Clinical prediction models to predict the risk of multiple binary outcomes

Martin, Glen P; Sperrin, Matthew; Snell, Kym I E; Buchan, Iain; Riley, Richard D

DOI:
10.1002/sim.8787

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (Harvard):
Martin, GP, Sperrin, M, Snell, KIE, Buchan, I & Riley, RD 2021, 'Clinical prediction models to predict the risk of multiple binary outcomes: a comparison of approaches', *Statistics in Medicine*, vol. 40, no. 2, pp. 498-517. https://doi.org/10.1002/sim.8787

Received: 21 January 2020 | Revised: 25 August 2020 | Accepted: 7 October 2020

DOI: 10.1002/sim.8787

RESEARCH ARTICLE

Statistics in Medicine WILEY

Clinical prediction models to predict the risk of multiple binary outcomes: a comparison of approaches

Glen P. Martin[1] | **Matthew Sperrin**[1] | **Kym I. E. Snell**[2] | **Iain Buchan**[3] | **Richard D. Riley**[2]

[1]Division of Informatics, Imaging and Data Science, Faculty of Biology, Medicine and Health, Manchester Academic Health Science Centre, University of Manchester, Manchester, UK

[2]Centre for Prognosis Research, School of Primary, Community and Social Care, Keele University, Staffordshire, UK

[3]Institute of Population Health Sciences, Faculty of Health and Life Sciences, University of Liverpool, Liverpool, UK

Correspondence
Glen P. Martin, Health Data Science Division of Informatics, Imaging and Data Science, Faculty of Biology, Medicine and Health, University of Manchester, Vaughan House, Manchester M13 9GB, UK.
Email: glen.martin@manchester.ac.uk

Funding information
Medical Research Council, Grant/Award Number: MR/T025085/1; National Institute for Health Research

Abstract

Clinical prediction models (CPMs) can predict clinically relevant outcomes or events. Typically, prognostic CPMs are derived to predict the risk of a single future outcome. However, there are many medical applications where two or more outcomes are of interest, meaning this should be more widely reflected in CPMs so they can accurately estimate the joint risk of multiple outcomes simultaneously. A potentially naïve approach to multi-outcome risk prediction is to derive a CPM for each outcome separately, then multiply the predicted risks. This approach is only valid if the outcomes are conditionally independent given the covariates, and it fails to exploit the potential relationships between the outcomes. This paper outlines several approaches that could be used to develop CPMs for multiple binary outcomes. We consider four methods, ranging in complexity and conditional independence assumptions: namely, probabilistic classifier chain, multinomial logistic regression, multivariate logistic regression, and a Bayesian probit model. These are compared with methods that rely on conditional independence: separate univariate CPMs and stacked regression. Employing a simulation study and real-world example, we illustrate that CPMs for joint risk prediction of multiple outcomes should only be derived using methods that model the residual correlation between outcomes. In such a situation, our results suggest that probabilistic classification chains, multinomial logistic regression or the Bayesian probit model are all appropriate choices. We call into question the development of CPMs for each outcome in isolation when multiple correlated or structurally related outcomes are of interest and recommend more multivariate approaches to risk prediction.

KEYWORDS

binary outcomes, clinical prediction model, multiple outcomes, multivariate modeling, regression, risk prediction

1 | INTRODUCTION

Clinical prediction models (CPMs) aim to predict the probability that clinically relevant outcomes are present (diagnostic prediction) or will occur in the future (prognostic prediction) for an individual, given information known about them at the time of prediction.[1-3] CPMs are predominately derived in a multivariable regression framework (eg, logistic regression for binary outcomes), which combine estimated associations between multiple predictors (risk or prognostic factors) and an outcome of interest.

Generally, different CPMs are developed in isolation, where each model considers only a single outcome. However, there are many medical applications where two or more outcomes are of interest. As such, this should be more widely reflected in CPMs so they can accurately estimate the joint risk of multiple outcomes simultaneously.[4,5] For example, clinical teams consider mortality, morbidity, and quality of life in their decision-making for performing cardiovascular surgery, but surgical risk models (which are widely considered integral to surgical practice) are usually developed to predict single outcomes.[6,7] Another motivating example is in predicting likely outcomes during and after pregnancy, which often requires a multivariate perspective.[8] As a final motivating example, individuals are increasingly developing multiple diseases over their lifetime (ie, multimorbidity), but the plethora of CPMs developed to predict risks of common noncommunicable diseases such as cardiovascular disease,[9] types of cancer,[10] and chronic kidney disease[11] are usually developed in isolation. In each of these motivating examples, a multivariate (multiple outcome) approach to prediction is required for any CPM to be of maximum clinical value. For instance, within the multimorbidity example, managing patients with multimorbidity is not the same as managing each disease separately because of treatment side-effects, tolerances, and interactions. Hence, for CPMs to assist in multimorbidity resource planning and management, one needs to be able to estimate the (joint) risk of different combinations of conditions co-occurring,[12,13] which is only possible from taking a multivariate approach to prediction.

A naïve approach to estimating joint risk of multiple outcomes is to multiply the predicted risks from the univariate CPMs of each outcome. However, this approach will only lead to reliable estimates of the joint risk if the outcomes are conditionally independent given the covariates. Additionally, it fails to exploit the potential relationships between the outcomes, which could improve inference.[14,15] A common alternative approach is to create a composite outcome (defined as occurrence of any of the individual outcomes); for example, the definition of major adverse cardiovascular events.[16] While this approach simplifies the modeling, it makes it problematic to estimate marginal risks of each outcome separately (or different combinations of the outcomes co-occurring), and leads to a loss of important information.

Regression approaches that would allow multiple outcomes to be modeled simultaneously have a long history in the statistical literature.[14,15,17,18] Additionally, some prediction models have used multistate models to predict the risk of patients moving between different states (of diseases/conditions/pathways) pertaining to a combination of different outcomes through time and accounting for competing risks.[19-21] Similarly, problems in text classification and annotation have spawned machine learning techniques based on multilabel classification/learning.[22,23] These include binary relevance,[24] ensemble of classifier chains,[25] multilabel decision trees[22] and multilabel neural networks.[26]

Nonetheless, multiple outcome prediction methods are rarely utilized within the predictive modeling field,[4] and the effects of ignoring dependency between outcomes on predictive performance has received little attention.[5,27] In this study, we propose a variety of approaches for developing prognostic CPMs for multiple binary outcomes and compare their performance through a simulation study and real-world example. We consider estimation of both marginal and joint probabilities of outcomes and compare each method's ability to estimate these under different scenarios.

The remainder of the paper is structured as follows: in Section 2 we outline notation and present current univariate approaches to developing CPMs (ie, those that rely on conditional independence); we provide an overview of several methods to develop prognostic CPMs for multiple binary outcomes in Section 3; in Section 4 we describe the design and results of a simulation study comparing the methods, while in Section 5 we apply the methods to a real-world critical care example; finally, in Section 6 we discuss our findings and present directions for future work.

