Practice: Impact of Using Prediction Rules To Make Decisions". In: *Annals of Internal Medicine* 144.3 (Feb. 2006), p. 201. DOI: 10.7326/0003-4819-144-3-200602070-00009.

[34] J. Hippisley-Cox, C. Coupland, and P. Brindle. "Development and validation of QRISK3 risk prediction algorithms to estimate future risk of

cardiovascular disease: Prospective cohort study". In: *BMJ (Online)* 357 (May 2017). DOI: 10.1136/bmj.j2099.

[35] *Quality statement 1: Full formal risk assessment using QRISK2 — Cardiovascular risk assessment and lipid modification — Quality standards — NICE*. 2015.

[36] D. M. Lloyd-Jones. "Cardiovascular risk prediction: Basic concepts, current status, and future directions". In: *Circulation* 121.15 (Apr. 2010), pp. 1768–1777. DOI: 10.1161/CIRCULATIONAHA.109.849166.

[37] S. A. Nashef et al. "European system for cardiac operative risk evaluation (EuroSCORE)". In: *European Journal of Cardio-Thoracic Surgery* 16.1 (July 1999), pp. 9–13. DOI: 10.1016/S1010-7940(99)00134-7.

[38] P. W. Wilson et al. "Prediction of Coronary Heart Disease Using Risk Factor Categories". In: *Circulation* 97.18 (May 1998), pp. 1837–1847. DOI: 10.1161/01.CIR.97.18.1837.

[39] V. Sharma et al. "Adoption of clinical risk prediction tools is limited by a lack of integration with electronic health records". In: *BMJ Health Care Informatics* 28 (2021), p. 100253. DOI: 10.1136/bmjhci-2020-100253.

[40] *QRISK3*. 2017.

[41] T. Panch, H. Mattie, and L. A. Celi. "The "inconvenient truth" about AI in healthcare". In: *npj Digital Medicine* 2.1 (Aug. 2019), pp. 1–3. DOI: 10.1038/s41746-019-0155-4.

[42] M. Sendak et al. "A Path for Translation of Machine Learning Products into Healthcare Delivery". In: *EMJ Innovations* (2020). DOI: 10.33590/emjinnov/19-00172.

[43] M. J. Sweeting et al. "The use of repeated blood pressure measures for cardiovascular risk prediction: a comparison of statistical models in the ARIC study". In: *Statistics in Medicine* 36.28 (Dec. 2017), pp. 4514–4528. DOI: 10.1002/sim.7144.

[44] Ö. Asar et al. "Joint modelling of repeated measurement and time-to-event data: an introductory tutorial". In: *International Journal of Epidemiology* 44.1 (Feb. 2015), pp. 334–344. DOI: 10.1093/IJE/DYU262.

[45] L. M. Bull et al. "Harnessing repeated measurements of predictor variables for clinical risk prediction: a review of existing methods". In: *Diagnostic and Prognostic Research* 4.1 (Dec. 2020), p. 9. DOI: 10.1186/s41512-020-00078-z.

[46] A. Maier et al. "A gentle introduction to deep learning in medical image processing". In: *Zeitschrift für Medizinische Physik* 29.2 (May 2019), pp. 86–101. DOI: 10.1016/J.ZEMEDI.2018.12.003. arXiv: 1810.05401.

[47] I. Spasic and G. Nenadic. "Clinical Text Data in Machine Learning: Systematic Review". In: *JMIR Med Inform 2020;8(3):e17984 https://medinform.jmir.org/20* 8.3 (Mar. 2020), e17984. DOI: 10.2196/17984.

[48] E. Christodoulou et al. "A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models". In: *Journal of Clinical Epidemiology* 110 (June 2019), pp. 12–22. DOI: 10.1016/J.JCLINEPI.2019.02.004.

[49] T. Van Der Ploeg, P. C. Austin, and E. W. Steyerberg. "Modern modelling techniques are data hungry: A simulation study for predicting dichotomous endpoints". In: *BMC Medical Research Methodology* 14.1 (Dec. 2014), pp. 1–13. DOI: 10.1186/1471-2288-14-137/FIGURES/13.

[50] R. Little and D. Rubin. *Statistical Analysis with Missing Data, Second Edition.* Second Edi. John Wiley & Sons, Inc., 2002. DOI: 10.1002/9781119013563.

[51] S. Greenland and W. D. Finkle. "A Critical Look at Methods for Handling Missing Covariates in Epidemiologic Regression Analyses". In: *American Journal of Epidemiology* 142.12 (Dec. 1995), pp. 1255–1264. DOI: 10.1093/OXFORDJOURNALS.AJE.A117592.

[52] K. G. Moons et al. "Using the outcome for imputation of missing predictor values was preferred". In: *Journal of Clinical Epidemiology* 59.10 (Oct. 2006), pp. 1092–1101. DOI: 10.1016/J.JCLINEPI.2006.01.009.

[53] G. Molenberghs, C. Beunckens, and C. Sotto. "Every Missingness Not at Random Model Has a Missingness at Random Counterpart with Equal Fit". In: *Journal of the Royal Statistical Society* 70.2 (2008), pp. 371–388.

[54] S. van Buuren. *Flexible Imputation of Missing Data. Second Edition.* CRC/Chapman & Hall, FL: Boca Raton, 2018.

[55] S. Haneuse and M. Daniels. “A General Framework for Considering Selection Bias in EHR-Based Studies: What Data Are Observed and Why?” In: *EGEMS (Washington, DC)* 4.1 (2016), p. 1203. DOI: 10.13063/2327-9214.1203.

[56] A. Tsvetanova et al. “Missing data was handled inconsistently in UK prediction models: a review of method used”. In: *Journal of Clinical Epidemiology* 140 (Sept. 2021), pp. 149–158. DOI: 10.1016/J.JCLINEPI.2021.09.008.

[57] I. R. White, P. Royston, and A. M. Wood. “Multiple imputation using chained equations: Issues and guidance for practice”. In: *Statistics in Medicine* 30.4 (Feb. 2011), pp. 377–399. DOI: 10.1002/SIM.4067/FORMAT/PDF.

[58] J. Henry et al. *Adoption of Electronic Health Record Systems among U.S. Non-Federal Acute Care Hospitals: 2008-2015*. 2016.

[59] P. Coorevits et al. “Electronic health records: new opportunities for clinical research”. In: *Journal of Internal Medicine* 274.6 (Dec. 2013), pp. 547–560. DOI: 10.1111/joim.12119.

[60] J. F. Ludvigsson and H. O. Adami. “The urgency to embrace Big Data opportunities in medicine”. In: *Journal of Internal Medicine* 283.5 (May 2018), pp. 479–480. DOI: 10.1111/JOIM.12749.

[61] R. D. Riley et al. “External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges”. In: *BMJ* 353 (June 2016), p. i3140. DOI: 10.1136/bmj.i3140.

[62] B. A. Goldstein et al. “Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review”. In: *Journal of the American Medical Informatics Association* 24.1 (Jan. 2017), pp. 198–208. DOI: 10.1093/jamia/ocw042.

[63] T. Botsis et al. “Secondary Use of EHR: Data Quality Issues and Informatics Opportunities.” In: *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science* 2010 (Mar. 2010), pp. 1–5.

[64] S. C. Denaxas and K. I. Morley. “Big biomedical data and cardiovascular disease research: opportunities and challenges”. In: *European Heart Journal - Quality of Care and Clinical Outcomes* 1.1 (July 2015), pp. 9–16. DOI: 10.1093/EHJQCCO/QCV005.

[65] W. Raghupathi and V. Raghupathi. “Big data analytics in healthcare: promise and potential”. In: *Health Information Science and Systems* 2.1 (Dec. 2014). DOI: 10.1186/2047-2501-2-3.

[66] M. Fleming, B. Kirby, and K. I. Penny. “Record linkage in Scotland and its applications to health research”. In: *Journal of Clinical Nursing* 21.19 (Oct. 2012), pp. 2711–2721. DOI: 10.1111/J.1365-2702.2011.04021.X.

[67] F. S. Collins et al. “PCORnet: turning a dream into reality”. In: *Journal of the American Medical Informatics Association : JAMIA* 21.4 (2014), p. 576. DOI: 10.1136/AMIAJNL-2014-002864.

[68] M. R. Cowie et al. “Electronic health records to facilitate clinical research”. In: *Clinical Research in Cardiology* 106.1 (Jan. 2017), p. 1. DOI: 10.1007/S00392-016-1025-6.

[69] S. Haneuse, D. Arterburn, and M. J. Daniels. “Assessing Missing Data Assumptions in EHR-Based Studies: A Complex and Underappreciated Task”. In: *JAMA Network Open* 4.2 (Feb. 2021), e210184–e210184. DOI: 10.1001/JAMANETWORKOPEN.2021.0184.

[70] S. Shilo, H. Rossman, and E. Segal. “Axes of a revolution: challenges and promises of big data in healthcare”. In: *Nature Medicine 2020 26:1* 26.1 (Jan. 2020), pp. 29–38. DOI: 10.1038/s41591-019-0727-5.

[71] N. G. Weiskopf, A. Rusanov, and C. Weng. “Sick patients have more data: the non-random completeness of electronic health records.” In: *AMIA Annual Symposium proceedings*. Vol. 2013. American Medical Informatics Association, 2013, pp. 1472–7.

[72] E. M. Pullenayegum and L. S. Lim. “Longitudinal data subject to irregular observation: A review of methods with a focus on visit processes, assumptions, and study design”. In: *Statistical Methods in Medical Research* 25.6 (2014). DOI: 10.1177/0962280214536537.

[73] B. A. Goldstein et al. “Controlling for Informed Presence Bias Due to the Number of Health Encounters in an Electronic Health Record”. In: *American Journal of Epidemiology* 184.11 (Dec. 2016), pp. 847–855. DOI: 10.1093/aje/kww112.

[74] A. Sharafoddini et al. “A new insight into missing data in intensive care unit patient profiles: Observational study”. In: *Journal of Medical Internet Research* 21.1 (Jan. 2019). DOI: 10.2196/11605.

[75] National Institute for Health and Care Excellence (NICE). *Type 2 diabetes in adults: management.* Dec. 2015.

[76] G. Moscelli et al. “Socioeconomic inequality of access to healthcare: Does choice explain the gradient?” In: *Journal of Health Economics* 57 (Jan. 2018), pp. 290–314. DOI: 10.1016/J.JHEALECO.2017.06.005.

[77] P. Barlow et al. “Area-level deprivation and geographic factors influencing utilisation of General Practitioner services”. In: *SSM - Population Health* 15 (Sept. 2021), p. 100870. DOI: 10.1016/J.SSMPH.2021.100870.

[78] J. M. Neuhaus, C. E. McCulloch, and R. D. Boylan. “Analysis of longitudinal data from outcome-dependent visit processes: Failure of proposed methods in realistic settings and potential improvements”. In: *Statistics in Medicine* 37.29 (Dec. 2018), pp. 4457–4471. DOI: 10.1002/sim.7932.

[79] J. Gruger, R. Kay, and M. Schumacher. “The Validity of Inferences Based on Incomplete Observations in Disease State Models”. In: *Biometrics* 47.2 (June 1991), p. 595. DOI: 10.2307/2532149.

[80] A. Gasparini et al. “Mixed-effects models for health care longitudinal data with an informative visiting process: A Monte Carlo simulation study”. In: *Statistica Neerlandica* (2019). DOI: 10.1111/stan.12188. arXiv: 1808.00419.

[81] M. Phelan, N. A. Bhavsar, and B. A. Goldstein. “Illustrating Informed Presence Bias in Electronic Health Records Data: How Patient Interactions with a Health System Can Impact Inference”. In: *EGEMS (Washington, DC)* 5.1 (2017). DOI: 10.5334/egems.243.

[82] J. L. Schafer and J. W. Graham. “Missing data: Our view of the state of the art”. In: *Psychological Methods* 7.2 (2002), pp. 147–177. DOI: 10.1037/1082-989X.7.2.147.

[83] A. R. T. Donders et al. “Review: A gentle introduction to imputation of missing values”. In: *Journal of Clinical Epidemiology* 59.10 (Oct. 2006), pp. 1087–1091. DOI: `10.1016/J.JCLINEPI.2006.01.014`.

[84] J. Carpenter and M. Kenward. *Missing data in randomised controlled trials–a practical guide.* 2007.

[85] R. Little and D. Rubin. *Statistical analysis with missing data.* New York: John Wiley and Sons, 1987.

[86] J. A. C. Sterne et al. “Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls”. In: *BMJ (Clinical research ed.)* 338 (June 2009), b2393. DOI: `10.1136/bmj.b2393`.

[87] J. C. Jakobsen et al. “When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts”. In: *BMC Medical Research Methodology* 17.1 (Dec. 2017), p. 162. DOI: `10.1186/s12874-017-0442-1`.

[88] A. Marshall et al. “Comparison of techniques for handling missing covariate data within prognostic modelling studies: A simulation study”. In: *BMC Medical Research Methodology* 10 (2010). DOI: `10.1186/1471-2288-10-7`.

[89] M. Sperrin et al. *Missing data should be handled differently for prediction than for description or causal explanation.* Sept. 2020. DOI: `10.1016/j.jclinepi.2020.03.028`.

[90] M. Sperrin and G. P. Martin. “Multiple imputation with missing indicators as proxies for unmeasured variables: Simulation study”. In: *BMC Medical Research Methodology* 20.1 (July 2020), p. 185. DOI: `10.1186/s12874-020-01068-x`.

[91] S. Fletcher Mercaldo and J. D. Blume. “Missing data and prediction: the pattern submodel”. In: *Biostatistics* (Sept. 2018). DOI: `10.1093/biostatistics/kxy040`.

[92] R. H. Groenwold et al. “Missing covariate data in clinical research: When and when not to use the missing-indicator method for analysis”. In: *CMAJ* 184.11 (Aug. 2012), pp. 1265–1269. DOI: `10.1503/cmaj.110977`.

[93] R. H. H. Groenwold. “Informative missingness in electronic health record systems: the curse of knowing”. In: *Diagnostic and Prognostic Research* 4.1 (Dec. 2020), p. 8. DOI: 10.1186/s41512-020-00077-0.

[94] E. W. Steyerberg and Y. Vergouwe. *Towards better clinical prediction models: Seven steps for development and an ABCD for validation.* Aug. 2014. DOI: 10.1093/eurheartj/ehu207.

[95] J. Hoogland et al. “Handling missing predictor values when validating and applying a prediction model to new patients”. In: *Statistics in Medicine* 39.25 (Nov. 2020), pp. 3591–3607. DOI: 10.1002/sim.8682.

[96] S. W. J. Nijman et al. “Real-time imputation of missing predictor values improved the application of prediction models in daily practice”. In: *Journal of Clinical Epidemiology* 134 (June 2021), pp. 22–34. DOI: 10.1016/j.jclinepi.2021.01.003.

[97] M. van Smeden, R. H. Groenwold, and K. G. Moons. “A cautionary note on the use of the missing indicator method for handling missing data in prediction research”. In: *Journal of Clinical Epidemiology* 125 (Sept. 2020), pp. 188–190. DOI: 10.1016/j.jclinepi.2020.06.007.

[98] J.-H. Lin and P. J. Haug. “Exploiting missing clinical data in Bayesian network modeling for predicting medical problems”. In: *Journal of Biomedical Informatics* 41.1 (Feb. 2008), pp. 1–14. DOI: 10.1016/J.JBI.2007.06.001.

[99] D. Rizopoulos. *Joint models for longitudinal and time-to-event data : with applications in R.* Chapman and Hall/CRC, July 2012.

[100] G. L. Hickey et al. “Joint modelling of time-to-event and multivariate longitudinal outcomes: Recent developments and issues”. In: *BMC Medical Research Methodology* 16.1 (2016), pp. 1–15. DOI: 10.1186/s12874-016-0212-5.

[101] J. G. Ibrahim, H. Chu, and L. M. Chen. “Basic concepts and methods for joint models of longitudinal and survival data”. In: *Journal of clinical oncology* 28.16 (June 2010), pp. 2796–2801. DOI: 10.1200/JCO.2009.25.0654.

[102] R. Henderson, P. Diggle, and A. Dobson. “Joint modelling of longitudinal measurements and event time data”. In: *Biostatistics* 1.4 (Dec. 2000), pp. 465–480. DOI: 10.1093/BIOSTATISTICS/1.4.465.

[103] A. Tsiatis and M. Davidian. “Joint modeling of longitudinal and time-to-event data: an overview”. In: *Statistica Sinica* 14.3 (2004), pp. 809–834.

[104] H. C. van Houwelingen. “Dynamic Prediction by Landmarking in Event History Analysis”. In: *Scandinavian Journal of Statistics* 34.1 (Mar. 2007), pp. 70–85. DOI: 10.1111/J.1467-9469.2006.00529.X.

[105] J. C. van Houwelingen and H. Putter. *Dynamic prediction in clinical survival analysis*. CRC Press, 2012.

[106] L. Liu, X. Huang, and J. O’Quigley. “Analysis of longitudinal data in the presence of informative observational times and a dependent terminal event, with application to medical cost data”. In: *Biometrics* 64.3 (Sept. 2008), pp. 950–958. DOI: 10.1111/j.1541-0420.2007.00954.x.

[107] L. Qu, L. Sun, and X. Song. “A Joint Modeling Approach for Longitudinal Data with Informative Observation Times and a Terminal Event”. In: *Statistics in Biosciences* 10.3 (Dec. 2018), pp. 609–633. DOI: 10.1007/s12561-018-9221-8.

[108] A. M. Alaa, S. Hu, and M. Schaar. “Learning from Clinical Judgments: Semi-Markov-Modulated Marked Hawkes Processes for Risk Prognosis”. In: *International Conference on Machine Learning (ICML)*. International Conference on Machine Learning (ICML), July 2017, pp. 60–69.

[109] D. Agniel, I. S. Kohane, and G. M. Weber. “Biases in electronic health record data due to processes within the healthcare system: retrospective observational study”. In: *BMJ* 361 (Apr. 2018), k1479. DOI: 10.1136/BMJ.K1479.

[110] A. S. Levey et al. “Definition and classification of chronic kidney disease: A position statement from Kidney Disease: Improving Global Outcomes (KDIGO)”. In: *Kidney International* 67.6 (June 2005), pp. 2089–2100. DOI: 10.1111/J.1523-1755.2005.00365.X.

[111] A. K. Cheung et al. "KDIGO 2021 Clinical Practice Guideline for the Management of Blood Pressure in Chronic Kidney Disease". In: *Kidney International* 99.3 (Mar. 2021), S1–S87. DOI: 10.1016/J.KINT.2020.11.003/ATTACHMENT/F431878B-95B5-4839-B871-E9587652AE0A/MMC1.PDF.

[112] B. Bikbov et al. "Global, regional, and national burden of chronic kidney disease, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017". In: *The Lancet* 395.10225 (Feb. 2020), pp. 709–733. DOI: 10.1016/S0140-6736(20)30045-3/ATTACHMENT/234A6931-5886-48B10.1016/S0140-6736(20)30045-3.

[113] J. C. Lv and L. X. Zhang. "Prevalence and Disease Burden of Chronic Kidney Disease". In: *Advances in Experimental Medicine and Biology* 1165 (2019), pp. 3–15. DOI: 10.1007/978-981-13-8871-2_1.

[114] S. Liabeuf et al. "Guideline attainment and morbidity/mortality rates in a large cohort of European haemodialysis patients (EURODOPPS)". In: *Nephrology Dialysis Transplantation* 34.12 (Dec. 2019), pp. 2105–2110. DOI: 10.1093/NDT/GFZ049.

[115] M. W. Kang et al. "An independent validation of the kidney failure risk equation in an Asian population". In: *Scientific Reports 2020 10:1* 10.1 (July 2020), pp. 1–10. DOI: 10.1038/s41598-020-69715-3.

[116] R. H. Whitlock et al. "Validation of the Kidney Failure Risk Equation in Manitoba." In: *Canadian journal of kidney health and disease* 4 (Apr. 2017), p. 2054358117705372. DOI: 10.1177/2054358117705372.

[117] R. W. Major et al. "The Kidney Failure Risk Equation for prediction of end stage renal disease in UK primary care: An external validation and clinical impact projection cohort study". In: *PLOS Medicine* 16.11 (2019), e1002955. DOI: 10.1371/JOURNAL.PMED.1002955.

[118] C. S. Lennartz et al. "External Validation of the Kidney Failure Risk Equation and Re-Calibration with Addition of Ultrasound Parameters". In: *Clinical Journal of the American Society of Nephrology* 11.4 (Apr. 2016), pp. 609–615. DOI: 10.2215/CJN.08110715.

[119] N. Tangri et al. "A Dynamic Predictive Model for Progression of CKD". In: *American Journal of Kidney Diseases* 69.4 (Apr. 2017), pp. 514–520. DOI: 10.1053/j.ajkd.2016.07.030.

[120] J. A. J. G. van den Brand et al. “Predicting kidney failure from longitudinal kidney function trajectory: A comparison of models”. In: *PLOS ONE* 14.5 (May 2019). Ed. by P. Bjornstad, e0216559. DOI: `10.1371/journal.pone.0216559`.

Chapter 2

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

The work in this chapter follows on from the general introduction and literature review, by providing a systematic and comprehensive review of the state of the existing literature around informative presence/observation and clinical prediction modelling. The findings of this work form the basis of the plan for the remainder of this thesis, as we identify areas that require further study and development as well as methodological frameworks that could be used in our clinical exemplar.

The protocol for this review was registered with the Open Science Framework (OSF) prior to commencing the search process (`https://osf.io/rtqsg/`). The protocol and search strategy are provided at the end of the chapter as supplementary materials, as well as a record of any changes made to the search strategy after it began. The protocol version presented in the supplementary materials of this chapter has been edited slightly (from the registered version) for readability. The main text of this chapter is published in the Journal of the American Informatics Association (JAMIA) in its current form. The supplementary materials of the published work (available online with the open-access version of the manuscript) contains elements of the supplement presented here, such as further details of the search strategy and the paper-level summary table. These methodological details of the search strategy are important to report, but were omitted from the main text of the published article due to restrictions on word count imposed by the journal.

The citation for the published version of this article is as follows (with initials added as referenced in text): Rose Sisk (RS), Lijing Lin (LL), Matthew Sperrin (MS), Jessica K Barrett (JKB), Brian Tom (BT), Karla Diaz-Ordaz (KDO), Niels Peek (NP), Glen P Martin (GPM), Informative presence and observation in routine health data: A review of methodology for clinical risk prediction, Journal of the American Medical Informatics Association, Volume 28, Issue 1, January 2021, Pages 155–166, `https://doi.org/10.1093/jamia/ocaa242`

2.1 Abstract

2.1.1 Objectives

Informative presence (IP) is the phenomenon whereby the presence/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, where 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 aimed at handling these issues. This review aims to synthesise such existing methodology for applied researchers, thereby helping identify an agenda for future methodological work.

2.1.2 Methods

A systematic literature search was conducted by two independent reviewers using pre-specified keywords.

2.1.3 Results

Thirty-six papers were included. We categorised the methods presented within as: derived predictors (including some representation of the measurement process as a predictor in the model); modelling under IP; and latent structures. Including missing indicators/summary measures as predictors is the most commonly

presented approach amongst the included studies (24/36 papers).

2.1.4 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 modellers to use these methods.

2.1.5 Conclusion

A growing recognition of IP and IO exists within the literature, and methodology is increasingly becoming available to leverage these phenomena for predictive gain. 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/IO, and therefore we recommend that applied health researchers consider incorporating these methods in their work.

2.2 Background & 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, amongst other uses.

Widespread adoption of electronic health records (EHRs) facilitates the development of CPMs [1, 2], since 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 pre-specified as they would be in a typical research study (e.g. in a prospective cohort study with scheduled follow-up visits). Instead, data are collected opportunistically, where patient/clinician decisions directly dictate whether we observe clinical biomarkers and patient information [3]. For example, GP visits occur more frequently during periods of ill health [4], 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 2.1).

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 two key properties that an observation process may have, when presence of data is informative:

1. Informative presence (IP) (Figure 2.1 a) – the presence or absence of a patient's data at any given time point carries information about their health status that is not available elsewhere in the observed data.
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 that is not available elsewhere in the observed data.. See Figure 2.1 b for an example.

Informative presence is challenging from a statistical perspective as it implies a missing not at random (MNAR) 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. Informative presence has previously been defined elsewhere by Goldstein et al. [6], Phelan, Bhavsar, and Goldstein [7], with Phelan, Bhavsar, and Goldstein [7] discussing how interactions contained within electronic health records are informative with respect to patient health.

Informative observation is the continuous time generalisation of informative presence: a longitudinal Visiting (at time t) Not at Random (VNAR) process, defined as "given data recorded up to time t, visiting at time t is not independent of outcome at time t". [8] By generalising the definition of informative presence 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/IO [1]. Moreover, much of the existing methodological literature in this area has focussed on IP/IO only in the context of effect estimation (i.e. in causal or associational studies)

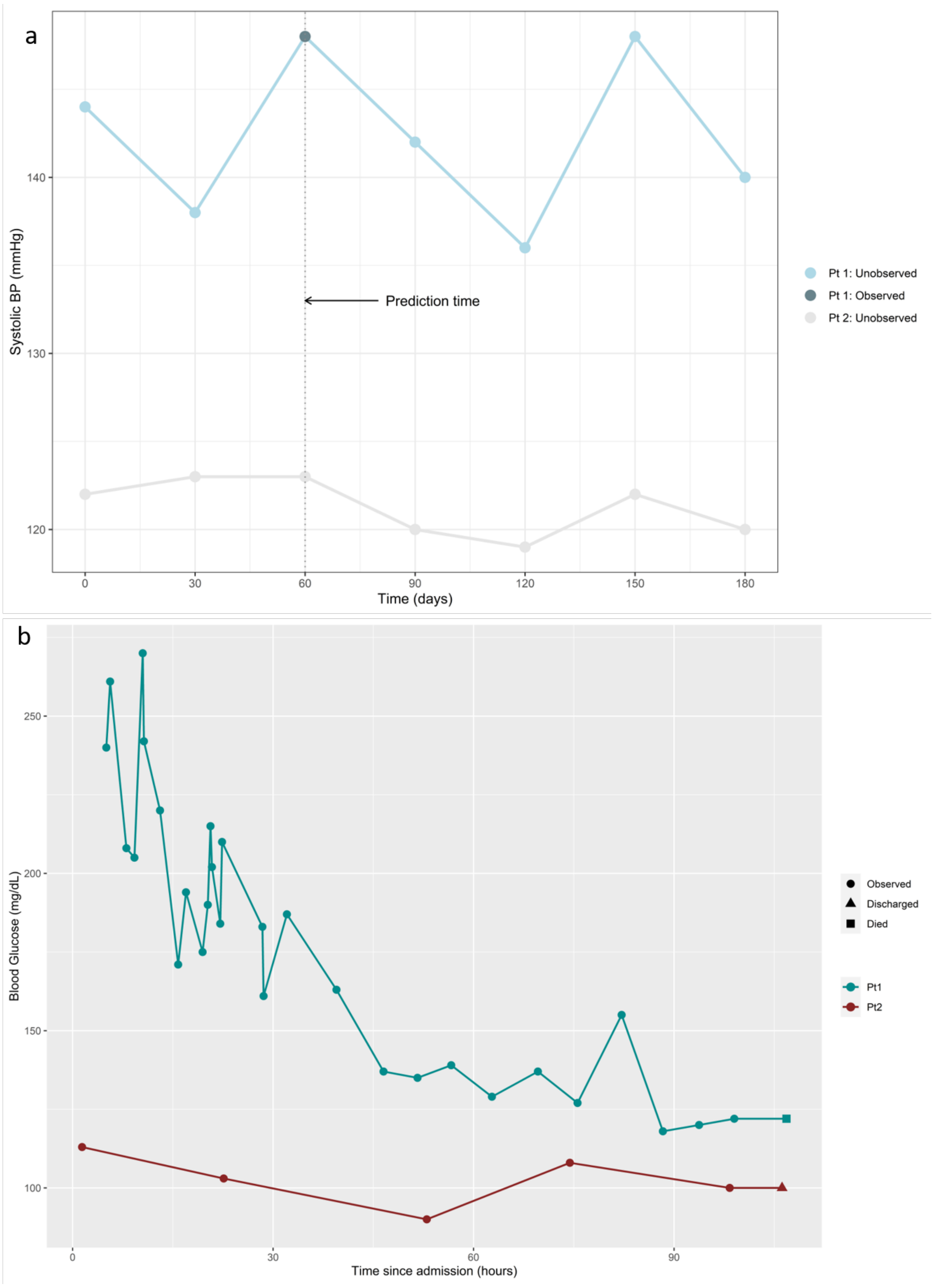


Figure 2.1: (A) An illustration of informative presence and how this could impact the information available at prediction time. We see the longitudinal pattern of blood pressure for 2 patients over time (since registration with their GP practice), with their observed and unobserved values shown. Patient 1 has a single observed value of systolic blood pressure (BP), measured when their BP was at its highest. Patient 2 has no observed values, but their BP remains in the normal range. (B) An illustration of informative observation, taken from the MIMIC (Medical Information Mart for Intensive Care) dataset.[5] 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. Pt2 was discharged safely, Pt1 died at the end of their admission.

[9–15], and has generally viewed it as a “nuisance” – i.e. 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 paper we focus on informative measurement patterns in the predictors, and we do not discuss presence or absence of outcome data. Agniel, Kohane, and Weber [16] demonstrated how the timing of a lab test better predicts mortality than the actual result of the test. Others have illustrated how incorporation of the presence or absence of a particular test for an individual into a CPM can be harnessed for predictive gain [17–19].

2.3 Objectives

This article aims to review the literature on methodology allowing CPMs to utilise 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 prioritised. Finally, we summarise existing software packages capable of implementing the methodology.

2.4 Materials & Methods

The strategy employed in this review loosely follows a scoping review framework [20]. Our protocol has been registered on the Open Science Framework [21]. Full details on the methods employed and search strategy can be found in the supplementary materials of this chapter (Section 2.8).

2.4.1 Search strategy

We searched MEDLINE, Embase and Web of Science for relevant articles using pre-specified search terms. Further details of the full search strategy (including search terms and an additional snowballing stage) can be found in the Supplementary materials (Section 2.8) and the published protocol [21].

2.4.2 Study selection

We had the following inclusion criteria: any paper presenting a method that allows CPMs to incorporate IP or IO. We excluded: papers that applied existing methods that had already been published elsewhere, and included those earlier publications instead, non-medical areas of application (or motivating examples), IP/IO in outcome measures, and methods that handle sample selection bias, imputation or censoring only. See the Supplementary material 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 (RS & LL) conducted a two-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/exclusion criteria, in consultation with a third reviewer (GPM).

Primarily, we extracted information regarding the modelling 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.

2.5 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 papers through forward and backward citation searching, giving a final set of 36 included papers (Figure 2.2). All of the validation papers (listed in the supplementary materials) were discovered by the search strategy. Throughout this section, we will illustrate each method with the notation described in this section. Note, however, that the literature search was not restricted to binary outcomes and methods that can model time-to-event and continuous outcomes are also included.

Consider a binary outcome $Y(t)$ (or Y if only observed once) for patients $i = 1, \ldots, n$, at time t, where $Y(t) = 1$ denotes that the event occurred, with marginal

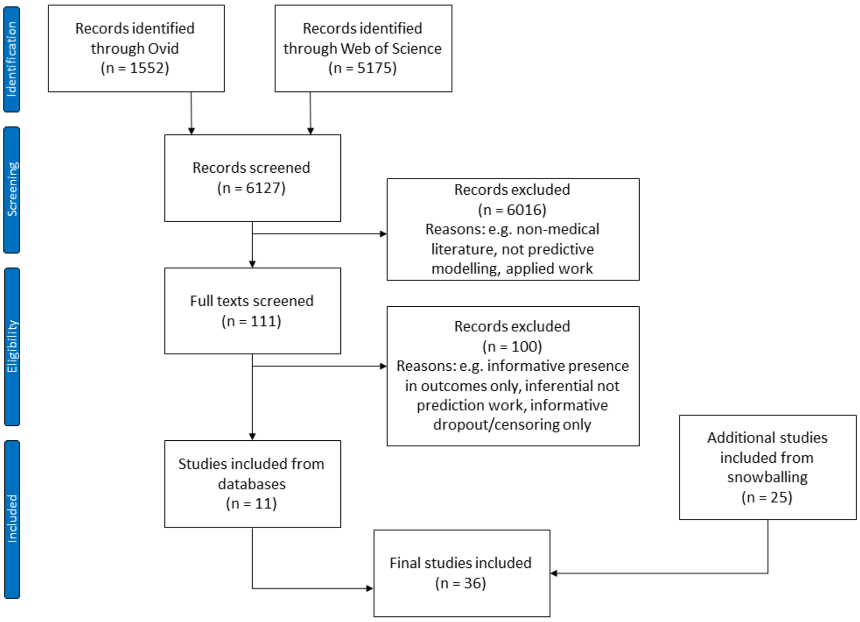

Identification
Screening
Eligibility
Included
Records identified through Ovid (n = 1552)
Records identified through Web of Science (n = 5175)
Records screened (n = 6127)
Records excluded (n = 6016) Reasons: e.g. non-medical literature, not predictive modelling, applied work
Full texts screened (n = 111)
Records excluded (n = 100) Reasons: e.g. informative presence in outcomes only, inferential not prediction work, informative dropout/censoring only
Studies included from databases (n = 11)
Additional studies included from snowballing (n = 25)
Final studies included (n = 36)

Figure 2.2: PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram showing the various screening stages and reasons for ex- clusion at each stage

probability $P[Y=1]$. Define a potentially time-varying continuous covariate process $X(t)$, with potential realisations 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 jth realisation of $X(t)$ is $t_{ij} \in R^+$, where R^+ denotes a real-valued positive number. Denote $R=1$ if $X(t)$ is ever observed at any time t, and $R=0$ if not. Define $r_{ij}=1$ if the covariate process is observed (more specifically) at time t_{ij}. We assume that Z is a completely observed time-invariant covariate. $g(.)$ represents a link function, e.g. the logit function.

Broadly, the methods in this paper cover the three scenarios described in Table 2.1. To illustrate the prediction scenarios and methods, we consider a simplified version of the Sequential Organ Failure Assessment (SOFA) score, [22] used to predict mortality in critical care, assuming that the only predictors in the model are bilirubin and blood pressure. Of these two predictors, we assume that blood pressure is completely observed for all patients, and bilirubin is informatively observed. Depending on the specific scenario, it may be a one-time point observation, or a longitudinal process [18].

There exists a breadth of methodological literature covering Scenario S2 (without accounting for IP/IO), which has recently been synthesised by Bull et al. [23]

Table 2.1: A description of different prediction scenarios, covering cross-sectional vs longitudinal predictors and outcomes.

Scenario	Scenario Name	Description	Example (SOFA)
S1	Cross-sectional prediction	Interest lies in obtaining a single prognostic estimate (prediction) using a single value for each predictor.	Use single values of bilirubin and Blood pressure (BP) obtained upon ICU admission to predict in-hospital survival (binary).
S2	Cross-sectional prediction with longitudinal predictor measurements	Interest lies in obtaining a single prognostic estimate but using the longitudinal history of predictor values.	Use all repeated lab tests obtained throughout inpatient admission (prior to ICU admission onset) for bilirubin and BP to predict in-hospital survival at the time of ICU admissioin.
S3	Longitudinal prediction with longitudinal predictors and outcomes	Interested in updating prognostic estimates at multiple prediction time points, potentially using the (updated) longitudinal history of predictor values.	Use all repeated measures of BP and bilirubin obtained throughout inpatient and ICU admission to predict survival, and update these predictors as new predictor values are observed throughout the admission.

We therefore focus on modelling strategies that have specifically been proposed or extended to accommodate IP or IO.

2.5.1 Identified Approaches to Handle Informative Presence and Observation

We identified three broad categories of method based on the included papers: (i) methods that incorporate IP/IO through derived predictors; (ii) methods for modelling under informative presence; and (iii) methods that incorporate IP/IO using latent structures. Within these three categories, we identified eight modelling strategies. A summary of the methods can be found in Table 2.2. Table 2.3 summarises the advantages, disadvantages, software, and assumptions for each method – here, the reported advantages and disadvantages were inferred by the research team since they are not consistently mentioned in the included literature. A summary table at paper-level can be found in the supplementary materials (Table 2.9).

Table 2.2: Descriptive summary table of methods, detailing when each method may be appropriate and how it would work with the running example of a simplified SOFA score.

Modelling approach	Broad category	Refs	Scenario(s)	IP or IO	Description	Example
Missing indicators and Separate class	Derived predictors	[17, 18, 24–30]	S1	IP	Creating a binary indicator, representing presence or absence of a predictor at a given time point or in a given window	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 at model development, and use information on the recording of bilirubin at the point of prediction.

Summary measures	Derived predictors	[16, 24, 31–44]	S2	IO	Summarising the observation process into a single variable, e.g. counting visits, rates of visits over a window, weighted counts	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 (at development), and derive this count for use as a predictor at the point of prediction.
Pattern-specific models	Modelling under informed presence	[45, 46]	S1	IP	Derive separate models for each missingness pattern	Develop models for: bilirubin and BP observed, and only BP observed. Predict using the appropriate model depending on the set of information available at the point of prediction.
Likelihood-based methods	Modelling under informed presence	[47, 48]	S1	IP	Incorporating missingness mechanism into maximum-likelihood estimation of parameter estimates	Bilirubin is missing not at random. Estimate model parameters using method-of-weights and EM algorithm at development. Model is applied as usual at the point of prediction.

Similarity measures	Derived predictors	[49]	S2	IO	Calculate similarity between target patient and all others, based on predictor values and measurement timings. Develop models separately for ”similar” groups of patients.	Develop separate models amongst cohorts of patients with similar bilirubin, BP and timings of those measures. For new patients, use a similarity metric that can identify the most appropriate model to apply to the out-of-sample patient, based on observed information prior to the point of prediction.
Latent variable	Latent structures	[50, 51]	S1, S3	IP	Outcome can be partially latent, and the observation process infers the latent state.	The occurrence of a bilirubin measurement is used to infer patient state in a hierarchical model at development, and this information is again used at deployment similarly to a ”missing indicator”.

HMMs	Latent structures	[52, 53]	S3	IO	Outcome is a partially latent process, and the observation process infers the state at any time.	The intensity of bilirubin measurements over the course of a patient's admission infers their severity at any time point. At deployment, information on past intensity of measurement is used to estimate risk.

Joint modelling and shared random effects	Latent structures	[54–56]	S2, S3	IP and IO	Model the outcome, predictor and observation processes separately, but join them through random effects shared across the models.	At model development, 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 one shared random effect across the models. Random effects can be estimated for new individuals using the Best Linear Unbiased Predictor (BLUP) assuming some predictor information has already been observed prior to the point of prediction.

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 generalisability 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/Separate Class The missing indicator approach [17, 18, 24–30] 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. 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.

Since 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). Most commonly (in both prediction modelling and descriptive or causal work), the missing predictor is set to 0 [26], but others have also explored the use of more sophisticated imputation techniques to handle predictor missingness, such as hot deck imputation [18] and multiple imputation [46, 59]. The missing indicator approach results in a model of the form:

$$g(P[Y=1|X,Z,R]) = \beta_0 + \beta_1 X + \beta_2 Z + \gamma R \tag{2.1}$$

for continuous predictors within cross-sectional prediction (S1).

Similarly, for a categorical predictor x_i with k categories, then the missing indicator approach would set $x_i \in \{Cat_1, ..., Cat_k, Missing\}$ and our model would be

$$g(P[Y=1|X,Z]) = \beta_0 + \beta_1 X + \beta_2 Z \tag{2.2}$$

The above two 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 [28–30].

Including a missing indicator or separate class is straightforward and has demonstrated improved predictive performance over models omitting them [18]. 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.[16, 24, 31–44] Examples include a count of the number of measurements (of e.g. throughout a critical care admission), [37] weighted counts,[42] combined missing indicators, [31] missingness rates over time, [32] time intervals between measures,[33–35], embedding vectors that represent missing values,[36] or information relating to hospital processes.[38, 39]

In some cases, combined missing indicators and time intervals also alter the relationship between a predictor and outcome. Che et al. [24] 's 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 towards a mean value.

Piecewise-Constant Intensity Models (PCIMs) have also been proposed to handle informatively observed predictors [40, 41]. PCIMs 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, e.g. a count of the number of times $X(t)$ (whether continuous or categorical) has been observed: $Q = m_i$. For cross-sectional prediction with a time-varying covariate, we then have:

$$g(P[Y=1|X,Z,Q]) = \beta_0 + \beta_1 X + \beta_2 Z + \gamma Q \tag{2.3}$$

where X is a summary of $X(t)$ deemed to have predictive value, e.g. 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 generalisability where measurement processes vary across locations [18, 39]. Simple summary measures such as counts may also fail to capture the complex relationship between the observation process and outcome.

Category 2: Modelling under informative presence

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

Pattern-specific models The pattern-specific approach [45, 46] derives separate models for each missingness pattern, generalising 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:

$$g(P[Y=1 \mid R=1, X, Z]) = \beta_{0,1} + \beta_{1,1}X + \beta_{2,1}Z \tag{2.4}$$

$$g(P[Y=1 \mid R=0, X, Z]) = \beta_{0,2} + \beta_{2,2}Z \tag{2.5}$$

Where Z is completely observed. Note that Equations (2.4) and (2.5) could also be combined by including interaction terms with the missing indicator, illustrating how this approach extends the missing indicator method.

Similar submodels could be derived for categorical and continuous predictors. Saar-Tsechansky [45] propose using all available data to train each submodel, whereas Fletcher Mercaldo and Blume [46] recommend that only individuals in each observed pattern be used in the derivation of that pattern's submodel (also illustrated by Janssen et al. [60]). The latter approach does not require knowledge of the missingness mechanism.

The pattern-specific approach is flexible, as it can be applied to any form of prediction algorithm. However, a practical limitation is that the number of candidate submodels becomes intractable as the number of predictors increases.

Likelihood-based methods A different approach assumes that missingness in the predictors is non-ignorable, and incorporates this into parameter estimates via likelihood-based methods[47, 48]. The model formulation would take, e.g. the same form as Equation (2.1), with parameter estimates obtained according to estimation procedures detailed in the following examples. Escarela, Ruiz-de-Chavez, and Castillo-Morales [47] assume a bivariate copula-based probability function for the missing covariates and the missingness mechanism. Kirkham [48] instead applies the "method of weights", which assumes a parametric model for the missingness mechanism and incorporates this into the maximum likelihood estimation of parameter estimates.

Escarela, Ruiz-de-Chavez, and Castillo-Morales [47] describe how their MNAR model can also be used to impute missing values. However, this does not remove the need to make untestable assumptions on the missing data mechanism.

Category 3: Latent structures

Similarity measures Patient similarity measures apply a sequencing algorithm to establish the alignment of two sequences of patient data, e.g. longitudinal EHR data. Sha [49] presents a novel similarity measure, which recognises 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 Sharafoddini, Dubin, and Lee [61]). 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, where 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 rederiving multiple models, and requiring access to the training data at prediction time to train a model using similar patients.

Latent variables A simple way of representing a latent clinical condition is to use a single (partially) latent binary variable, representing (e.g.) one of two states. This approach was used by Coley et al. [50] and Hubbard et al. [51], where 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 one of two patient states, where only one state is entirely observed. In Coley's [50] 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_i within the hierarchical model.

$$R|Y^L, Z \sim Bern(P[R=1|Y^L, Z, \beta]) \tag{2.6}$$

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 where the measurement process influences predictions compared to a model that ignores IP/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. [52] and Alaa, Hu, and Schaar [53] use HMMs to capture IO, but the way they incorporate the observation process differs. HMM-based prediction models incorporate IO by allowing the measurement frequency or rate to infer the clinical state at any given time.

Alaa, Hu, and Schaar [53] propose 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, i.e. $P[Y(\infty) = 4]$. The observation process (i.e. timings) of $X(t)$ is used to infer this clinical state, where 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. Informative observation is captured through state-specific intensity functions for the monitoring frequency as follows:

$$\lambda(t|Y^L(t) = 1) = \lambda_1 + \alpha_1 \sum_{\tau < t_m < t} e^{-\beta_1(t-t_m)} \tag{2.7}$$

...

$$\lambda(t|Y^L(t) = 4) = \lambda_4 + \alpha_4 \sum_{\tau < t_m < t} e^{-\beta_4(t-t_m)} \tag{2.8}$$

$\lambda_1, ..., \lambda_4$, $\alpha_1, ..., \alpha_4$ and $\beta_1, ..., \beta_4$ are state-specific parameters to be estimated. t_m is the time of the last measure of $X(t)$. τ 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 paper.

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 since there are multiple components to be estimated simultaneously.

Joint modelling Joint modelling has been developed extensively within the prediction context, particularly for dynamic prediction, i.e. incorporating time-updated variables (S2, Table 2.1) [62–64]. Joint modelling can be extended to handle IP and IO, by linking the outcome to the observation process via a shared random effect [54, 55], which can be seen as an alternative approach to modelling 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 two models, the outcome and observation processes are linked. Liang, Li, and Zhang [54] and Choi et al. [55] 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. Since this effect also appears in the model for the outcome, the visit rate indirectly affects the prediction for the outcome.

The method outlined by Zhang, Chen, and Zou [56] 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, where 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_ij. Assuming that measurement times are regular, i.e. $t_{ij} = t_j \forall\ i,j$.

$$X = \alpha_0 + \alpha_1 Z + \alpha_2 t + U \tag{2.9}$$

$$g(P[Y=1|Z,U,V]) = \beta_1 Z + \beta_2 U + \beta_3 V \tag{2.10}$$

$$(P[R_j = 1|U,V,Z]) = \delta_0 U + V + \delta_1 R_{j-1} + \delta_2 Z \tag{2.11}$$

Here U and V are independent subject-specific random effects, and $g(.)$ and $h(.)$ are link functions. β_2 and δ_1 control the relationships between the longitudinal predictor and the outcome, and the longitudinal predictor and the observation process respectively. β_3 controls the association between the outcome and the missingness process. Missingness at time t depends on missingness at the previous measurement time.

The listed examples illustrate the flexibility of joint modelling, as the models for both the observation outcome processes can take different functional forms. Complex dependencies between the processes can be specified. However, fitting these models can be computationally intensive, and the interpretation of random effects in a prediction model can be challenging, especially for end users. [54]

Table 2.3: Summary of (subjective assessments of) advantages, disadvantages, software and assumptions for each method described in this review

Modelling approach	Advantages	Disadvantages	Software	Assumptions
Missing indicators & Separate class	Straightforward, Flexible, Low computational cost, Easy to communicate	Potentially doubles no. of predictors, Too simplistic for complex relationships between missingness and outcome, Assumes discrete time intervals	Easily applied in common statistical software	Assumes absence is a proxy for some unmeasured patient feature, Linear relationship with outcome
Summary measures	Straightforward, Flexible, Low computational cost, Easy to communicate	Generalisability of models across centres may be questioned, May fail to capture complex relationships between observation process and outcome	Easily applied in common statistical software	Assumes observation process is a proxy for some unmeasured patient feature, Largely assumes linear relationship with outcome
Pattern-specific models	Straightforward, Flexible	Number of models becomes large as no. of predictors increases	Easily applied in common statistical software	No assumptions placed on how missingness relates to observed or unobserved variables, Assumes same functional form for all pattern-specific models

Likelihood-based methods	Also allows for imputation	Computationally intensive	None provided	Assumes absence is related to the unobserved value
Similarity measures	Flexible	Computationally intensive	None provided	None provided
Latent variable	Flexible	Computationally intensive	R code provided by Coley and Hubbard	Association between outcome and observation process is captured through latent variable and other predictors
HMMs	Using a Hawkes process allows for time-varying intensity	Complex and computationally intensive	None provided	Assumes longitudinal predictors are normally distributed
Joint modelling/shared random effects	Flexible to different forms of outcome and observation process	Complex, Computationally intensive, Often requires independence assumption between processes given random effects	Frailtypack in R, WinBUGS, merlin in STATA for flexible user-defined models.	Assumes processes (outcome, observation) are independent conditional on random effects, Existing methods assume constant intensity of observation

2.6 Discussion

This study has identified three broad categories of approaches to incorporate IP and/or IO into clinical prediction models: derived predictors; modelling under informed presence; and latent structures. This is a growing area of research, and much of the included literature illustrates that informative presence and informative observation can be incorporated into clinical prediction models in a meaningful way. Where missing data and non-random visit processes have been seen as a nuisance in effect estimation, a more positive outlook is possible when the goal is prediction. Although methodology allowing CPMs to accommodate IP and IO are emerging, further challenges remain, which will be discussed later.

Pullenayegum and Lim [8] and Neuhaus, McCulloch, and Boylan [10] have previously reviewed methods for handling informative observation in studies where the primary aim is to recover unbiased effect estimates. Both articles assume that the outcome is informatively observed, which differs from the focus of our work where we assume informatively measured predictors. Phelan, Bhavsar, and Goldstein [7] present a set of design considerations for EHR-based studies that could help to attenuate issues caused by IP and IO by carefully considering and defining the population of interest, e.g. in which part of the care system patient interactions occur, and how health status could affect patient interactions. None of these articles explicitly discuss prediction, where we anticipate the most appropriate methods will differ from those for effect estimation.

Empirical studies [37, 65] have compared methods capable of handling repeatedly measured predictors in CPMs, and many of these methods can be extended to accommodate IO, such as summarising the process into a single measure (e.g. the mean or maximum - derived predictors), or more complex latent process methods. Both studies found that joint modelling provided little benefit in predictive performance when compared to simple summary measures, but care should be taken in selecting an appropriate summary measure suited to the clinical context. Bull et al. [23] also recommend three key considerations when choosing the most suitable method for harnessing a longitudinally measured predictor: the type and amount of information available at prediction time, how the CPM can benefit from the longitudinal information and the validity of assumptions for the particular application. We expect that these considerations will also be relevant to selecting the most appropriate means of incorporating IO.

To our knowledge, this is the first attempt at synthesising the methodology available to handle IP and IO specifically for prediction purposes. We have achieved this through a systematic search of the literature. A potential limitation is that only the health and biomedical literature was considered; as such, our search potentially did not capture methods that have been developed for use in other fields. Defining relevant terminology around IP and IO is challenging, since the nomenclature differs across the literature. This is illustrated by the fact that a minority (11/36) of included papers were discovered directly through database searches. However, this is a common challenge with methodological reviews [20, 66]. It is possible that methods were missed as a result, but we aimed to mitigate against this by conducting a backward and forward citation search on papers identified through the search strategy and on a set identified as relevant a priori.

Many of the methods discussed herein remain underdeveloped and future studies should investigate the degree to which these methodological choices matter for prediction contexts. We have identified multiple avenues for further research. Missing indicators are the most common approach (in terms of number of studies included) to incorporating the observation process. Although this method is straightforward and adaptable to any type of prediction model, key challenges remain, including but not limited to the requirement to impute missing values when developing and applying the model. Under most prediction frameworks, a value must be entered for any predictor in the model when a prediction is made. The impact of using different imputation techniques at model development and prediction time should be established.

Pattern-specific models present a promising extension to the missing indicator approach, and do not require imputation at either model development or application. Further development should explore ways to borrow strength across models, or pool together sets of patterns, to overcome the issue of developing models with few data points for rarely observed missingness patterns.

Most methods capable of handling informative observation fall under the "summary measures" category (16 papers). The simplicity of this approach is 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 (e.g. modelling 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 modelling. These more complex measures should be compared with both joint modelling 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 modelling for prediction, particularly covering scenario S2 (incorporating longitudinal predictors). Such methods have also recently been extended to functional data [67], allowing them to accommodate complex structures in longitudinal predictors. Joint models have also been proposed to handle IO under an inferential framework [9, 10, 68, 69], 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, since the association between the observation process and outcome is unlikely to be causal, this relationship may not generalise well to different settings. For example, clinicians' monitoring behaviours 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, Hu, and Schaar [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 behaviour 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 [70], or by explicitly modelling 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 informative presence/observation, whereas others illustrate the added benefit of incorporating informative measurement patterns. Missing data has typically been seen as a threat to the estimation of parameters, but since 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.

2.7 Conclusion

We have demonstrated that there is a growing recognition of both informative presence and informative observation within prediction research. Although parallels exist with missing data, informative presence should not be considered the same way, especially within the context of prediction and routinely collected data where there is no pre-specified observation process. By synthesising the available methods and software that could be applied to incorporate IO/IP into CPMs, this paper 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.

References

[1] B. A. Goldstein et al. “Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review”. In: *Journal of the American Medical Informatics Association* 24.1 (Jan. 2017), pp. 198–208. DOI: 10.1093/jamia/ocw042.

[2] R. D. Riley et al. “External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges”. In: *BMJ* 353 (June 2016), p. i3140. DOI: 10.1136/bmj.i3140.

[3] S. Haneuse and M. Daniels. “A General Framework for Considering Selection Bias in EHR-Based Studies: What Data Are Observed and Why?” In: *EGEMS (Washington, DC)* 4.1 (2016), p. 1203. DOI: 10.13063/2327-9214.1203.

[4] N. G. Weiskopf, A. Rusanov, and C. Weng. “Sick patients have more data: the non-random completeness of electronic health records.” In: *AMIA Annual Symposium proceedings*. Vol. 2013. American Medical Informatics Association, 2013, pp. 1472–7.

[5] A. E. W. Johnson et al. “MIMIC-III, a freely accessible critical care database.” In: *Scientific data* 3 (May 2016), p. 160035. DOI: `10.1038/sdata.2016.35`.

[6] B. A. Goldstein et al. “Controlling for Informed Presence Bias Due to the Number of Health Encounters in an Electronic Health Record”. In: *American Journal of Epidemiology* 184.11 (Dec. 2016), pp. 847–855. DOI: `10.1093/aje/kww112`.

[7] M. Phelan, N. A. Bhavsar, and B. A. Goldstein. “Illustrating Informed Presence Bias in Electronic Health Records Data: How Patient Interactions with a Health System Can Impact Inference”. In: *EGEMS (Washington, DC)* 5.1 (2017). DOI: `10.5334/egems.243`.

[8] E. M. Pullenayegum and L. S. Lim. “Longitudinal data subject to irregular observation: A review of methods with a focus on visit processes, assumptions, and study design”. In: *Statistical Methods in Medical Research* 25.6 (2014). DOI: `10.1177/0962280214536537`.

[9] A. Gasparini et al. “Mixed effects models for healthcare longitudinal data with an informative visiting process: a Monte Carlo simulation study”. In: *Statistica Neerlandica* 74.1 (2018), pp. 5–23. DOI: `10.1111/stan.12188`.

[10] J. M. Neuhaus, C. E. McCulloch, and R. D. Boylan. “Analysis of longitudinal data from outcome-dependent visit processes: Failure of proposed methods in realistic settings and potential improvements”. In: *Statistics in Medicine* 37.29 (Dec. 2018), pp. 4457–4471. DOI: `10.1002/sim.7932`.

[11] B. A. Goldstein et al. “How and when informative visit processes can bias inference when using electronic health records data for clinical research”. In: *Journal of the American Medical Informatics Association* 26.12 (Dec. 2019), pp. 1609–1617. DOI: `10.1093/jamia/ocz148`.

[12] C. E. McCulloch, J. M. Neuhaus, and R. L. Olin. “Biased and unbiased estimation in longitudinal studies with informative visit processes”. In: *Biometrics* 72.4 (Dec. 2016), pp. 1315–1324. DOI: `10.1111/biom.12501`.

[13] L. Liu, X. Huang, and J. O'Quigley. "Analysis of longitudinal data in the presence of informative observational times and a dependent terminal event, with application to medical cost data". In: *Biometrics* 64.3 (Sept. 2008), pp. 950–958. DOI: 10.1111/j.1541-0420.2007.00954.x.

[14] K. S. Tan, B. French, and A. B. Troxel. "Regression modeling of longitudinal data with outcome-dependent observation times: extensions and comparative evaluation". In: *Statistics in Medicine* 33.27 (Nov. 2014), pp. 4770–4789. DOI: 10.1002/sim.6262.

[15] J. Sun et al. "Semiparametric Regression Analysis of Longitudinal Data with Informative Observation". In: *Journal of the American Statistical Association* 100.471 (2005), pp. 882–889. DOI: 10.1198/016214505000000060.

[16] D. Agniel, I. S. Kohane, and G. M. Weber. "Biases in electronic health record data due to processes within the healthcare system: retrospective observational study." In: *BMJ (Clinical research ed.)* 361 (Apr. 2018), k1479. DOI: 10.1136/BMJ.K1479.

[17] J.-H. Lin and P. J. Haug. "Exploiting missing clinical data in Bayesian network modeling for predicting medical problems". In: *Journal of Biomedical Informatics* 41.1 (Feb. 2008), pp. 1–14. DOI: 10.1016/J.JBI.2007.06.001.

[18] A. Sharafoddini et al. "A new insight into missing data in intensive care unit patient profiles: Observational study". In: *Journal of Medical Internet Research* 21.1 (Jan. 2019). DOI: 10.2196/11605.

[19] M. Sperrin, E. Petherick, and E. Badrick. "Informative Observation in Health Data: Association of Past Level and Trend with Time to Next Measurement". In: *Studies in Health Technology and Informatics* 235 (2017), pp. 261–265. DOI: 10.3233/978-1-61499-753-5-261.

[20] G. P. Martin et al. "Towards a Framework for the Design, Implementation and Reporting of Methodology Scoping Reviews". In: *Journal of Clinical Epidemiology* 127 (July 2020), pp. 191–197. DOI: 10.1016/j.jclinepi.2020.07.014.

[21] R. Sisk et al. *Scoping review of informative observation in clinical prediction models: protocol.* 2019. DOI: https://osf.io/rtqsg/.

[22] J. L. Vincent et al. "The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure". In: *Intensive Care Medicine* 22.7 (1996), pp. 707–710. DOI: 10.1007/BF01709751.

[23] L. M. Bull et al. "Harnessing repeated measurements of predictor variables for clinical risk prediction: a review of existing methods". In: *Diagnostic and Prognostic Research* 4.1 (Dec. 2020), p. 9. DOI: 10.1186/s41512-020-00078-z.

[24] Z. Che et al. "Recurrent Neural Networks for Multivariate Time Series with Missing Values". In: *Scientific Reports* 8.2018 (2018). DOI: 10.1038/s41598-018-24271-9.

[25] E. Helander et al. "Time-series modeling of long-term weight self-monitoring data". In: *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*. Vol. 2015-Novem. Institute of Electrical and Electronics Engineers Inc., Nov. 2015, pp. 1616–1620. DOI: 10.1109/EMBC.2015.7318684.

[26] Z. C. Lipton et al. "Modeling Missing Data in Clinical Time Series with RNNs". In: *Machine Learning for Healthcare*. 2016.

[27] D. Jarrett, J. Yoon, and M. van der Schaar. "Dynamic Prediction in Clinical Survival Analysis using Temporal Convolutional Networks". In: *IEEE Journal of Biomedical and Health Informatics* 24.2 (July 2020), pp. 424–436. DOI: 10.1109/jbhi.2019.2929264.

[28] L. M. Barclay, J. L. Hutton, and J. Q. Smith. "Chain Event Graphs for Informed Missingness". In: *Bayesian Analysis* 9.1 (2014), pp. 53–76. DOI: 10.1214/13-BA843.

[29] B. E. T. H. Twala, M. C. Jones, and D. J. Hand. "Good methods for coping with missing data in decision trees". In: *Pattern Recognition Letters* 29.7 (2008), pp. 950–956. DOI: 10.1016/j.patrec.2008.01.010.

[30] Y. Ding and J. S. Simonoff. "An investigation of missing data methods for classification trees applied to binary response data". In: *Journal of Machine Learning Research* 11 (2010), pp. 131–170.

[31] F. J. Rodenburg, Y. Sawada, and N. Hayashi. "Improving RNN Performance by Modelling Informative Missingness with Combined Indicators". In: *Applied Sciences* 9.8 (Apr. 2019), p. 1623. DOI: 10.3390/app9081623.

[32] Q. Li and Y. Xu. "VS-GRU: A Variable Sensitive Gated Recurrent Neural Network for Multivariate Time Series with Massive Missing Values". In: *Applied Sciences* 9.15 (July 2019), p. 3041. DOI: `10.3390/app9153041`.

[33] A. Sengupta et al. "Prediction and imputation in irregularly sampled clinical time series data using hierarchical linear dynamical models". In: *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*. Institute of Electrical and Electronics Engineers Inc., Sept. 2017, pp. 3660–3663. DOI: `10.1109/EMBC.2017.8037651`.

[34] N. Du et al. "Recurrent Marked Temporal Point Processes: Embedding Event History to Vector". In: *ACM SIGKDD Conference on Knowledge Discovery*. 2016. DOI: `10.1145/2939672.2939875`.

[35] S. Wu et al. "Modeling asynchronous event sequences with RNNs". In: *Journal of Biomedical Informatics* 83 (July 2018), pp. 167–177. DOI: `10.1016/j.jbi.2018.05.016`.

[36] A. Ghorbani and J. Y. Zou. "Embedding for Informative Missingness: Deep Learning with Incomplete Data". In: *2018 56th Annual Allerton Conference on Communication, Control, and Computing, Allerton 2018*. Institute of Electrical and Electronics Engineers Inc., Feb. 2019, pp. 437–445. DOI: `10.1109/ALLERTON.2018.8636008`.

[37] B. A. Goldstein et al. "A comparison of risk prediction methods using repeated observations: an application to electronic health records for hemodialysis". In: *Statistics in Medicine* 36.17 (July 2017), pp. 2750–2763. DOI: `10.1002/sim.7308`.

[38] M. A. Badgeley et al. "Deep learning predicts hip fracture using confounding patient and healthcare variables". In: *NPJ Digital Medicine* 2.1 (Dec. 2019), p. 31. DOI: `10.1038/s41746-019-0105-1`.

[39] Z. Zhang et al. "Healthcare processes of laboratory tests for the prediction of mortality in the intensive care unit: A retrospective study based on electronic healthcare records in the USA". In: *BMJ Open* 9.6 (June 2019). DOI: `10.1136/bmjopen-2018-028101`.

[40] J. Fauber and C. R. Shelton. *Modeling "Presentness" of Electronic Health Record Data to Improve Patient State Estimation*. Tech. rep. 2018, pp. 1–13.

[41] K. T. Islam et al. "Marked Point Process for Severity of Illness Assessment". In: *Machine Learning for Healthcare*. 2017.

[42] J. Zhao et al. "Handling Temporality of Clinical Events for Drug Safety Surveillance". In: *AMIA Annual Symposium proceedings*. Vol. 2015. 2015, pp. 1371–1380.

[43] M. Zabihi, S. Kiranyaz, and M. Gabbouj. "Sepsis Prediction in Intensive Care Unit Using Ensemble of XGboost Models". In: *Computing in Cardiology (CinC)*. 2019, pp. 1–4.

[44] F. Bagattini et al. "A classification framework for exploiting sparse multivariate temporal features with application to adverse drug event detection in medical records". In: *BMC Medical Informatics and Decision Making* 19.1 (Dec. 2019), p. 7. DOI: 10.1186/s12911-018-0717-4.

[45] M. Saar-Tsechansky. *Handling Missing Values when Applying Classification Models*. Tech. rep. 2007, pp. 1625–1657.

[46] S. Fletcher Mercaldo and J. D. Blume. "Missing data and prediction: the pattern submodel". In: *Biostatistics* 21.2 (Sept. 2020), pp. 236–252. DOI: 10.1093/biostatistics/kxy040.

[47] G. Escarela, J. Ruiz-de-Chavez, and A. Castillo-Morales. "Addressing missing covariates for the regression analysis of competing risks: Prognostic modelling for triaging patients diagnosed with prostate cancer". In: *Statistical Methods in Medical Research* 25.4 (Aug. 2016), pp. 1579–1595. DOI: 10.1177/0962280213492406.

[48] J. J. Kirkham. "A comparison of hospital performance with non-ignorable missing covariates: An application to trauma care data". In: *Statistics in Medicine* 27.27 (Nov. 2008), pp. 5725–5744. DOI: 10.1002/sim.3379.

[49] Y. Sha, J. Venugopalan, and M. D. Wang. "A Novel Temporal Similarity Measure for Patients Based on Irregularly Measured Data in Electronic Health Records". In: *ACM BCB* 2016 Oct (2016), pp. 337–344. DOI: 10.1145/2975167.2975202.

[50] R. Y. Coley et al. "A Bayesian hierarchical model for prediction of latent health states from multiple data sources with application to active surveillance of prostate cancer". In: *Biometrics* 73.2 (2017). DOI: 10.1111/biom.12577.

[51] R. A. Hubbard et al. "A Bayesian latent class approach for EHR-based phenotyping". In: *Statistics in Medicine* 38.1 (Jan. 2019), pp. 74–87. DOI: 10.1002/sim.7953.

[52] K. Zheng et al. "Resolving the Bias in Electronic Medical Records". In: *ACM SIGKDD Conference on Knowledge Discovery.* Vol. 10. 2017, pp. 2171–2180. DOI: 10.1145/3097983.3098149.

[53] A. M. Alaa, S. Hu, and M. Schaar. "Learning from Clinical Judgments: Semi-Markov-Modulated Marked Hawkes Processes for Risk Prognosis". In: *International Conference on Machine Learning (ICML).* International Conference on Machine Learning (ICML), July 2017, pp. 60–69.

[54] Y. Liang, Y. Li, and B. Zhang. "Bayesian nonparametric inference for panel count data with an informative observation process". In: *Biometrical Journal* 60.3 (May 2018), pp. 583–596. DOI: 10.1002/bimj.201700176.

[55] Y.-H. Choi et al. "Joint nested frailty models for clustered recurrent and terminal events: An application to colonoscopy screening visits and colorectal cancer risks in Lynch Syndrome families". In: *Statistical Methods in Medical Research* 29.5 (July 2019), pp. 1466–1479. DOI: 10.1177/0962280219863076.

[56] N. Zhang, H. Chen, and Y. Zou. "A joint model of binary and longitudinal data with non-ignorable missingness, with application to marital stress and late-life major depression in women". In: *Journal of Applied Statistics* 41.5 (May 2014), pp. 1028–1039. DOI: 10.1080/02664763.2013.859235.

[57] R. H. H. Groenwold. "Informative missingness in electronic health record systems: the curse of knowing". In: *Diagnostic and Prognostic Research* 4.1 (Dec. 2020), p. 8. DOI: 10.1186/s41512-020-00077-0.

[58] M. van Smeden, R. H. Groenwold, and K. G. Moons. "A cautionary note on the use of the missing indicator method for handling missing data in prediction research". In: *Journal of Clinical Epidemiology* 125 (Sept. 2020), pp. 188–190. DOI: 10.1016/j.jclinepi.2020.06.007.

[59] M. Sperrin and G. P. Martin. “Multiple imputation with missing indicators as proxies for unmeasured variables: Simulation study”. In: *BMC Medical Research Methodology* 20.1 (July 2020), p. 185. DOI: 10.1186/s12874-020-01068-x.

[60] K. J. M. Janssen et al. “Dealing with Missing Predictor Values When Applying Clinical Prediction Models”. In: *Clinical chemistry* 55.5 (2009), pp. 994–1001. DOI: 10.1373/clinchem.2008.115345.

[61] A. Sharafoddini, J. A. Dubin, and J. Lee. “Patient Similarity in Prediction Models Based on Health Data: A Scoping Review”. In: *JMIR Medical Informatics* 5.1 (Mar. 2017), e7. DOI: 10.2196/medinform.6730.

[62] D. Rizopoulos. “Dynamic Predictions and Prospective Accuracy in Joint Models for Longitudinal and Time-to-Event Data”. In: *Biometrics* 67.3 (Sept. 2011), pp. 819–829. DOI: 10.1111/j.1541-0420.2010.01546.x.

[63] G. L. Hickey et al. “Joint modelling of time-to-event and multivariate longitudinal outcomes: Recent developments and issues”. In: *BMC Medical Research Methodology* 16.1 (2016), pp. 1–15. DOI: 10.1186/s12874-016-0212-5.

[64] M. Alsefri et al. “Bayesian joint modelling of longitudinal and time to event data: A methodological review”. In: *BMC Medical Research Methodology* 20.1 (Apr. 2020), p. 94. DOI: 10.1186/s12874-020-00976-2.

[65] M. J. Sweeting et al. “The use of repeated blood pressure measures for cardiovascular risk prediction: a comparison of statistical models in the ARIC study”. In: *Statistics in Medicine* 36.28 (Dec. 2017), pp. 4514–4528. DOI: 10.1002/sim.7144.

[66] D. O. Lawson, L. Thabane, and L. Mbuagbaw. “A call for consensus guidelines on classification and reporting of methodological studies”. In: *Journal of Clinical Epidemiology* 121 (2020), pp. 109–116. DOI: 10.1016/j.jclinepi.2020.01.017.

[67] K. Li and S. Luo. “Dynamic predictions in Bayesian functional joint models for longitudinal and time-to-event data: An application to Alzheimer’s disease”. In: *Statistical Methods in Medical Research* 28.2 (Feb. 2019), pp. 327–342. DOI: 10.1177/0962280217722177.

[68] R. Miao, X. Chen, and L. quan Sun. “Analyzing longitudinal data with informative observation and terminal event times”. In: *Acta Mathematicae Applicatae Sinica* 32.4 (Oct. 2016), pp. 1035–1052. DOI: `10.1007/s10255-016-0624-3`.

[69] L. Qu, L. Sun, and X. Song. “A Joint Modeling Approach for Longitudinal Data with Informative Observation Times and a Terminal Event”. In: *Statistics in Biosciences* 10.3 (Dec. 2018), pp. 609–633. DOI: `10.1007/s12561-018-9221-8`.

[70] M. Sperrin et al. “Using marginal structural models to adjust for treatment drop-in when developing clinical prediction models”. In: *Statistics in Medicine* 37.28 (Dec. 2018), pp. 4142–4154. DOI: `10.1002/sim.7913`.

[71] G.-J. Geersing et al. “Search filters for finding prognostic and diagnostic prediction studies in Medline to enhance systematic reviews.” In: *PloS one* 7.2 (2012), e32844. DOI: `10.1371/journal.pone.0032844`.

[72] B. J. Ingui and M. A. Rogers. “Searching for clinical prediction rules in MEDLINE.” In: *Journal of the American Medical Informatics Association : JAMIA* 8.4 (2001), pp. 391–7. DOI: `10.1136/jamia.2001.0080391`.

2.8 Supplementary Materials for Chapter 2

2.8.1 Protocol for the Scoping Review

Overview

'Informative observation' is an inherent problem in electronic health records (EHRs) as it results from the way that records are created. Each observation in an EHR is a result of a patient engaging with health services, most likely due to their ill health. The data collected within EHRs therefore contains complex and potentially informative patterns of measurement.

Informative observation presents various challenges in prediction modelling and is likely to result in biased predictions if not handled correctly. There are, however, opportunities to exploit the way in which patients interact with health services to improve predictive performance, by drawing information from the frequency and timing of the data collected within the EHR.

Definitions

We broadly define the terms relevant to this review:

- Informative presence – the presence or absence of a single observation carries meaning about a patient's health status.
- Informative observation – the longitudinal pattern of observation acts as a proxy for a latent process of their underlying health status. This could refer to frequency and/or rate of measurement.

Methodology

Scope of review This review aims to collate the current methodology around allowing clinical prediction models to adjust for and potentially learn from informative observation/informative presence. That is, developing models that remove the bias introduced by informative observation, and potentially improve the predictive performance of models by incorporating information on the frequency or intensity of observation.

The review will focus on the handling of informative observation in clinical prediction problems. This is not necessarily limited to prediction using EHRs, however it is expected that much of the resulting work will address informative observation within EHRs as the problem occurs most regularly through non-random contact with healthcare services. We do not plan to include methodology in fields other than medical/clinical research, as the primary interest of this review is to explore the way in which a patient's medical history can be used to aid prediction. However, following this review, further work should be conducted to explore whether methodology has been proposed in other fields which relate to the same problem. There may be potential to apply methods developed in other fields to a medical setting.

Specifically, we wish to identify methodology that allows the development of prognostic models to handle and incorporate the informative presence of observations. We will summarise the available methods and highlight any remaining methodological challenges in the area.

As we wish to focus on identifying methodology, we do not intend to include papers that only consider applications with no proposed novel methodology. However, we do intend to pick up applications as part of our early searches, as there may be methodological development 'hidden' within a primarily application-based article. These articles will be screened during the first stage of our review, and excluded if they are deemed to be purely application-based.

A secondary aim of the review will to be to collate terminology which is commonly used to refer to concepts relating to informative observation/presence.

There are various aspects of the design of this review that are pre-specified to follow an iterative process; the authors do not have a clear idea of the types of work that will be discovered during this review and therefore we wish to allow some flexibility in some of our strategies and definitions. We have identified which aspects of the design may follow an iterative process within this protocol, and the end of this protocol records additional screening stages or decisions that were made once the screening process had begun. The research team, particularly those conducting the searches, will meet regularly throughout the review process to discuss any changes or additions to this protocol.

Search terms

A pre-defined list of search terms will be used to interrogate selected databases. An initial set of terms related to observation processes and informed presence has been identified from literature, and these will form the basis of our search strategy.

Table 2.4: Search terms to be used in Ovid and Web of Science. † This search will be performed on title only due to the vast quantity of literature around missing data mechanisms.

Terms
Inform* presence
Inform* observ*
Observ* process
Inform* missing*
Inform* follow up
Inform* sampl*
Irregular* sampl*
Non random sampl*
Non random completeness
Inform* completeness
Inconsistently collected data
Selection bias†
Information bias
Admixture bias
Visit* process
Visit* pattern
MNAR
Missing not at random†
Non ignorable missingness

Table 2.5: Web of Science search strategy

Web of Science Search Terms – related to Informative Observation
TS = ((Informative* OR Informed OR Nonrandom OR Nonignorabl* OR Non-random OR Non-ignorabl*) NEAR/5 (Observation* OR Presence OR Absence OR Missing* OR Follow-up OR ”follow up” OR completeness OR sampl* OR nonresponse OR non-response OR drop-out OR dropout))
TS = (”Observation process”)
TS = (“Visit* process”)
TS = (“Visit* pattern”)
TS = (”Admixture bias”)
TS = (”Information bias”)
TS = (”Inconsistently collected data”)
TS = (MNAR)
TI = (”Missing not at random”)

Web of Science Search Strategy Web of Science covers a much broader range of topic areas, as an initial test, we translated the exact strategy used in Ovid into Web of Science syntax. This resulted in an unmanageable quantity of articles to review (380,000+). A more targeted search will therefore be performed in Web of Science, taking advantage of the ‘NEAR’ syntax. Where the Ovid search would pick up any articles containing both words ‘Informative’ and ‘Presence’, the Web of Science strategy has been limited to only find cases where the two terms occur within 5 words of each other.

Geersing filter Each of the above terms will be combined with a search string proposed by Geersing et al [71], which was specifically designed for use in reviews of clinical prediction research. The Geersing string is based on the search terms proposed by Ingui et al [72]. It includes an additional update which has been shown to provide excellent sensitivity in picking up articles related to clinical prediction model development. We therefore anticipate that it should provide a comprehensive means of finding papers related to methodological development in CPMs. Both the Ovid and Web of Science search strategies above will be combined using the ‘AND’ Boolean operator with the Geersing string below.

Ingui CPM Search Strategy + Geersing Update
(Validat* OR Predict_.ti. OR Rule_) OR (Predict* AND (Outcome* OR Risk* OR Model_)) OR ((History OR Variable_ OR Criteria OR Scor* OR Characteristic* OR Finding* OR Factor_) AND (Predict_ OR Model* OR Decision* OR Identif* OR Prognos_)) OR (Decision_ AND (Model* OR Clinical* OR Logistic Models)) OR (Prognostic AND (History OR Variable* OR Criteria OR Scor* OR Characteristic* OR Finding* OR Factor* OR Model*)) OR Stratification OR ROC Curve OR Discrimination OR Discriminate OR c-statistic OR c statistic OR Area under the curve OR AUC OR Calibration OR Indices OR Algorithm OR Multivariable

Information sources

We will use the above search terms to extract potentially relevant articles from: - MEDLINE (via Ovid) - Ovid MEDLINE(R) and Epub Ahead of Print, In-Process & Other Non-Indexed Citations, Daily and Versions(R - Embase (via Ovid) - Web of Science

Validation set

The validation set will be used to test the set of terms which address the concept of informative observation/informative presence, prior to applying the Geersing filter. These articles will be used to test the main body of our search terms, i.e. we hope to find all of the articles listed below using at least one of the terms around 'informative observation', prior to applying the Geersing filter which limits the results to only prediction-related studies. These have been taken from the body of work that the research team is already aware of prior to beginning the review.