2 | PREDICTION APPROACHES UNDER CONDITIONAL INDEPENDENCE

2.1 | Notation and preliminaries

In all notations, we denote random variables with capital letters and observations of the random variable with corresponding lowercase letters. Throughout, we assume that the modeler has access to individual participant data (IPD) on a

population of interest. The IPD includes N independent observations of P predictor variables, arranged in an $N \times P$ matrix $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_P)$, with the (i, p)th element of $\boldsymbol{X}$ denoted as $x_{i,p}$. Additionally, each observation within the IPD has a set of K unique (but potentially related) binary outcomes, which we denote as $Y_{i1}, \ldots, Y_{iK}$, where $Y_{ij} = 1$ if observation i had the jth outcome event, with $Y_{ij} = 0$ otherwise. We assume that occurrence of one outcome does not preclude the occurrence of any of the others (eg, excluding death). For ease of exposition we describe each of the methods in the case where $K = 2$, but extensions to $K > 2$ follow naturally.

Unless otherwise stated, each of the methods described below (in Sections 2 and 3) are fitted using maximum likelihood estimation (MLE), and we assume that a suitable CPM development strategy is employed,[1,2,28] which may include adjustment for overfitting. For example, many of the methods described below could equally be fit using penalized regression, such as LASSO or fitting through Bayesian inference with penalizing priors, to help minimize overfitting.[2,29-32]

2.2 | Univariate CPMs

The naïve approach to estimating joint risk of developing multiple outcomes is to develop a univariate CPM for each outcome separately, for example, using logistic regression. Specifically, we have that

$$P(Y_{ij} = 1|\boldsymbol{X}_i) = [1 + \exp(-(\beta_{0,j} + \boldsymbol{X}_i\boldsymbol{\beta}_j))]^{-1}, \tag{1}$$

for $j = 1, \ldots, K$, where $\beta_{0,j}$ is the intercept and $\boldsymbol{\beta_j} = (\beta_{1,j}, \ldots, \beta_{P,j})$ is a vector of coefficients, which represent the conditional prognostic effects of the predictors on the jth outcome. Here, $\beta_{0,j} + \boldsymbol{X}_i\boldsymbol{\beta}_j$ is referred to as the linear predictor for individual i.

At the time of making a prediction for a new individual with covariates $\boldsymbol{X}_i^*$, the marginal predicted risk for the jth outcome is $P(Y_{ij} = 1|\boldsymbol{X}_i^*)$, while the joint probability $P(Y_{i1} = 1, \ldots, Y_{iK} = 1|\boldsymbol{X}_i^*)$ is $\prod_{j=1}^{K} P(Y_{ij} = 1|\boldsymbol{X}_i^*)$, meaning the approach relies on conditional independence of the outcomes to be valid.

2.3 | Stacked regression

One way to extend the univariate approach is based on the stacked regression literature,[32,33] which allows the model for one of the outcomes to exploit the information (ie, predictor-outcome associations) contained within the other outcome models, thereby improving marginal risk prediction. This approach can also be used in a setting where there are existing univariate CPMs available in the literature.[32,34,35] Stacked regression is a two-stage approach to model fitting, whereby individual CPMs are fit to each outcome independently using the IPD (or obtained from the literature), the linear predictors of which are then used in a second stage as covariates in a stacked regression model for each outcome.

Specifically, with $K = 2$, in the first stage we have that $\widehat{f}_1(\boldsymbol{X}) = \beta_{01} + \boldsymbol{X}\boldsymbol{\beta}_1$ and $\widehat{f}_2(\boldsymbol{X}) = \beta_{02} + \boldsymbol{X}\boldsymbol{\beta}_2$ (each estimated using unpenalized MLE of Equation (1), or obtained from the literature[32,34,35]). Then, in the second stage, we fit the following models in the IPD:

$$P(Y_{i1} = 1|\boldsymbol{X}_i) = \left[1 + \exp\left(-\left(\widehat{\eta}_{0,1} + \widehat{\eta}_{1,1}\widehat{f}_1(\boldsymbol{X_i}) + \widehat{\eta}_{2,1}\widehat{f}_2(\boldsymbol{X_i}) + \sum_{p=1}^{P}\widehat{\delta}_{p,1}x_{i,p}\right)\right)\right]^{-1}, \tag{2}$$

and

$$P(Y_{i2} = 1|\boldsymbol{X}_i) = \left[1 + \exp\left(-\left(\widehat{\eta}_{0,2} + \widehat{\eta}_{1,2}\widehat{f}_1(\boldsymbol{X_i}) + \widehat{\eta}_{2,2}\widehat{f}_2(\boldsymbol{X_i}) + \sum_{p=1}^{P}\widehat{\delta}_{p,2}x_{i,p}\right)\right)\right]^{-1}. \tag{3}$$

The unknown parameters $\widehat{\eta}_{1,1}, \widehat{\eta}_{2,1}, \widehat{\eta}_{1,2}, \widehat{\eta}_{2,2}, \widehat{\delta}_{p,1}$ and $\widehat{\delta}_{p,2}$, in Equations (2) and (3) are estimated by maximizing a lasso penalized likelihood,[36] as previously described.[32] A penalized likelihood is recommended here to help handle the highly colinear covariates in Equations (2) and (3), and to perform predictor selection for the final summations in these equations. The final summations in Equations (2) and (3) allow the individual predictor effects to differ dependent on the inclusion of $\widehat{f}_1(\boldsymbol{X_i})$ and $\widehat{f}_2(\boldsymbol{X_i})$ in each model.

At the time of prediction, the predicted marginal risks for the jth outcome for a new individual with covariates $\boldsymbol{X}_i^*$, are obtained by calculating $\widehat{f}_1(\boldsymbol{X}_i^*)$ and $\widehat{f}_2(\boldsymbol{X}_i^*)$, using the models obtained in the first stage, and then inserting these into the appropriate stacked regression model from the second stage (ie, Equations (2) or (3)). We consider this approach since it is a previously proposed alternative to univariate CPMs, but note that the joint probabilities for all outcomes, $P(Y_{i1} = 1, \ldots, Y_{iK} = 1|\boldsymbol{X_i})$ is still computed as $\prod_{j=1}^{K} P(Y_{ij} = 1|\boldsymbol{X}_i)$.

3 | PREDICTION APPROACHES ACCOUNTING FOR CONDITIONAL DEPENDENCE

In this section, we describe four approaches that can readily be applied to real-world data that relax the conditional independence assumption to enable joint outcome risk estimation. In all the approaches, we allow for differences in the predictors for each of the outcomes since some elements of $\boldsymbol{\beta}_j$ could be estimated (or fixed) to be zero.