- Goldstein, Benjamin A., Nrupen A. Bhavsar, Matthew Phelan, and Michael J. Pencina. 2016. "Controlling for Informed Presence Bias Due to the Number of Health Encounters in an Electronic Health Record." American Journal of Epidemiology 184(11): 847–55.
- Lin, Jau-Huei, and Peter J. Haug. 2008. "Exploiting Missing Clinical Data in Bayesian Network Modeling for Predicting Medical Problems." 2018. Journal of Biomedical Informatics 41(1): 1–14.

- Sun J, Park D, Sun L, Zhao X, Un JS, Ark DP, Un LS, Hao XZ. “Semiparametric Regression Analysis of Longitudinal Data With Informative Observation Times Semiparametric Regression” Journal of the American Statistical Association 100(471): 882-889
- Informative Observation in Health Data: Association of Past Level and Trend with Time to Next Measurement. Stud. Health Technol. Inform 2017;235:261–265.
- Phelan, Matthew, Nrupen A Bhavsar, and Benjamin A Goldstein. “Illustrating Informed Presence Bias in Electronic Health Records Data: How Patient Interactions with a Health System Can Impact Inference.” 2017. EGEMS 5(1): 22
- Lin H, Scharfstein DO, Rosenheck RA. Analysis of Longitudinal Data with Irregular, Informative Follow-Up. 2004. Journal of the Royal Statistical Society (B) 66(3) 791-813
- Pullenayegum EM, Lim LS. Longitudinal data subject to irregular observation: A review of methods with a focus on visit processes, assumptions, and study design. 2016. Statistical Methods in Medical Research 25(6): 2992-3014

Snowballing

In addition to the key word searched, we will use a ‘snowballing’ approach, whereby we look for any papers that cite, or are cited by, those identified as relevant in our initial review. In doing so we hope to ensure our search is inclusive, robust and no important articles are missed. This is in acknowledgment of the lack of consensus around the language used to define

We also have a separate ‘snowballing set’ of articles which we know to demonstrate either informed presence or informative observation. These papers do not necessarily offer any methodological solutions to the issue, however they are likely to have been cited by articles that do. We will therefore search through articles that have cited the following papers (as well as the references for each article listed below) to look for others which could be relevant to the review.

A key difference between the snowballing and validation sets is that the snowballing set articles do not necessarily have informative observation as the primary focus, and may address other data quality issues more broadly. We therefore cannot guarantee that they would be picked up by the search terms. Note that some of these papers fit the scope of our review and will therefore be expected to appear in the results.

- Weiskopf, Nicole G, Alex Rusanov, and Chunhua Weng. 2013. "Sick Patients Have More Data: The Non-Random Completeness of Electronic Health Records." AMIA Annual Symposium proceedings 2013: 1472–77
- Rusanov A, Weiskopf NG, Wang S, Weng C. Hidden in plain sight: bias towards sick patients when sampling patients with sufficient electronic health record data for research. BMC Med Inform Decis Mak. 2014;14(1):51
- Sperrin, Matthew, Emily Petherick, and Ellena Badrick. 2017. "Informative Observation in Health Data: Association of Past Level and Trend with Time to Next Measurement." Studies in health technology and informatics 235: 261–65
- Agniel D, Kohane IS, Weber GM. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. BMJ. 2018;361:k1479.
- Pivovarov R, Albers DJ, Sepulveda JL, Elhadad N. Identifying and mitigating biases in EHR laboratory tests. J Biomed Inform. 2014;51:24-34.
- Haneuse S, Daniels M. A General Framework for Considering Selection Bias in EHR-Based Studies: What Data Are Observed and Why? EGEMS (Washington, DC). 2016;4(1):1203.
- Alaa AM, Hu S, Schaar M. Learning from Clinical Judgments: Semi-Markov-Modulated Marked Hawkes Processes for Risk Prognosis. In: International Conference on Machine Learning (ICML). International Conference on Machine Learning (ICML); 2017:60-69.
- Goldstein BA, Navar AM, Pencina MJ, Ioannidis JPA. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J Am Med Informatics Assoc. 2017;24(1):198-208.
- Fletcher Mercaldo S, Blume JD. Missing data and prediction: the pattern submodel. Biostatistics. September 2018.
- All articles from the Validation Set

Further searching of articles from known groups

We are aware of several research groups that have published work in the area of informative observation. We will therefore manually search for any relevant publications from these groups that we may not be aware of prior to conducting the review.

- Benjamin Goldstein, Matthew Phelan et. al at Duke
- Ahmed M Alaa, Mihaela van de Schaar et. al (UCLA & Cambridge)
- Nicole Weiskopf, Alex Rusanov et. al (Columbia University)

Re-iteration of search terms following initial results

It may become evident during our initial searches that we have missed some potentially useful search terms, or that the strategy should be adapted to better target relevant articles. If this is the case, a new search will be constructed using these new terms and performed after the first screening run.

Inclusion/Exclusion criteria

We have attempted to identify all relevant inclusion/exclusion criteria ahead of article screening, however it may become apparent during the review that more criteria are required in order to better define the scope of the review. This is due to the uncertainty in the type of work that we will find during this review.

Inclusion

- Papers that either:
  - Propose methodology which controls for the bias associated with informative observation AND/OR
  - Propose methodology which uses the observation processes (or presence vs absence of measures) to inform predictions.

Exclusion

- Applications without any novel methodological development
- Studies that only account for informative censoring and not the observation process/presence vs absence
- Studies that only propose ways of handling missing data in MAR and MCAR settings
- Studies that only handle selection bias
- Non-English language

- No mention of prediction or forecasting for out-of-sample participants
- Methodology relating to a field other than medical/clinical prediction

Two-stage search process

A two-stage screening process will be applied to select relevant articles. First, the titles and abstracts of all results from the initial set of search terms will be screened for possible inclusion in the review. Those articles that are marked as relevant during the first stage will then be read in their entirety to further establish their relevance to the review.

Two reviewers (RS and LL) will independently review all searches and consult on their resulting set of articles. Any discrepancies in articles included/excluded will be discussed and resolved prior to summarising the findings of the review. If the two main reviewers are unable to come to an agreement on the inclusion on some papers, then a third party (GPM or NP) will also independently review articles to decide on their inclusion/exclusion.

Information to extract from resulting papers

A pre-defined set of information will be extracted from each included article in the review. Given the uncertainty around the types of work we will find, this list is subject to change following any emerging patterns of relevant information in the resulting studies. Initially, we will record the following information in an Excel spreadsheet for each relevant paper:

- Title, authors, year of publication
- Methodology: details of the methods used to incorporate informative presence/informative observation. ‘Classic’ statistics or Machine Learning.
- Area of application (if any)
- Has the methodology been applied outside of the original study/motivating example?
- Advantages: any reported strengths of the method – possibly related to performance measures or advantages over existing methods (e.g. simulation studies – include comparators)
- Disadvantages: any challenges associated with the methodology, or limitations of the proposed methods.

- Terminology: any relevant terms used in the article that are linked to informative presence/observation, or observation processes more generally. These terms may be used to construct further searches.

Record of changes made after the screening process began

This section has been added following protocol registration, and after the start of the screening process. No changes were made to the inclusion/exclusion criteria, so those presented in the main text of this chapter represent the final set of criteria used to screen the literature.

Changes to search strategy

Removed “Selection bias” and “Information bias” from terms – these were not producing any relevant literature, but instead introduced a large number of irrelevant articles to be screened.

Final search strategy

The final set of search terms used to identify literature (account for changes throughout the screening process) for Ovid is shown in Table 2.7 and for Web of Science in Table 2.8 And for Web of Science:

Additional sub-screening stage Following the first screening stage (titles and abstracts), the two reviewers ended up with a much larger than expected quantity of papers to review ‘in full’. It was noted that many of these would likely not be directly relevant, but had been kept in out of caution. Due to time constraints, the reviewers (RS and LL) decided to add in an additional screening stage, which would first allow for a skim-read of papers retained after the first stage. This is to identify which papers are clearly not relevant to prediction modelling problems, as this is something which is difficult to establish from a title/abstract, but fairly obvious from a glance at the full paper.

2.8.2 Additional results: description of all included papers

Table 2.9 contains a paper-by-paper summary of all literature included in this review.

Table 2.7: Final set of search terms to be used in Ovid.

Terms
Inform* presence
Inform* observ*
Observ* process
Inform* missing*
Inform* follow up
Inform* sampl*
Irregular* sampl*
Non random sampl*
Non random completeness
Inform* completeness
Inconsistently collected data
Visit* process
Visit* pattern
MNAR
Missing not at random†
Non ignorable missingness

Table 2.8: Web of Science search strategy

Web of Science Search Terms – related to Informative Observation
TS = ((Informative* OR Informed OR Nonrandom OR Nonignorabl* OR Non-random OR Non-ignorabl*) NEAR/5 (Observation* OR Presence OR Absence OR Missing* OR Follow-up OR ”follow up” OR completeness OR sampl* OR nonresponse OR non-response OR drop-out OR dropout))
TS = (”Observation process”)
TS = (“Visit* process”)
TS = (“Visit* pattern”)
TS = (”Inconsistently collected data”)
TS = (MNAR)
TI = (”Missing not at random”)

Table 2.9: Paper-by-paper summary of inclued literature

Author(s)	Year	Title	Broad group	Category	IP or IO	Description
Liang et al.	2018	Bayesian nonparametric inference for panel count data with an informative observation process.	Latent structures	Joint modelling	Both	- Cumulative count outcome e.g. recurrent events (tumour recurrence) - predicting future disease recurrences. - Bivariate joint model for panel count data when observation process and event processes are dependent. - Nonhomogeneous Poisson processes for event process and observation process - Stationary Gaussian processes for baseline functions of the two processes - Processes linked via correlated frailty terms, following a bivariate lognormal distribution.
Che et al.	2018	Recurrent Neural Networks for Multivariate Time Series with Missing Values.	Derived predictors	Missing indicator, summary measures	IO	-Takes multivariate time series (longitudinal predictors) data to predict diagnoses and mortality, as binary outcomes. - Uses a form of Recurrent Neural Network called the Gated Recurrent Unit. - Uses both presence/absence of predictors ('masking') and time intervals between measures as inputs in RNN. - Also allows the influence of predictors to decay over time when they have been missing for a while. - Allows for different decay rates for each predictor, to be learned from the data.

Coley et al.	2017	A Bayesian hierarchical model for prediction of latent health states from multiple data sources with application to active surveillance of prostate cancer.	Latent structures	Latent variable	IP	- Bayesian hierarchical model that predicts an individual's underlying health state via joint modelling of repeated PSA measures and biopsies. - Predictions are informed by a subset of patients for whom the true state is actually observed (those who underwent prostatectomy). Therefore prediction target 'cancer state' is partially latent. - PSA (continuous predictor) is modelled using a multilevel model, with random effects (intercept and age effect) varying across latent states. - Biopsy occurrence modelled as logistic regression (binary indicator of biopsy vs no biopsy) within regular time intervals.
Sengupta et al.	2017	Prediction and imputation in irregularly sampled clinical time series data using hierarchical linear dynamical models.	Derived predictors	Summary measures	IO	- Authors develop Kalman filters that explicitly model the time difference between two measures, capturing the dependency between clinical variables and the measurement times. - The state at a given time is allowed to depend on the previous state and the time instant at which the previous observation was made. - Outcomes are all continuous physiological variables.

Zhang et al.	2014	A joint model of binary and longitudinal data with non-ignorable missingness, with application to marital stress and late-life major depression in women	Latent structures	Joint modelling	Both	- Predicting binary primary endpoint: probability of having Major Depressive Disorder (MDD), given individual trajectory of marital stress and an informative missing data mechanism. - Three components of the Shared Parameter Model: 1) Linear Mixed Model for longitudinal measures of marital stress, 2) GLM for binary primary endpoint (MDD), and 3) Shared parameter logistic regression model for the missingness mechanism - Subject-specific random effect shared across all models. - Include missingness at the previous visit as a predictor in missingness at current visit to account for dependence on prior missingness.
Escarela et al.	2016	Addressing missing covariates for the regression analysis of competing risks: Prognostic modelling for triaging patients diagnosed with prostate cancer	Modelling under informed presence	Likelihood-based methods	IP	- Likelihood-based method for estimating parameters under MAR and MNAR missingness in two categorical covariates. - Competing risks outcome, so mixture model is used. - They use a copula formulation for the covariate model and missing data mechanism.

Helander et al.	2015	Time-series modeling of long-term weight self-monitoring data.	Derived predictors	Missing indicator/summary measures	IO	- The goal is to predict future Weight, given a set of past weight data (time-series data). - The authors note that absence of weight data on a previous day predicts absence of data on the next day. - They build an ARIMA model to predict future weight, and incorporate absence flags for the M previous days in the model. M was varied between 0 and 15 days, and the best value chosen on the basis of AIC. For one subject a value of M = 3 was selected, for the other, a value of M = 9.
Barclay et al.	2014	Chain Event Graphs for Informed Missingness	Derived predictors	Separate class	IP	- A form of tree-based method, which incorporates missingness as a separate 'event' in the Chain Event Graph, allowing for it to be informative of outcome. - By exploring predictions made under MAR and MNAR assumptions, the method allows us to assess plausibility of the MAR assumption.
Kirkham	2008	A comparison of hospital performance with non-ignorable missing covariates: An application to trauma care data	Modelling under informed presence	Likelihood-based methods	IP	- Outcome is 30-day survival following trauma as a dichotomous variable. - The method used to handle missing covariates is the 'method of weights' in generalized linear models. - They adapt the work of Joseph Ibrahim, who proposed a ML based approach using the EM algorithm assuming a nonignorable missing mechanism. - The author anticipates that under many settings, missingness is related to the condition of the patient and therefore nonignorable (NMAR), and failure to observe depends on the values that would have been observed.

Alaa et al.	2017	Learning from clinical judgments: semi-Markov-modulated marked Hawkes processes for risk prognosis	Latent structures	Hidden (semi-) Markov Models	IO		- Method designed to account for an incorporate the fact that the frequency of patient monitoring in inpatient care is dependent on the patient’s latent clinical state. - They propose representing the monitoring scheme as a marked Hawkes process. Intensity of the point process is defined by intensity parameters which depend on patient state (latent), and the observed physiological data are modelled using a switching multi-task Gaussian process. - Patient latent clinical states are represented as an absorbing semi-Markov jump process (absorbing to reflect the fact that episodes are informatively censored). - The target of prediction, and a patient’s risk score at any time, is taken to be the probability of eventual absorption into a ’clinical deterioration’ state.

Zheng et al.	2017	Resolving the bias in electronic medical records	Latent structures	Hidden (semi-) Markov Models	IO		- This method recognises that EMR data are essentially an irregular time series, with irregular visiting times and different diagnoses/tests recorded at each visit. The goal is to transform it into a regular time series which is easier to analyse. - A multivariate time-series with regular intervals is created, and the hidden condition at each regular time point is learned, but uses the informative observation process to infer the hidden states. The goal is to then use methods developed for regular time series on the transformed series. - Authors define their specific type of bias as the fact that 1) patients visit hospital more often when sick and 2) doctors order lab tests that are likely to be abnormal. - A ’hidden condition’ is defined at each time point, which is inferred by how and whether patients with particular conditions are observed frequently. - They define the ’observation rate’ as ’the probability of one medical feature being observed at a time point, based on its actual condition (e.g. present/absent, negative/normal).’

Islam et al.	2017	Marked Point Processes for Severity of Illness Assessment	Derived predictors	Summary measures	IO	- Prediction of mortality in the ICU from noisy, incomplete, heterogeneous, unevenly sampled patient data. - This paper fits a Piecewise Constant Conditional Intensity Model - a non-Markovian marked point process to model irregular observation streams in continuous time. - The PCIM point process can be expressed as decision tree, with internal nodes (e.g. 'time between t-1 and t-5?') the binary test functions, and leaves as the 'states', which define the intensity rate (easiest to visualise this - see diagrams in paper). - They learn separate PCIM models for patients who died, and those who did not. The log odds of the two models is then used as a severity score feature for individuals, and entered into a SVM classifier as a feature.

Lipton et al.	2016	Modeling Missing Data in Clinical Time Series with RNNs	Derived predictors	Missing indicator/summary measures	IO	- Goal of prediction model is multilabel classification; choosing from a set of 128 possible diagnoses, where each patient can experience more than one. - They aim to model missingness directly as a binary indicator, as the authors note that which tests were ordered can be more predictive than the results of said tests. - They also compare missingness indicators in combination with Zero Imputation and Forward Filling (LOCF) imputation techniques. The justification for LOCF is that items are likely to be measured when clinicians believe there has been a change in the value, and remain the same otherwise. - They compute a range of features related to missingness of individual items for use in the logistic regression model only, since a linear model 'can only learn hard substitution rules,' - They find that all models improve when either indicators or the manually computed features are included in the model, but this improvement is more modest in the logistic regression case.
Ghorbani & Zou	2018	Embedding for Informative Missingness: Deep Learning With Incomplete Data	Derived predictors	Summary measures	IP	- The aim is to provide a general framework for training neural network predictors when the training data has missing features. - The authors propose a flexible embedding method that learns a representation for missingness directly from the data. - The method does not require any imputation, and can handle informative missingness.

Li & Xu	2019	VS-GRU: A Variable Sensitive Gated Recurrent Neural Network for Multivariate Time Series with Massive Missing Values	Derived predictors	Missing indicator/summary measures	IO	- The proposed method is called variable sensitive GRU (VS-GRU). It considers the missingness rates of each variable individually rather than as a whole. - For each variable at each time point, create a missingness indicator to differentiate between imputed and observed values. Also calculate the missing rate of each predictor. - The missing indicators for each individual variable as used as inputs/features in the VS-GRU model, as well as the missing factor (defined in the next column).
Rodenburg et al.	2019	Improving RNN Performance by Modelling Informative Missingness with Combined Indicators	Derived predictors	Missing indicator/summary measures	Both	- The method proposes summing missing indicators to avoid the issue of potentially doubling the number of predictors in a model where using a single missing indicator for each predictor.
Sharafoddini et al.	2019	A New Insight Into Missing Data in Intensive Care Unit Patient Profiles: Observational Study	Derived predictors	Missing indicator/summary measures	IP	- Uses simple missing indicators to predict patient mortality in the ICU. - Each patient's data was summarised over every day of their admission, with indicators representing which lab tests were ordered in a particular day. - The missing indicators were added to the predictor matrix to create an augmented dataset. Missing values were imputed using Hot Deck and predictive mean matching single imputation techniques. - Feature selection methods were employed to select the most informative missing data indicators. - They attempt fitting a model on missing indicators alone, and find fairly good predictive performance. However they note that these models would not be sufficient for use in clinical practice.

Lin & Haug	2008	Exploiting missing clinical data in Bayesian network modeling for predicting medical problems	Derived predictors	Missing indicator/separate class	IP	- The method explicitly represents missing items in a clinical decision system to improve predictive performance. - All methods are a form of Bayesian Network used to predict diagnoses; a naïve Bayes structure, a human-composed network structure, and two networks based on structural learning algorithms. - They compare different ways of incorporating (or ignoring) information in the missingness, and find that those methods explicitly modelling missing items perform best. - Missingness is represented as either a separate class or a separate indicator variable.
Badgeley et al.	2019	Deep learning predicts hip fracture using confounding patient and healthcare variables	Derived predictors	Summary measures	IP	- Hospital process variables (related to image acquisition) are added as predictors in a model. - Variables considered are: department, scanner model, scanner manufacturer, laterality, study date (and day of week), order priority, technician, radiologist, radiation dose, time from image order to acquisition, time from image acquisition to initial interpretation, time from image acquisition to final interpretation. - They fit: logistic regression models and convolutional neural networks. - Most hospital process vars were found to be statistically significantly associated with fracture (p ¡ 0.05) - Missing items were imputed.

Zhang et al.	2019	Healthcare processes of laboratory tests for the prediction of mortality in the intensive care unit: a retrospective study based on electronic healthcare records in the USA	Derived predictors	Summary measures	IO	- Similar to the Badgeley et al paper; this time predicting mortality in the ICU using variables related to the collection of lab tests. - Process variables this time are: the clock hour, the number of measurements and the measurement time from ICU admission. - GLMs (logistic regression) are fitted with hospital mortality as the outcome. - AUROC increased with addition of the process variables.
Sha et al.	2016	A Novel Temporal Similarity Measure for Patients Based on Irregularly Measured Data in Electronic Health Records	Latent structures	Similarity measures	IO	- Authors create a patient similarity measure which incorporates the ordering and time intervals between lab tests. - They hypothesise that the timestamps, order, and frequency of measurements in addition to the results could carry meaningful information about patient condition. - Their similarity is novel since it incorporates time-varying information, as well as information on e.g. time intervals between tests. - The measure takes only the 10 most commonly ordered lab tests from each dataset used (MIMIC-II and CHOA). - Their novel similarity measure is compared against two non-temporal similarity measures, and find improved predictive performance. - The similarity measure is used to define patient cohorts within which to develop separate models.

Hubbard et al.	2018	A Bayesian latent class approach for EHR-based phenotyping	Latent structures	Latent variable	IP	- Develops a method that can handle informatively missing predictors using a Bayesian latent class approach in a phenotyping context. - This is an unsupervised learning method, where the true gold standard phenotype is not available for any patients. - Method assumes that true disease state is unavailable, but may influence which data are available for an individual. - The approach appears to perform well, even under MNAR. - Prior knowledge about classification accuracy of biomarkers and codes can be incorporated through suitable choice of priors.
Goldstein et al.	2017	A comparison of risk prediction methods using repeated observations: an application to electronic health records for hemodialysis	Derived predictors	Summary measures	IO	- This paper primarily compares methods that generally allow for repeated biomarker measurements in a prediction model. - However, they explore the incorporation of informed presence, by adding in the number of times a measurement is taken as a predictor. - Authors comment on the predictability of the number of measurements taken as a simple summary statistic. However they note that this is only the case for the vital signs, as these are not measured on a scheduled basis as labs would be in this setting.

Fletcher Mercaldo & Blume	2018	Missing data and prediction: the pattern submodel	Modelling under informed presence	Pattern-specific models	IP	- Develops separate models for each missingness pattern: the pattern submodel (PS) - Therefore accommodates missing data at both model development and prediction time, and does not require any imputation at development or prediction time. - The key difference with regular pattern mixture models is that only data in the observed patterns are used to develop the models; this means a that no assumptions must be placed on the missing data mechanism - The paper focuses on assessing performance at prediction time, comparing the pattern submodel with commonly applied imputation techniques.
Fauber & Shelton	2018	Modeling "Presentness" of Electronic Health Record Data to Improve Patient State Estimation	Derived predictors	Point processes	IO	- Uses Piecewise-Constant Intensity Models to build a generative model of observation times and values. - The model is used to predict future values of vital signs based on the history of these events. - They note that data are rarely MAR in medical settings, and instead that the frequency or absence of events should be used to estimate patient state. - An existing PCIM model is extended to incorporate not only the rate of events, but also values.

Zabihi et al.	2019	Sepsis Prediction in Intensive Care Unit Using Ensemble of XGboost Models	Derived predictors	Summary measures	IO	- Authors note that the pattern of missing data may convey useful information, and should therefore be used to aid prediction. - Missing data are first imputed and summary measures are computed to be used as features in the model. - Authors define sequence abstraction: each sequence is defined as a set of consecutive measures where values are either missing or present, e.g. SBP measures for a 6 hour period (1 hour intervals): {NA, 122, 98, NA, NA, 123}. Based on their definition, we have four sequences of {NA}, {122,98}, {NA, NA} and {123}. Sequence abstraction calculates and uses as features: 1) Mean and variance of the lengths of sequences along each covariate. 2) Summation and variance of the lengths of sequences with only valid values (no missing) along each covariate, and 3) Mean and variance of the lengths of sequences along each observation, in the last 5 hours. - These features representing different aspects of the missingness patterns are entered into a classifier.

Du et al.	2016	Recurrent Marked Temporal Point Processes: Embedding Event History to Vector	Derived predictors	Marked point processes	IO	- Proposes ’Recurrent Marked Temporal Point Processes’ (RMTPP) to simultaneously model event timings and markers. - They aim to predict the time and type of future events from the history of a sequence of many events. - The key idea of the approach is to view the intensity function of a temporal point process as a nonlinear function of the history of the process, and parameterize the function using a recurrent neural network. - Using our model, event history is embedded into a compact vector representation which can be used for predicting the next event time and marker type. - Based on the hidden unit of RNN, we are able to learn a unified representation of the dependency over the history.

Choi et al	2019	Joint nested frailty models for clustered recurrent and terminal events: An application to colonoscopy screening visits and colorectal cancer risks in Lynch Syndrome families	Latent structures	Joint modelling	IO	- Authors introduce and use a joint nested frailty model (JNFM) to predict risk of colorectal cancer (CRC). - They incorporate the visit process of screening visits as a recurrent events process and cancer occurrence (prediction target) as a terminal event. - Model provides dynamic predictions, allowing predictions to update as new information becomes available. - They allow for an individual-specific frailty which links the processes. - Their data also contains clustering at the family level, which is included in the model via another frailty term for each family. - The number of screening visits per individual is highly irregular, and the timing between visits varies both within and between individuals.

Jarrett et al	2019	Dynamic Prediction in Clinical Survival Analysis using Temporal Convolutional Networks	Derived predictors	Missing indicators	Both	- Authors are proposing Match-Net: a Missingness-Aware Temporal Convolutional Hitting-time Network. Designed to capture temporal dependencies and heterogeneous interactions in covariate trajectories and patterns of missingness. - The model makes no assumptions regarding the underlying longitudinal or time-to-event processes. - The model can provide dynamically updated survival predictions, as well as accommodating informative patterns of missingness. - The model explicitly accounts for informative missingness by learning correlations between patterns of missingness and disease progression. - The model accounts for potential informativeness of both irregular sampling (intervals between consecutive visits and measures may vary) and asynchronous sampling (not all features are measured at the same time or same frequency).

Saar-Tsechansky	2007	Handling Missing Values when Applying Classification Models	Modelling under informed presence	Pattern-specific models	IP	- Key method of interest here is the ’reduced models’ approach, where separate models are developed for different missingness patterns (as described later by Fletcher Mercaldo). -Proposes developing a separate model for each missingness pattern, but using all available data, NOT just those observed within the pattern (as with Fletcher-Mercaldo’s more recent paper). - Authors also propose a workaround for the possibility of having to develop huge numbers of models when p is large - develop models for ’important’ patterns, and using ’lazy learning’ or imputation for less important patterns.
Ding & Simonoff	2010	An Investigation of Missing Data Methods for Classification Trees Applied to Binary Response Data	Derived predictors	Separate class	IP	- Authors find that ’separate class’ (adding in an additional category for missing values) is the best method to use when the training set contains missing values and missingness is related to the outcome. - All methods here are considered under a classification tree framework, but also extended to logistic regression. Predictors must be categorised in tree-based methods, so the separate class works well here. - They also study different methods in a logistic regression model: missing indicator method, separate models for data with/without missing data (by-group method), imputing missing values with mean/mode and complete case. Missing indicator and separate models observations with/without missing values are the same as the separate class method in tree methods.

Bagattini et al	2019	A classification framework for exploiting sparse multi-variate temporal features with application to adverse drug event detection in medical records	Derived predictors	Summary measures	IO	- Provides a framework for using multivariate time series data to detect adverse drug events, considering that the sparsity in the available data may be useful in determining the existence of an ADE. - Proposes and compares three different methods for handling sparsity, one of which explicitly exploits it.
Wu et al	2018	Modeling Asynchronous Event Sequences with RNNs	Derived predictors	Missing indicators/summary measures/time intervals	IO	- Discusses different ways of measuring time, and of incorporating this into RNNs, .e.g time between events, time since a landmark event, burstiness of events. - Then establishes how this information should be used in RNNs; either concatenated into the predictor matrix, or used to mediate the importance of an input. i.e. the longer something has been unobserved, the less important it is.

Zhao et al	2015	Handling Temporality of Clinical Events for Drug Safety Surveillance	Derived predictors	Summary measures	IO	- This method handles informative observation (longitudinally measured predictors) by proposing different ways of counting the number of measures (or clinical events) that occur. - The setting is in detecting Adverse Drug Events (ADEs), which are not necessarily recorded in the patient record. - The first method (Bag of events - BE) simply counts the number of times a measure occurs within D days. - Bag of Binned Events (BBE) counts the number of occurrences of each x in each day within D days. So each day has a separate feature calculated. - Bag of Weighted Events (BWE) assigns different weights to event x that occurred at different days d, and takes into account the weights when counting the number of occurrences of x. Weights are assigned according to the time distance between the event and the target ADE (prediction target). Those further away from the target ADE receive proportionally less weight.
Twala et al	2008	Good methods for coping with missing data in decision trees	Derived predictors	Separate class	IP	- Proposes 'missingness incorporated in attributes' and compares against competing methods. - Method is very similar to separate class, but has also been extended for use in continuous predictors, where missingness can be used as the basis of a split in a tree-based model.

Agniel et al	2018	Biases in electronic health record data due to processes within the healthcare system: retrospective observational study	Derived predictors	Summary measures	Both	Explores the predictive ability of time of day, day of the week, and time between measures on mortality in inpatient admissions. Shows that the timing is a more accurate predictor of mortality than the result itself of some blood tests.

Chapter 3

Imputation and Missing Indicators for handling missing data in the development and implementation of prediction models: a simulation study

The work in this chapter was conceptualised as a result of the findings of Chapters 1 (general introduction and literature review) and 2 (scoping review of informative presence/observation). The existing literature does not properly address how informatively observed/missing predictors should be handled in the context of clinical prediction modelling. There also remain challenges in the implementation of existing missing data handling strategies across the development and implementation stages of the model pipeline.

The work uncovered in Chapter 2 found that missing indicators have been proposed as a method of incorporating informative presence into clinical prediction models, but this work has not addressed how they should be combined with either regression or multiple imputation, or the effect of using them when missing data is not allowed at the point of prediction.

The work in this chapter therefore seeks to provide recommendations on how (informative) missing data should be approached at both model development and deployment, and whether missing indicators can provide gains in predictive performance under informative presence.

3.1 Abstract

3.1.1 Background

Existing guidelines for handling missing data are generally not consistent with the goals of clinical prediction modelling, where missing data can occur at any stage of the model pipeline; development, validation or application. Multiple imputation (MI), often heralded as the gold standard approach, can be challenging to apply in the clinic. Clearly, the outcome cannot be in the imputation model, as recommended under MI, at prediction time. Regression imputation (RI) is an alternative strategy, that involves using a fitted model to impute a single value of missing values based on observed ones. RI could offer a pragmatic alternative in the context of prediction, that is simpler to apply at the point of prediction. Moreover, the use of missing indicators has been proposed to handle informative missingness, but it is currently unknown how well this method performs in the context of CPMs.

3.1.2 Methods

We performed an extensive simulation study where data were generated under various missing data mechanisms to compare the predictive performance of CPMs developed using both imputation methods. We consider deployment scenarios where missing data is permitted or prohibited, and develop models that use or omit the outcome during imputation and include or omit missing indicators. We assume that the missingness mechanism remains constant between the stages of the model pipeline.

3.1.3 Results

When complete data must be available at deployment, our findings were generally in line with widely used recommendations; that the outcome should be used to impute development data when using MI, yet omitted if using RI. When imputation is applied at deployment, developing a model that instead omits the outcome from imputation at development was preferred. Missing indicators were found to improve model performance in some specific cases, but found to be harmful when missingness is dependent on the outcome.

3.1.4 Conclusion

We provide evidence that commonly taught principles of handling missing data via MI may not apply to CPMs, particularly when data can be missing at deployment. In such settings, RI and missing indicator methods can (marginally) outperform MI. As shown, the performance of the missing data handling method must be evaluated on a study-by-study basis, and should be based on whether missing data are allowed at deployment. Some guidance is provided.

3.2 Background

Clinical Prediction Models (CPMs) can be used to guide clinical decision making and facilitate conversations about risk between care providers and patients [1]. A CPM is a mathematical tool that takes patient and clinical information (predictors) as inputs and, most often, produces an estimated risk that a patient currently has (diagnostic model) or will develop (prognostic model) a condition of interest [2]. A common challenge in the development, validation and deployment of CPMs is the handling of missing data on predictors and outcome data. The methods used to handle missing data in CPM development and validation are usually complete case analysis or multiple imputation (MI) approaches [1, 3], the latter of which is often heralded as the gold standard in handling missing data. This topic has recently received renewed interest, with authors arguing the basis on which MI is used relies too heavily on principles relevant to causal inference and descriptive research, which are less relevant when the goal is to provide accurate predictions [4].

The objectives of prediction research differ from those of descriptive or causal inference studies. For the latter, missing data should be handled in such a way that minimises bias in the estimation of key parameters, and generally this is achieved through multiple imputation of missing data. In the development of prediction models, however, unbiased parameter estimates are not necessarily the ones that optimise predictive performance. Moreover, in prediction research we must distinguish between handling missing data across the entire model pipeline; model development, model validation, and model deployment (or prediction time), and anticipate whether missing data shall be allowed at deployment. Ideally, all predictors considered for inclusion in a CPM should be either readily available, or easily measured, at the point of prediction. There exist, however, notable examples that allow missingness at the point of prediction [3]; the QRisk3 [5] and QKidney [6] algorithms are examples of such models that allow users to make a prediction in the absence of clinical predictors (such as cholesterol) that may not be available, or easily measured, at the time of prediction.

Best practice states that the outcome should be used in the imputation model when applying MI [7], creating a congenial imputation model. Clearly the outcome is unknown at prediction time, and applying imputation without the outcome would violate the assumption of congeniality. Since model validation should evaluate predictive performance under the same missing data handling strategy to be used in practice, the outcome should be omitted from any imputation model at validation, potentially resulting in less accurate imputations since predictors are normally predictive of the outcome. We therefore define "performance under no missingness", where we assume all predictors are always available (or easily obtained) at prediction time, and "performance under missingness", assuming missing data is allowed and will be imputed at deployment.

Regression imputation (RI) could provide a more pragmatic alternative to MI in the context of prediction. RI is a form of single imputation that fits a model to impute missing predictors using observed data. The key difference between RI and MI is that RI is based on a single equation, and produces one imputed value (deterministic process), whereas MI is a stochastic sampling process that involves repeatedly sampling from a distribution. For RI to be applied in practice, only the imputation model(s) needs to be available alongside the full prediction model, as opposed to MI which generally also requires access to the development dataset. Existing literature has, however, demonstrated several pitfalls of RI in the context of causal estimation - it is highly sensitive to model misspecification, can increase correlation between predictors and underestimate variability in parameter estimates [8]. Although these issues may therefore persist within the prediction context, they may not apply since the recovery of unbiased parameter estimates is no longer of direct concern. RI may also overcome some of the previously mentioned issues related to predictive modelling with MI, since inclusion of the outcome in the imputation model is not recommended [1]. To our knowledge these issues and challenges have not been studied to date.

Both MI and RI are techniques devised under the assumption that data are missing at random, i.e. missingness does not depend on unobserved values. The validity of the MAR assumption within health data is often dubious, especially when using routinely collected data [9, 10], however these definitions were created with the goal of recovering unbiased parameter estimates in mind and therefore may be less relevant to the prediction modelling context [4]. Within routinely collected data, the recording of key clinical markers is often driven by the needs of the patient and clinical judgments of the care provider [10]. Missingness is therefore potentially informative with respect to a patient's current or future condition, and including information about the way an individual has been observed into a prediction model has the potential to improve its predictive performance [11]. A commonly used, effective approach to achieve this is through the inclusion of missing indicators as predictors in a CPM.

This study therefore aims to explore the use of missing indicators as model predictors alongside both regression and multiple imputation. We explore the effect of omitting/including the outcome from each imputation model at development, and imputing data without the outcome at validation (and therefore deployment). We compare the two imputation strategies under each development/validation strategy. Our results will inform recommendations on the handling of missing data during model development and deployment that will be especially relevant to applied researchers developing clinical prediction models.

3.3 Methods

We performed an extensive simulation study in which we evaluated a range of different missingness mechanisms. Our study has been designed according to best practice and reported according to the ADEMP structure (modified as appropriate for a prediction-focused study), proposed by Morris, White, and Crowther [12].

3.3.1 Aims

To compare MI and RI approaches in imputing missing data when the primary goal is in developing and deploying a prediction model, under a range of missing data mechanisms (MCAR, MAR, MNAR), with/without a missing indicator and with/without the outcome included in the imputation model. Each of these will be examined both allowing for and prohibiting missing data at deployment, and performance will be estimated separately for each of these two scenarios.

Throughout this study, we assume that both the missingness mechanism and handling strategy will remain the same across validation and deployment, and therefore validation is a valid replication of model deployment and our performance estimates are reliable estimates of model performance at deployment. The only case where this is not true is when we impute data using the outcome at validation, which will be discussed in more detail in the following sections.

3.3.2 Data-generating mechanisms

We focus on a logistic regression-based CPM to predict a binary outcome, Y, that is assumed to be observed for all individuals (i.e. no missingness in the outcome) during development and validation of the model. Without loss of generality, we assume that the data-generating model contains up to three predictors, X_1, X_2 and U, where X_1 is partially observed and potentially informatively missing (depending on simulation scenario, as outlined below), X_2 is fully observed and U is unobserved. We denote missingness in X_1 with binary indicator R_1, where $R_1 = 1$ if X_1 is missing, and $R_1 = 0$ if it is observed.

We construct four separate DAGs depicted in Figure 3.1, each representing different missingness structures covering: Missing Completely at Random (MCAR), Missing at Random (MAR), Missing Not at Random dependent on X_1 (MNAR-X) and Missing Not At Random dependent on Y (MNAR-Y). MNAR-Y is often instead referred to as “MAR-Y”, however we felt that in this instance, classifying this mechanism as missing not at random is more appropriate since the outcome (Y) is not observed at the point of prediction. This mechanism could therefore be classified differently according to which stage of the CPM pipeline we are referring to: at model development (and equally at model validation) it may be more appropriate to refer to this as “MAR-Y” since information on observed outcomes is available at this stage. We adopt a single name across this study, however, to ease presentation and interpretation of results.

The DAGs further illustrate how missingness in X_1 is related to X_1 or X_2. In each of the DAGs, X_1^* represents the observed part of X_1 and R_1 is the missing indicator. In order to reconstruct these DAGs in simulated data, we stipulate the following parameter configurations:

X_1 and X_2 are drawn from a bivariate normal distribution to allow moderate correlation between the two predictors, such that:

$\mathbf{X} \sim MVN(\mu, \Sigma)$

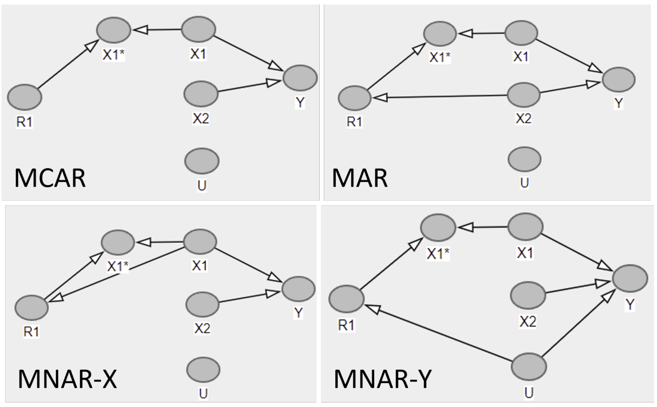


Figure 3.1: Directed Acyclic Graphs for four missingness structures, constructed via our data generating mechanisms

Where $\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

and $\mathbf{\Sigma} = \begin{bmatrix} 1 & 0.4 \\ 0.4 & 1 \end{bmatrix}$

- $R_1 \in \{0,1\}$, and $P[R_{1i}=1] = expit(\beta_0 + \beta_{X_1}X_{1i} + \beta_{X_2}X_{2i} + \beta_Y Y_i)$ i.e. missingness in X_1 can depend on X_1, and/or X_2, and/or Y.

- β_{X_1}, β_{X_2} and β_Y were varied across $\{0, 0.5, 1\}$

- Y is binary, with $P[Y_i = 1] = expit(\gamma_0 + \gamma_{X_1}X_{1i} + \gamma_{X_2}X_{2i} + \gamma_{X_1X_2}X_{1i}X_{2i})$

- γ_{X_1} and γ_{X_2} were varied across $\{0.5, 0.7\}$

- $P[Y=1] = \pi_Y$ was fixed to be 0.1.

- $P[R_1 = 1] = \pi_{R1} \in \{0.1, 0.25, 0.5, 0.75\}$

- $\gamma_{X_1X_2}$ can take values $\{0, 0.1\}$.

- β_0 and γ_0 are calculated empirically as required to set the desired level of π_{R1} and π_Y.
  U was not directly simulated, but the MNAR-Y scenario generated via the inclusion of $\beta_Y \neq 0$.

The parameter values above were selected to represent what might be observed in real-world data, assuming that X_1 and X_2 can represent some summary of a set of model predictors. The γ coefficients (specifying the relationship between predictors and the model outcome) are strong, as we wish to illustrate the impact of (potentially informative) missingness in a very important predictor (or set of predictors).

Datasets will be generated with $n = 10000$ records from the DGMs described above, and split 50/50 into development and validation sets. The development and validation sets will therefore contain 5000 records each. This is chosen as a suitably large size that should be sufficient to estimate underlying parameters, and avoid overfitting. It is also in line with what we would expect to see in electronic health record data.

Each simulated DGM will be repeated for 200 iterations. The parameter values listed above result in a total of 864 parameter configurations. 200 iterations was selected as optimal to balance the requirement to obtain reliable estimates of key performance metrics (by using a sufficient number of repetitions) with the size of the study and computational requirements of repeatedly running multiple imputation over a large number of simulated scenarios.

Our simulation procedure first generates data under the DGMs described Figure 3.1, according to the above parameter configurations. We then take a split-sample approach to assessing model performance (we recognise that this is a statistically inefficient approach to use in model development applications, but our simulated sample size is sufficiently large that this should not pose an issue)[13]. We randomly separate the data into 50% development and 50% validation, fit the models on the development data, and calculate performance measures on the derived models applied to the validation set. The full simulation procedure is illustrated in Figure 3.2. Note that in this instance, we fit the imputation models separately in the development and validation sets. Since the DGMs and missingness mechanisms remain constant between the two datasets, we assume that the fitted imputation model would not change and this is therefore a valid approach to take. In a real-world setting, however this would not be feasible as only a single patient's data would be available at the point of prediction, and we would want to use the same imputation model as was used/developed in the development data.

3.3.3 Missing data handling strategies

We consider two main methods for handling missing data at the development and implementation stages of the CPM pipeline: multiple imputation and regression imputation. Multiple imputation can be applied with relative ease at the model development stage, specifying an imputation model and method for every predictor with missing data (in this case just X_1), conditional on other data available at model development/implementation. Multiple draws are then made from the imputation model, resulting in multiple completed datasets. The relevant CPMs are then fit separately to each resulting imputed dataset, and the model coefficients pooled according to Rubin's rules to obtain a single set of model coefficients. For regression imputation, we follow a similar process in fitting a model to the missing predictor(s) based on observed data. The key difference between MI and RI is that we obtain only a single completed dataset under RI. We can then fit the analysis model to this new complete data to obtain the CPM's parameter estimates. Both methods can be implemented using the `mice` package in R (amongst others), but more flexible user-defined imputation models for regression imputation could be fit with relative ease using alternative modelling packages. In this study, we consider non-stochastic RI, as we only have a single missing predictor so there is no requirement to apply sampling procedures to handle more complex patterns of missingness in the model predictors.

Applying multiple imputation to incomplete data for new individuals at the point of prediction is more challenging as it is not generally possible to extract the final imputation model from the output provided by standard statistical software. In order to "fix" the imputation model for new individuals, it has therefore instead been proposed that the new individual's data should first be appended to the original (imputed) development data, and the imputation re-run on the new stacked dataset [14–16]. Regression imputation, on the other hand, is easier to implement at the point of prediction, since flexible models can be defined for each (potentially) missing predictor, and these models can be stored and used to impute at the point of prediction for new individuals. Ideally, model validation should follow the same steps as model deployment in order to properly quantify how the model will perform in practice. However, since validation is usually completed for a large cohort of individuals at the same time (as opposed to a single individual), it is likely that missing data imputation would take place as a completely separate exercise, with the imputation model depending solely on the validation data.

In this study, we take a split-sample approach to model validation, whereby the imputation is run (and therefore separate imputation models fit) within the development and validation data. This would not be possible in practice in order to predict for new individuals, since only data for a single individual would be available at the point of prediction. We expect, however, that the fitted imputation model(s) would remain the same between the development and validation datasets under this simulation framework, since the DGM and missing data mechanisms remain constant between the two datasets.

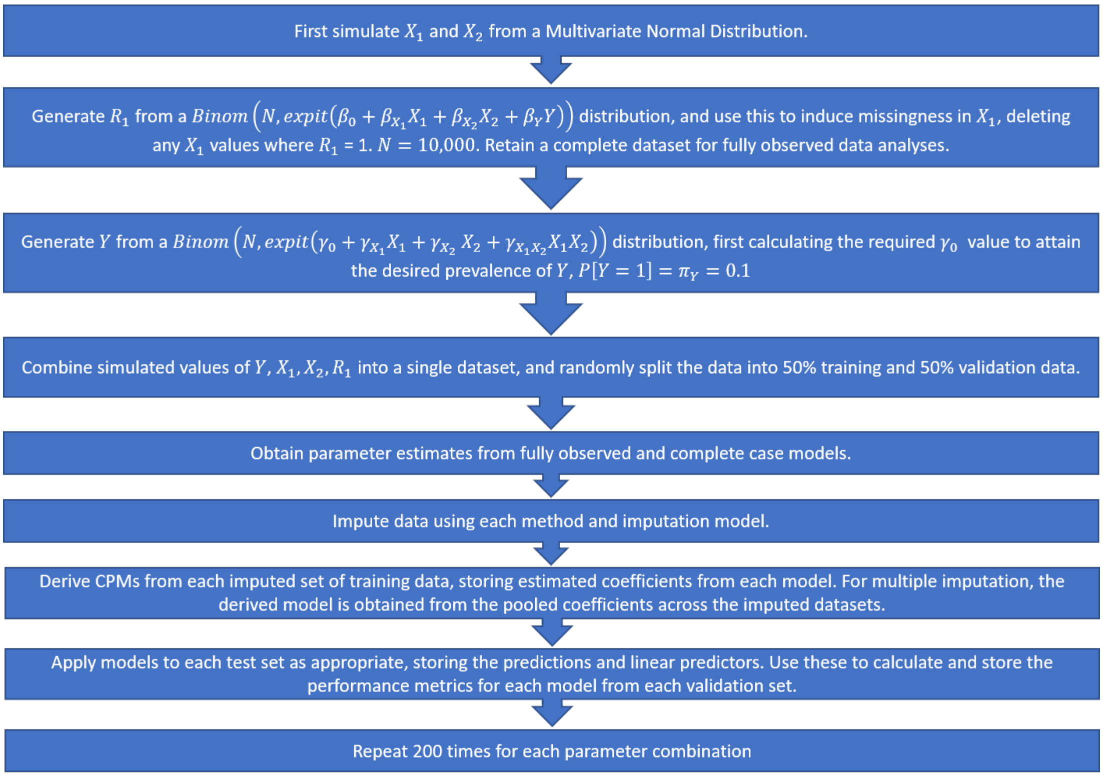
As an alternative to the split-sample approach, it would have been possible to use re-sampling techniques such as cross-validation or bootstrapping. In the case of cross-validation, since only a small proportion of the data is used to estimate model performance, it would make more sense to run the imputation using an alternative approach (for multiple imputation). This would involve appending the fold-specific test set to the (imputed) development data, and re-running the multiple imputation this way. If instead using bootstrapping, a separate imputation model could be fit to the bootstrap sample (as was done here), or alternatively the same approach as described above for cross-validation. The stack-then-impute approach would be preferred in a real-world setting when there is only a single CPM to be fit, i.e. when not working within a simulation study as this is more computationally intensive, but feasible when the study is not being repeated under multiple simulated DGMs.

3.3.4 Fitted CPMs

We fit three possible CPMs to the development data (under each different imputation method), firstly with a simple model including both derived predictors and their interaction. We then fit models incorporating missing indicators, as well as considering a model with an interaction between the missing covariate X_1 and its missing indicator R_1 [17]:

- Predictors and their interaction only: $P[Y_i = 1] = expit(\gamma_0 + \gamma_{X_1} X_{1i} + \gamma_{X_2} X_{2i} + \gamma_{X_1 X_2} X_{1i} X_{2i})$

- Inclusion of an additional missing indicator: $P[Y_i = 1] = expit(\gamma_0 + \gamma_{X_1} X_{1i} + \gamma_{X_2} X_{2i} + \gamma_{X_1 X_2} X_{1i} X_{2i} + \gamma_{R_1} R_{1i})$



First simulate X_1 and X_2 from a Multivariate Normal Distribution.

Generate R_1 from a $Binom\left(N, expit(\beta_0 + \beta_{X_1}X_1 + \beta_{X_2}X_2 + \beta_Y Y)\right)$ distribution, and use this to induce missingness in X_1, deleting any X_1 values where R_1 = 1. $N = 10{,}000$. Retain a complete dataset for fully observed data analyses.

Generate Y from a $Binom\left(N, expit(\gamma_0 + \gamma_{X_1}X_1 + \gamma_{X_2}\, X_2 + \gamma_{X_1X_2}X_1X_2)\right)$ distribution, first calculating the required γ_0 value to attain the desired prevalence of Y, $P[Y = 1] = \pi_Y = 0.1$

Combine simulated values of Y, X_1, X_2, R_1 into a single dataset, and randomly split the data into 50% training and 50% validation data.

Obtain parameter estimates from fully observed and complete case models.

Impute data using each method and imputation model.

Derive CPMs from each imputed set of training data, storing estimated coefficients from each model. For multiple imputation, the derived model is obtained from the pooled coefficients across the imputed datasets.

Apply models to each test set as appropriate, storing the predictions and linear predictors. Use these to calculate and store the performance metrics for each model from each validation set.

Repeat 200 times for each parameter combination

Figure 3.2: Simulation procedure, step-by-step

- Inclusion of an interaction between the X_1 and R_1 terms in the outcome model: $P[Y_i = 1] = expit(\gamma_0 + \gamma_{X_1}X_{1i} + \gamma_{X_2}X_{2i} + \gamma_{X_1X_2}X_{1i}X_{2i} + \gamma_{R_1}R_{1i} + \gamma_{R_1X_1}R_{1i}X_{1i})$

Each model will be derived on incomplete data using both MI and RI, with and without the outcome in the imputation model. The derived models will then be applied to the validation set according to the strategies listed in Table 3.1.

The underlying imputation model (for missing predictor X_1) will be the same for both regression imputation and multiple imputation, as they are both based on a linear regression model. The key difference between the two methods (for the implementation used in this study) is that multiple imputation allows for multiple draws from this same imputation model, allowing the parameter estimates to account for the uncertainty associated with the missing data handling strategy.

The "passive imputation" approach is used here to account for the $X_1 : X_2$ interaction in the analysis model, whereby a value (X_1^*) is imputed for X_1, and

Table 3.1: Imputation and validation strategies

Strategy	Description	Missingness Allowed at Deployment
$DA+VA$	All data, before inducing missingness, at development and validation	No
$DY+VY$	With Y in imputation model at development and validation	No
$D\bar{Y} + V\bar{Y}$	Without Y in imputation model at development and validation	Yes
$DY+V\bar{Y}$	With Y in imputation model at development, but not at validation	Yes
$DY+VA$	With Y in imputation model at development, all (completed) data required at validation	No
$D\bar{Y}+VA$	Without Y in imputation model at development, all (completed) data requied at validation	No

$X_1 : X_2$ is calculated as $X_1^* : X_2$. This results in an imputation model that is not congenial with the analysis model which can result in biased parameter estimates. Due to the lack of availability of the outcome Y at the point of prediction, any imputation model (based on MI) will also be uncongenial at the point of prediction, therefore achieving congeniality was not a primary consideration in the implementation of MI in this particular case. Other methods could, however, have been considered to mitigate any bias introduced by passively imputing the $X_1 : X_2$ interaction term such as the “just another variable” (JAV) approach, which separately imputes a value for $X_1 : X_2$ as part of the imputation procedure. The JAV approach has been shown to perform slightly better than passive imputation under some circumstances.

3.3.5 Development/Validation scenarios

We apply each of the imputation and validation strategies described above in Table 3.1, $DA+VA$ to $D\bar{Y}+VA$. MI is performed using Bayesian linear regression as the underlying form of the imputation model (as implemented by mice.impute.norm in the mice package in R) and 20 imputed datasets. Parameter estimates are pooled across the the imputed datasets according to Rubin's rules, and, similarly, we take the "pooled performance" approach to validation [18] whereby imputation-specific predictions are obtained in the multiply imputed validation datasets, the predictive performance of each imputed dataset is calculated, and then these estimates of model performance are pooled across the imputed datasets. An alternative strategy would be to first pool the predictions obtained from the validation set, obtaining a single measure of predictive performance for each validation dataset (so-called "pooled predictions"). The chosen "pooled performance" strategy was based on recommendations from Wood, Royston, and White [18], as "pooled predictions" can at times over-estimate model performance. It is, however, perhaps unrealistic to expect that multiple predictions would be obtained when applying a model in practice and therefore the "pooled predictions" method may make more sense when trying to validate the model as it will be used in a real-world setting.

$DY+VA$ and $D\bar{Y}+VA$ can be considered estimates of "performance under no missingness", and to estimate this we retain the fully observed validation data before missing data are induced. $DY+VY$ will be classed as "approximated performance under no missingness", since it attempts to estimate performance assuming no missingness at deployment, but with missing data in the validation set. Note, however, that this strategy could not realistically be applied in a real-world setting (at prediction/implementation time) due to the inclusion of Y during imputation of the validation data.

$D\bar{Y}+V\bar{Y}$ and $DY+V\bar{Y}$ are both strategies that could be applied in practice when missingness is allowed, with the key difference between the two being that Y is omitted from the imputation model at validation and deployment. They therefore correspond to measures of "performance under missingness", assuming this imputation strategy could reasonably be applied at the point of prediction. For this approach, we do not have a true estimand that we are targeting, so the methods will be compared against each other to establish the optimal missing data handling strategy.

The fully observed data strategy in $DA+VA$ will be considered to be the reference approach, since this is equivalent to the data-generating model prior to applying any missing data strategy, and will be used as a comparator for other methods. Strategies that do not allow missingness at deployment will be directly compared against this approach, as will $DY + VY$ since it aims to approximate performance under no missingness.

In strategies $DA+VA$, $DY+VY$, $D\bar{Y}+V\bar{Y}$, we assume that the missingness mechanism, the missing data strategy, and proportion of missing data remain constant between model development and validation, which is perhaps a strong assumption in practice. For strategies that allow missingness at deployment, we assume that the missingness mechanism and proportion of missingness remain constant across the pipeline.

3.3.6 Target and Performance Measures

Our key target is an individual's predicted risk, and we compare each method's ability to estimate this using the following metrics of predictive performance, covering both calibration and discrimination [1, 19]:

- Calibration-in-the-large (CITL) - the intercept from a logistic regression model fitted to the observed outcome with the linear predictor as an offset

- Calibration slope - the model coefficient of the linear predictor from a model fitted to the observed outcome with the linear predictor as the only explanatory variable

- Discrimination (Concordance/C-statistic) - a measure of discriminative ability of the fitted model. Defined as the probability that a randomly selected individual who experienced the outcome has a higher predicted probability than a patient that did not experience the outcome

- Brier score - a measure of overall predictive accuracy, equivalent to the mean squared error of predicted probabilities

We assume that the estimates of the above measures are valid representations of performance at model deployment, based on the following assumptions: 1) when missingness is allowed at deployment, it will be imputed in the same way as performed in our validation set, 2) the missingness mechanism will not change between validation and deployment, and 3) when missingness is not allowed at deployment, we assume that the data-generating mechanism remains constant across validation and deployment.

We also extract the obtained parameter estimates and any associated bias from each fitted CPM, as these will likely provide insight into the performance of the models.

3.3.7 Software

All analyses are performed using R version 3.6.0 or greater [20]. The pROC library [21] was used to calculate C-statistics and the mice package (Buuren and Groothuis-Oudshoorn [22]) was used for all imputations. Code to replicate the simulation can be found in the following GitHub repository: `https://github.com/rosesisk/regrImpSim`.

3.4 Results

Select parameter combinations have been chosen to highlight important results within this chapter, but full results for all combinations are made available in a rShiny dashboard at `https://rosesisk.shinyapps.io/regrimpsim`.

3.4.1 Predictive Performance

For simplicity, we restrict our results to a single parameter configuration for each missingness mechanism. The following parameters remain fixed throughout this section: $\gamma_{X1} = \gamma_{X2} = 0.7$, $\gamma_{X1X2} = 0.1$, $\pi_{R1} = 0.5$.

Inclusion of the Outcome in the Imputation Model

Figure 3.3 summarises the estimated Brier Scores for each imputation strategy (defined in Table 3.1) for both imputation methods and all fitted outcome models, and calibration slopes are presented in Figure 3.4 The imputation strategies have been split according to whether or not they allow missingness at model deployment.

When missingness is allowed at deployment (i.e. imputation will be applied at the point of prediction), we primarily want to know whether imputation should be performed with or without the outcome at development, since at deployment it must be omitted from the imputation model. We observe that predictions

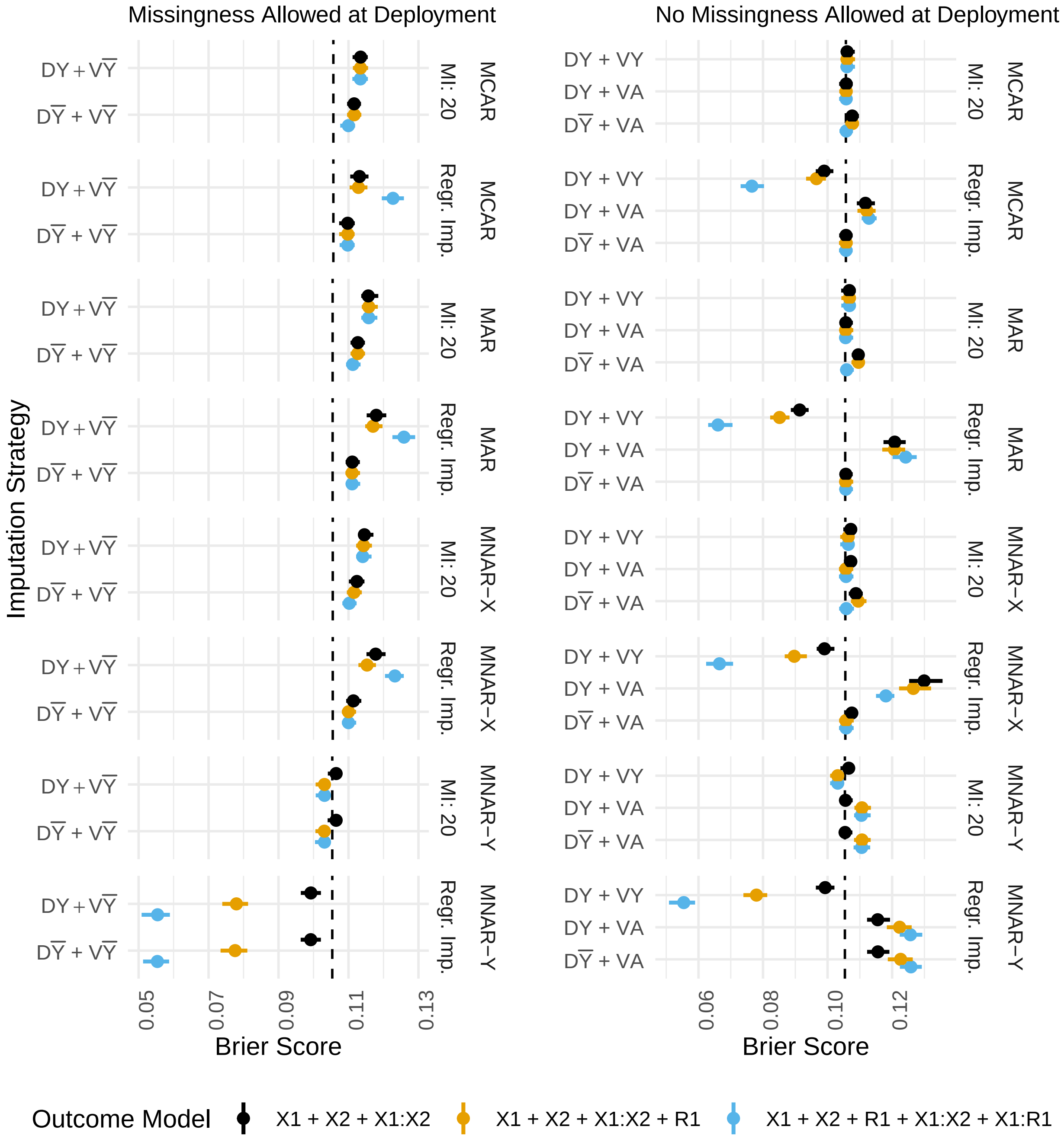


Figure 3.3: Brier Score estimates across Development/Validation scenarios, imputation methods and missingess mechanisms. The vertical dashed lines represent estimates from the complete data scenario (DA + VA)

in the validation set are far better calibrated under $D\bar{Y} + V\bar{Y}$ than $DY + V\bar{Y}$, i.e. when the imputation model remains consistent between development and validation/deployment (Figure 3.4). This difference can also be seen in marginal differences in the Brier Score (Figure 3.3).

When complete data is required at deployment, RI still performs better under $D\bar{Y} + VA$ than $DY + VA$ in terms of calibration (Figure 3.4), and this result becomes more pronounced when data are MAR or MNAR-X at development. MI, on the other hand, favours retention of Y in the imputation model at development in terms of both Brier Score and model calibration (under no missingness at deployment). Figure 3.5 summarises the parameter estimates from the fitted models, and we can see that effect estimates are less biased for RI: No Y and MI: Y, which is in line with the predictive performance estimates.

A notable result is that the $DY + VY$ strategy fails to recover the performance under no missingness ($DY + VA$) under both methods, i.e. when Y is used to impute at both development and validation. Under MI, this strategy should be a valid means of estimating the performance under complete data at deployment, but there are marginal differences in the calibration slope between MI: $DY + VY$ and MI: $DY + VA$. Moreover, the performance of RI considerably breaks down with the inclusion of Y in the imputation model under all missingness mechanisms and regardless of whether missingness is allowed at deployment. This same result is evident in Figure 3.5): Parameter estimation, where RI: Y consistently fails to recover unbiased parameter estimates.

Comparison of Imputation Methods

Overall, the performance estimates from MI are considerably more stable than those from RI; the differences in Brier Score between the various imputation strategies are smaller for MI as can be seen from Figure 3.3). When RI performs poorly, the poorer model tends to be even worse than the worst MI model, in terms of both Brier Score and calibration. RI does, however, often perform at least as well as, or better than, MI when the preferred imputation model is applied (i.e. omitting the outcome when applying RI). For example, both methods perform comparably under $D\bar{Y} + V\bar{Y}$. With missingness permitted at deployment, performance is comparable between MI: $DY + VA$ and RI: $D\bar{Y} + VA$.

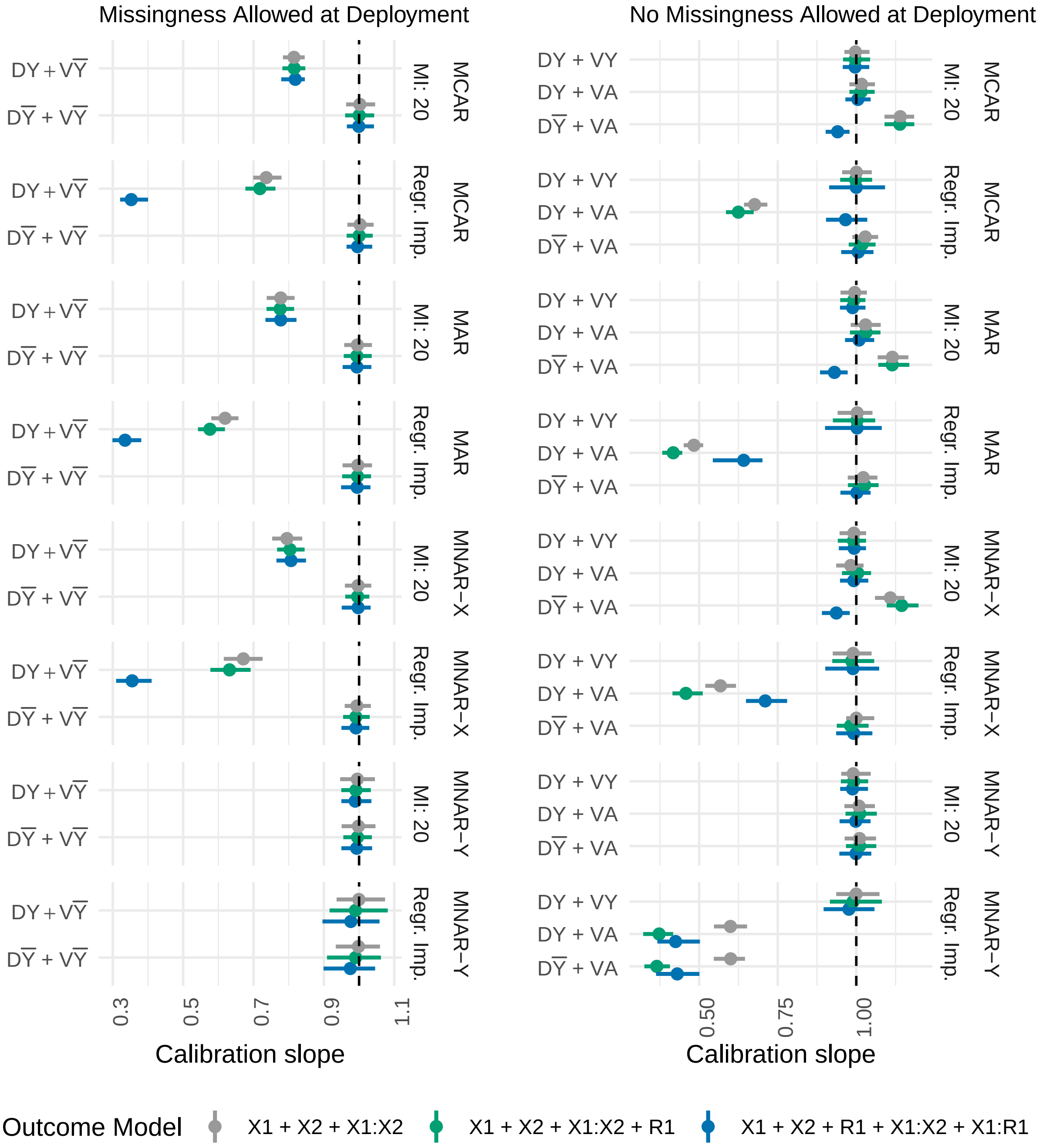


Figure 3.4: Calibration Slope estimates across imputation strategies, imputation methods and missingness mechanisms. Vertical dashed lines are placed at 1.

Inclusion of a Missing Indicator

The inclusion of a missing indicator appears to have minimal impact on the Brier Score and calibration under most methods and imputation strategies, with a few notable exceptions.

Missingness allowed at deployment Under MNAR-Y and missingness allowed at deployment, inclusion of a missing indicator in the outcome model provides reductions in the Brier Score, and improvements in the C-statistic (for both imputation methods, C-statistics presented in the supplementary material Figure 3.7), performing even better than the complete data model, since the inclusion of the indicator allows the CPM to extract additional information about the outcome that is not available from only observed data. We can further see from Figure 3.5 (parameter estimates) how the estimates of the intercept ($\hat{\gamma}_0$) are less biased for both methods when the indicator is included under this mechanism.

Under this same missingness mechanism (MNAR-Y), inclusion of the indicator and its interaction with $X1$ produce marginally overfit models for RI (Calibration slope < 1, Figure 3.4). This result is explored further in Figure 3.8 in the supplementary materials, where we present plots of the predicted risk distributions - we see that this method produces predictions that are very close to 0 and 1.

No missingness allowed at deployment Inclusion of the indicator corrects the CITL for both MI: $DY + VA$ and RI: $D\bar{Y} + VA$ under MAR and MNAR-X structures when missingness is not allowed at deployment (Figure 3.6 in Supplementary material). Further improvements in the calibration slope can be achieved through inclusion of the additional $X_1 : R_1$ term under the preferred imputation model for both methods. Conversely, under MNAR-Y and fully observed data at deployment, inclusion of a missing indicator results in underestimated average predicted risk (CITL > 0) for both imputation methods, and severe overfitting for RI (calibration slope < 1, Figure 3.4). Models developed using MI under MNAR-Y, are, however generally still well calibrated (slope close to 1) whether or not missing data are allowed at deployment.

Interestingly, inclusion of the $X_1 : R_1$ interaction term where only complete data will be used at deployment appears to (marginally) improve the calibration slope for MI under all missingness mechanisms. Clearly, this term would be 0 for all new individuals however its inclusion at development seems to aid model performance at deployment.

3.4.2 Parameter Estimation

We further present results for the CPM parameter estimates for selected scenarios presented in Figure 3.1. The same parameter configurations as specified in previous sections are used here. Presented are the coefficients obtained from

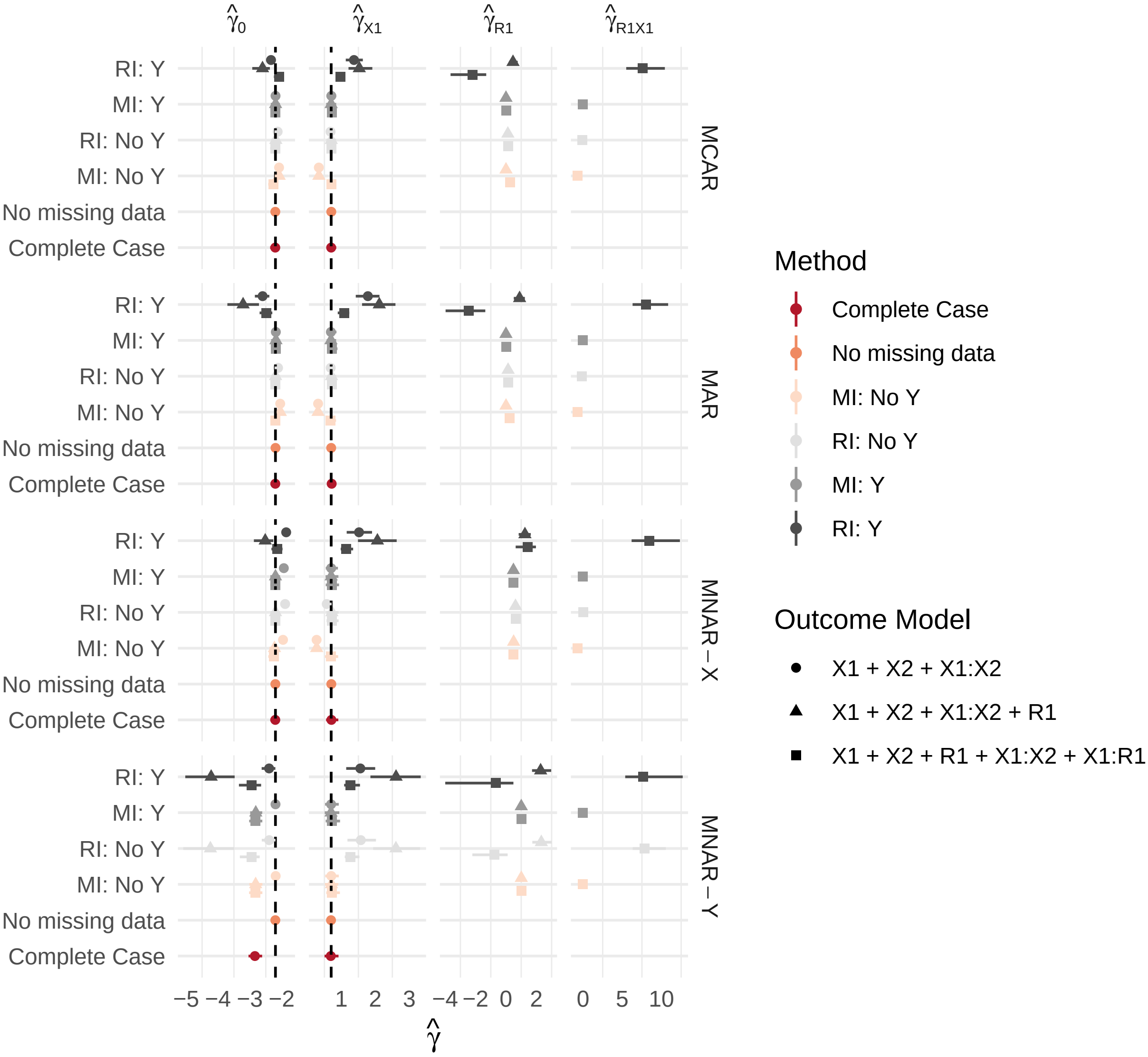


Figure 3.5: Parameter estimates across all missingness mechanisms. Missingness is fixed at 50%. The dashed lines represent the true parameter values under the data generating mechanism.

fitting each model within the development data. For MI, we present coefficients pooled according to Rubin's rules. RI using the outcome in the imputation model consistently produces parameter estimates that are both biased and much larger in magnitude than any other method, and this is reflected in the predictive performance estimates - this method fails to produce models that are well calibrated or have good overall accuracy (Brier Score). We frequently observe Brier Scores (Figure 3.3) that are too extreme under $DY + VY$, and at times calibration slopes that suggest the predicted risks are too extreme. Both of these are indicative of overfitting and likely due to the relative size of the parameter estimates compared to other methods. RI: $D\bar{Y}$ generally estimates parameters with minimal bias, with the exception of under MNAR-Y where, similarly, parameter estimates are large and generally very biased. This is further reflected in the Brier Scores (Figure 3.3) and predicted risk distribution (Figure 3.8 in supplementary material).

Perhaps as expected, MI: $D\bar{Y}$ fails to recover unbiased parameter estimates under all missingness mechanisms, whereas MI: Y generally recovers the true parameter values well. It can therefore be observed that biased parameters do not necessarily result in worse predictive performance, as models developed under MI: No Y, and missing data imputed in this same way at validation/deployment were favoured over models developed using the outcome at development.

The inclusion of a missing indicator reduces bias in the estimate of the effect of $X1$ on Y (γ_{X1}) under both MNAR structures for MI: Y.

3.5 Discussion

In this study, we have assessed model performance for multiple imputation and regression imputation, with and without the use of missing indicators across a range of missingness mechanisms. We considered how/when the outcome should be used in the imputation model for missing covariates, whether RI could offer a more practical and easier to implement solution than MI, and how the inclusion of missing indicators affects model predictive performance. All of these questions were considered in relation to whether or not missing data will be allowed once the model is deployed in practice. We have provided a concise list of recommendations in Table 3.2.

In the context of recovering unbiased parameter estimates, the literature advocates the use of the observed outcome in the imputation model for MI [7]. In the context of predictive performance, we found that RI consistently performed better when Y is instead omitted from the imputation. This strategy is recommended by Steyerberg [23] for RI, where the author notes that including the outcome in the imputation model artificially strengthens the relationship between the predictors and outcome. MI overcomes this issue by introducing an element of randomness to the imputation procedure.