3.1 | Probabilistic classifier chains

If the risks of multiple outcomes are related to each other (after conditioning on the covariates), then one approach to modeling dependence is to condition sequentially on each outcome. Consider the (randomly indexed) sequence of outcomes $Y_{i1}, \ldots, Y_{iK}$, then one can relax the conditional independence assumption by conditioning on $Y_{i1}, \ldots, Y_{ij-1}$ when predicting Y_{ij} instead of only the covariates. Since the order of the indexing (of $Y_{i1}, \ldots, Y_{iK}$) will affect inference, an iterative approach is used, whereby all the permutations of the ordering of $Y_{i1}, \ldots, Y_{iK}$ are considered. As K increases, so too does the number of "permutations"; here, one could pick a random sample of permutations, rather than fitting models on all $K!$ permutations—this resembles the ensembles of classifier chains approach.[25,33]

Specifically, where $K = 2$, the first "permutation" (denoted by a superscript (1)) is such that

$$\mathrm{P}(Y_{i1} = 1|\boldsymbol{X}_i) = \pi_{i1}^{(1)} = [1 + \exp(-(\beta_{0,1}^{(1)} + \boldsymbol{X_i}\boldsymbol{\beta}_1^{(1)}))]^{-1}$$

$$\mathrm{P}(Y_{i2} = 1|\boldsymbol{X}_i, Y_{i1}) = \pi_{i2}^{(1)} = [1 + \exp(-(\beta_{0,2}^{(1)} + \boldsymbol{X_i}\boldsymbol{\beta}_2^{(1)} + \gamma_1^{(1)}Y_{i1}))]^{-1},$$

while the second "permutation" (denoted by a superscript (2)) is such that

$$\mathrm{P}(Y_{i2} = 1|\boldsymbol{X}_i) = \pi_{i2}^{(2)} = [1 + \exp(-(\beta_{0,2}^{(2)} + \boldsymbol{X_i}\boldsymbol{\beta}_2^{(2)}))]^{-1}$$

$$\mathrm{P}(Y_{i1} = 1|\boldsymbol{X}_i, Y_{i2}) = \pi_{i1}^{(2)} = [1 + \exp(-(\beta_{0,1}^{(2)} + \boldsymbol{X_i}\boldsymbol{\beta}_1^{(2)} + \gamma_2^{(2)}Y_{i2}))]^{-1}.$$

All models are fitted separately using MLE (ie, the models in the first "permutation" are fitted independently of the models in the second "permutation"). This approach is based on ensemble probabilistic classification chains from the multilabel classification literature.[25,37]

Importantly, the conditioning on "preceding" outcomes allows us to derive analytical expressions for the joint probabilities using Bayes' rule and by taking an average ensemble across the permutation models. For example, with $K = 2$ we have the following (omitting the conditions on $\boldsymbol{X}_i$ for notational brevity):

$P(Y_{i1} = 1, Y_{i2} = 1) = \frac{1}{2}[\mathrm{P}(Y_{i2} = 1|Y_{i1} = 1)\mathrm{P}(Y_{i1} = 1) + \mathrm{P}(Y_{i1} = 1|Y_{i2} = 1)\mathrm{P}(Y_{i2} = 1)]$
$= \frac{1}{2}[\pi_{i2}^{(1)}\pi_{i1}^{(1)} + \pi_{i1}^{(2)}\pi_{i2}^{(2)}]$,
$P(Y_{i1} = 1, Y_{i2} = 0) = \frac{1}{2}[\mathrm{P}(Y_{i2} = 0|Y_{i1} = 1)\mathrm{P}(Y_{i1} = 1) + \mathrm{P}(Y_{i1} = 1|Y_{i2} = 0)\mathrm{P}(Y_{i2} = 0)]$,
$P(Y_{i1} = 0, Y_{i2} = 1) = \frac{1}{2}[\mathrm{P}(Y_{i2} = 1|Y_{i1} = 0)\mathrm{P}(Y_{i1} = 0) + \mathrm{P}(Y_{i1} = 0|Y_{i2} = 1)\mathrm{P}(Y_{i2} = 1)]$,
$P(Y_{i1} = 0, Y_{i2} = 0) = \frac{1}{2}[\mathrm{P}(Y_{i2} = 0|Y_{i1} = 0)\mathrm{P}(Y_{i1} = 0) + \mathrm{P}(Y_{i1} = 0|Y_{i2} = 0)\mathrm{P}(Y_{i2} = 0)]$,

from which one can calculate the associated marginal probabilities of $P(Y_{i1} = 1|\boldsymbol{X}_i)$ and $P(Y_{i2} = 1|\boldsymbol{X}_i)$.

When predicting for a new individual (where clearly their outcomes are unknown), one would use the derived models to calculate the joint probabilities for all possible outcome combinations $Y_{ij} \in \{0, 1\}$ for $j = 1, \ldots, K$, which by definition sum to one.

3.2 | Multinomial logistic regression

A second approach to modeling outcome dependence, which allows calculation of the predicted risk of different combinations of $Y_{ij} \in \{0, 1\}$ for $j = 1, \ldots, K$, is to use multinomial logistic regression, where the 2^K combinations are each treated as a nominal outcome category. For example, with two outcomes ($K = 2$), we fit the following models:

$$\log\left(\frac{P(Y_{i1}=1, Y_{i2}=1)}{P(Y_{i1}=0, Y_{i2}=0)}\right) = \beta_{0,1} + \boldsymbol{X}_i\boldsymbol{\beta}_1.$$
$$\log\left(\frac{P(Y_{i1}=1, Y_{i2}=0)}{P(Y_{i1}=0, Y_{i2}=0)}\right) = \beta_{0,2} + \boldsymbol{X}_i\boldsymbol{\beta}_2.$$
$$\log\left(\frac{P(Y_{i1}=0, Y_{i2}=1)}{P(Y_{i1}=0, Y_{i2}=0)}\right) = \beta_{0,3} + \boldsymbol{X}_i\boldsymbol{\beta_3}.$$

These models are estimated using iterative procedures to find numerical optimization of the parameters; practically, such models can be fit using R with the package nnet.[38]

At the time of prediction for a new individual with covariates $\boldsymbol{X}_i^*$, we use the fact that the probabilities must sum to one, to allow us to explicitly obtain each joint probability. For the case with $K = 2$ we have

$$P(Y_{i1}=1, Y_{i2}=1|\boldsymbol{X}_i^*) = \frac{\exp(\beta_{0,1} + \boldsymbol{X}_i^*\boldsymbol{\beta}_1)}{1 + \sum_{k=1}^{3}\exp(\beta_{0,k} + \boldsymbol{X}_i^*\boldsymbol{\beta}_k)},$$
$$P(Y_{i1}=1, Y_{i2}=0|\boldsymbol{X}_i^*) = \frac{\exp(\beta_{0,2} + \boldsymbol{X}_i^*\boldsymbol{\beta}_2)}{1 + \sum_{k=1}^{3}\exp(\beta_{0,k} + \boldsymbol{X}_i^*\boldsymbol{\beta}_k)},$$
$$P(Y_{i1}=0, Y_{i2}=1|\boldsymbol{X}_i^*) = \frac{\exp(\beta_{0,3} + \boldsymbol{X}_i^*\boldsymbol{\beta}_3)}{1 + \sum_{k=1}^{3}\exp(\beta_{0,k} + \boldsymbol{X}_i^*\boldsymbol{\beta}_k)},$$
$$P(Y_{i1}=0, Y_{i2}=0|\boldsymbol{X}_i^*) = \frac{1}{1 + \sum_{k=1}^{3}\exp(\beta_{0,k} + \boldsymbol{X}_i^*\boldsymbol{\beta}_k)}.$$