We further observed that the performance of a model with inconsistent imputation models between development and validation $(DY + V\bar{Y})$ performed worse than one where the imputation model remained consistent. For instance, omitting the observed outcome at both stages resulted in better predictive performance, even when using MI. Although we have also observed that the inclusion of Y in the imputation model helps in recovering unbiased effect estimates, others have recommended a more considered approach when targeting a model that allows for missing data at prediction time. In a simulation study conducted by Fletcher Mercaldo and Blume [15], multiple imputation including the outcome produced larger overall prediction error than omitting the outcome entirely, as the out-of-sample imputations were biased by attempting to use the imputation model derived during development (with Y) to impute in the test set (where Y is unobserved). This imputation bias carried through to the overall performance of the model. This is in line with our findings, whereby a consistent imputation model between development and deployment resulted in stronger performance overall, at the cost of slightly biased parameter estimates. An interesting result to note here is that a model with unbiased model parameters is not necessarily the one that predicts best, especially when data will be imputed again at deployment.

We have demonstrated that RI could offer a practical alternative to multiple imputation within the context of prediction. As discussed above, there are several challenges associated with applying multiple imputation during deployment of a CPM, including but not limited to: requiring access to the development data and the availability of computational power and time. Recent developments, have, however proposed methods that potentially mitigate these requirements [16]. RI also overcomes both of these major issues, in that only the deterministic imputation models would be required to produce imputations during model implementation. We emphasize, however, that RI consistently showed poor performance when the observed outcome was included in the imputation model, and this method should therefore only be used when other model covariates are used to impute missing ones. MI, on the other hand, proved to be more stable across

Table 3.2: Table of recommendations for the use of multiple imputation, regression imputation and missing indicators in the development and deployment of clinical prediction models

Recommendations
Assuming no missing data will be present at deployment, multiple imputation (including the outcome) is recommended as the best strategy
Where missingness is allowed at deployment, and multiple imputation is impossible at the point of prediction, regression imputation can be used as an alternative.
Always omit the outcome from the imputation model under regression imputation.
Where data are assumed to be MNAR-X or MNAR-Y, inclusion of a missing indicator can offer marginal improvements in model performance, and does not harm performance under MCAR or MAR mechanisms
The use of missing indicators under MNAR-Y can harm model performance when missingness is not allowed at deployment, and is not recommended

a range of scenarios and imputation models.

The careful use of missing indicators has also proven to be beneficial in specific cases. For example, under MNAR-X, multiple imputation has marginally stronger performance in both imputed and complete deployment data when a missing indicator is included in the outcome model. Under incomplete data at deployment, inclusion of an indicator further provided small improvements in overall predictive accuracy to MI under MNAR-Y, but resulted in some overfitting for models developed and applied using RI. Since MI is only assumed to recover unbiased effects under MAR, the indicator appears to correct this bias under informative missingness patterns. We observe further (marginal) improvements in model performance for MI with the inclusion of an interaction between X_1 and R_1 when missingness is not allowed at deployment, and when Y is omitted from the imputation model at development ($D\bar{Y} + VA$). However, we noted some surprising results in the use of indicators when data are MNAR-Y; specifically when missing data are not allowed at deployment the inclusion of the indicator is harmful and resulted in small increases in the Brier Score, and poor Calibration-in-the-large. Related literature under a causal inference framework by Sperrin et al. [4] and Groenwold et al. [24] has found that the inclusion of missing indicators is not recommended under MCAR, and can lead to biased parameter estimates under this missingness structure. Smeden, Groenwold, and Moons [25] discuss at length how missing indicators should be approached with caution in predictive modelling - inclusion of a missing indicator introduces an additional assumption that the missingness mechanism remains stable across the CPM pipeline; an assumption that is generally dubious, but especially within routinely collected health data. The propensity to measure certain predictors is likely to vary across care providers, settings and over time as clinical guidelines and practices change. This in turn potentially changes the relationship between the missing indicator and the outcome and could have implications for model performance. As others have highlighted, the strategy to handle missing data should be devised on an individual study basis, taking into consideration the potential drivers of missingness, how stable these are likely to be, and how/whether missing data will be handled once the model has been deployed.

We recommend that the strategy for handling missing data during model validation should mimic that to be used once the model is deployed, and that measures of predictive performance be computed in either complete or imputed data depending on whether missingness will be allowed in the anticipated deployment setting or not. For example, complex model applications integrated into electronic health record systems are better suited to applying imputation strategies, whereas simple models that must be filled in at the point-of-care are more likely to require a complete set of predictors. The difference between performance allowing for and prohibiting missing data at deployment may also be of interest in assessing any drop in performance related to the handling of missing data. Interestingly, we have observed somewhat different results depending on whether missingness is allowed at deployment or not. We may therefore wish to optimize a model for either one of these use cases, resulting in a different model (and hence different coefficients) dependent on whether we envisage complete data at implementation.

Although we have considered a wide range of simulated scenarios, a key limitation to this study is that we only considered a relatively simple CPM with two covariates, where only one was allowed to be missing. This was to restrict the complexity and size of the work, as only a limited set of scenarios can realistically be presented. We do, however, expect that the fundamental findings would generalise to more complex models, since we could consider each of the two model predictors to represent some summary of multiple missing and observed predictors. With more predictors in the model, there would not be any additional missingness mechanisms and we therefore anticipate that such complex models would not provide any additional insight. A further possible limitation is that this work has been restricted to the study of a single binary outcome, although we would not expect the results to change in the context of e.g. continuous or time-to-event outcomes. We accompany this work with a rShiny dashboard allowing readers to explore our results in more detail across the entire range of parameter configurations.

Avenues for further work would include exploring the impact of more complex patterns of missingness in multiple predictor variables. As the number of incomplete predictors increases, so does the number of potential missing indicators eligible for inclusion in the outcome model, which could introduce issues of overfitting [26, 27], and variable selection becomes challenging. Here we have also limited our studies to scenarios where the missingness mechanism remains constant between development and deployment, however it would be interesting to explore whether these same results hold if the mechanism were to change between the two stages.

3.6 Conclusion

We have conducted an extensive simulation study, and found that when no missingness is allowed at deployment, existing guidelines on how to impute missing data at development are generally appropriate. However, if missingness is to be allowed when the model is deployed into practice, the missing data handling strategy at each stage should be more carefully considered, and we found that omitting the outcome from the imputation model at all stages was optimal in terms of predictive performance.

We have found that RI performs at least as well as multiple imputation when the outcome is omitted from the imputation model, but tends to result in more unstable estimates of predictive performance. Missing indicators can offer marginal improvements in predictive performance under MAR and MNAR-X structures, but can harm model performance under MNAR-Y.

We recommend that if missing data is likely to occur both during the development and implementation of a CPM, that RI be considered as a more practical alternative to multiple imputation, and that if either imputation method is to be applied at deployment, the outcome be omitted from the imputation model at both stages. Model performance should be assessed in such a way that reflects how missing data will occur and be handled at deployment, as the most appropriate strategy may depend on whether missingness will be allowed once the model is applied in practice. We also advocate for the careful use of missing indicators in the outcome model if MNAR-X can safely be assumed, but this should be assessed on a study-by-study basis since the inclusion of missing indicators also has the potential to reduce predictive performance, especially when missingness is not permitted at deployment.

References

[1] E. W. Steyerberg and M. van Veen. *Imputation is beneficial for handling missing data in predictive models.* Sept. 2007. DOI: `10.1016/j.jclinepi.2007.03.003`.

[2] M. van Smeden et al. “Clinical prediction models: diagnosis versus prognosis”. In: *Journal of Clinical Epidemiology* 132 (Apr. 2021), pp. 142–145. DOI: `10.1016/J.JCLINEPI.2021.01.009`.

[3] A. Tsvetanova et al. “Missing data was handled inconsistently in UK prediction models: a review of method used”. In: *Journal of Clinical Epidemiology* 140 (Sept. 2021), pp. 149–158. DOI: `10.1016/J.JCLINEPI.2021.09.008`.

[4] M. Sperrin et al. *Missing data should be handled differently for prediction than for description or causal explanation.* Sept. 2020. DOI: `10.1016/j.jclinepi.2020.03.028`.

[5] J. Hippisley-Cox, C. Coupland, and P. Brindle. “Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: Prospective cohort study”. In: *BMJ (Online)* 357 (May 2017). DOI: `10.1136/bmj.j2099`.

[6] J. Hippisley-Cox and C. Coupland. “Predicting the risk of Chronic Kidney Disease in Men and Women in England and Wales: prospective derivation and external validation of the QKidney®Scores”. In: *BMC Family Practice 2010 11:1* 11.1 (June 2010), pp. 1–13. DOI: `10.1186/1471-2296-11-49`.

[7] K. G. Moons et al. “Using the outcome for imputation of missing predictor values was preferred”. In: *Journal of Clinical Epidemiology* 59.10 (Oct. 2006), pp. 1092–1101. DOI: `10.1016/J.JCLINEPI.2006.01.009`.

[8] S. van Buuren. *Flexible Imputation of Missing Data. Second Edition.* CRC/Chapman & Hall, FL: Boca Raton, 2018.

[9] S. Haneuse and M. Daniels. "A General Framework for Considering Selection Bias in EHR-Based Studies: What Data Are Observed and Why?" In: *EGEMS (Washington, DC)* 4.1 (2016), p. 1203. DOI: 10.13063/2327-9214.1203.

[10] N. G. Weiskopf, A. Rusanov, and C. Weng. "Sick patients have more data: the non-random completeness of electronic health records." In: *AMIA Annual Symposium proceedings*. Vol. 2013. American Medical Informatics Association, 2013, pp. 1472–7.

[11] R. Sisk et al. "Informative presence and observation in routine health data: A review of methodology for clinical risk prediction". In: *Journal of the American Medical Informatics Association* 28.1 (Jan. 2021), pp. 155–166. DOI: 10.1093/jamia/ocaa242.

[12] T. P. Morris, I. R. White, and M. J. Crowther. "Using simulation studies to evaluate statistical methods". In: *Statistics in Medicine* 38.11 (May 2019), pp. 2074–2102. DOI: 10.1002/sim.8086.

[13] F. E. Harrell, K. L. Lee, and D. B. Mark. "Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors". In: *Statistics in Medicine* 15 (1996), pp. 361–387. DOI: 10.1002/(SICI)1097-0258(19960229)15:4.

[14] K. J. M. Janssen et al. "Dealing with Missing Predictor Values When Applying Clinical Prediction Models". In: *Clinical chemistry* 55.5 (2009), pp. 994–1001. DOI: 10.1373/clinchem.2008.115345.

[15] S. Fletcher Mercaldo and J. D. Blume. "Missing data and prediction: the pattern submodel". In: *Biostatistics* (Sept. 2018). DOI: 10.1093/biostatistics/kxy040.

[16] S. W. J. Nijman et al. "Real-time imputation of missing predictor values improved the application of prediction models in daily practice". In: *Journal of Clinical Epidemiology* 134 (June 2021), pp. 22–34. DOI: 10.1016/j.jclinepi.2021.01.003.

[17] M. Sperrin and G. P. Martin. “Multiple imputation with missing indicators as proxies for unmeasured variables: Simulation study”. In: *BMC Medical Research Methodology* 20.1 (July 2020), p. 185. DOI: 10.1186/s12874-020-01068-x.

[18] A. M. Wood, P. Royston, and I. R. White. “The estimation and use of predictions for the assessment of model performance using large samples with multiply imputed data”. In: *Biometrical Journal* 57.4 (July 2015), pp. 614–632. DOI: 10.1002/bimj.201400004.

[19] E. W. Steyerberg and Y. Vergouwe. *Towards better clinical prediction models: Seven steps for development and an ABCD for validation.* Aug. 2014. DOI: 10.1093/eurheartj/ehu207.

[20] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing. Vienna, Austria, 2020.

[21] X. Robin et al. “pROC: an open-source package for R and S+ to analyze and compare ROC curves”. In: *BMC Bioinformatics* 12 (2011), p. 77. DOI: 10.1186/1471-2105-12-77.

[22] S. van Buuren and K. Groothuis-Oudshoorn. “{mice}: Multivariate Imputation by Chained Equations in R”. In: *Journal of Statistical Software* 45.3 (2011), pp. 1–67. DOI: 10.18637/jss.v045.i03.

[23] E. Steyerberg. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating.* Second. Springer, 2019. DOI: 10.1007/978-0-387-77244-8.

[24] R. H. Groenwold et al. “Missing covariate data in clinical research: When and when not to use the missing-indicator method for analysis”. In: *CMAJ* 184.11 (Aug. 2012), pp. 1265–1269. DOI: 10.1503/cmaj.110977.

[25] M. van Smeden, R. H. Groenwold, and K. G. Moons. “A cautionary note on the use of the missing indicator method for handling missing data in prediction research”. In: *Journal of Clinical Epidemiology* 125 (Sept. 2020), pp. 188–190. DOI: 10.1016/j.jclinepi.2020.06.007.

[26] R. D. Riley et al. “Calculating the sample size required for developing a clinical prediction model”. In: *BMJ* 368 (Mar. 2020). DOI: 10.1136/BMJ.M441.

[27] M. van Smeden et al. “Sample size for binary logistic prediction models: Beyond events per variable criteria”. In: *Statistical Methods in Medical Research* 28.8 (Aug. 2019), pp. 2455–2474. DOI: 10.1177/0962280218784726.

3.7 Supplementary Materials for Chapter 3

Here we present the remaining performance metrics that were omitted from the main text, C-statistic and Calibration-in-the-large. We further present some additional post-hoc analysis in an attempt to explain some of the unexpected results.

The results from the entire range of parameter estimates can be explored via an interactive dashboard at: `https://rosesisk.shinyapps.io/regrimpsim`.

3.7.1 Calibration-in-the-large (calibration intercept)

We first present the computed performance metrics not included in the main manuscript: Calibration-in-the-large (calibration intercept) and the C-statistic. The calibration intercept is defined as the intercept from a model fitted to the observed outcome with the linear predictor as an offset, and has a target value of 0. As in the main manuscript, the following parameters remain fixed throughout this section: $\gamma_{X1} = \gamma_{X2} = 0.7$, $\gamma_{X1X2} = 0.1$, $\pi_{R1} = 0.5$.

The results in Figure 3.6 are in line with the results of the other performance metrics - the calibration intercept is further away from 0 for models that use inconsistent imputation models between development and deployment. The inclusion of a missing indicator also appears to improve calibration when missingness is not allowed at deployment under MAR mechanisms, but can be harmful under MNAR-Y mechanisms.

3.7.2 C-statistic

Figure 3.7 shows the results for the C-statistic, using the same fixed parameter values as in the main text. Here we see that missing indicators can offer improvements in model discrimination under MNAR-Y (as expected) when used in combination with multiple imputation, but can cause some over-fitting when used in combination with regression imputation in this same context, as suggested by the calibration slopes that are slightly < 1 and the increased C-statistic from these models (MNAR-Y, regression imputation, missingness allowed at deployment).

3.7.3 Predicted risk distribution

In 3.8 we present the distribution of the predicted risks, obtained from a single simulated dataset. The models were generated and validated omitting Y from

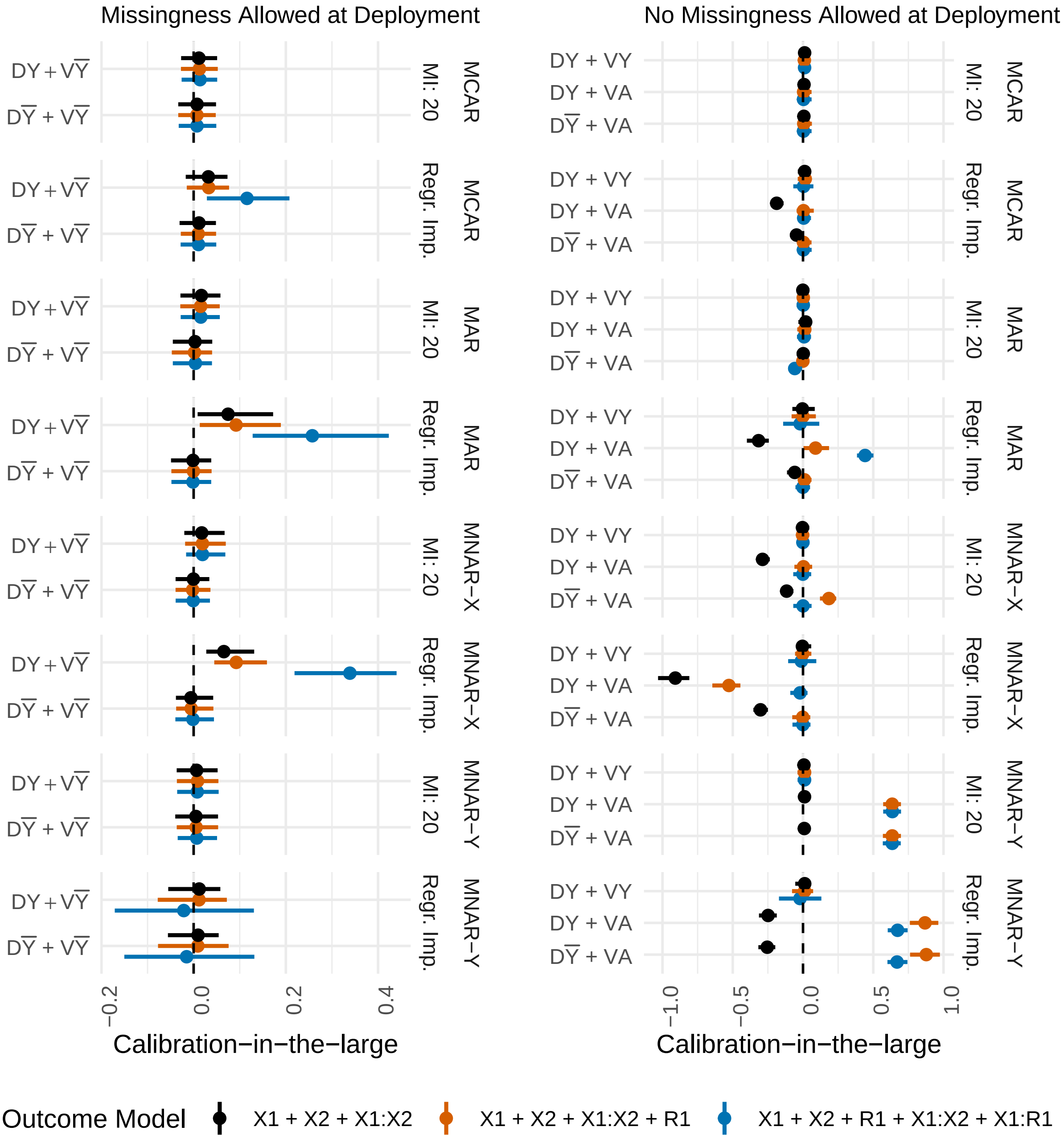


Figure 3.6: Calibration-in-the-large estimates across imputation strategies, imputation methods and missingness mechanisms

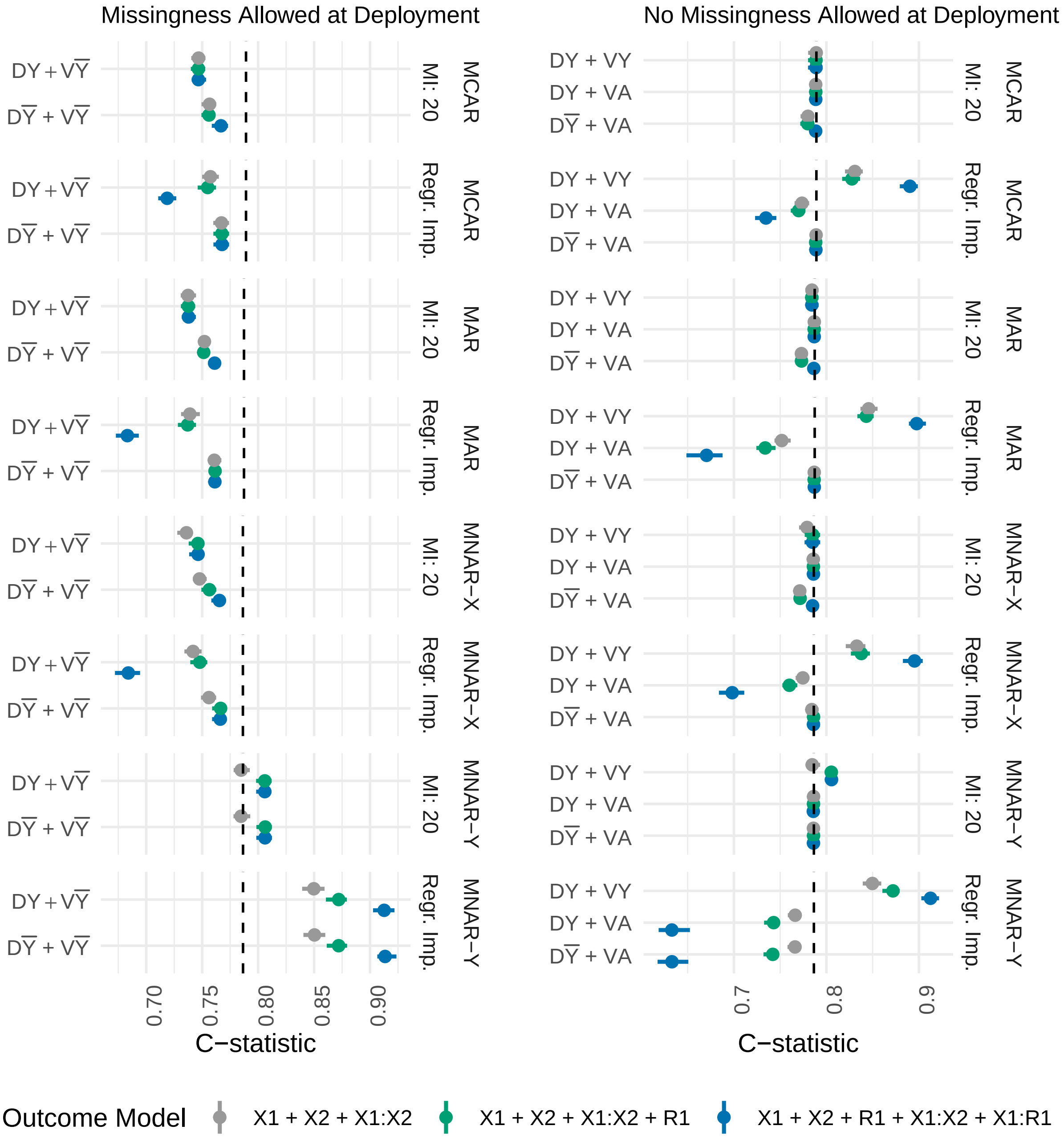


Figure 3.7: C-statistic estimates across imputation strategies, imputation methods and missingness mechanisms

the imputation model at both stages.

This was a post-hoc exploratory analysis, to explore the reason behind the seemingly optimistic performance estimates within this mechanism and missingness strategy. We can see from the plots that the predicted risks are quite extreme, with many very close to 0, and for the model containing the $X_1 : R_1$ term, another peak towards 1. This would suggest that the model is overfit.

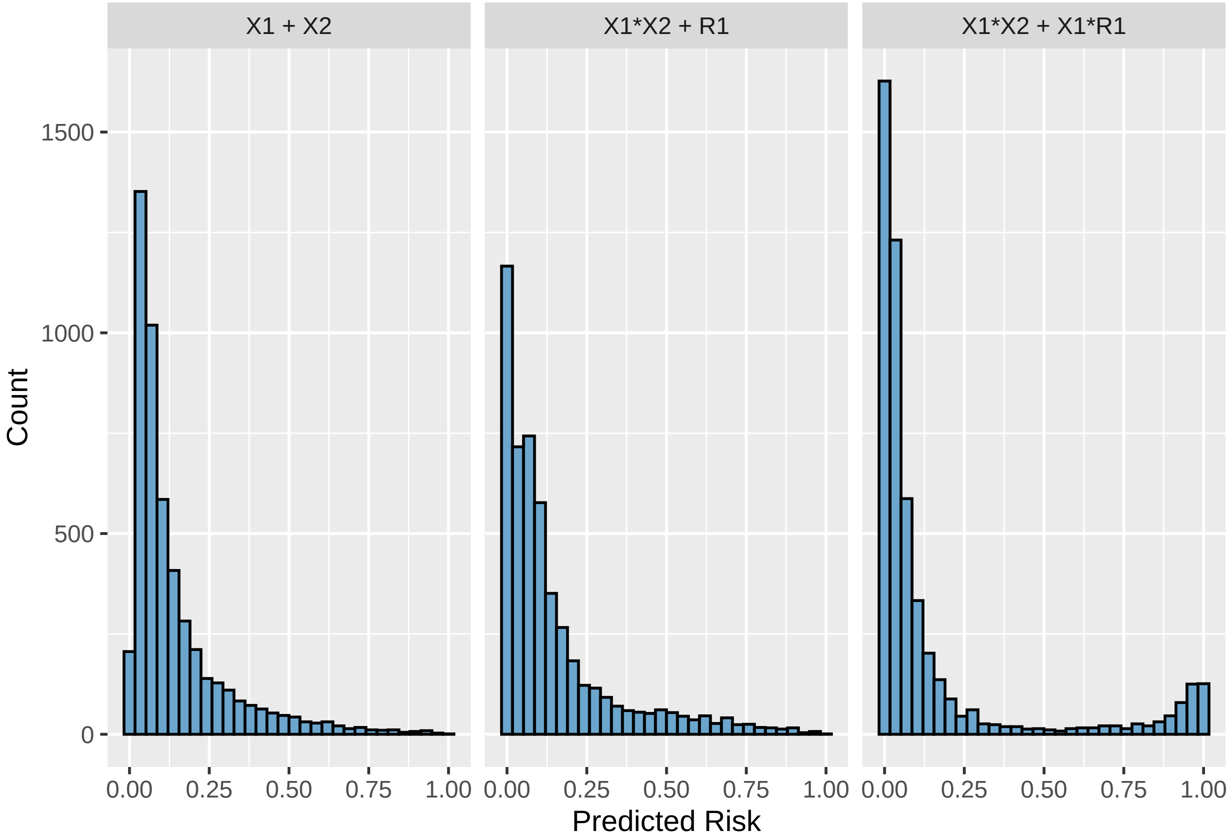


Figure 3.8: Predicted risk distribution from model developed using regression imputation under MNAR-Y, missingness allowed and imputed at deployment, imputation model omitting the outcome at both stages

Chapter 4

Harnessing informative patterns of eGFR measurement for improved prediction of key outcomes in chronic kidney disease

The work in this chapter applies and compares methodology falling under the "derived predictors" category described in the scoping review in Chapter 2. We demonstrate a range of different ways of summarising the observation process into a single measure that can be included as a predictor in a landmarking framework, and we quantify any gains in predictive performance as a result of doing so. The work also draws on ideas from the joint modelling literature, where informative observation processes have been modelled as a separate "process" within the joint model via a recurrent events model. One of the summary measures derived in this work is the estimated frailty term from a recurrent events model fit to the gaptimes between observations, which fits nicely into the two-stage landmarking framework.

4.1 Abstract

4.1.1 Background

Assessing risk of chronic kidney disease (CKD) progression is a key part of the clinical decision-making process. A widely used tool for this task is the Kidney Failure Risk Equation. Although well accepted and validated, the tool is limited to cross sectional information only. There is the potential for the model to learn from not only the longitudinal evolution of eGFR, but its patterns of measurement over time, so called “informative observation”.

We therefore adapt the existing model to firstly incorporate repeated eGFR measurements, and secondly to learn from the timings and frequency of repeated eGFR measures. The goal of this work is to apply and empirically compare different ways of summarising potentially informative measurement processes in the context of clinical prediction using EHR data, through the clinical exemplar of chronic kidney disease progression.

4.1.2 Methods

We adopt a two-stage landmarking approach - at the first stage, the longitudinal eGFR process is modelled, and at the second, predictions from the eGFR model are used to predict time to kidney failure/death. We also derive representations of the observation process that can be incorporated into prediction models with ease in contexts where measurement patterns are potentially informative with respect to patient condition. We develop and validate Cox models for the prediction of kidney failure and death using data observed in Grampian, Scotland, UK between 2009 and 2014, and estimate model performance metrics via bootstrapping. We compare C-statistics, Brier Scores and calibration slopes across models.

4.1.3 Results

We observe some improvements in model discrimination by incorporating informative measurement patterns into models for the prediction of death (C-statistic 0.689 [0.686 0.697] vs. 0.735 [0.732, 0.738] at landmark time 1.5 years), but no major differences in model calibration. There appears to be no benefit or harm from incorporating informative observation into models for the prediction of kidney failure, but models for this outcome were based on a very small number of

events.

4.1.4 Conclusion

We have demonstrated how clinical prediction models could learn from informative observation within a landmarking framework. We have observed small gains in model discrimination as a result of doing so, and illustrated a broad range of different possible parametrisations that could be applied to other clinical settings. We recommend working alongside clinical experts to identify the most informative elements of the observation process with respect to risk for possible inclusion in prediction models.

4.2 Background

Chronic Kidney Disease (CKD) is associated with a high risk of progression to kidney failure and mortality, and the prevalence (and therefore burden) of CKD is increasing globally [1–3]. Establishing prognosis on an individual-level is therefore essential to make informed decisions about required care, reducing this burden on both patients and healthcare systems globally. Assessing the risk of someone developing kidney failure enables timely referral to nephrologists and planning for interventions such as kidney replacement therapy. Various clinical prediction models (CPMs) have been developed to assess the future risk of kidney failure within patients with Chronic Kidney Disease, such as the Kidney Failure Risk Equation (KFRE) [4].

These models typically take the most recent observation of each risk factor as an input, for example estimated glomerular filtration rate (eGFR), and estimate risk of kidney failure based on this static snapshot of an individual's status. The use of a static measure lacks face validity and is potentially missing out on useful information that would improve prediction. Furthermore, clinicians typically evaluate the trajectory of blood test measures over time to separate those who appear to be have a rapid decline in kidney function ("rapid progressors") from those who have stable or slow decline ("slow progressors"). In addition, a clinician and patient will typically re-evaluate their shared decisions in light of new test results as they become available. These dynamic aspects of kidney function evaluation over time have not been correctly incorporated into current validated tools for kidney failure prediction.

Patients are only monitored as required to support routine care based on the perceived requirements of the patient and judgements of healthcare professionals, introducing additional complexity into the development of CPMs from routinely collected health data such as electronic health records (EHRs). In this instance, data are not recorded with a pre-defined regularity, but instead at irregular intervals according to clinical need or indication. This results in the phenomenon known as "informative observation", whereby the frequency and potentially timing of patient observations are informative with respect to their health status [5, 6]. In the context of kidney failure, this could mean that key lab tests are ordered more frequently when there are other indications of poor kidney function, or ill health more broadly. Or alternatively, individuals with relatively normal and stable kidney function are seen much less frequently.

The development of clinical prediction models from rich, longitudinal EHR data presents numerous opportunities for improved performance of such models. Firstly, longitudinal trajectories that may be informative with respect to patient prognosis can be incorporated as model predictors. The potential of integrating such (more complex "dynamic") models into existing clinical computer systems allow the use of models that would previously have been too difficult to use in the clinic [7], due to the requirement for clinicians to manually enter the longitudinal predictor information. Second, it seems plausible that the features of an informative observation process can be exploited to improve the predictive performance of a CPM [6]. We hypothesise this will be true if the observation process acts as a proxy for unmeasured prognostic factors (observed by the healthcare provider, but not reported in the EHR) that trigger clinicians to monitor more intensively. For example, the patient may report mild symptoms that cause their care provider some concern, or they may present with comorbidities that are not directly recorded in the record. Even if the resulting tests come back normal, the increase in monitoring frequency could allow us to infer a period of ill health (or "clinician concern"), and in turn capture the effect this has on the individual's prognosis.

Two methods dominate the literature in the development of prediction models that allow repeatedly measured predictor values: joint models [8, 9] and landmarking [10]. Both methods have both been proposed as a methodologically sound means of allowing CPMs to update as new information becomes available. Asar et al. [11] have provided a tutorial on the joint modelling approach using the same clinical exemplar as this study: prediction of Renal Replacement Therapy in Chronic Kidney Disease patients using longitudinally measured eGFR values, however they did not consider informative measurement times. Indeed, the impact of incorporating informative observation into prediction models more generally has not yet been explored in real-world data.

Choi et al. [12] propose an extension to the joint modelling framework that allows a separate recurrent events process to model the timings of repeat visits. A similar approach has been discussed by Gasparini et al. [13], whereby informative measurements are jointly modelled with the predictor process via a nonhomogeneous Poisson process. However, JMs carry a heavy computational burden in fitting and applying them. This is especially challenging in the case of EHR data where large sample sizes are common. Conversely, most existing implementations of landmarking only consider the most recently observed value of a longitudinal predictor, and assumes that it is measured without error. In head-to-head comparisons between joint modelling and landmarking, joint models have often been observed to perform better than landmark models[14], however improvements tend to be marginal, and these have mainly been observed in relatively straightforward simulated scenarios. In this study, we therefore adopt the two-stage landmarking approach discussed by Paige et al. [15] and Keogh et al. [16] that separately models the longitudinal biomarker process within landmark time specific cohorts, and uses the first model to predict the biomarker value at each landmark time.

Counting the number of measurements of an informatively observed predictor has been proposed as a method of allowing CPMs to learn from informative observation [17], as well as more complex summaries that directly model the gaptimes between observations within a joint modelling framework [12, 13]. But these parametrisations of the observation process have never been compared through empirical analyses. We therefore further extend the two-stage landmark approach to include and compare different parametrisations that summarise the intensity of eGFR monitoring over time for an individual. We further allow the model to learn from the longitudinal trajectory, capturing differences in prognosis between rapid and slow progressors.

Using kidney failure as a clinical exemplar, the aims of this study are threefold: firstly to demonstrate how informative observation can be incorporated into a two-stage landmarking framework; second to investigate whether directly modelling the observation process within the CPM can lead to improved predictive performance, and finally to explore whether more complex representations of the observation process provide additional gains in predictive performance compared to simpler approaches. Our goal is not to develop a new CPM for clinical use, but rather explore the methods using this setting as an example, the results of which could then be incorporated into future CPM development studies.

4.3 Methods

4.3.1 Data & Study Population

To illustrate our methods, we use the functional form of the established Kidney Failure Risk Equation (KFRE) [4], a model designed to predict chronic kidney disease progression. We use the same four predictors from the simplified version of the KFRE: Age, Sex, eGFR, Urine Albumin to Creatinine Ratio (UACR). In this context, we aim to demonstrate the added value of incorporating longitudinally and informatively observed eGFR data. We therefore explicitly model the longitudinal trajectory of this variable (eGFR), and use a last-observation carried forward approach to modelling UACR. Where missing, UACR will be set to 0, since it is common to only request a formal lab test following an abnormal dipstick test in the clinic, and 0 can be viewed as a "normal" result (i.e. there is no protein in the urine). Under the LOCF approach, this will work as follows: if a patient has never had a UACR recording (prior to the landmark time), then this predictor will be set to 0 until a UACR observation is made.

Our dataset consists of all patients with blood tests taken in Grampian (region of Scotland, UK) in primary or secondary care, between 2009 and 2014. Only those with incident CKD stages G3 and G4 will be included in our cohort to reflect the intended users of the KFRE (patients already in stage G5 at baseline will be excluded). CKD stages G3 and G4 will be defined as persistent eGFR $< 60ml/min/1.72m^2$ for at least 90 days, and patients will be followed up from the first recorded date of CKD within the data. To ensure we only capture new incidences of CKD (as opposed to prevalent cases), we only include patients with a first recording of CKD from 2011 onwards, omitting those patients that first showed evidence of stage G3/G4 during the first two years of follow-up. This "burn-in" period aims to distinguish between incident and prevalent cases.

The procedure for identifying onset of CKD is as follows:

1. Extract all observed eGFR values $< 60ml/min/1.72m^2$. These are classed as the index observations.

2. Define a 3 month "look-forward" window for each index observation - 90 days after the date of the index observation.

3. Check for any normal ($>= 60ml/min/1.72m^2$) eGFR observations within the look-forward window. If there are any, then the index observation should not be classed as CKD onset. If there are no normal eGFR observations within the look-forward window, proceed to the next step.

4. Check that there is at least one measurement of eGFR ($< 60ml/min/1.72m^2$) after the end of the look-forward window. If there is, then the index meets the criteria for CKD onset. If not, the index observation does not qualify as CKD onset.

5. Any patient with at least one index observation meeting the criteria should be included in the cohort, and their study entry date (time 0) will be taken as the first observation $< 60ml/min/1.72m^2$ that qualifies as CKD onset, i.e. the date from which the sustained drop was observed. Patients can contribute multiple discontiguous periods of observation, however they will only be classed as "on study" once they meet the definition of CKD, as defined in the previous steps.

This process is illustrated in Figure 4.1, where a set of hypothetical patient trajectories are shown. Onset of kidney failure will be defined in a similar way, but using a threshold of eGFR $< 60ml/min/1.72m^2$. Any patients with evidence of kidney failure at or prior to onset of CKD will be excluded from the cohort.

The definition of CKD onset (as the first presentation of abnormal eGFR) could introduce immortal time bias, as it depends on patients being alive and uncensored for at least 3 months from the index observation for the drop in kidney function to be classed as "sustained". This definition was adopted as it is the widely accepted definition of CKD onset. An alternative definition that may mitigate the issue of immortal time bias would be to class the beginning of follow-up as the end of the 90-day window.

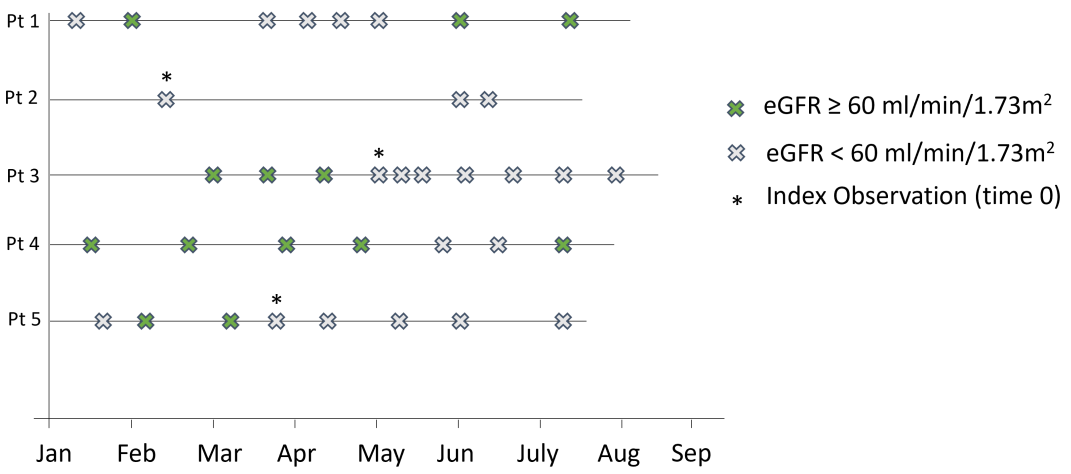


Figure 4.1: An illustration of how eGFR trajectories were assessed for inclusion as CKD diagnoses. Patients 2, 3 and 5 in the figure below would be selected for inclusion based on the sample trajectories, whereas patients 1 and 4 would not, since they do not meet the criteria for a sustained drop in eGFR for 3 months. The observations marked with * are the index observations - the dates of these would be used as each patient's time 0.

4.3.2 Notation

Let $Y_i(t_{ij})$ denote the observed value of eGFR for patient $i = 1,...,n$, at time t_{ij} where $j = 1,...m_i$, thereby allowing irregular measurement times across patients, and a different number of eGFR measurements per patient. Throughout this study, time is measured as "time since first presentation of CKD" (in years), where $t_{ij} = 0$ is the time at which patient i first met the criteria for CKD.

Assume that we wish to make predictions at a set of landmark times $\mathcal{L}$, using all information observed prior to each landmark time. For this study, define the set of landmark times as $\mathcal{L} \in \{0, 0.25, 0.5, 0.75, 1, 1.25, 1.5\}$, measured in years. These times have been chosen to reflect the fact that CKD is often monitored in 3-month windows: a definition for onset of CKD or kidney failure is based on a sustained drop in kidney function for at least 3 months, so these are clinically relevant times at which a care provider might want to update their prognostic predictions.

Define the information on event and survival time as $S_i = (T_i, D_i)$ where T_i is the time from index date(time at which first met CKD criteria) until the first of administrative censoring, kidney failure or death. $D_i \in \{0, 1, 2\}$, where $D_i = 0$ if patient i was censored, $D_i = 1$ if they experienced kidney failure at T_i and $D_i = 2$

if they died.

The primary outcome of interest will be kidney failure, defined as: a sustained eGFR of $< 15ml/min/1.73m^2$, commencement of Renal Replacement Therapy (RRT) or kidney transplant (whichever occurs first). Follow-up was defined as the calendar time from time 0 for each patient (as defined above) until the first occurrence of the following: kidney failure (any one of a sustained eGFR of $< 15ml/min/1.73m^2$, RRT or kidney transplant), death or the end of the study period (administrative censoring). Where kidney failure is diagnosed based on a drop in eGFR, this is done in the same way as the cohort entry point, i.e. using the timing of the first observation of eGFR $< 15ml/min/1.73m^2$ that is followed by a sustained drop of $< 15ml/min/1.73m^2$. Where multiple possible kidney failure events happen to the same patient (e.g. a sustained drop in eGFR followed by commencement of RRT), the timing of the first of these is used as that patient's event time. Death is a competing risk for kidney failure, so we account for this via cause-specific modelling, and additionally model the hazard of all-cause mortality prior to kidney failure.

Define $n_i(t)$ to be the cumulative count of the number of eGFR observations for patient i upto time t, i.e. $n_i(t) = \#\{j|t_{ij} \leq t\}$.

Denote the last-observation carried forward observations of UACR and eGFR at landmark time l as x_{il} and y_{il} respectively.

4.3.3 Two-stage landmarking approach

We first describe the two-stage landmarking approach at a high level, and provide further detail on the mathematical notation and modelling process in following sections.

Landmarking approaches provide a more computationally efficient approach to incorporating longitudinal predictors than joint modelling, yet the classical version of landmarking only takes the last-observed value of a repeatedly measured predictor in the time-to-event prediction, which fails to acknowledge that the repeated predictor may be measured with error. We therefore adopt a slightly more sophisticated "Two-stage Landmarking" approach that allows for measurement error in the longitudinal predictor (in this case eGFR). We first fit a mixed effects model to the evolution of eGFR over time, then use this model to predict eGFR at each landmark time. This prediction will then be used as a predictor in the landmark time time-to-event model. This approach is further outlined by

Paige et al. [15], Keogh et al. [16] and Rizopoulos, Molenberghs, and Lesaffre [14].

We define the set of landmark times l relative to each patient's onset of CKD (0 to 1.5 years) at which we might want to make a prediction. At each of these landmark times, we define the landmark risk set cohort, defined as: any individual included in the study that has not yet progressed to kidney failure, died or been censored at time l. Any individual can fall into one or more landmark risk sets. Mixed effects models are fit separately within each landmark risk set, including eGFR measurements observed upto and including the landmark time. Those in the risk sets are then censored at the prediction horizon of interest $t_{hor} = 2$ (in line with the kidney failure risk equation), as recommended by Houwelingen and Putter [18]. We then fit a "superlandmarking" model to the stacked landmark time specific datasets for the time-to-event outcome [10]. The models are estimated with robust standard errors, to adjust for the fact that patients can be present multiple times in the stacked dataset. The baseline hazard is stratified by landmark time. Note that the KFRE is also used with a prediction horizon of 5 years, but limited follow-up data does not allow for prediction at 5 years. We compare the two-stage landmark models against a classic landmarking approach that uses the last observation carried forward (LOCF) for repeated predictor measurements, to identify and quantify any predictive gains from the two-stage approach. All models that include a representation of the observation process will be fit using the two-stage approach.

4.3.4 Statistical analyses: fitted models

Mixed effects models

We assume that the longitudinal evolution of eGFR can be described by a linear mixed effects model such that:

$$Y_i(t_{ij}) = \mu_i(t_{ij}) + \epsilon_i(t_{ij})$$

$$\mu_i(t_{ij}) = \beta_0 + \beta_1 g(t_{ij}) + \beta_2 \text{Location}_{ij} + \beta_3 \text{Gender}_i + \beta_4 \text{AgeCKD}_i + b_{0i} + b_{1i} g(t_{ij})$$

$$\epsilon_{ij} \sim N(0, \sigma^2),\ \mathbf{b_i} = (b_{0i}, b_{1i})^T \sim N(0, \Sigma)$$

As described above, this mixed effects model is fit separately within each landmark risk set, restricting the observations used in the model fitting to $t_{ij} \leq l$.

We include a categorical predictor representing the location that the eGFR measurement was initiated, since we have observed that eGFR values tend to vary based on the clinical setting they were measured in. The possible locations are summarised into Community (GP or outpatient) or Inpatient, with the reference category set to Community. AgeCKD$_i$ is the age of patient i at their first presentation of CKD, i,e. age at baseline.

Multiple parametrisations of time are considered (i.e. for $g(t_{ij})$), to accommodate a possible non-linear trajectory of eGFR over time. A linear term for time ($g(t_{ij}) = t_{ij}$) is compared against a model with time modelled using natural cubic splines. For the spline models, a range of options for the number of knots will be tested and compared using model fit statistics and graphical summaries of the fits. All confirmatory analyses to establish the best fitting form of the mixed model are performed using the entire patient cohort, and all repeated measures of eGFR. The favoured model formulation is then adopted across all landmark time specific models.

Time-to-event models

The hazard for progression to kidney failure or death can be modelled by:

$$h_{1,k}(t) = h_{0,k,l}(t)exp(\gamma_{\mathbf{1}}{}^T Z_i + \gamma_{\mathbf{2}}{}^T W_i(l) + \gamma_{\mathbf{3}}{}^T \rho_i(l))$$

where $k = 1$ for kidney failure, and $k = 2$ for death. t is time since landmark time, and l is the landmark time.

Z_i is the vector of time-invariant predictors (age of CKD onset and gender), $W_i(l)$ is the set of time-varying predictors (UACR and eGFR), $\rho_i(l)$ is some parametrisation of the observation process upto time l, and the γ terms are vectors of model coefficients. Note that the specific forms of $W_i(l)$ and $\rho_i(l)$ vary depending on the model and are defined more specifically in Table 4.1. The UACR-component of $W_i(l)$ is always the last-observation carried forward at landmark time l, but the form of eGFR varies depending on the fit.

We assume that the current value ($\mu_i(t)$), and time-varying rate of change ($\mu_i'(t)$) in eGFR are both risk factors for the occurrence of kidney failure and death and are estimated at each landmark time. These have been replaced by the estimates ($\hat{\mu}_i(t)$ and $\hat{\mu}_i'(t)$) obtained from the linear mixed model at the previous stage within the two-stage models. BLUPs can be estimated based on observed data upto the landmark time for each individual, and can be obtained for new individuals assuming they have contributed at least one eGFR measurement before the landmark time, or two for the time-varying slope $\hat{\mu}_i'(t)$.

Parametrisations of the observation process (informative observation)

We consider a range of parametrisations for the observation process of eGFR, described by Models 2a - 4b in Table 4.1. The most simple approach (Model 2a) is to include a (rolling) count of the number of times that eGFR has been observed over the past year ($n_i(t)$). Clinical insight suggests that a history of inpatient admissions can also be a driver of risk, so Model 2b adapts the previous model to only include inpatient observations.

A more sophisticated representation of the observation process can be obtained by fitting a recurrent events model to the gaptimes between observations (as proposed within the joint modelling framework [13], and including the individual-level estimated frailty term as a model predictor (Models 3a and 3b). Model 3b has been included for closer comparison with Model 2a, the count model, which counts observations over the past year. Finally, clinicians may choose to monitor certain patients more closely than others during periods of concern, and therefore any sudden changes in the observation process could indicate a change in their risk. Models 4a and 4b therefore include the variance of the gaptimes between eGFR measures to capture such inconsistency in the way a patient may have been observed. Model 4b combines this variance parameter with a count to additionally account for the rate of eGFR measurement.

The fitted recurrent events model used to estimate the frailty terms in Models 3a and 3b takes the form:

$$r(\tilde{t}_{ij}) = r_0(\tilde{t}_{ij}) w_i exp(\phi_1 AgeCKD + \phi_2 Location_{ij} + \phi_3 Gender + \lambda \mu_i(t))$$

A Weibull model is adopted and fitted to the gaptimes $\tilde{t}_{ij} = t_{ij+1} - t_{ij}$, and the random effect term w_i is assumed to come from a Gamma distribution. This frailty term can be estimated for new patients as a function of the observed gaptimes and predictors, as illustrated in Munda, Rotolo, and Legrand [19].

Table 4.1 below summarises all of the parametrisations adopted to model the informative nature of eGFR observations.

Table 4.1: Summary of fitted models. The general form of the fitted models is: $h_1(t) = h_{0,k,l}(t) exp({\gamma_{\mathbf{1}}}^T Z_i + {\gamma_{\mathbf{2}}}^T W_i(l) + {\gamma_{\mathbf{3}}}^T \rho_i(l))$ where $W_i(l)$ represents the form of the longitudinal eGFR trajectory, and $\rho_i(l)$ is the parametrisation of the (informative) observation process at landmark time l.

Model	Name	$W_i(l)$	$\rho_i(l)$	Description
1a	Null: LOCF	$x_{il} + y_{il}$	0	Null standard landmarking model
1b	Null: Lmm	$x_{il} + \hat{\mu}_i(l) + {\hat{\mu}_i}'(l)$	0	Null two-stage landmarking model
2a	Count	$x_{il} + \hat{\mu}_i(l) + {\hat{\mu}_i}'(l)$	$n_i(l) = \#\{j : l - 1 \leq t_{ij} \leq l\}$	Number of eGFR observations over the past year as a predictor
2b	Inpatient count	$x_{il} + \hat{\mu}_i(l) + {\hat{\mu}_i}'(l)$	$n_i^*(l) = \#\{j : l - 1 \leq t_{ij} \leq l\}$	Number of inpatient eGFR observations over the past year as a predictor
3a	Frailty	$x_{il} + \hat{\mu}_i(l) + {\hat{\mu}_i}'(l)$	$\hat{w}_i$	Estimated frailty term from recurrent events model fitted to gaptimes as a predictor
3b	Frailty (past year)	$x_{il} + \hat{\mu}_i(l) + {\hat{\mu}_i}'(l)$	$\hat{w_i^*}$	Estimated frailty term using only the past year of data from recurrent events model fitted to gaptimes as a predictor
4a	Var. gaptimes	$x_{il} + \hat{\mu}_i(l) + {\hat{\mu}_i}'(l)$	$var(\tilde{t}_i), t_{ij} < l$	Variance of gaptimes (between eGFR measurements) as a predictor
4b	Var. gaptimes + Count	$x_{il} + \hat{\mu}_i(l) + {\hat{\mu}_i}'(l)$	$n_i(l) + var(\tilde{t}_i), t_{ij} < l$	Rolling observation count + Variance of gaptimes (between eGFR measurements) as a predictor

4.3.5 Prediction for new individuals

To obtain a prediction for the risk of kidney failure or death at landmark time l, the following procedure can be applied to obtain risk estimates of kidney failure.

1) Obtain the required values of UACR and eGFR, depending on the model formulation ($W_i(l)$). For Models 1b - 4b, use the observed eGFR data before l and parameters estimated from the landmark time-specific mixed effects model to generate BLUPs for the random effects, using these to estimate $\hat{\mu}_i(l)$ and $\hat{\mu}_i'(l)$.

2) For Models 3a, and 3b, use $\hat{\mu}_i(t)$ alongside the remaining predictors in the recurrent events model to estimate the new patient's frailty term.

3) Input estimates from stages 1 and 2 to get a prediction for the 2-year risk of kidney failure and death, conditional on survival upto landmark time l. The general form of the formula used to estimate this risk is as follows: $1 - P(T > l+2|T > l) = 1 - \hat{S}_0(l+2|l)^{exp(\gamma_1{}^T Z_i + \gamma_2{}^T W_i(l) + \gamma_3{}^T \rho_i(l))}$

$\hat{S}_0(l+2|l)$ is the Breslow estimate of being event free at time $l+2$ conditional on being event-free at l. Following the model-fitting process, two formulae of this form will be available - one for kidney failure, and one for death (with a separate set of γ coefficients, and separate baseline survival estimates for each outcome), so the one relating to the outcome of interest should be used to predict for that outcome.

Note that risk estimates were obtained in this way to follow the approach used in the development of the original KFRE, whereby competing risks were handled by simply censoring patients that died (for the outcome of kidney failure) at their time of death, and treating kidney failure as the only outcome of interest. This is not, however, a valid approach to properly accounting for competing risk of death. The resulting estimand under this approach is “the risk of kidney failure assuming nobody dies”. Instead, the cumulative incidence function should be estimated, that properly accounts for the fact that patients can die prior to experiencing kidney failure, and the risk of death is explicitly used in the estimation of the risk of kidney failure [20] (and vice versa in the prediction of death).

While we choose to match the approach taken in the original KFRE model (for comparison with that model), future work will need to use the cumulative incidence – exploring the impact of this on the result is beyond the scope of this thesis, but will be considered prior to any publication of this chapter.

4.3.6 Assessment of model performance

Each model will be compared using calibration (the agreement between observed and expected event rate) and discrimination (the ability of the model to differentiate between those with and without kidney failure). These will be estimated and summarised at each individual landmark time. Calibration will be assessed via the calibration slope, discrimination via Harrell's C-statistic, and a measure of overall model performance via the Brier Score.

Performance measures are estimated via bootstrapping: a sample of the same size as the full cohort is drawn (with replacement) from the original cohort. The entire model fitting process is repeated within this bootstrap sample, including fitting the mixed models for eGFR trajectories. Performance is then calculated in the original data. A total of 100 bootstrap samples are drawn, and performance is summarised across these bootstrap samples.

4.3.7 Software

All analyses were performed in R version 4.0.3. Mixed effects models were fit using the `nlme` and `splines` packages. Frailty models were fit using the `frailtypack` package. Brier scores and C-statistics were calculated using the `ipred` and `Hmisc` packages respectively. `reReg` was used to produce Figure 4.5 (with some customization via `ggplot2` and `cowplot`). The `survival` package was used to fit superlandmark models.

4.4 Results

A total of 8033 patients met the criteria for incident Chronic Kidney Disease stage G3 or G4 at some point during the follow-up period. The total size of the stacked superlandmark dataset was 49078, as some patients were present in the risk sets at multiple landmark times. See Figure 4.2 for a flowchart of the full cohort derivation, including all inclusions and exclusions, and Table 4.2 presents a summary of the cohort at landmark time 0. The number of patients in each landmark time risk set ranges from 8033 at $l = 0$ to 5512 at $l = 1.5$. These numbers

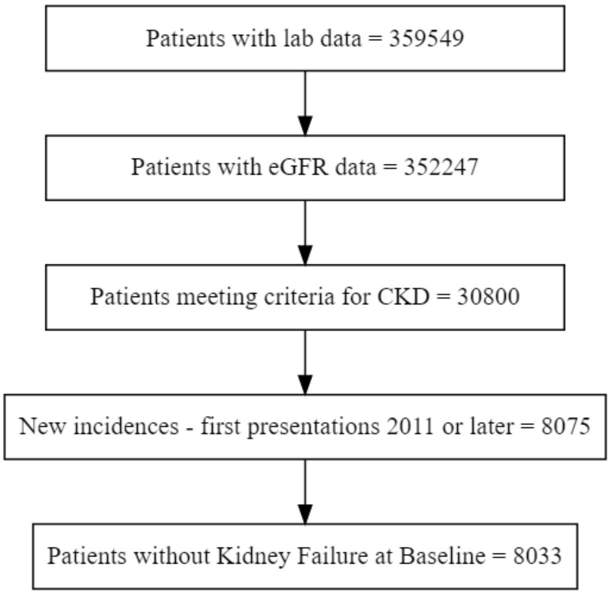

Patients with lab data = 359549
Patients with eGFR data = 352247
Patients meeting criteria for CKD = 30800
New incidences - first presentations 2011 or later = 8075
Patients without Kidney Failure at Baseline = 8033

Figure 4.2: A flowchart of the cohort derivation

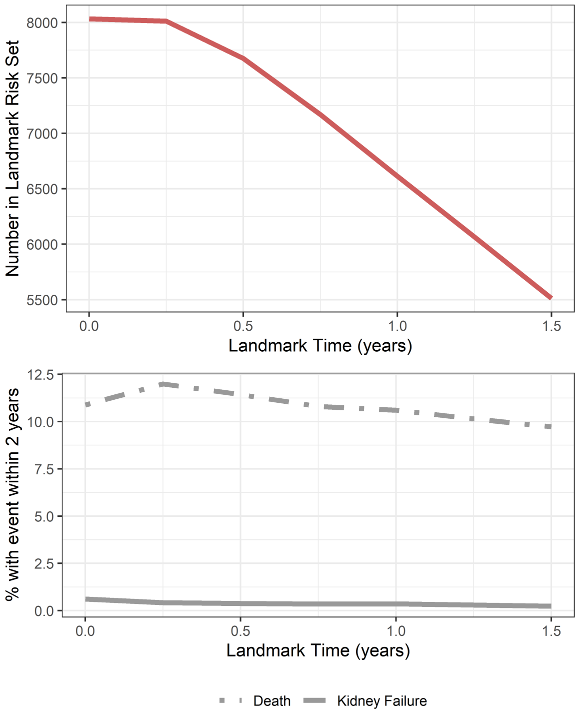


Figure 4.3: A summary of the risk sets and events (within 2 years) across landmark times

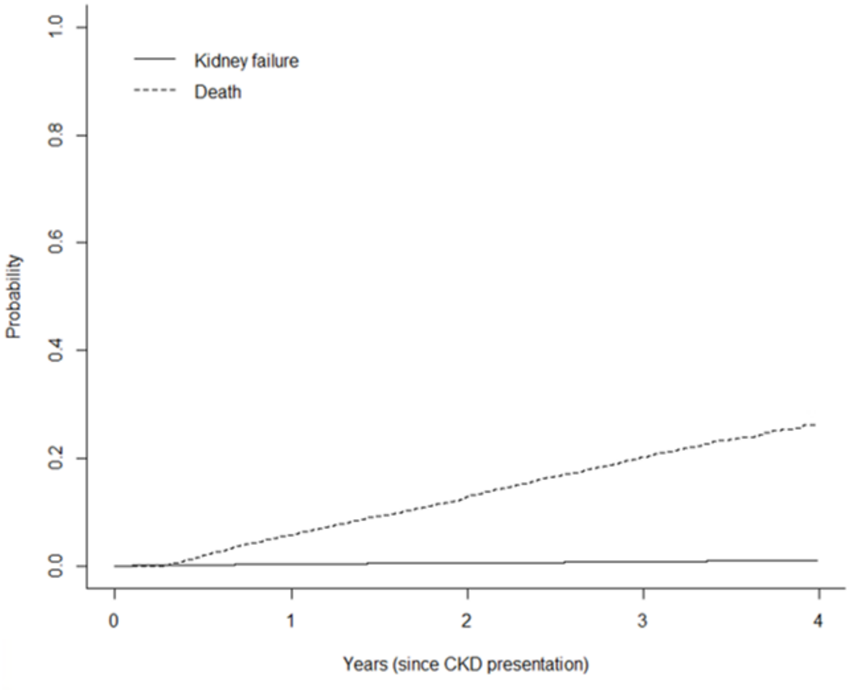


Figure 4.4: Cumulative incidence plot for the occurrence of kidney failure and death, summarised on the 'time on study' timescale.

are plotted in Figure 4.3 below, as well as how many patients died or experienced kidney failure within 2 years of the landmark time. As expected, there is a gradual decrease in the size of the risk set as landmark time increases. The number of kidney failure events remains relatively low and constant, with a small peak at the first landmark time (0). Cumulative incidence curves are presented Figure 4.4. We present results from confirmatory analyses to determine the best fitting form of the linear mixed models in the supplementary materials (section 4.7).

4.4.1 Patterns of eGFR measurement

To gain insight into the patterns of observation in this cohort, we visualized a random subset of the patient journeys (60 patients) in Figure 4.5. We randomly sampled 20 patients with each outcome at the end of their follow-up: kidney failure (top plot), death (middle plot) or censored (bottom plot). Figure 4.5 shows that patients in the first two plots (kidney failure and death samples) had more inpatient eGFR measures than the censored group, and there appears to be a slight increase in the overall frequency of eGFR measures in those that died or progressed to kidney failure.

Table 4.2: Summary of cohort at landmark time 0 years (i.e. at cohort entry)

	Mean/N	SD/%	N. observed
Female	4722	0.59	8033
Male	3311	0.41	8033
Age at CKD presentation	75.72	10.62	8033
eGFR	51.61	8.78	8033
UACR	12.88	53.15	2821
Median Follow-up (IQR)	2.10	(1.28, 2.97)	8033
Kidney failure events (N)	59		8033
Deaths (N)	1240		8033

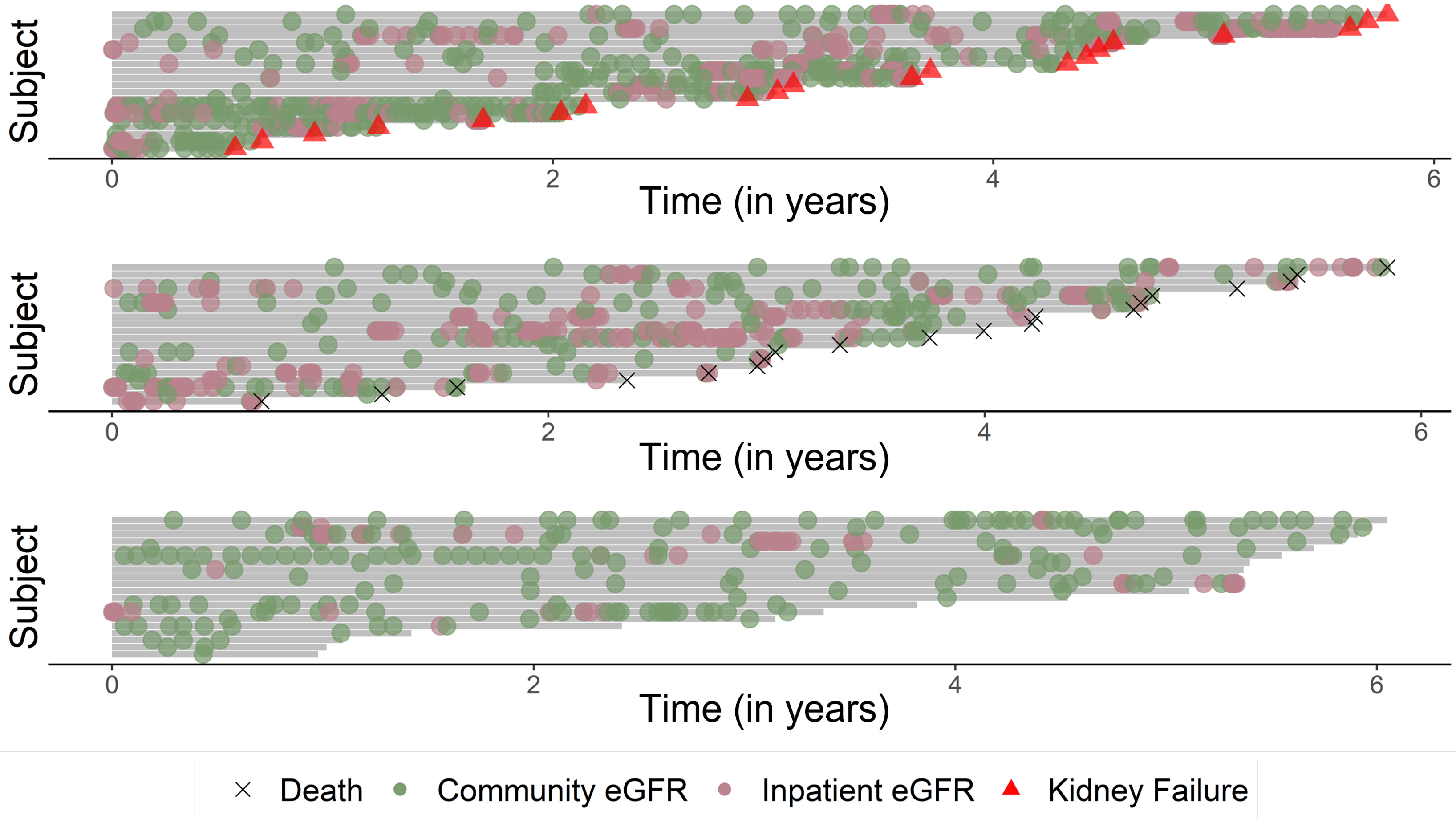


Figure 4.5: Visualisation of the observation processes, split by outcome status (Kidney failure, death, or censored). The green and pink circular marks correspond to times of eGFR measurements. These are classified into inpatient and community (GP or outpatient). Kidney failure and death are denoted by triangles and circles respectively. Note that for the purpose of this plot, we have perturbed the actual measurement times by a random draw from a normal distribution with mean 0 and standard deviation of 2 days, to reduce the risk of disclosure of individual patients. The monitoring frequencies are, however, the same as in the original data.

Table 4.3: Medians and interquartile ranges (IQRs) of the annual rate of observation of eGFR measurements, broken down by outcome status and measurement setting

	Censored		Kidney Failure		Death	
Setting	Median	IQR	Median	IQR	Median	IQR
Community	2.26	(1.61, 3.36)	4.89	(2.66, 12.61)	2.64	(1.75, 4.38)
Inpatient	1.39	(0.54, 3.38)	11.20	(4.42, 19.06)	4.67	(2.02, 9.55)

We further summarised the median and interquartile ranges (IQR) of the rates of eGFR observation, broken down by outcome status and care setting in Table 4.3. Patients that experienced kidney failure had considerably higher rates of observation compared to the censored and death groups, especially within inpatient settings. Patients that died had marginally higher rates of community and inpatient observation compared to the censored group.

4.4.2 Model performance

Harrell's C statistic, Brier Score and Calibration slope are plotted across landmark times in Figure 4.6 for both outcomes, and summarised in Tables 4.4 - 4.9 for a subset of the landmark times (0, 0.5, 1 and 1.5 years). Confidence intervals are derived from the bootstrap samples as the 2.5th and 97.5th percentiles, and CIs for all other landmark times (not presented here in the main text) are presented in the supplementary materials. We further summarised the proportion of bootstrap samples within which each model outperformed others, e.g. "Model 2a outperformed Model 1b at landmark time $l = 0$ in 100% of the bootstrap samples". Parameter estimates for all fitted models (linear mixed models and time-to-event models) are presented in the supplementary materials.

Overall model performance

For both outcomes, there was considerable variability in predictive performance across time. Models for both outcomes appear to be quite poorly calibrated, as reflected by the calibration slope in the third row of Figure 4.6, but this improves as landmark time increases for both outcomes. For both outcomes, predicted risks were generally too extreme (calibration slope < 1) until around landmark time 1 year, after which point the calibration slope is closer to 1. The kidney

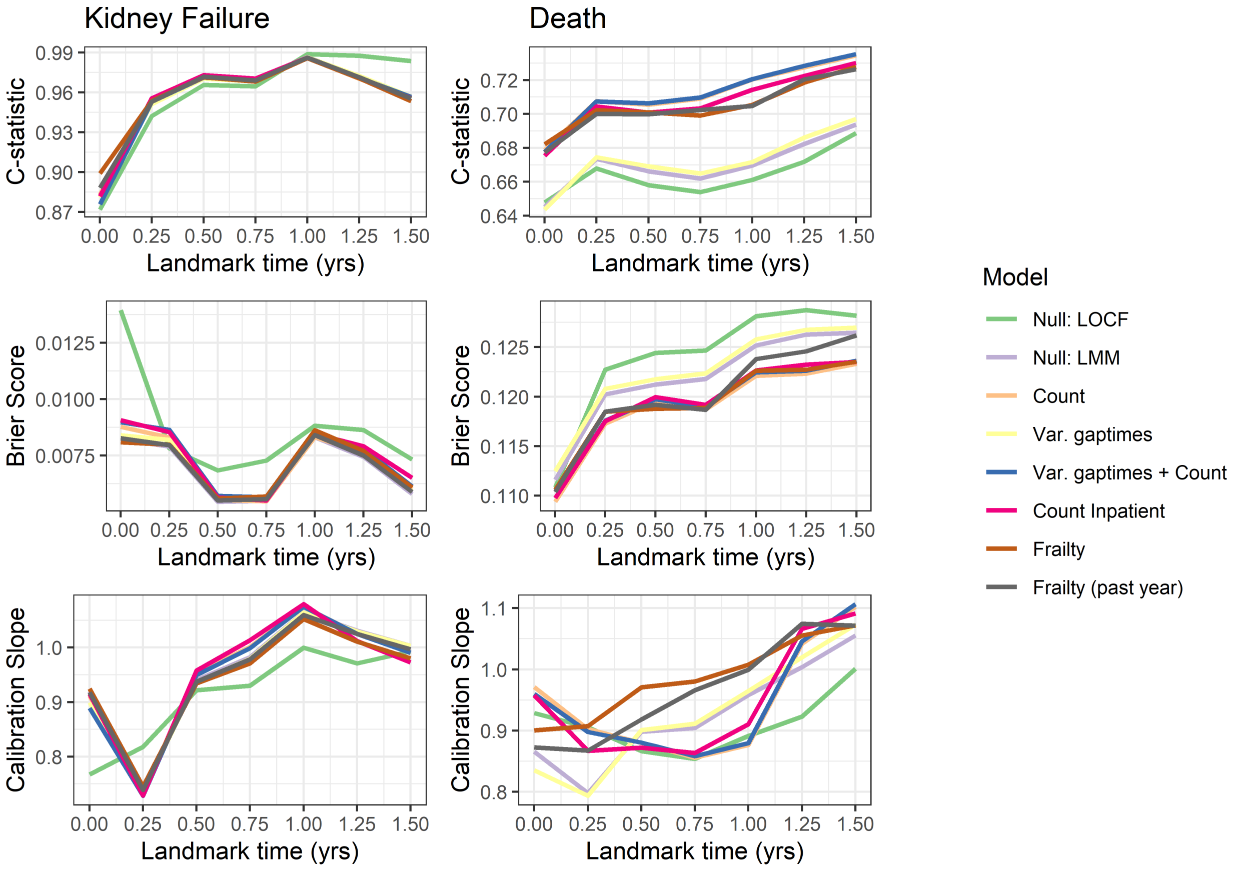


Figure 4.6: A summary of model performance metrics for the prediction of both kidney failure (left column) and death (right column). All metrics have been calculated in the landmark specific risk sets for the prediction of an event within 2 years. The lines represent the mean over the 100 bootstrap samples, by landmark time. The left hand column contains all plots corresponding to predictive performance for kidney failure, and death in the right hand column.

failure models appeared slightly better calibrated than the mortality ones, but still fairly variable across the landmark times.

For the prediction of both death and kidney failure, the two-stage landmark model appeared to offer slight improvements in all three performance metrics compared to the last observation carried forward model. These gains are, however, marginal.

Impact of informative observation

For the prediction of kidney failure, there were no apparent differences (neither benefit nor harm) between the performance of the null models (Models 1a and 1b), and those incorporating informative observation (Models 2a - 4b).

Discrimination In the prediction of death, we observed some considerable improvements in discrimination (of around 3-4% in Harrell's C) as a result of including the observation process as a predictor in the landmark model, indicating that there may be additional value in the measurement frequency, above and beyond the realised values. In all 100 bootstrap samples, the C-statistic for non-null models (Model 2a - 4b) was greater than those for the null models. The confidence intervals displayed in Table 4.5 (C-statistics) also do not overlap, and therefore there is considerable evidence to support this improvement in model discrimination. The "count" and "inpatient count" models appeared to offer the largest gains in discrimination, but the "var. gaptimes" model did not appear to offer any improvement over the null model.

Brier Scores The Brier scores for the prediction of death were also smaller for the models accommodating informative observation (Model 2a - 4b). The count model (Model 2a) resulted in a lower Brier score than the null model (Model 1b) across all 100 bootstrap samples at all landmark times. The frailty model's Brier Score was less than the null model's Brier Score in all 100 bootstrap samples, apart from at landmark time 0 where it was smaller in 96/100 of the samples. There is therefore strong evidence to suggest that these two models provide gains in overall model fit compared to the null model. This improvement does, however, appear to be very small. The mean difference in Brier scores between Models 2a and 1b was 0.022, or 2% at landmark time 0 years, 0.0031 or 2.4% at landmark time 1 year, and 0.0031 or 2.5% at 1.5 years.

Calibration The models based on frailty terms (Models 3a and 3b) for the prediction of death offered some improvements in model calibration compared with the null models, and at later time points. The calibration slope estimates were, however, highly variable both within and between landmark times, as can be seen in Tables 4.8 and 4.9, but most confidence intervals (after landmark time 0) contain 1. Since all confidence intervals across the different models overlap, it is not possible to conclude that any model offers improvements over the null ones for this metric. There is strongest evidence of an improvement in calibration (death models) from the informative observation models at earlier landmark times (0, 0.25) where the count and frailty models' calibration slopes were closer to 1 than those from the null model in more than 85% of the bootstrap samples. There is also no strong evidence that the frailty models result in larger gains in calibration than the count models at any landmark time.