3.3 | Multivariate logistic regression

Our third approach, which has previously been described in the context of modeling correlated binary outcomes, makes explicit use of a multivariate logistic distribution.[39] For ease of exposition, we again describe the case when $K = 2$ and readers refer to the literature for the more general case.[39] Explicitly, we use the bivariate logistic distribution proposed by Gumbel[40] to set

$$P(Y_{i1}=1, Y_{i2}=1|\boldsymbol{X}_i) = F_{i1}F_{i2} + \rho\sqrt{F_{i1}S_{i1}F_{i2}S_{i2}}$$

where $F_{ij} = P(Y_{ij}=1|\boldsymbol{X}_i)$ for $j = 1, 2$ (as defined in Equation (1)) and $S_{ij} = 1 - F_{ij}$. Here, ρ estimates the residual correlation between the outcomes. Similarly, we have the following:

$$P(Y_{i1}=1, Y_{i2}=0|\boldsymbol{X}_i) = F_{i1}S_{i2} - \rho\sqrt{F_{i1}S_{i1}F_{i2}S_{i2}},$$

$$P(Y_{i1}=0, Y_{i2}=1|\boldsymbol{X}_i) = S_{i1}F_{i2} - \rho\sqrt{F_{i1}S_{i1}F_{i2}S_{i2}},$$

$$P(Y_{i1}=0, Y_{i2}=0|\boldsymbol{X}_\mathrm{i}) = S_{i1}S_{i2} + \rho\sqrt{F_{i1}S_{i1}F_{i2}S_{i2}}.$$

We maximize the following (unpenaliszd) log-likelihood to estimate the parameters $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ and ρ:

$$l(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \rho) = \sum_{i=1}^{N} y_{i1}y_{i2}\log(p_{11i}) + y_{i1}(1-y_{i2})\log(p_{10i}) + (1-y_{i1})y_{i2}\log(p_{01i}) + (1-y_{i1})(1-y_{i2})\log(p_{00i}),$$

where $p_{abi} = P(Y_{i1} = a, Y_{i2} = b|\boldsymbol{X}_i)$. There are no closed-form solutions to maximize the derivatives of this log-likelihood and so numerical optimization is required. At the time of prediction, the joint and marginal probabilities of each outcome can be obtained directly for each new individual.

Of note is that the residual correlation parameter, ρ, is constrained by the marginal probabilities, and cannot therefore take values in the full $[-1, 1]$ range[39]; previous studies have shown this to impact the degrees of dependency between the outcomes that the approach can handle.[41,42]

3.4 | Multivariate Bayesian probit CPM

Our final approach follows naturally from the case of modeling multiple continuous outcomes through multivariate linear regression.[15] Limiting again to the case for $K = 2$ for ease of exposition, let Z_{i1} and Z_{i2} denote two latent variables for each outcome, such that

$$Y_{ij} = \begin{cases} 1 & \text{if } Z_{ij} > 0, \\ 0 & \text{if } Z_{ij} \leq 0 \end{cases}$$

where $Z_{ij} = \beta_{0j} + \boldsymbol{X_i\beta_j} + \epsilon_{ij}$. Dependency between the outcomes is induced by assuming a joint distribution on ϵ_{ij} for $j = 1$, 2. Given the benefits of specifying dependency through a multivariate normal distribution, we propose to fit a multivariate probit model on Y, by assuming

$$\begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{12} \\ \rho_{12} & 1 \end{pmatrix}\right),$$

where we take a common variance of one for model identifiability reasons.[43,44] We propose to fit this model using Bayesian inference, since parameter estimation can be obtained naturally using MCMC methods (which is necessary for $K > 2$). In this study we set the prior distributions to $\rho_{12} \sim \text{Unif}(-1, 1)$ and

$$\begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_1 & \boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2 \\ \boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2 & \boldsymbol{\Sigma}_2 \end{pmatrix}\right),$$

where $\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2 = \mathbf{0}$ and $\boldsymbol{\Sigma}_j = \text{diag}(10)$, for $j = 1$, 2. In most medical applications, it might be more appropriate to set $\rho_{12} \sim \text{Unif}(0, 1)$, since the correlation would usually be positive (or could be constructed to be such through transformation of the outcomes prior to modeling).[45]

At time of prediction for a new individual with covariates $\boldsymbol{X}_i^*$, estimates of $P(Y_{i1} = 1, Y_{i2} = 1|\boldsymbol{X}_i^*)$ can be obtained through the cumulative distribution function of the bivariate standard normal distribution, Φ, as $\Phi(\boldsymbol{X}_i^*\boldsymbol{\beta_1}, \boldsymbol{X}_i^*\boldsymbol{\beta_2}, \boldsymbol{\rho_{12}})$. Similarly, estimates of $P(Y_{i1} = 1, Y_{i2} = 0|\boldsymbol{X}_i^*)$ and $P(Y_{i1} = 0, Y_{i2} = 1|\boldsymbol{X}_i^*)$ can be obtained through $\Phi(\boldsymbol{X}_i^*\boldsymbol{\beta_1}, -\boldsymbol{X}_i^*\boldsymbol{\beta_2}, -\boldsymbol{\rho_{12}})$ and $\Phi(-\boldsymbol{X}_i^*\boldsymbol{\beta_1}, \boldsymbol{X}_i^*\boldsymbol{\beta_2}, -\boldsymbol{\rho_{12}})$, respectively.

In this study, we implemented this model using JAGS (Just Another Gibbs Sampler), using the R package rjags.[46] We took 10 000 posterior samples of each parameter and summaried them over the final 5000 samples (ie, 5000 burn-in).

4 | SIMULATION STUDY

4.1 | Aim

We designed a simulation study to investigate the effects on predictive performance of developing CPMs that model multiple outcomes using the aforementioned approaches, compared with modeling each outcome separately through univariate analyses. We designed and report the simulation in line with best practice.[47]