Table 4.4: C-statistics and confidence intervals for prediction of Kidney Failure. The columns represent a subset of the landmark times l = 0, l = 0.5 years, l = 1 year and l = 1.5 years

	l = 0		l = 0.5		l = 1		l = 1.5	
Model	Mean	CI	Mean	CI	Mean	CI	Mean	CI
1a: Null: LOCF	0.872	(0.857, 0.887)	0.966	(0.954, 0.973)	0.989	(0.986, 0.991)	0.984	(0.980, 0.986)
1b: Null: LMM	0.888	(0.860, 0.911)	0.972	(0.965, 0.977)	0.986	(0.983, 0.989)	0.956	(0.948, 0.961)
2a: Count	0.883	(0.859, 0.908)	0.973	(0.965, 0.978)	0.986	(0.981, 0.989)	0.955	(0.948, 0.961)
2b: Inpatient count	0.882	(0.854, 0.907)	0.973	(0.966, 0.978)	0.986	(0.984, 0.989)	0.956	(0.948, 0.961)
3a: Frailty	0.899	(0.856, 0.922)	0.971	(0.964, 0.976)	0.986	(0.983, 0.989)	0.953	(0.945, 0.962)
3b: Frailty (past year)	0.888	(0.856, 0.917)	0.972	(0.964, 0.977)	0.986	(0.981, 0.989)	0.955	(0.948, 0.962)
4a: Var. gaptimes	0.881	(0.847, 0.910)	0.970	(0.955, 0.976)	0.986	(0.983, 0.989)	0.957	(0.947, 0.965)
4b: Var. gaptimes + Count	0.876	(0.845, 0.906)	0.971	(0.957, 0.978)	0.986	(0.981, 0.989)	0.957	(0.947, 0.965)

Table 4.5: C-statistics and confidence intervals for prediction of Death. The columns represent a subset of the landmark times l = 0, l = 0.5 years, l = 1 year and l = 1.5 years

	l = 0		l = 0.5		l = 1		l = 1.5	
Model	Mean	CI	Mean	CI	Mean	CI	Mean	CI
1a: Null: LOCF	0.648	(0.646, 0.650)	0.658	(0.657, 0.659)	0.661	(0.659, 0.666)	0.689	(0.686, 0.692)
1b: Null: LMM	0.645	(0.643, 0.648)	0.666	(0.664, 0.668)	0.670	(0.667, 0.673)	0.694	(0.691, 0.697)
2a: Count	0.679	(0.676, 0.682)	0.706	(0.702, 0.710)	0.720	(0.716, 0.725)	0.735	(0.732, 0.738)
2b: Inpatient count	0.675	(0.672, 0.679)	0.701	(0.698, 0.704)	0.714	(0.710, 0.718)	0.730	(0.727, 0.734)
3a: Frailty	0.682	(0.681, 0.684)	0.701	(0.699, 0.703)	0.705	(0.703, 0.708)	0.728	(0.725, 0.731)
3b: Frailty (past year)	0.677	(0.676, 0.679)	0.700	(0.697, 0.702)	0.705	(0.702, 0.707)	0.726	(0.724, 0.729)
4a: Var. gaptimes	0.644	(0.640, 0.646)	0.669	(0.667, 0.671)	0.672	(0.669, 0.675)	0.697	(0.694, 0.700)
4b: Var. gaptimes + Count	0.678	(0.674, 0.681)	0.706	(0.703, 0.710)	0.721	(0.717, 0.725)	0.735	(0.732, 0.739)

Table 4.6: Brier Scores and confidence intervals for prediction of Kidney Failure. The columns represent a subset of the landmark times l = 0, l = 0.5 years, l = 1 year and l = 1.5 years

	l = 0		l = 0.5		l = 1		l = 1.5	
Model	Mean	CI	Mean	CI	Mean	CI	Mean	CI
1a: Null: LOCF	0.0139	(0.0112, 0.0180)	0.0068	(0.0052, 0.0098)	0.0088	(0.0063, 0.0132)	0.0073	(0.0050, 0.0113)
1b: Null: LMM	0.0082	(0.0070, 0.0102)	0.0054	(0.0045, 0.0071)	0.0083	(0.0063, 0.0122)	0.0058	(0.0044, 0.0087)
2a: Count	0.0088	(0.0072, 0.0104)	0.0055	(0.0046, 0.0071)	0.0083	(0.0062, 0.0121)	0.0060	(0.0044, 0.0089)
2b: Inpatient count	0.0091	(0.0074, 0.0111)	0.0056	(0.0046, 0.0073)	0.0085	(0.0063, 0.0124)	0.0065	(0.0046, 0.0096)
3a: Frailty	0.0081	(0.0069, 0.0102)	0.0056	(0.0046, 0.0077)	0.0086	(0.0066, 0.0127)	0.0061	(0.0046, 0.0091)
3b: Frailty (past year)	0.0083	(0.0069, 0.0103)	0.0055	(0.0045, 0.0078)	0.0084	(0.0064, 0.0125)	0.0059	(0.0043, 0.0090)
4a: Var. gaptimes	0.0084	(0.0071, 0.0106)	0.0056	(0.0045, 0.0082)	0.0085	(0.0063, 0.0124)	0.0059	(0.0044, 0.0088)
4b: Var. gaptimes + Count	0.0090	(0.0073, 0.0116)	0.0057	(0.0046, 0.0084)	0.0085	(0.0063, 0.0125)	0.0061	(0.0045, 0.0090)

Table 4.7: Brier Scores and confidence intervals for prediction of Death. The columns represent a subset of the landmark times l = 0, l = 0.5 years, l = 1 year and l = 1.5 years

	l = 0		l = 0.5		l = 1		l = 1.5	
Model	Mean	CI	Mean	CI	Mean	CI	Mean	CI
1a: Null: LOCF	0.111	(0.110, 0.112)	0.124	(0.123, 0.126)	0.128	(0.127, 0.130)	0.128	(0.127, 0.130)
1b: Null: LMM	0.112	(0.111, 0.113)	0.121	(0.120, 0.123)	0.125	(0.124, 0.127)	0.126	(0.125, 0.128)
2a: Count	0.109	(0.108, 0.111)	0.119	(0.118, 0.121)	0.122	(0.121, 0.124)	0.123	(0.122, 0.125)
2b: Inpatient count	0.110	(0.109, 0.111)	0.120	(0.119, 0.121)	0.123	(0.121, 0.124)	0.124	(0.122, 0.125)
3a: Frailty	0.111	(0.109, 0.112)	0.119	(0.118, 0.120)	0.123	(0.121, 0.124)	0.124	(0.122, 0.125)
3b: Frailty (past year)	0.110	(0.109, 0.112)	0.119	(0.118, 0.121)	0.124	(0.122, 0.126)	0.126	(0.125, 0.128)
4a: Var. gaptimes	0.112	(0.111, 0.114)	0.122	(0.121, 0.123)	0.126	(0.125, 0.127)	0.127	(0.126, 0.129)
4b: Var. gaptimes + Count	0.110	(0.109, 0.111)	0.120	(0.119, 0.121)	0.122	(0.121, 0.124)	0.124	(0.122, 0.125)

Table 4.8: Calibration Slopes and confidence intervals for prediction of Kidney Failure. The columns represent a subset of the landmark times l = 0, l = 0.5 years, l = 1 year and l = 1.5 years

	l = 0		l = 0.5		l = 1		l = 1.5	
Model	Mean	CI	Mean	CI	Mean	CI	Mean	CI
1a: Null: LOCF	0.77	(0.64, 0.87)	0.92	(0.43, 1.13)	1.00	(0.44, 1.25)	0.99	(0.46, 1.18)
1b: Null: LMM	0.92	(0.85, 0.98)	0.94	(0.44, 1.16)	1.06	(0.52, 1.23)	1.00	(0.49, 1.16)
2a: Count	0.91	(0.85, 0.98)	0.95	(0.44, 1.19)	1.07	(0.54, 1.24)	0.99	(0.50, 1.15)
2b: Inpatient count	0.91	(0.84, 0.99)	0.96	(0.45, 1.19)	1.08	(0.55, 1.25)	0.97	(0.50, 1.12)
3a: Frailty	0.92	(0.84, 0.99)	0.93	(0.45, 1.15)	1.05	(0.53, 1.22)	0.98	(0.49, 1.14)
3b: Frailty (past year)	0.92	(0.83, 0.98)	0.94	(0.44, 1.16)	1.06	(0.52, 1.23)	1.00	(0.49, 1.16)
4a: Var. gaptimes	0.90	(0.75, 0.98)	0.94	(0.44, 1.16)	1.06	(0.52, 1.23)	1.00	(0.48, 1.16)
4b: Var. gaptimes + Count	0.89	(0.72, 0.98)	0.95	(0.44, 1.18)	1.07	(0.53, 1.24)	0.99	(0.49, 1.15)

Table 4.9: Calibration Slopes and confidence intervals for prediction of Death. The columns represent a subset of the landmark times l = 0, l = 0.5 years, l = 1 year and l = 1.5 years

	l = 0		l = 0.5		l = 1		l = 1.5	
Model	Mean	CI	Mean	CI	Mean	CI	Mean	CI
1a: Null: LOCF	0.93	(0.86, 1.02)	0.87	(0.52, 1.03)	0.89	(0.52, 1.07)	1.00	(0.53, 1.23)
1b: Null: LMM	0.87	(0.80, 0.95)	0.90	(0.59, 1.09)	0.96	(0.65, 1.14)	1.06	(0.68, 1.28)
2a: Count	0.97	(0.88, 1.05)	0.88	(0.65, 1.04)	0.88	(0.67, 1.00)	1.10	(0.82, 1.26)
2b: Inpatient count	0.96	(0.88, 1.03)	0.87	(0.63, 1.01)	0.91	(0.67, 1.04)	1.09	(0.78, 1.24)
3a: Frailty	0.90	(0.84, 0.96)	0.97	(0.71, 1.09)	1.01	(0.77, 1.11)	1.07	(0.79, 1.19)
3b: Frailty (past year)	0.87	(0.81, 0.93)	0.92	(0.68, 1.06)	1.00	(0.75, 1.15)	1.07	(0.83, 1.21)
4a: Var. gaptimes	0.84	(0.78, 0.91)	0.90	(0.60, 1.09)	0.96	(0.66, 1.14)	1.07	(0.69, 1.30)
4b: Var. gaptimes + Count	0.96	(0.88, 1.04)	0.88	(0.65, 1.03)	0.88	(0.68, 1.01)	1.11	(0.83, 1.26)

4.5 Discussion

In this study we have extended the landmarking framework to accommodate and learn from repeatedly and informatively measured clinical predictors, specifically eGFR, in the context of predicting kidney failure and death in patients with Chronic Kidney Disease. We found that there may be small gains in predictive performance, in terms of both model discrimination and Brier Score, from doing so, especially in the prediction of death which is often a difficult task within this population. The prediction of kidney failure onset was not improved by allowing the model to learn from informative observation, however these models already have very strong discriminative ability. We did not observe any additional benefit from complex representations such as estimated frailty terms from recurrent events models over simply counting the number of observed measurements. We have further illustrated how prediction models in other clinical areas may be able to benefit from informative observation.

This is the first study (to our knowledge) to compare the predictive performance of multiple representations of informative observation in real world data. We obtained access to rich longitudinal data, derived from real world care processes, that contain every observed eGFR measurement for all patients in our cohort across community, inpatient and outpatient locations, allowing us to model the full trajectory of each patient's kidney function. We have also provided a novel way of visualising observation processes that can reveal potentially informative elements of the process, and inform how informative observation could be harnessed for gains in predictive performance.

A limitation of this study is that our cohort definition included new incidences of chronic kidney disease, defined as a sustained drop in eGFR (below 60ml/min/1.72m^2) for more than 3 months. There is, however, no guidance on how best to determine this sustained drop using routinely collected data. Our "strict" definition did not allow any normal eGFR measures (60ml/min/1.72m^2 or greater) within the 3 month period, whereas a more lenient definition would have been to include anybody with a mean or median of < 60 within the 3-month window. The choice of cohort definition could have implications for the severity of illness of these patients, and therefore impact model performance.

Due to the limited follow-up time available in our dataset, we were not able to explore the additional prediction horizon of 5-years, at which the Kidney Failure Risk Equation is commonly used. We would have observed many more events (especially kidney failure events) were this extended follow-up period available, and therefore different results may have been observed for this outcome at the longer prediction horizon.

By applying landmarking as opposed to joint models, our models could be applied and tested using readily available software and computational power. Extending the joint model in a similar way would require fitting trivariate joint models, increasing the complexity and computational burden to such a level that likely becomes intractable in large sample sizes often found within routinely collected data. We attempted fitting such trivariate models in this cohort, but were unsuccessful using readily available software and computational power. By instead modelling the longitudinal process and time-to-event outcomes separately, we were able to add additional flexibility and complexity to the form of the underlying models that would be more difficult to achieve in existing software. We observed some small gains in predictive performance from the two-stage landmarking approach compared to the classic last observation carried forward landmarking method. It would, however, be of interest to see how the joint modelling approach proposed by Choi et al. [12] compares to our landmarking approach, as others have found that joint modelling often outperforms landmarking [14]. A key limitation of the landmarking approach is that uncertainty in the predictions of the longitudinal marker do not propagate through to the time-to-event model(s), which may be a challenge in this case since the assumption of linearity in repeated eGFR values is perhaps dubious, and therefore any error in the predictions of eGFR might need to be accounted for in the time-to-event model.

All models fit in this study allow for repeated measures of eGFR, and consider the entire trajectory of eGFR values in the prediction of CKD progression rather than a single observation. We adopted a mixed effects model with random intercept and slope to model the longitudinal eGFR values, whilst fitting a natural cubic spline to the effect of time and allowing the coefficients of the spline terms to enter the random effects. This fairly complex parametrisation offered some small gains in predictive performance over the LOCF approach, but came with the attached additional computational cost of fitting mixed models at each landmark time. We have observed complex patterns of non-linearity in this cohort, that seem to vary between patients, so it is plausible that our chosen model did not fit the longitudinal progression of eGFR well. For the purpose of this study, our primary goal was to illustrate and compare the methods of incorporating an informative observation process, so adopting a common model across the different parametrisations of the observation process seemed appropriate. Diggle, Sousa, and Asar [21] proposed a Gaussian Process model, where individual variability away from linearity is captured by integrated Brownian motion. The authors have also provide an R package (now archived) to apply their proposed modelling strategy, so this would be an interesting avenue to explore further. Interestingly, Ferrer, Putter, and Proust-Lima [22] observed that landmarking is less sensitive to incorrect specification of the longitudinal marker process when compared against joint modelling, therefore if our eGFR model were to be misspecified, the effects of this on predictive performance may be mitigated by the choice of methodological framework (landmarking).

We compared a range of parametrisations of the (informative) observation process here, and this is the first study to conduct such comparisons head-to-head witin the context of prediction modelling. Goldstein et al. [17] observed improved predictive performance from including a count of the number of vital sign observations made over time, but did not explore any alternative parametrisations of the observation process. Gasparini et al. [13] compared a joint model that includes a recurrent events process for the gaptimes between observations against a simple linear mixed model that includes a cumulative count of the number of observations. They found a reduction in bias in the estimates of model parameters from the joint model, but did not quantify measures of predictive performance which is of primary interest here. Their setting also differed as their primary outcome was the same variable that was longitudinally observed, i.e. model eGFR based on the informative observation process of eGFR. In simulation settings, they found that the model that uses the “total number of observations” as a model predictor performs the worst, but that in their empirical analysis of longitudinal eGFR values, all compared models perform similarly and produce similar estimated trajectories. They note that the poor performance of their count model could be due to the fact that it “conditions on the future” since they condition on the total number of measurements, which is not known at earlier time points. This issue is avoided in our context whereby the landmarking framework dictates that only information observed prior to each landmark time can be used in the modelling process.

Our finding of no added benefit from the more complex representations of IO is promising for applied researchers, as again the count model can be more easily derived and applied in commonly used software packages compared to the frailty model, as well as being more interpretable to end-users than the latent frailty terms. This finding is in line with the literature in dynamic modelling, where often simple approaches to handling repeated measures of predictor variables have been found to perform as have more complex ones within the context of prediction [17, 23]. It would be possible to explore alternative model forms for the frailty model that may have provided additional gains in performance, if it was found that the form we adopted here was ill-fitting.

There are additional considerations to be factored into the implementation of models that learn from the observation process, especially where risk scores are updated over time. Once the model exists, clinicians are likely to adapt their behaviour based on the resulting predictions [24]. Under the CKD clinical exemplar, this would mean that patients at higher risk of CKD progression (as identified by the KFRE model) undergo more intensive eGFR monitoring, and it seems feasible that the model might be used in this way. This could in turn cause the model to break down, as the association between the observation process parameter and the outcome will likely change [25]. Therefore any such model should be continuously monitored to identify any decrease in its performance. Further research is required to develop methods that can continuously update in line with any changes to the observation process resulting from the implementation of the model. An additional avenue for further methodological research is to establish any gains in predictive performance from learning from informative observation within a joint modelling context rather than a landmarking one.

We recommend that applied researchers considering incorporating informative observation into a prediction model should first consider using simple parametrisations of the observation process, such as a count or rate of past measurements. Where appropriate, using such measures will aid the interpretation of the final model for end users. The effect of incorporating informative observation should be quantified by comparing against models that omit this element to establish whether this is appropriate in their particular context. These decisions are also clinical as well as statistical - we recommend working closely with clinical experts to identify the most informative elements of the observation process (i.e. count of all measurements, or only certain types/locations), and to further visualise these processes to understand their relationship with the outcome.

4.6 Conclusion

This is the first study that has assessed the impact of incorporating informative measurement processes in EHRs into an existing clinical prediction model. We applied and compared 6 different approaches to summarising informative measurement processes, and incorporated them in models for predicting kidney failure and death in patients with CKD. We found modest gains in model performance by doing so, but found that more complex methods offer no added value over simpler ones.

References

[1] A. S. Go et al. "Chronic Kidney Disease and the Risks of Death, Cardiovascular Events, and Hospitalization". In: *New England Journal of Medicine* 351.13 (Oct. 2009), pp. 1296–1305. DOI: 10.1056/NEJMOA041031.

[2] B. Bikbov et al. "Global, regional, and national burden of chronic kidney disease, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017". In: *The Lancet* 395.10225 (Feb. 2020), pp. 709–733. DOI: 10.1016/S0140-6736(20)30045-3/ATTACHMENT/234A6931-5886-48B10.1016/S0140-6736(20)30045-3.

[3] J. C. Lv and L. X. Zhang. "Prevalence and Disease Burden of Chronic Kidney Disease". In: *Advances in Experimental Medicine and Biology* 1165 (2019), pp. 3–15. DOI: 10.1007/978-981-13-8871-2_1.

[4] N. Tangri et al. "A predictive model for progression of chronic kidney disease to kidney failure". In: *JAMA - Journal of the American Medical Association* 305.15 (Apr. 2011), pp. 1553–1559. DOI: 10.1001/jama.2011.451.

[5] E. M. Pullenayegum and L. S. Lim. "Longitudinal data subject to irregular observation: A review of methods with a focus on visit processes, assumptions, and study design". In: *Statistical Methods in Medical Research* 25.6 (2014). DOI: 10.1177/0962280214536537.

[6] R. Sisk et al. “Informative presence and observation in routine health data: A review of methodology for clinical risk prediction”. In: *Journal of the American Medical Informatics Association* 28.1 (Jan. 2021), pp. 155–166. DOI: 10.1093/jamia/ocaa242.

[7] V. Sharma et al. “Adoption of clinical risk prediction tools is limited by a lack of integration with electronic health records”. In: *BMJ Health Care Informatics* 28 (2021), p. 100253. DOI: 10.1136/bmjhci-2020-100253.

[8] A. Tsiatis and M. Davidian. “Joint modeling of longitudinal and time-to-event data: an overview”. In: *Statistica Sinica* 14.3 (2004), pp. 809–834.

[9] D. Rizopoulos. *Joint models for longitudinal and time-to-event data : with applications in R*. Chapman and Hall/CRC, July 2012.

[10] H. C. van Houwelingen. “Dynamic Prediction by Landmarking in Event History Analysis”. In: *Scandinavian Journal of Statistics* 34.1 (Mar. 2007), pp. 70–85. DOI: 10.1111/J.1467-9469.2006.00529.X.

[11] Ö. Asar et al. “Joint modelling of repeated measurement and time-to-event data: an introductory tutorial”. In: *International Journal of Epidemiology* 44.1 (Feb. 2015), pp. 334–344. DOI: 10.1093/IJE/DYU262.

[12] Y.-H. Choi et al. “Joint nested frailty models for clustered recurrent and terminal events: An application to colonoscopy screening visits and colorectal cancer risks in Lynch Syndrome families”. In: *Statistical Methods in Medical Research* (July 2019), p. 096228021986307. DOI: 10.1177/0962280219863076.

[13] A. Gasparini et al. “Mixed effects models for healthcare longitudinal data with an informative visiting process: a Monte Carlo simulation study”. In: *Statistica Neerlandica* 74.1 (2018), pp. 5–23. DOI: 10.1111/stan.12188.

[14] D. Rizopoulos, G. Molenberghs, and E. M. Lesaffre. “Dynamic predictions with time-dependent covariates in survival analysis using joint modeling and landmarking”. In: *Biometrical Journal* 59.6 (Nov. 2017), pp. 1261–1276. DOI: 10.1002/BIMJ.201600238.

[15] E. Paige et al. “Landmark Models for Optimizing the Use of Repeated Measurements of Risk Factors in Electronic Health Records to Predict Future Disease Risk”. In: *American Journal of Epidemiology* 187.7 (July 2018), p. 1530. DOI: 10.1093/AJE/KWY018.

[16] R. H. Keogh et al. “Dynamic Prediction of Survival in Cystic Fibrosis: A Landmarking Analysis Using UK Patient Registry Data”. In: *Epidemiology* 30.1 (Jan. 2019), pp. 29–37. DOI: 10.1097/EDE.0000000000000920.

[17] B. A. Goldstein et al. “A comparison of risk prediction methods using repeated observations: an application to electronic health records for hemodialysis”. In: *Statistics in Medicine* 36.17 (July 2017), pp. 2750–2763. DOI: 10.1002/sim.7308.

[18] J. C. van Houwelingen and H. Putter. *Dynamic prediction in clinical survival analysis*. CRC Press, 2012.

[19] M. Munda, F. Rotolo, and C. Legrand. “parfm: Parametric Frailty Models in R”. In: *Journal of Statistical Software* 51.1 (Nov. 2012), pp. 1–20. DOI: 10.18637/JSS.V051.I11.

[20] H. Putter, M. Fiocco, and R. B. Gekus. “Tutorial in biostatistics: Competing risk and multi-state models”. In: *Statistics in Medicine* 26.11 (May 2007), pp. 2389–2430. DOI: 10.1002/sim.2712.

[21] P. J. Diggle, I. Sousa, and Ö. Asar. “Real-time monitoring of progression towards renal failure in primary care patients”. In: *Biostatistics* 16.3 (July 2015), pp. 522–536. DOI: 10.1093/BIOSTATISTICS/KXU053.

[22] L. Ferrer, H. Putter, and C. Proust-Lima. “Individual dynamic predictions using landmarking and joint modelling: Validation of estimators and robustness assessment”. In: *Statistical Methods in Medical Research* 28.12 (Dec. 2019), pp. 3649–3666. DOI: 10.1177/0962280218811837. arXiv: 1707.03706.

[23] M. J. Sweeting et al. “The use of repeated blood pressure measures for cardiovascular risk prediction: a comparison of statistical models in the ARIC study”. In: *Statistics in Medicine* 36.28 (Dec. 2017), pp. 4514–4528. DOI: 10.1002/sim.7144.

[24] M. C. Lenert, M. E. Matheny, and C. G. Walsh. “Prognostic models will be victims of their own success, unless…” In: *Journal of the American Medical Informatics Association* 26.12 (Dec. 2019), pp. 1645–1650. DOI: 10.1093/JAMIA/OCZ145.

[25] A. M. Alaa, S. Hu, and M. Schaar. “Learning from Clinical Judgments: Semi-Markov-Modulated Marked Hawkes Processes for Risk Prognosis”. In: *International Conference on Machine Learning (ICML)*. International Conference on Machine Learning (ICML), July 2017, pp. 60–69.

[26] A. R. Smith et al. “Estimated GFR Trajectories in Pediatric and Adult Nephrotic Syndrome: Results From the Nephrotic Syndrome Study Network (NEPTUNE)”. In: *Kidney Medicine* 2.4 (July 2020), pp. 407–417. DOI: 10.1016/J.XKME.2020.03.006.

[27] M. Weldegiorgis et al. “Longitudinal Estimated GFR Trajectories in Patients With and Without Type 2 Diabetes and Nephropathy”. In: *American journal of kidney diseases* 71.1 (Jan. 2018), pp. 91–101. DOI: 10.1053/J.AJKD.2017.08.010.

4.7 Supplementary materials for Chapter 4

4.7.1 Summary of confirmatory analyses for linear mixed models

Here we provide additional information on the confirmatory analyses conducted to decide on the most appropriate form of the linear mixed models.

We first fit a series of possible models to the entire dataset including all eGFR measurements for all landmark times. We explored different parametrisations of the fixed effects, random effects and the correlation structure of the residuals. All possible parametrisations were justified by clinical intuition and background knowledge.

Fixed effects

The main effects to be included in the mixed model were: age (at CKD onset), location of test (inpatient vs community), gender, and time on study. We are predominantly interested in the most appropriate functional form for the effect of time, since non-linear progressions of eGFR over time have frequently been observed [21, 26, 27]. We therefore propose either a linear model, or a natural cubic spline model (with the number of knots to be decided). These models were fit using the `nlme` and `splines` packages in R. The models were initially fit using maximum likelihood (as opposed to restricted maximum likelihood (REML)) for comparison via Akaike's Information Criterion (AIC).

Based on the considerable improvement in apparent model fit from the spline terms, we settled on this parametrisation for the fixed effects (Table 4.10). The optimal number of knots was chosen to be 3, since the improvement from including an additional knot is minimal.

Random effects

The structure of the random effects was also to be decided. Due to reported individual-level variability in the rate of decline of eGFR, a "random slope" model was initially compared against a model with only random intercept terms. We then explored the idea of allowing the coefficients of the spline model to vary randomly at the individual-level, i.e. fitting a model with "spline random effects". AICs for all of these models are displayed in Table 4.10 below.

We also considered adding structure to the residuals, fitting a continuous autoregressive (CAR1) process to the residuals to allow for the fact that observations that are closers in time are more likely to be more highly correlated. We obtained unexpected results in the AIC for this parametrisation, so explored the actual model fits further by plotting the predictions from each model (See Figure 4.7).

Table 4.10: AICs from fitted linear mixed models

Model	DF	AIC
Random Intercept	7	70711.6
Random Intercept and Slope	9	42111.15
Continuous AR1 process (residuals)	10	-56600.96
Random Intercept, Slope and Spline fixed effect	11	36514.41
Random Intercept, Slope and Spline fixed and random effects	18	21119.02

It is apparent that the “spline random effects” model appears to provide the best fit in terms of AIC. It also appears to capture individual-level variability well, as can be seen in Figure 4.7 of the predicted values from each model. This parametrisation was therefore adopted across each of the landmark time-specific linear mixed effects models.

Figure 4.7 further illustrates one of the key challenges in modelling eGFR in CKD patients; CKD patients are at risk of acute kidney injury (AKI), where kidney function suddenly and temporarily drops, but then returns to normal shortly after. Patients 10 and 11 in Figure 4.7 have clear episodes of AKI throughout their follow-up, but do not progress to kidney failure at those times. This also illustrates the requirement for patients to experience a **sustained** drop in kidney function to be diagnosed with CKD or kidney failure. AKI can also occur in patients with otherwise normal kidney function (non-CKD patients).

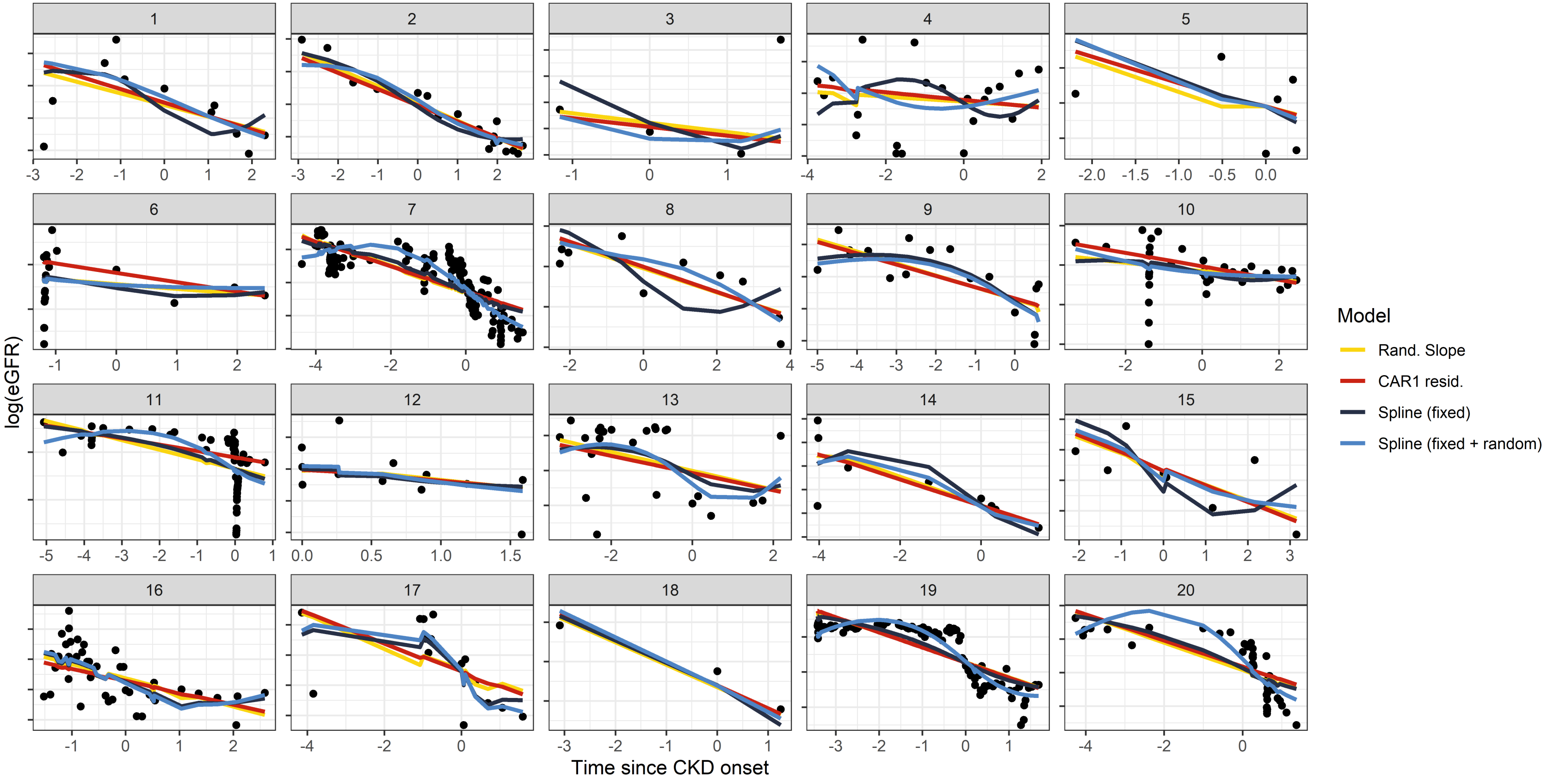


Figure 4.7: Observed eGFR trajectories from 20 randomly sampled patients, with lines showing the predicted values from each fitted model. 'Rand slope' model contains a random intercept and slope, 'CAR1 resid.' contains random intercept, slope and a continuous AR1 process on the residuals. 'Spline (fixed)' has a compound symmetry structure on the residuals, random intercept, slope and a natural cubic spline on the fixed effect of time. 'Spline (fixed + random)' is like 'Spline (fixed)' but with additional random coefficients for the spline terms

Custom code was written to predict eGFR from the fitted mixed models for out-of-sample individuals, using the estimated model coefficients and best linear unbiased predictors (BLUPs) for the random effects. This code was based on the source code for `IndvPred_lme` function which is part of the `JMbayes` package.

4.7.2 Model coefficients from linear mixed models

The "random spline" model was fit at each landmark time, including all eGFR data observed upto and including the landmark time, for all individuals still at risk at that landmark time (i.e. no event/censoring prior to time l). In Table 4.11 we summarise the model coefficients (fixed effects) for the linear mixed effects models, at each landmark time, fitted to the entire development dataset. We observe that the "test location" effect is significant at all landmark times, indicating that not only the frequency of observation, but also location of observation is associated with patient condition, perhaps unsurprisingly.

Table 4.11: Coefficients from the Linear Mixed Models fitted at each landmark time.

	l = 0		l = 0.25		l = 0.5		l = 0.75		l = 1		l = 1.25		l = 1.5	
Term	Est.	CI	Est.	CI	Est.	CI	Est.	CI	Est.	CI	Est.	CI	Est.	CI
Intercept	4.503	(4.478, 4.528)	4.488	(4.462, 4.514)	4.369	(4.343, 4.395)	4.285	(4.259, 4.311)	4.248	(4.222, 4.274)	4.232	(4.205, 4.258)	4.247	(4.219, 4.274)
Spline basis 1	-0.045	(-0.056, -0.034)	-0.053	(-0.065, -0.042)	-0.103	(-0.114, -0.091)	-0.159	(-0.171, -0.148)	-0.214	(-0.225, -0.203)	-0.254	(-0.265, -0.243)	-0.302	(-0.313, -0.290)
Spline basis 2	-0.449	(-0.485, -0.413)	-0.547	(-0.585, -0.509)	-0.296	(-0.334, -0.258)	-0.064	(-0.101, -0.026)	0.065	(0.029, 0.101)	0.113	(0.077, 0.148)	0.069	(0.034, 0.104)
Spline basis 3	-0.246	(-0.253, -0.239)	-0.387	(-0.396, -0.378)	-0.451	(-0.461, -0.441)	-0.423	(-0.434, -0.412)	-0.371	(-0.383, -0.360)	-0.327	(-0.338, -0.316)	-0.293	(-0.304, -0.282)
Gender: Male	-0.007	(-0.013, -0.002)	-0.005	(-0.011, 0.001)	-0.005	(-0.01, 0.001)	-0.004	(-0.01, 0.002)	-0.003	(-0.008, 0.003)	-0.001	(-0.007, 0.004)	-0.002	(-0.008, 0.005)
Age (CKD Onset)	-0.002	(-0.002, -0.002)	-0.002	(-0.002, -0.002)	-0.002	(-0.002, -0.002)	-0.002	(-0.002, -0.002)	-0.002	(-0.002, -0.002)	-0.002	(-0.002, -0.002)	-0.002	(-0.002, -0.002)
Test location: Inpatient	-0.004	(-0.007, -0.000)	-0.030	(-0.033, -0.026)	-0.035	(-0.038, -0.031)	-0.031	(-0.034, -0.027)	-0.028	(-0.032, -0.024)	-0.026	(-0.029, -0.022)	-0.025	(-0.028, -0.021)

Note:

The reference categories for Gender and Test Location are Female and Community respectively

4.7.3 Model coefficients from superlandmark models

We fit the 8 models described in Table 4.1 of the main text, where each non-null model allows a different representation of the observation process of eGFR measures over time. The estimated model coefficients for the models to predict kidney failure are displayed Table 4.12, and in Table 4.13 for the models to predict death. Note that these models were all fit with using robust standard errors, to account for the fact that each individual can contribute to multiple landmark risk sets, and therefore appears in the stacked landmark dataset multiple times.

Model coefficients for eGFR are for a 5-unit increase. All models have been fit on the entire development dataset.

All $\rho_i(l)$ parameters are statistically significant for the prediction of death, but not for the prediction of kidney failure. This is in line with our estimates of predictive performance, whereby informative patterns of eGFR observation provide gains in model performance for the prediction of death, but not for kidney failure.

Table 4.12: Model coefficients from the superlandmark models for kidney failure. Abbreviations: HR = Hazard Ratio, UACR = Urine to Albumin Creatinine Ratio

Model	Term	HR	SE	p.value
Model 1a: Null LOCF	eGFR (LOCF)	0.536	0.026	0.000
	Slope eGFR	1.002	0.001	0.208
	Age (CKD Onset)	0.953	0.004	0.000
	UACR	1.001	0.000	0.000
	Gender (Male, ref = Female)	2.691	0.160	0.000
Model 1b: Null LMM	eGFR (predicted)	0.510	0.029	0.000
	Age (CKD Onset)	0.953	0.004	0.000
	Slope eGFR	1.006	0.002	0.001
	UACR	1.002	0.000	0.000
	Gender (Male, ref = Female)	2.608	0.162	0.000
Model 2a: Count	eGFR (predicted)	0.499	0.032	0.000
	Age (CKD Onset)	0.952	0.004	0.000
	UACR	1.002	0.000	0.000
	Slope eGFR	1.005	0.002	0.008
	Gender (Male, ref = Female)	2.670	0.163	0.000
	Count	0.991	0.005	0.080

Model 3a: Frailty	eGFR (predicted)	0.517	0.029	0.000
	Age (CKD Onset)	0.952	0.004	0.000
	UACR	1.002	0.000	0.000
	Slope eGFR	1.006	0.002	0.000
	Gender (Male, ref = Female)	2.660	0.162	0.000
	Frailty	1.267	0.126	0.061
Model 3b: Frailty (past year)	eGFR (predicted)	0.513	0.030	0.000
	Age (CKD Onset)	0.953	0.004	0.000
	UACR	1.002	0.000	0.000
	Slope eGFR	1.006	0.002	0.000
	Gender (Male, ref = Female)	2.612	0.162	0.000
	Frailty (past year)	1.076	0.132	0.578
Model 2b: Inpatient Count	eGFR (predicted)	0.491	0.031	0.000
	Age (CKD Onset)	0.951	0.004	0.000
	UACR	1.002	0.000	0.000
	Slope eGFR	1.004	0.002	0.025
	Gender (Male, ref = Female)	2.751	0.163	0.000
	Inpatient Count	0.982	0.006	0.003
Model 4a: Var. gaptimes	eGFR (predicted)	0.509	0.029	0.000
	Age (CKD Onset)	0.953	0.004	0.000
	UACR	1.002	0.000	0.000
	Slope eGFR	1.006	0.002	0.000
	Gender (Male, ref = Female)	2.615	0.162	0.000
	Var. gaptimes	1.168	0.107	0.147
Model 4b: Var. gaptimes + Count	eGFR (predicted)	0.497	0.032	0.000
	Age (CKD Onset)	0.952	0.004	0.000
	UACR	1.002	0.000	0.000
	Slope eGFR	1.005	0.002	0.006
	Gender (Male, ref = Female)	2.676	0.163	0.000
	Count	0.991	0.005	0.090
	Var. gaptimes	1.155	0.109	0.187

Table 4.13: Model coefficients from the superlandmark models for death. Abbreviations: HR = Hazard Ratio, UACR = Urine to Albumin Creatinine Ratio

Model	Term	HR	SE	p.value
Model 1a: Null LOCF	eGFR (LOCF)	0.906	0.007	0.000
	Age (CKD Onset)	1.061	0.002	0.000
	UACR	1.001	0.000	0.000
	Slope eGFR	1.000	0.000	0.647
	Gender (Male, ref = Female)	1.428	0.028	0.000

Model 1b: Null LMM	eGFR (predicted)	0.931	0.011	0.000
	Age (CKD Onset)	1.066	0.002	0.000
	UACR	1.001	0.000	0.000
	Slope eGFR	0.992	0.001	0.000
	Gender (Male, ref = Female)	1.396	0.028	0.000
Model 2a: Count	eGFR (predicted)	0.951	0.009	0.000
	Age (CKD Onset)	1.069	0.002	0.000
	UACR	1.000	0.000	0.020
	Slope eGFR	0.995	0.001	0.000
	Gender (Male, ref = Female)	1.372	0.028	0.000
	Count	1.034	0.001	0.000
Model 3a: Frailty	eGFR (predicted)	0.929	0.010	0.000
	Age (CKD Onset)	1.067	0.002	0.000
	UACR	1.000	0.000	0.002
	Slope eGFR	0.995	0.001	0.000
	Gender (Male, ref = Female)	1.424	0.028	0.000
	Frailty	1.793	0.019	0.000
Model 3b: Frailty (past year)	eGFR (predicted)	0.919	0.010	0.000
	Age (CKD Onset)	1.067	0.002	0.000
	UACR	1.000	0.000	0.003
	Slope eGFR	0.995	0.001	0.000
	Gender (Male, ref = Female)	1.392	0.028	0.000
	Frailty (past year)	1.788	0.019	0.000
Model 2b: Inpatient Count	eGFR (predicted)	0.950	0.010	0.000
	Age (CKD Onset)	1.067	0.002	0.000
	UACR	1.000	0.000	0.002
	Slope eGFR	0.995	0.001	0.000
	Gender (Male, ref = Female)	1.370	0.028	0.000
	Inpatient Count	1.036	0.001	0.000
Model 4a: Var. gaptimes	eGFR (predicted)	0.938	0.011	0.000
	Age (CKD Onset)	1.066	0.002	0.000
	UACR	1.001	0.000	0.000
	Slope eGFR	0.991	0.001	0.000
	Gender (Male, ref = Female)	1.392	0.028	0.000
	Var. gaptimes	0.731	0.052	0.000
Model 4b: Var. gaptimes + Count	eGFR (predicted)	0.954	0.009	0.000
	Age (CKD Onset)	1.069	0.002	0.000
	UACR	1.000	0.000	0.024
	Slope eGFR	0.995	0.001	0.000
	Gender (Male, ref = Female)	1.370	0.028	0.000
	Count	1.033	0.001	0.000
	Var. gaptimes	0.872	0.046	0.003

4.7.4 Full set of performance metrics across all landmark times

In the main text of Chapter 4, we presented only a subset of the bootstrap validation results, but here we present the full set across all landmark times.

Table 4.14: C-statistics and confidence intervals for prediction of Kidney Failure

	l = 0		l = 0.25		l = 0.5		l = 0.75		l = 1		l = 1.25		l = 1.5	
Model	Mean C	CI	Mean C	CI	Mean C	CI	Mean C	CI	Mean C	CI	Mean C	CI	Mean C	CI
1a: Null: LOCF	0.872	(0.857, 0.887)	0.942	(0.928, 0.953)	0.966	(0.954, 0.973)	0.964	(0.953, 0.973)	0.989	(0.986, 0.991)	0.988	(0.986, 0.990)	0.984	(0.980, 0.986)
1b: Null: LMM	0.888	(0.860, 0.911)	0.953	(0.934, 0.966)	0.972	(0.965, 0.977)	0.970	(0.960, 0.975)	0.986	(0.983, 0.989)	0.972	(0.967, 0.975)	0.956	(0.948, 0.961)
2a: Count	0.883	(0.859, 0.908)	0.955	(0.934, 0.966)	0.973	(0.965, 0.978)	0.970	(0.961, 0.977)	0.986	(0.981, 0.989)	0.970	(0.964, 0.975)	0.955	(0.948, 0.961)
2b: Inpatient count	0.882	(0.854, 0.907)	0.956	(0.937, 0.966)	0.973	(0.966, 0.978)	0.970	(0.962, 0.976)	0.986	(0.984, 0.989)	0.971	(0.966, 0.975)	0.956	(0.948, 0.961)
3a: Frailty	0.899	(0.856, 0.922)	0.954	(0.934, 0.966)	0.971	(0.964, 0.976)	0.968	(0.958, 0.975)	0.986	(0.983, 0.989)	0.971	(0.966, 0.975)	0.953	(0.945, 0.962)
3b: Frailty (past year)	0.888	(0.856, 0.917)	0.953	(0.933, 0.965)	0.972	(0.964, 0.977)	0.969	(0.958, 0.976)	0.986	(0.981, 0.989)	0.971	(0.966, 0.975)	0.955	(0.948, 0.962)
4a: Var. gaptimes	0.881	(0.847, 0.910)	0.951	(0.930, 0.965)	0.970	(0.955, 0.976)	0.968	(0.956, 0.975)	0.986	(0.983, 0.989)	0.972	(0.967, 0.976)	0.957	(0.947, 0.965)
4b: Var. gaptimes + Count	0.876	(0.845, 0.906)	0.953	(0.929, 0.966)	0.971	(0.957, 0.978)	0.969	(0.957, 0.977)	0.986	(0.981, 0.989)	0.971	(0.965, 0.975)	0.957	(0.947, 0.965)

Table 4.15: C-statistics and confidence intervals for prediction of Death

	l = 0		l = 0.25		l = 0.5		l = 0.75		l = 1		l = 1.25		l = 1.5	
Model	Mean C	CI	Mean C	CI	Mean C	CI	Mean C	CI	Mean C	CI	Mean C	CI	Mean C	CI
1a: Null: LOCF	0.648	(0.646, 0.650)	0.668	(0.664, 0.672)	0.658	(0.657, 0.659)	0.654	(0.652, 0.658)	0.661	(0.659, 0.666)	0.672	(0.669, 0.676)	0.689	(0.686, 0.692)
1b: Null: LMM	0.645	(0.643, 0.648)	0.673	(0.671, 0.676)	0.666	(0.664, 0.668)	0.662	(0.660, 0.665)	0.670	(0.667, 0.673)	0.682	(0.680, 0.685)	0.694	(0.691, 0.697)
2a: Count	0.679	(0.676, 0.682)	0.708	(0.705, 0.711)	0.706	(0.702, 0.710)	0.709	(0.705, 0.714)	0.720	(0.716, 0.725)	0.727	(0.724, 0.731)	0.735	(0.732, 0.738)
2b: Inpatient count	0.675	(0.672, 0.679)	0.704	(0.703, 0.707)	0.701	(0.698, 0.704)	0.703	(0.700, 0.707)	0.714	(0.710, 0.718)	0.722	(0.720, 0.726)	0.730	(0.727, 0.734)
3a: Frailty	0.682	(0.681, 0.684)	0.702	(0.700, 0.704)	0.701	(0.699, 0.703)	0.699	(0.697, 0.702)	0.705	(0.703, 0.708)	0.719	(0.716, 0.721)	0.728	(0.725, 0.731)
3b: Frailty (past year)	0.677	(0.676, 0.679)	0.700	(0.699, 0.701)	0.700	(0.697, 0.702)	0.702	(0.699, 0.705)	0.705	(0.702, 0.707)	0.720	(0.718, 0.724)	0.726	(0.724, 0.729)
4a: Var. gaptimes	0.644	(0.640, 0.646)	0.674	(0.672, 0.677)	0.669	(0.667, 0.671)	0.665	(0.663, 0.668)	0.672	(0.669, 0.675)	0.686	(0.684, 0.689)	0.697	(0.694, 0.700)
4b: Var. gaptimes + Count	0.678	(0.674, 0.681)	0.707	(0.705, 0.711)	0.706	(0.703, 0.710)	0.710	(0.706, 0.715)	0.721	(0.717, 0.725)	0.728	(0.725, 0.732)	0.735	(0.732, 0.739)

Table 4.16: Brier Scores and confidence intervals for prediction of Kidney Failure

	l = 0		l = 0.25		l = 0.5		l = 0.75		l = 1		l = 1.25		l = 1.5	
Model	Brier	CI	Brier	CI	Brier	CI	Brier	CI	Brier	CI	Brier	CI	Brier	CI
1a: Null: LOCF	0.0139	(0.0112, 0.0180)	0.0078	(0.0061, 0.0106)	0.0068	(0.0052, 0.0098)	0.0073	(0.0055, 0.0106)	0.0088	(0.0063, 0.0132)	0.0086	(0.0059, 0.0134)	0.0073	(0.0050, 0.0113)
1b: Null: LMM	0.0082	(0.0070, 0.0102)	0.0079	(0.0063, 0.0107)	0.0054	(0.0045, 0.0071)	0.0055	(0.0045, 0.0075)	0.0083	(0.0063, 0.0122)	0.0074	(0.0053, 0.0111)	0.0058	(0.0044, 0.0087)
2a: Count	0.0088	(0.0072, 0.0104)	0.0083	(0.0065, 0.0113)	0.0055	(0.0046, 0.0071)	0.0055	(0.0045, 0.0075)	0.0083	(0.0062, 0.0121)	0.0076	(0.0053, 0.0113)	0.0060	(0.0044, 0.0089)
2b: Inpatient count	0.0091	(0.0074, 0.0111)	0.0085	(0.0065, 0.0119)	0.0056	(0.0046, 0.0073)	0.0055	(0.0044, 0.0075)	0.0085	(0.0063, 0.0124)	0.0079	(0.0055, 0.0119)	0.0065	(0.0046, 0.0096)
3a: Frailty	0.0081	(0.0069, 0.0102)	0.0080	(0.0065, 0.0112)	0.0056	(0.0046, 0.0077)	0.0057	(0.0046, 0.0082)	0.0086	(0.0066, 0.0127)	0.0077	(0.0056, 0.0113)	0.0061	(0.0046, 0.0091)
3b: Frailty (past year)	0.0083	(0.0069, 0.0103)	0.0080	(0.0064, 0.0109)	0.0055	(0.0045, 0.0078)	0.0056	(0.0045, 0.0081)	0.0084	(0.0064, 0.0125)	0.0075	(0.0054, 0.0112)	0.0059	(0.0043, 0.0090)
4a: Var. gaptimes	0.0084	(0.0071, 0.0106)	0.0082	(0.0064, 0.0121)	0.0056	(0.0045, 0.0082)	0.0057	(0.0045, 0.0083)	0.0085	(0.0063, 0.0124)	0.0076	(0.0053, 0.0111)	0.0059	(0.0044, 0.0088)
4b: Var. gaptimes + Count	0.0090	(0.0073, 0.0116)	0.0086	(0.0064, 0.0132)	0.0057	(0.0046, 0.0084)	0.0057	(0.0044, 0.0084)	0.0085	(0.0063, 0.0125)	0.0078	(0.0054, 0.0113)	0.0061	(0.0045, 0.0090)

Table 4.17: Brier Scores and confidence intervals for prediction of Death

	l = 0		l = 0.25		l = 0.5		l = 0.75		l = 1		l = 1.25		l = 1.5	
Model	Brier	CI	Brier	CI	Brier	CI	Brier	CI	Brier	CI	Brier	CI	Brier	CI
1a: Null: LOCF	0.111	(0.110, 0.112)	0.123	(0.122, 0.124)	0.124	(0.123, 0.126)	0.125	(0.123, 0.126)	0.128	(0.127, 0.130)	0.129	(0.127, 0.131)	0.128	(0.127, 0.130)
1b: Null: LMM	0.112	(0.111, 0.113)	0.120	(0.119, 0.122)	0.121	(0.120, 0.123)	0.122	(0.121, 0.123)	0.125	(0.124, 0.130)	0.126	(0.125, 0.128)	0.126	(0.125, 0.128)
2a: Count	0.109	(0.108, 0.111)	0.117	(0.116, 0.119)	0.119	(0.118, 0.121)	0.119	(0.118, 0.120)	0.122	(0.121, 0.120)	0.122	(0.121, 0.124)	0.123	(0.122, 0.125)
2b: Inpatient count	0.110	(0.109, 0.111)	0.118	(0.116, 0.119)	0.120	(0.119, 0.121)	0.119	(0.118, 0.121)	0.123	(0.121, 0.120)	0.123	(0.122, 0.125)	0.124	(0.122, 0.125)
3a: Frailty	0.111	(0.109, 0.112)	0.118	(0.117, 0.120)	0.119	(0.118, 0.120)	0.119	(0.118, 0.120)	0.123	(0.121, 0.120)	0.123	(0.122, 0.124)	0.124	(0.122, 0.125)
3b: Frailty (past year)	0.110	(0.109, 0.112)	0.118	(0.117, 0.120)	0.119	(0.118, 0.121)	0.119	(0.118, 0.120)	0.124	(0.122, 0.130)	0.125	(0.123, 0.127)	0.126	(0.125, 0.128)
4a: Var. gaptimes	0.112	(0.111, 0.114)	0.121	(0.120, 0.122)	0.122	(0.121, 0.123)	0.122	(0.121, 0.124)	0.126	(0.125, 0.130)	0.127	(0.125, 0.129)	0.127	(0.126, 0.129)
4b: Var. gaptimes + Count	0.110	(0.109, 0.111)	0.118	(0.117, 0.119)	0.120	(0.119, 0.121)	0.119	(0.118, 0.120)	0.122	(0.121, 0.120)	0.123	(0.121, 0.124)	0.124	(0.122, 0.125)

Table 4.18: Calibration Slopes and confidence intervals for prediction of Kidney Failure

	l = 0		l = 0.25		l = 0.5		l = 0.75		l = 1		l = 1.25		l = 1.5	
Model	Cal. Slope	CI	Cal. Slope	CI	Cal. Slope	CI	Cal. Slope	CI	Cal. Slope	CI	Cal. Slope	CI	Cal. Slope	CI
1a: Null: LOCF	0.77	(0.64, 0.87)	0.82	(0.37, 1.00)	0.92	(0.43, 1.13)	0.93	(0.43, 1.15)	1.00	(0.44, 1.25)	0.97	(0.41, 1.22)	0.99	(0.46, 1.18)
1b: Null: LMM	0.92	(0.85, 0.98)	0.74	(0.35, 0.91)	0.94	(0.44, 1.16)	0.98	(0.47, 1.18)	1.06	(0.52, 1.23)	1.03	(0.47, 1.21)	1.00	(0.49, 1.16)
2a: Count	0.91	(0.85, 0.98)	0.73	(0.35, 0.90)	0.95	(0.44, 1.19)	1.00	(0.48, 1.20)	1.07	(0.54, 1.24)	1.03	(0.48, 1.20)	0.99	(0.50, 1.15)
2b: Inpatient count	0.91	(0.84, 0.99)	0.73	(0.35, 0.90)	0.96	(0.45, 1.19)	1.01	(0.48, 1.22)	1.08	(0.55, 1.25)	1.01	(0.48, 1.19)	0.97	(0.50, 1.12)
3a: Frailty	0.92	(0.84, 0.99)	0.74	(0.35, 0.91)	0.93	(0.45, 1.15)	0.97	(0.48, 1.17)	1.05	(0.53, 1.22)	1.01	(0.48, 1.19)	0.98	(0.49, 1.14)
3b: Frailty (past year)	0.92	(0.83, 0.98)	0.74	(0.34, 0.91)	0.94	(0.44, 1.16)	0.98	(0.47, 1.17)	1.06	(0.52, 1.23)	1.03	(0.47, 1.20)	1.00	(0.49, 1.16)
4a: Var. gaptimes	0.90	(0.75, 0.98)	0.74	(0.34, 0.91)	0.94	(0.44, 1.16)	0.98	(0.47, 1.18)	1.06	(0.52, 1.23)	1.03	(0.47, 1.21)	1.00	(0.48, 1.16)
4b: Var. gaptimes + Count	0.89	(0.72, 0.98)	0.73	(0.34, 0.90)	0.95	(0.44, 1.18)	1.00	(0.47, 1.20)	1.07	(0.53, 1.24)	1.03	(0.47, 1.20)	0.99	(0.49, 1.15)

Table 4.19: Calibration Slopes and confidence intervals for prediction of Death

	l = 0		l = 0.25		l = 0.5		l = 0.75		l = 1		l = 1.25		l = 1.5	
Model	Cal. Slope	CI	Cal. Slope	CI	Cal. Slope	CI	Cal. Slope	CI	Cal. Slope	CI	Cal. Slope	CI	Cal. Slope	CI
1a: Null: LOCF	0.93	(0.86, 1.02)	0.91	(0.50, 1.09)	0.87	(0.52, 1.03)	0.85	(0.52, 1.01)	0.89	(0.52, 1.07)	0.92	(0.51, 1.12)	1.00	(0.53, 1.23)
1b: Null: LMM	0.87	(0.80, 0.95)	0.80	(0.54, 0.99)	0.90	(0.59, 1.09)	0.90	(0.62, 1.08)	0.96	(0.65, 1.14)	1.00	(0.66, 1.21)	1.06	(0.68, 1.28)
2a: Count	0.97	(0.88, 1.05)	0.90	(0.65, 1.09)	0.88	(0.65, 1.04)	0.86	(0.66, 0.99)	0.88	(0.67, 1.00)	1.04	(0.78, 1.20)	1.10	(0.82, 1.26)
2b: Inpatient count	0.96	(0.88, 1.03)	0.87	(0.61, 1.05)	0.87	(0.63, 1.01)	0.86	(0.65, 0.99)	0.91	(0.67, 1.04)	1.07	(0.76, 1.22)	1.09	(0.78, 1.24)
3a: Frailty	0.90	(0.84, 0.96)	0.91	(0.67, 1.04)	0.97	(0.71, 1.09)	0.98	(0.74, 1.09)	1.01	(0.77, 1.11)	1.06	(0.77, 1.18)	1.07	(0.79, 1.19)
3b: Frailty (past year)	0.87	(0.81, 0.93)	0.87	(0.66, 1.01)	0.92	(0.68, 1.06)	0.97	(0.72, 1.11)	1.00	(0.75, 1.15)	1.07	(0.79, 1.24)	1.07	(0.83, 1.21)
4a: Var. gaptimes	0.84	(0.78, 0.91)	0.79	(0.55, 0.97)	0.90	(0.60, 1.09)	0.91	(0.63, 1.09)	0.96	(0.66, 1.14)	1.02	(0.67, 1.23)	1.07	(0.69, 1.30)
4b: Var. gaptimes + Count	0.96	(0.88, 1.04)	0.90	(0.66, 1.08)	0.88	(0.65, 1.03)	0.86	(0.66, 0.99)	0.88	(0.68, 1.01)	1.05	(0.78, 1.20)	1.11	(0.83, 1.26)

Chapter 5

General Discussion

Each section of this thesis has contained chapter-specific discussions and conclusions, therefore this discussion relates more broadly to the aims of the thesis, how they have been addressed, and further areas of research that have been identified through the conduct of this work.

To recap, the aims of this thesis are:

1) Identify the extent to which IP and IO have been considered in the context of methodological prediction modelling research
2) Explore the use of missing indicators in combination with single and multiple imputation techniques to develop and apply CPMs under informative presence
3) Assess the added value of incorporating informative observation in a clinical prediction model.

5.1 Summary of findings

5.1.1 Aim 1: Existing methodology

We first aimed to identify the extent to which IP and IO have been considered in the context of methodological prediction modelling research. In order to identify and unify this body of literature, we proposed novel definitions of informative presence and informative observation that had not previously been identified. Although the concept of "informed presence" had been defined elsewhere [1], the proposed definition did not distinguish between cross-sectional and longitudinal measurement settings. We felt that the distinction is necessary, especially within the context of this thesis where the focus is on methodological considerations. The methodology to handle informative presence (cross-sectional) is different to those to handle informative observation (longitudinal). The review conducted in Chapter 2 of this thesis formally introduces this language. Indeed, methodology should be adapted to accommodate repeated measurements of predictor values in prediction modelling research. By introducing this language, we provided a unified way of considering observation processes within the EHR, and facilitate further discussions and research on these phenomena. Heterogeneity in language and the use of homonyms make it challenging for applied researchers to find relevant work, so by creating a unified language we aim to overcome this challenge.

Informative presence is often referred to as a missing data problem [2, 3]. Although there are clear parallels between the two, we argued that it is inappropriate to refer to "missing data" within EHR data. Since there is no expectation that particular observations are made, the focus should instead be on what is *present* (hence our choice of wording, informed *presence*) in the record rather than *missing* or absent. There are often multiple complex, and potentially interacting, processes that govern the way in which routinely collected patient data is observed [3], so we hope that by providing novel language this work can instigate a shift in the way observation processes are considered within EHR data, and especially in prediction modelling research with EHR data.

Note that this thesis has made reference to "missing data" in the traditional sense, to find the relevant methodology within our scoping review, and in the simulation in Chapter 3 to draw on existing definitions of missing data mechanisms. This was in acknowledgment of the fact that informative presence is most often considered as a missing data challenge, and of the parallels with missing data discussed above, to ensure that the work is picked up by the relevant audiences. EHR data represents a setting where the existing body of literature on missing data does not sufficiently capture the complexity in the underlying mechanisms and challenges, and therefore more specific work is needed to extend and build on existing guidelines (as we have done in Chapter 3).

This thesis has systematically assessed the extent to which informative presence and informative observation have been considered within the prediction modelling literature (Chapter 2), and how methods compare (Chapter 4). A significant, and growing, body of methodological literature exists in this area, yet many of the methods remain underdeveloped and little guidance on how to apply the existing methods has been provided to date. The most commonly proposed methodology simply involves including an additional predictor representing presence or absence of a predictor value [2, 4], or counting the number of measurements over time. We grouped similar strategies together into a "derived predictors" category. The missing indicator method is perhaps the most well known method under this category - it is relatively straightforward, and has repeatedly been shown to perform well in the context of prediction [2, 4, 5]. The work in Chapters 2 and 3 aids applied researchers in understanding and digesting the existing methodology, and provides recommendations on how and when to apply it.

Sharafoddini et al. [2] and Lipton et al. [4] have illustrated the added benefit of including missing indicators, and other predictors related to the observation process, in predicting outcomes in critical care. Their work focuses on improvements in discrimination (AUC), with no consideration of how well calibrated such models are. The discovery of such work therefore posed the question of how sensible (or well calibrated) the predictions from such models are, and how feasibly they could be applied in new patients.

5.1.2 Aim 2: Missing indicators and prediction

Chapter 3 performed a thorough investigation of the "missing indicator" method in the context of prediction, allowing the missing item(s) to be imputed via regression imputation (RI) or multiple imputation (MI). Since there are practical and methodological challenges associated with applying MI within prediction research [6], these were taken into consideration in the design and conduct of the study so that guidance could be developed dictating how best to apply these methods in practice. A notable finding from this work, that has not previously been identified, is that methods for handling missing data (and in turn informative presence) at the development stage of a CPM should be optimised according to whether or not missing data is allowed at the point of prediction. We can consider two extremes to illustrate the possible settings: firstly, a model that operates purely based on EHR data without input or additional information provided by users, with a built-in missing data handling strategy. At the other extreme would be a model that operates purely based on a form that must be manually completed, and does not allow any missing values.

Our findings are in line the existing literature to some extent, but previous guidance on missing data handling strategies advocates for inclusion of the outcome in the imputation model when applying MI [7, 8], without consideration of how this decision fits into the broader goal of the CPM, and how subsequent stages of the model pipeline could be affected. These two key findings therefore highlight the need to consider how prediction models should be developed with the entire model pipeline in mind, and how the impact of decisions made during the development stage could propagate through to validation and implementation.

A key finding throughout this thesis is that established best practices from descriptive or inferential frameworks are not always well suited to the aims of prediction research, and therefore methodological approaches to handling informative presence and informative observation should be tailored to the prediction modelling framework.

5.1.3 Aim 3: Empirical work

Chapter 4 of this thesis further explored the idea of including "derived predictors" as representations of the observation process in CPMs [9], this time in the context of informative observation and longitudinal predictor measurements. Using a clinical exemplar of Chronic Kidney Disease (CKD), we illustrate the flexibility of this method. There exist a broad range of possible representations of the observation process that can be incorporated into a CPM (based on essentially any underlying model type) with relative ease, and this chapter explores the use of simple counts as well as more complex model-derived parametrisations. In this work, we find that allowing the model to learn from not only the results of key lab tests, but their frequency too, can offer gains in predictive performance of the model in terms of both discrimination and overall model performance (Brier Score). We observed little to no predictive benefit in the more complex representations of the observation process compared to simpler options like a count or rate. This confirms our hypothesis that information held within the observation process can be informative above and beyond the data directly recorded in the EHR, and can be harnessed for predictive gain.

5.1.4 Summary

To sum up, although this thesis has a heavy focus on methodology, the requirements of the applied researcher have been at the forefront of the design and interpretation of all included work. The methodological review provides a useful summary of the current state of the art methodological development in this area, signposts readers for further reading on the specific methods, and provides an indication of where and when each method may be useful. The simulation in Chapter 3 provides guidance on how to select and apply well-known strategies within the context of CPM development/implementation under informative presence. Finally, the empirical work in Chapter 4 illustrates a range of possible ways of incorporating IO into a prediction model as well as highlighting ways of visualising and summarising informative observation processes, and provides recommendations on how to approach the development of a clinical prediction model using data subject to informative observation.

5.2 Future Work

The implications of this work extend to methodological researchers too, as we have identified priority areas for continued methodological development throughout this thesis. We have identified seven broad challenges that need further methodological research. Some of these have been discussed in Chapters 2-4, but we reiterate and summarise their relevance here in relation to the findings of this thesis.

An avenue of research that warrants further investigation is the use of joint models to incorporate informative observation in the development of clinical prediction models [10]. Joint models (JMs) have received considerable attention in the context of inferential and descriptive research [11], and they perform well in recovering unbiased association parameters within and between the underlying processes. They are, however, computationally intensive to fit and apply - a challenge that is especially relevant to the secondary analysis of electronic health record (EHR) data, as they can contain data on large numbers of patients and intensive monitoring frequencies. However, recent advances in statistical software packages allow users to fit flexible joint models e.g. in the `merlin` package in R and Stata. The use of joint models in incorporating IO in the development and implementation of CPMs, and how they compare to simpler methods like “derived predictors”, should be studied further. Simpler strategies for incorporating longitudinal information (such as summary measures or time-varying covariate models) often have comparable predictive performance to sophisticated joint modelling strategies in more typical dynamic prediction models that ignore IO [12], so similar results may be observed when JMs are extended to accommodate informative observation processes. An important trade-off exists between the potential gains in predictive performance of fitting a joint model, and the computational cost of doing so. This trade-off should be explored further - Chapter 4 of this thesis opted to use landmarking since JMs proved to be too much of a computational burden within this context.

CPMs that learn from informative observation processes inherently assume that the observation process (or its relationship with the outcome) does not change across the model pipeline, as with predictor-outcome relationships [13]. This may not, however, be a reasonable assumption if recording practices are highly variable across clinics, time periods or even clinicians. In this case, models that continuously monitor performance and update where necessary could be helpful to ensure model performance does not deteriorate as a result of changing measurement protocols [14, 15].

A more challenging issue arises if the existence of a model (that learns from IO) changes the way in which patients are observed. For example, a healthcare professional sees that a patient’s predicted mortality risk within the ICU is high, and in turn decides to monitor this patient more closely. This change in monitoring frequency then increases the predicted risk, creating a feedback loop. This is a key challenge in the implementation of models that learn from IO that could mean that their existence immediately degrades their performance. Alaa, Hu, and Schaar [16] propose an avenue of research that could overcome this challenge: by developing models that directly model the impact of previous risk estimates on the observation process, and also use continuous updating methods we have discussed here [14, 15]. A simpler, but similar, strategy to mitigate this issue is to include any previous predicted risks as a predictor in the model in interaction with the observation process, allowing it to account for possible changes in the underlying associations [16, 17]. It is becoming increasingly evident that any models that learn from informative presence or informative observation are at an increased risk of becoming “victims of their own success” [17, 18], and should be regularly (or even continuously) monitored and adapted in order to prevent this from happening.

This thesis has explored the idea of allowing CPMs to learn from informatively observed predictor measurements, however there is the potential to extend this to developing models that not only incorporate but also influence the observation process, so called "interactive measurement". When key predictors are missing at the point of prediction, certain biomarkers will likely be more informative with respect to an individual's (present or future) condition compared to others. So developing models that can identify which (if any) additional information should be collected in order to reduce the uncertainty in the predicted risk would be highly interesting. This idea has previously been explored in the context of diagnostic models under a "sequential testing" framework [19, 20], but the same ideas could be useful in prognostic modelling. Alternatively, models that update predictions over time could be used to trigger more intensive monitoring patterns once risk rises above a certain threshold. Explicit modelling of this process would be required to avoid the feedback loop discussed in the previous paragraph.

A related issue is that models that are widely accepted and used in practice often handle missing data (and therefore IP) inconsistently across different stages of the model pipeline [21]. This is most likely a pragmatic decision to allow models to predict in the absence of some of the model predictors, but without access to significant computing power in order to run the same techniques that were used at development (i.e. MI). The impact of this decision is currently unknown. Given that our work found that inconsistent imputation models between development and deployment resulted in reduced predictive performance, there is a strong possibility that using entirely different methods will have a similar effect. Commonly used combinations of missing data handling strategies should be compared to assess any change in model discrimination or calibration resulting from the inconsistency, and to identify combinations that tend to perform well together.

Broader challenges exist with the development of CPMs from EHR data, and further methodological research is warranted to extract their full value. We have considered the issue of allowing CPMs to learn from longitudinal biomarker trajectories in Chapter 4 of this thesis. Although this is becoming an increasingly common feature of CPMs [22], little guidance exists on some of the key challenges associated with learning from longitudinal data. Guidance on performing sample size calculations for the development of static CPMs have recently been developed

[23], but this is a remaining challenge in the context of dynamic CPMs. Both the number of patients, and number of repeated measures per patient form the total dataset size, and informative observation processes introduces variable measurement frequencies across patients. Guidance on calculating sufficient sample size and repeated measure frequency should be produced. Existing recommendations are tailored to the type of underlying model, so ideally recommendations should be made for each of the methodological frameworks presented by Bull et al. [22].

Finally, a relevant consideration in the analysis of EHR data, that we explored in Chapter 4, is the care setting within which predictor measurements are made, as the severity of a patient's condition will affect both the care setting they are treated within, and their resulting physiology. For example, a patient visiting their GP for a routine check-up is likely to have more normal, or at least controlled, blood pressure than patients presenting at an A&E department or during an inpatient stay. In the analysis in Chapter 4, we explicitly modelled this effect by allowing eGFR to be measured with error, and fitting a mixed effect model to its longitudinal progression. The care setting of each test was included as a predictor in this model to account for the heterogeneity across clinical settings, and was observed to be significantly associated with eGFR. Further work is needed in this area to explore this topic in other clinical areas. Measurement heterogeneity between development and validation cohorts can result in poor calibration and discrimination [24], but developing models that are applicable to multiple clinical settings, and exploiting information on setting could offer further benefits in predictive performance.

5.3 Strengths and Limitations

A key strength to this thesis is the provision of both simulated and empirical results to demonstrate and expand on methods for incorporating informative presence and observation in clinical prediction models. We have further provided clear recommendations to applied researchers on how best to approach informative presence and observation in the context of prediction model development based on the findings of this research. Through the conduct of this work, we have identified key areas that require further development and consideration from both the applied and methodological research communities, setting out an agenda for future work.

A limitation of this work is the lack of theoretical results or methodological development in this area, though these do not yet exist within the literature on informative observation. Throughout this work we uncovered a substantial body of existing literature and methodology that attempts to handle informative presence and observation, and therefore focused on improved understanding of existing methods as opposed to development of novel ones. There is, however, scope to develop novel methodology in this area or further expand on existing methods. A further limitation of this work is in the complexity of applying models that incorporate informative presence/observation into clinical practice, and how their performance may begin to degrade once implemented into practice. As discussed in the section on Future Work, it is necessary to establish the effect that such a model might have on its own future predictive performance if its existence changes monitoring behaviors. Although this was not explored within the context of this thesis, we have identified possible methodological frameworks that could mitigate these issues, and further study in this area should form the foundation of future work.

5.4 Practical implications

Based on our findings, we recommend that applied researchers tasked with developing a prediction model based on data subject to informative presence or observation should consider the following. First, establish the nature of the information in the observation process. This should ideally be done in consultation with clinical experts that can advise on exactly which aspect of the observation process is informative, such as the visiting process, clinician vs patient-initiated visits, or the type of information observed at each visit. As described in Chapter 4, visualising the observation process can be helpful here to identify key features that might be informative with respect to patient condition.

As explained in Chapter 3, for informatively observed predictors, it should further be established whether these should be allowed to be missing at the point of prediction and if so, how this missingness should or could be handled within the intended clinical application context. Chapter 3 of this thesis provides further guidance and more specific recommendations on this. In the context of repeatedly (and informatively) measured predictors, researchers should consider whether the longitudinal trajectory of the predictor has predictive value, whether predictions should be updated as new information becomes available, and therefore the most suitable methodological framework for allowing the model to achieve these goals (e.g. cross-sectional prediction vs dynamic prediction). Chapters 2 and 4 provide further guidance on selecting the most appropriate methodology, and applying this in practice to real data, but we recommend the use of simple derived predictors such as missing indicators and summary measures as a first option.

5.5 Conclusions

In conclusion, this thesis has provided considerable evidence that observation processes can be harnessed for predictive benefit during the development and implementation of clinical prediction models. We have provided guidance on how to apply and extend existing methods, and how to allow existing models to incorporate informative presence or informative observation.

The findings of this thesis demonstrate the need to carefully consider the drivers of observation processes within the EHR on a study-by-study basis, and in particular to consider **what** we observe and **why**, and how/whether this information can be harnessed within the development and implementation of CPMs. Multiple complex processes govern the way in which we observe data in the EHR, so working alongside clinicians or experts that understand the underlying care processes and patient pathways is essential. By identifying the relevant aspects of informative presence and informative observation, we have demonstrated how and when they can be exploited for predictive benefit.

References

[1] M. Phelan, N. A. Bhavsar, and B. A. Goldstein. "Illustrating Informed Presence Bias in Electronic Health Records Data: How Patient Interactions with a Health System Can Impact Inference". In: *EGEMS (Washington, DC)* 5.1 (2017). DOI: 10.5334/egems.243.

[2] A. Sharafoddini et al. "A new insight into missing data in intensive care unit patient profiles: Observational study". In: *Journal of Medical Internet Research* 21.1 (Jan. 2019). DOI: 10.2196/11605.

[3] S. Haneuse, D. Arterburn, and M. J. Daniels. "Assessing Missing Data Assumptions in EHR-Based Studies: A Complex and Underappreciated Task". In: *JAMA Network Open* 4.2 (Feb. 2021), e210184–e210184. DOI: 10.1001/JAMANETWORKOPEN.2021.0184.

[4] Z. C. Lipton et al. "Directly Modeling Missing Data in Sequences with RNNs: Improved Classification of Clinical Time Series". In: *Proceedings of the 1st Machine Learning for Healthcare Conference*. 2016, pp. 253–270.

[5] J.-H. Lin and P. J. Haug. "Exploiting missing clinical data in Bayesian network modeling for predicting medical problems". In: *Journal of Biomedical Informatics* 41.1 (Feb. 2008), pp. 1–14. DOI: 10.1016/J.JBI.2007.06.001.

[6] S. Fletcher Mercaldo and J. D. Blume. "Missing data and prediction: the pattern submodel". In: *Biostatistics* (Sept. 2018). DOI: 10.1093/biostatistics/kxy040.

[7] K. G. Moons et al. "Using the outcome for imputation of missing predictor values was preferred". In: *Journal of Clinical Epidemiology* 59.10 (Oct. 2006), pp. 1092–1101. DOI: 10.1016/J.JCLINEPI.2006.01.009.

[8] E. Steyerberg. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Second. Springer, 2019. DOI: 10.1007/978-0-387-77244-8.

[9] R. Sisk et al. "Informative presence and observation in routine health data: A review of methodology for clinical risk prediction". In: *Journal of the American Medical Informatics Association* 28.1 (Jan. 2021), pp. 155–166. DOI: 10.1093/jamia/ocaa242.

[10] Y.-H. Choi et al. "Joint nested frailty models for clustered recurrent and terminal events: An application to colonoscopy screening visits and colorectal cancer risks in Lynch Syndrome families". In: *Statistical Methods in Medical Research* (July 2019), p. 096228021986307. DOI: 10.1177/0962280219863076.

[11] E. M. Pullenayegum and L. S. Lim. "Longitudinal data subject to irregular observation: A review of methods with a focus on visit processes, assumptions, and study design". In: *Statistical Methods in Medical Research* 25.6 (2014). DOI: 10.1177/0962280214536537.

[12] M. J. Sweeting et al. "The use of repeated blood pressure measures for cardiovascular risk prediction: a comparison of statistical models in the ARIC study". In: *Statistics in Medicine* 36.28 (Dec. 2017), pp. 4514–4528. DOI: 10.1002/sim.7144.

[13] M. van Smeden, R. H. Groenwold, and K. G. Moons. "A cautionary note on the use of the missing indicator method for handling missing data in prediction research". In: *Journal of Clinical Epidemiology* 125 (Sept. 2020), pp. 188–190. DOI: 10.1016/j.jclinepi.2020.06.007.

[14] D. A. Jenkins et al. "Dynamic models to predict health outcomes: current status and methodological challenges". In: *Diagnostic and Prognostic Research 2018 2:1* 2.1 (Dec. 2018), pp. 1–9. DOI: 10.1186/S41512-018-0045-2.

[15] D. A. Jenkins et al. "Continual updating and monitoring of clinical prediction models: time for dynamic prediction systems?" In: *Diagnostic and Prognostic Research 2021 5:1* 5.1 (Jan. 2021), pp. 1–7. DOI: 10.1186/S41512-020-00090-3.

[16] A. M. Alaa, S. Hu, and M. Schaar. "Learning from Clinical Judgments: Semi-Markov-Modulated Marked Hawkes Processes for Risk Prognosis". In: *International Conference on Machine Learning (ICML)*. International Conference on Machine Learning (ICML), July 2017, pp. 60–69.

[17] M. C. Lenert, M. E. Matheny, and C. G. Walsh. "Prognostic models will be victims of their own success, unless..." In: *Journal of the American Medical Informatics Association* 26.12 (Dec. 2019), pp. 1645–1650. DOI: 10.1093/JAMIA/OCZ145.

[18] M. Sperrin et al. "Explicit causal reasoning is needed to prevent prognostic models being victims of their own success". In: *Journal of the American Medical Informatics Association : JAMIA* 26.12 (Nov. 2019), p. 1675. DOI: 10.1093/JAMIA/OCZ197.

[19] M. A. Gleser and M. F. Collen. "Towards automated medical decisions". In: *Computers and Biomedical Research* 5.2 (Apr. 1972), pp. 180–189. DOI: 10.1016/0010-4809(72)90080-8.

[20] U. Arslan et al. “Sequential test selection by quantifying of the reduction in diagnostic uncertainty for the diagnosis of proximal caries.” In: *Balkan medical journal* 30.2 (June 2013), pp. 142–6. DOI: 10.5152/balkanmedj.2012.103.

[21] A. Tsvetanova et al. “Missing data was handled inconsistently in UK prediction models: a review of method used”. In: *Journal of Clinical Epidemiology* 140 (Sept. 2021), pp. 149–158. DOI: 10.1016/J.JCLINEPI.2021.09.008.

[22] L. M. Bull et al. “Harnessing repeated measurements of predictor variables for clinical risk prediction: a review of existing methods”. In: *Diagnostic and Prognostic Research* 4.1 (Dec. 2020), p. 9. DOI: 10.1186/s41512-020-00078-z.

[23] R. D. Riley et al. “Calculating the sample size required for developing a clinical prediction model”. In: *BMJ* 368 (Mar. 2020). DOI: 10.1136/BMJ.M441.

[24] K. Luijken et al. “Impact of predictor measurement heterogeneity across settings on the performance of prediction models: A measurement error perspective”. In: *Statistics in Medicine* 38.18 (Aug. 2019), pp. 3444–3459. DOI: 10.1002/SIM.8183. arXiv: 1806.10495.