4.2 | Data-generating mechanisms

Throughout all simulations we assume that we have an IPD that includes 5000 individuals, on which one is interested in developing CPMs for two binary outcomes of interest. The IPD includes two continuous covariates that we generate as $\boldsymbol{X}_1 \sim N(0, 1)$ and $\boldsymbol{X}_2 \sim N(0, 1)$. Additionally, each observation within the IPD has two (potentially dependent) binary outcomes, $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$, which we simulate according to established methods.[48,49] Specifically, we generated the binary outcomes such that the marginal probabilities satisfied

$$P(Y_{i1} = 1|\boldsymbol{X}_i) = [1 + \exp(-(\beta_{01} + \beta_{11}X_{i1} + \beta_{21}X_{i2}))]^{-1},$$

$$P(Y_{i2} = 1|\boldsymbol{X}_i) = [1 + \exp(-(\beta_{02} + \beta_{12}X_{i1} + \beta_{22}X_{i2}))]^{-1},$$

where we fixed $\boldsymbol{\beta}_1 = (\beta_{01}, \beta_{11}, \beta_{21}) = (-1, \log(2), 0)$ and $\boldsymbol{\beta}_2 = (\beta_{02}, \beta_{12}, \beta_{22}) = (-1.5, 0, \log(3))$, meaning that $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ only predicted $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$, respectively. These coefficient values give (marginal) outcome proportions of 29% and 23% for $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$, respectively. We also considered a sensitivity analysis where β_{01} was set to -3 and β_{02} was set to -3.5, which results in lower (marginal) outcome proportions of 6% and 5% for $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$, respectively. Dependency between the outcomes was induced (while satisfying the above marginal probabilities) by generating two latent variables Z_{i1} and Z_{i2}, from a multivariate standard normal distribution, with correlation ρ. We then applied a probability transform such that $\epsilon_{i1} = \text{logit}(\Phi(Z_{i1}))$ and $\epsilon_{i2} = \text{logit}(\Phi(Z_{i2}))$, where logit(.) is the inverse logistic function, and $\Phi(.)$ is the cumulative distribution function of the standard normal distribution. The two binary outcomes were then generated such that

$$Y_{ij} = I(\epsilon_{ij} \leq (\beta_{0j} + \beta_{1j}X_{i1} + \beta_{2j}X_{i2})); j = 1, 2,$$

with $I(.)$ being the indicator function. Within this data generating process, ρ controls the level of dependence (after conditioning on the covariates) between the two outcomes; the outcomes are conditionally independent when $\rho = 0$. Hence, ρ was varied across simulation scenarios through values of {0, 0.25, 0.50, 0.75, 0.95}. Note that for $\rho > 0$ the correlation between $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ is less than ρ due to the nonlinear probability transform applied to Z_1 and Z_2 (eg, Reference[50]).

Finally, to test the predictive performance of all the analysis methods (outlined below), we generated an independent set of 10 000 observations, to serve as a validation set. This was generated using the exact same data-generating mechanisms as the IPD.

4.3 | Methods considered

Within each generated IPD set, we fit the following analysis models using MLE or MCMC, as appropriate: (i) two independent CPMs, one for each outcome (Section 2.2), (ii) stacked regression (Section 2.3), (iii) probabilistic classification chains (Section 3.1), (iv) multinomial logistic regression (Section 3.2), (v) multivariate logistic regression (Section 3.3), and (vi) multivariate Bayesian probit (Section 3.4). All analysis models included both $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ (ie, no variable selection).

4.4 | Target predictions

The main target outputs/predictions of interest were the predicted marginal probabilities $P(Y_{1i} = 1)$ and $P(Y_{2i} = 1)$, along with the predicted joint probability of both outcomes co-occurring: $P(Y_{1i} = 1, Y_{2i} = 1)$ where the conditionals on $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ have been omitted for brevity. As secondary outputs, we also consider $P(Y_{1i} = 1, Y_{2i} = 0)$, and $P(Y_{1i} = 0, Y_{2i} = 1)$, in order to evaluate each methods ability to predict all combinations of joint risk. Details of how each method estimates these joint and marginal probabilities were described in Sections 2 and 3.

4.5 | Performance measures

The data-generating mechanisms were repeated across 100 iterations for each simulation scenarios (ie, for all values of ρ). We used the validation sets, generated separately in each iteration, to assess the CPMs in terms of calibration (agreement

TABLE 1 Empirical results of the correlation between $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ for each value of ρ in the simulation, along with the observed joint outcome event rates for the main simulation

ρ	$\text{Corr}(Y_1, Y_2)$	$P(Y_1 = 1, Y_2 = 1)$	$P(Y_1 = 1, Y_2 = 0)$	$P(Y_1 = 0, Y_2 = 1)$
0	0.000	0.065	0.222	0.162
0.25	0.111	0.086	0.201	0.142
0.50	0.234	0.110	0.178	0.118
0.75	0.371	0.136	0.152	0.092
0.95	0.503	0.161	0.127	0.067

between the expected event rate and the observed event rate, across the full risk range), discrimination (ability of the model to separate cases from controls) and mean squared error (MSE) of the predicted risks compared with the "true" (data-generating) risks. The methods to estimate these differ across marginal or joint risk evaluation, as follows:

For marginal risk performance, we calculate the MSE as $n^{-1}\sum_{i=1}^{n}(\pi_i - \widehat{\pi}_i)^2$, where $n = 10\,000$ in the validation set, π_i is the data-generating (marginal) risk for observation i and $\widehat{\pi}_i$ is the corresponding estimated risk from each model. Calibration was quantified with the calibration-in-the-large (ideal value 0) and slope (ideal value 1), which was estimated by fitting a logistic regression model in the validation set for each observed marginal outcome and with the logit of the estimated marginal risk (from each model) as the only covariate; this covariate was used as an offset for estimation of calibration-in-the-large.[51] The discrimination of all models at predicting marginal risk was estimated with the area under the receiver operating characteristic curve.

For joint risk performance, we follow the methods documented previously.[5,52-55] The MSE was defined across the multivariate outcomes (ie, across each joint-outcome-combination: $\{Y_1 = 1, Y_2 = 1\}$, $\{Y_1 = 1, Y_{2i} = 0\}$, $\{Y_1 = 0, Y_2 = 1\}$ and $\{Y_1 = 0, Y_2 = 0\}$) as $n^{-1}\sum_{k=1}^{4}\sum_{i=1}^{n}(\pi_{i,k} - \widehat{\pi}_{i,k})^2$, where $n = 10\,000$ in the validation set, $\pi_{i,k}$ is the data-generating risk for observation i and joint-outcome-combination k, with $\widehat{\pi}_{i,k}$ being the corresponding estimated risk from each model.[5] The calibration of the joint outcomes was estimated using multinomial methods, as previously described in detail[52,53] (see supplementary methods in Appendix S1). Discrimination of the models at predicting joint risk was assessed using the polytomous discrimination index (PDI), where we report both overall PDI and joint-outcome-combination-specific PDIs (see Van Calster et al[54,55] for details). All performance measures were averaged across the 100 iterations and associated 95% confidence intervals (CIs) calculated.

4.6 | Software

The simulation was implemented in R version 4.0.2,[56] along with the following packages: tidyverse,[57] pROC,[58] rjags,[46] coda,[59] pbivnorm,[60] glmnet,[61] VGAM,[62-64] and nnet.[38] The code was written by the lead author and is available on GitHub at https://github.com/GlenMartin31/Multivariate-Binary-CPMs.

4.7 | Simulation results

We here present the simulation results for the case where the marginal outcome proportions were 29% and 23% for $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$, respectively. Quantitatively similar results were found for the sensitivity analysis where the marginal outcome proportions were lowered to 6% and 5% for $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$, respectively (simulation results available through the GitHub page).

Table 1 shows the empirical relationships between ρ and the correlation between the binary outcomes. The observed joint probability of both outcomes (averaged across all iterations) ranged from 6.5% for $\rho = 0$ to 16.1% for $\rho = 0.95$. A similar table for the sensitivity simulation (lower marginal outcome proportions of 6% and 5% for $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$, respectively) is given in Table S1.

When $\rho = 0$, all models returned estimates of the overall joint probabilities that were calibrated well with observed probabilities in the validation data (Figure 1**)**. However, as ρ increased, the calibration-in-the-large for $P(Y_{1i} = 1, Y_{2i} = 1)$ increased above 0 for univariate CPMs and stacked regression, indicating that these models (which ignore conditional dependency of the outcomes) underestimate the joint risk. In contrast, for all methods that account for dependence in the outcomes (ie, probabilistic classification chains, multinomial logistic regression, multivariate logistic regression and multivariate Bayesian probit regression), the calibration-in-the-large were consistently close to 0. However,

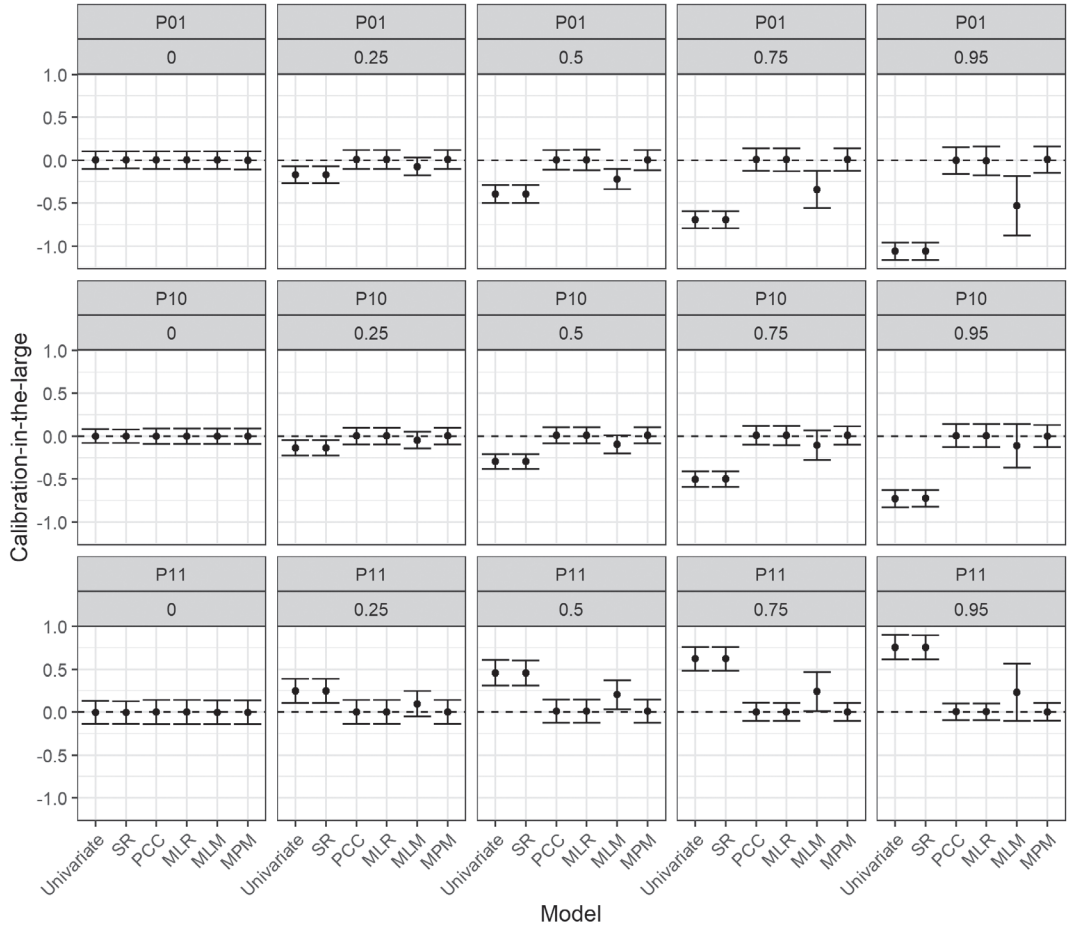

FIGURE 1 Calibration-in-the-large for each model across all simulation scenarios, upon validation. Bars represent the 95% confidence interval for the calibration-in-the-large across simulation iterations. Each column of plots corresponds to a value of ρ, while each row of plots is a joint outcome as follows: $P11$ denotes $P(Y_{1i} = 1, Y_{2i} = 1)$, $P10$ denotes $P(Y_{1i} = 1, Y_{2i} = 0)$, and $P01$ denotes $P(Y_{1i} = 0, Y_{2i} = 1)$. The dashed horizontal lines show the reference value for the calibration-in-the-large of 0. The models are as follows and as described in the methods section: Univariate, two independent clinical prediction models; SR, stacked regression; PCC, probabilistic classification chains; MLR, multinomial logistic regression; MLM, multivariate logistic model; MPM, multivariate Bayesian probit model

for higher values of ρ, the calibration-in-the-large deviated from 0 for multivariate logistic regression, especially when estimating $P(Y_{1i} = 0, Y_{2i} = 1)$; this is expected, as discussed in Section 3.3. In terms of marginal outcome risk, the calibration-in-the-large was sufficiently close to 0 for all methods across all values of ρ, indicating the overall expected marginal event rates matched the observed marginal event rates for all methods, as expected (Figure S1).

Similar findings were observed for the calibration slope for both the joint outcome risk (Figure 2) and marginal outcome risk (Figure S2). Specifically, predicted risks for each outcome were well calibrated for the two observed marginal probabilities for all models (Figure S2). For the joint outcome probabilities, only probabilistic classification chains, multinomial logistic regression and multivariate Bayesian probit regression (ie, the methods that account for dependence in the outcomes across the full range of ρ) had a calibration slope close to 1 for increasing ρ (Figure 2). Multivariate logistic regression was miscalibrated for predicting $P(Y_{1i} = 0, Y_{2i} = 1)$ for $\rho > 0.5$. The naïve approaches had calibration slope estimates below one when the outcomes where positively correlated, implying that the difference between the lowest joint risk and the highest joint risk was too extreme.

For discrimination, upon validation as ρ increased, the overall PDI for predicting joint outcomes was higher for probabilistic classification chains, multinomial logistic regression, and multivariate Bayesian probit regression CPMs, compared with univariate CPMs or stacked regression. Differences between the methods that account for dependence in the outcomes were modest, except for the multivariate logistic model (Figure 3, Figure S3). The discriminative ability of each method to predict the marginal outcomes were similar (Figure S4).

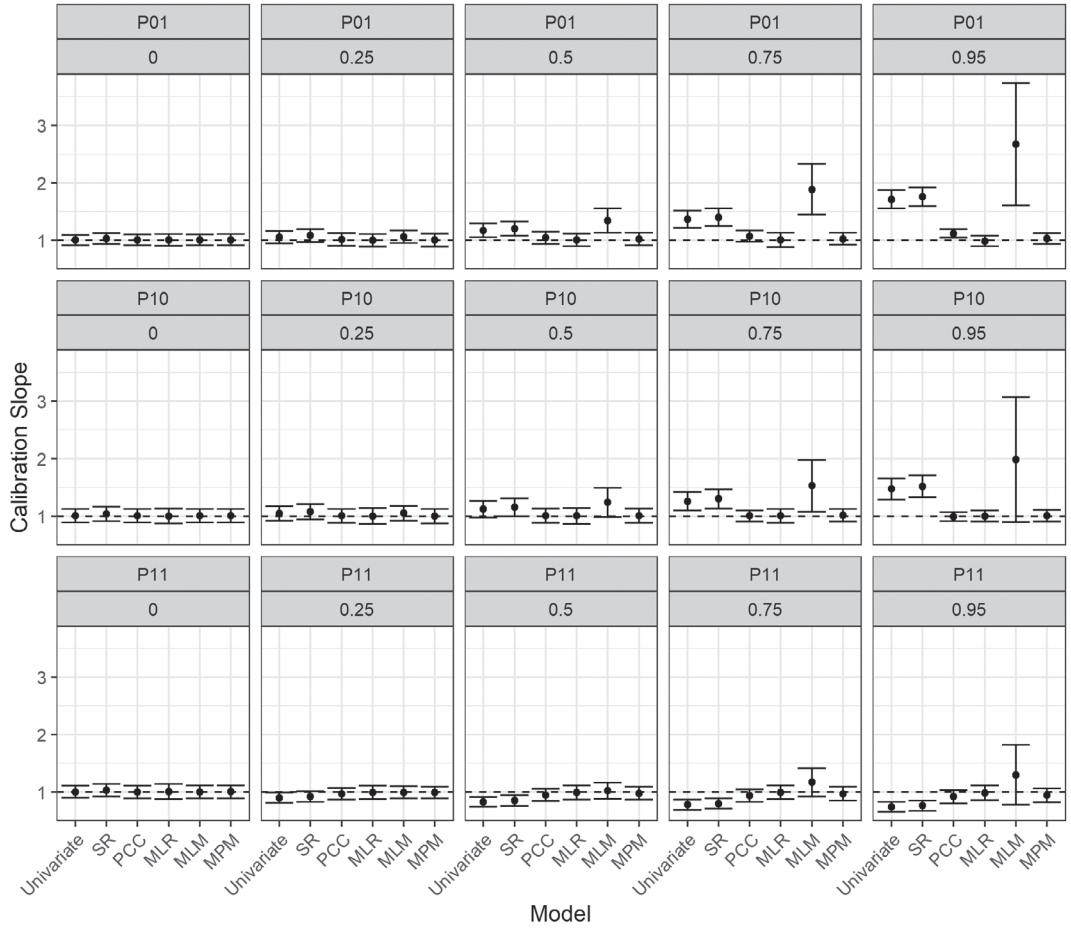

FIGURE 2 Calibration slope for each model across all simulation scenarios, upon validation. Bars represent the 95% confidence interval for the calibration slope across simulation iterations. Each column of plots corresponds to a value of ρ, while each row of plots is a joint outcome as follows: $P11$ denotes $P(Y_{1i} = 1, Y_{2i} = 1)$, $P10$ denotes $P(Y_{1i} = 1, Y_{2i} = 0)$, and $P01$ denotes $P(Y_{1i} = 0, Y_{2i} = 1)$. The dashed horizontal lines show the reference value for the calibration slope of 1. The models are as follows and as described in the methods section: Univariate, two independent clinical prediction models; SR, stacked regression; PCC, probabilistic classification chains; MLR, multinomial logistic regression; MLM, multivariate logistic model; MPM, multivariate Bayesian probit model

Finally, the MSE showed that all the methods consistently estimated the risks of the marginal outcomes (Figure S5), but the CPMs developed using probabilistic classification chains, multinomial logistic regression and multivariate Bayesian probit regression had much lower (ie, better) MSE for the joint outcome risks as ρ increased, compared with the methods that ignored the conditional dependency of the outcomes or multivariate logistic regression (Figure 4).

5 | EMPIRICAL STUDY

5.1 | Data source, study population, and outcomes

Data were obtained from the Medical Information Mart for Intensive Care III (MIMIC-III), which contains freely available and de-identified critical care data from the Beth Israel Deaconess Medical Center in Boston, Massachusetts, between 2001 and 2012.[65]

For the purposes of this empirical study of the methods, we considered the prediction of a binary indication of acute kidney injury (AKI) occurring within 48 hours after an ICU admission, and a binary indication of a total length of stay (LOS) on ICU of over 5 days. AKI is the most common cause of organ dysfunction in critically ill adults, and long LOS outcome captures overall ICU severity[66]; therefore, while this example is for illustrative purposes only, this could be of

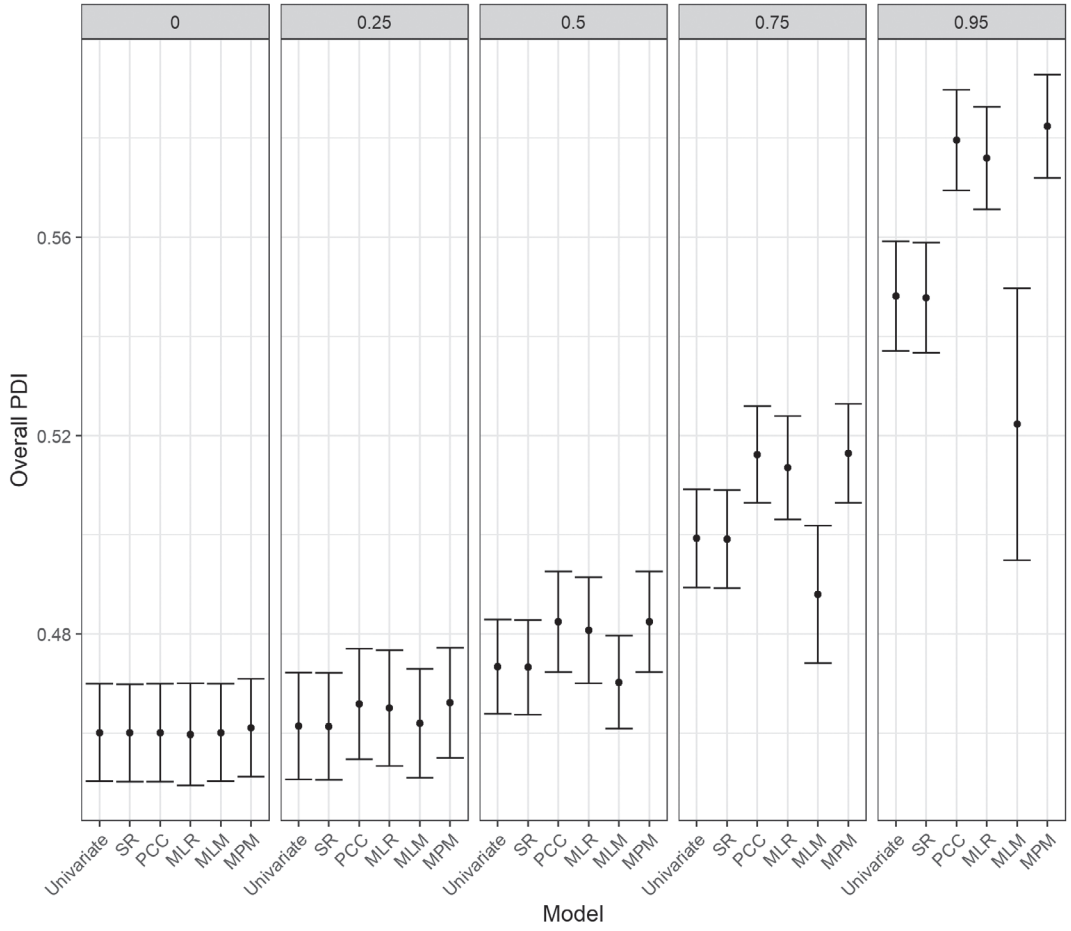

FIGURE 3 Overall polytomous discrimination index (PDI) for each model across all simulation scenarios, upon validation. Bars represent the 95% confidence interval for the PDI across simulation iterations. Each column of plots corresponds to a value of ρ. Note random performance on the PDI scale is 0.25 with four joint-outcome combinations (ie, $P11$, $P10$, $P01$, and $P00$). The models are as follows and as described in the methods section: Univariate, two independent clinical prediction models; SR, stacked regression; PCC, probabilistic classification chains; MLR, multinomial logistic regression; MLM, multivariate logistic model; MPM, multivariate Bayesian probit model

clinical interest in the context of ICU by aiding in early intervention to minimize risks of AKI and associated long LOS, both of which are important from a patient treatment and resource planning perspective.[66]

In this study, we defined an ICU admission to be any admission that lasted at least 24 hours, and we took the end of day 1 on ICU as our prediction time (the time point at which a prediction is made); hence, the AKI outcome was determined at the end of day 3. Ideally, LOS should be analyzed as a continuous or count outcome, but we considered LOS as binary here for illustrative purposes.

We extracted a cohort of patients over 18 years of age from the MIMIC-III database who were admitted to ICU for any cause for at least 24 hours. We extracted information on patients' age, gender, ethnicity, type of admission, and vital signs and lab results over the first 24 hours of each ICU admission (summarized as minimum, mean, and maximum values).

To define AKI, we extracted the maximum creatinine value for each patient collected between 24 and 72 hours after initial ICU admission. AKI was defined as present (coded as 1) if the maximum creatinine within 48 hours after the prediction time was either: (i) more than 1.5 times the minimum day 1 creatinine value, or (ii) over 0.3 mg/dL greater than the minimum day 1 creatinine value; AKI was coded 0 (absent) otherwise. This definition follows published clinical guidance.[67]

We excluded any ICU admission with indication of reduced kidney function within the first 24 hours, by excluding those with an estimated glomerular filtration rate GFR (eGFR) less than 60 mL/min/1.73 m^2 at baseline. The eGFR was calculated by the MDRD study equation,[68] using each patient's minimum creatinine value within the first 24 hours.

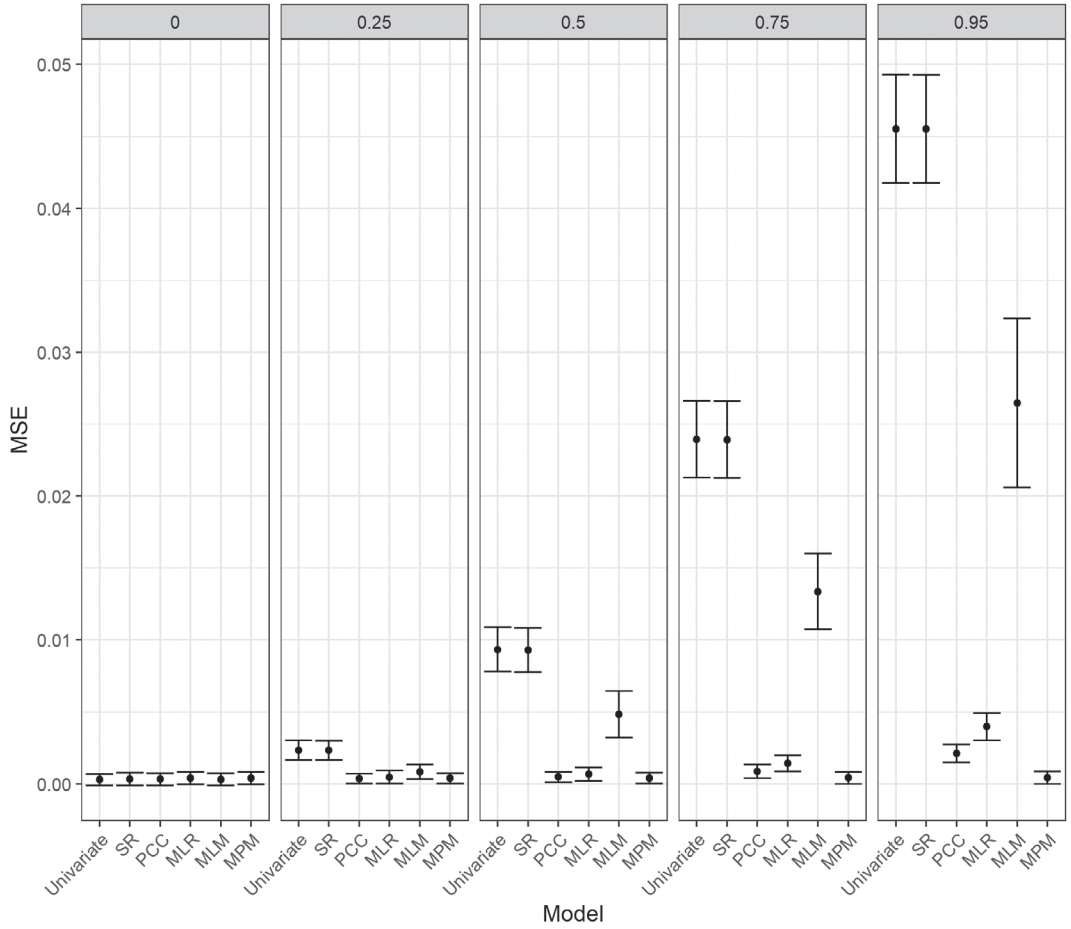

FIGURE 4 Multivariate mean squared error (MSE) of joint risk estimation, for each model across all simulation scenarios (lower is better), upon validation. Bars represent the 95% confidence interval for the MSE across simulation iterations. Each column of plots corresponds to a value of ρ. The models are as follows and as described in the methods section: Univariate, two independent clinical prediction models; SR, stacked regression; PCC, probabilistic classification chains; MLR, multinomial logistic regression; MLM, multivariate logistic model; MPM, multivariate Bayesian probit model

Patients with missing outcomes or who died within the hospitalization of their ICU stay were also excluded. We only included a patient's first ICU admission for a given hospitalization.

We developed CPMs on the extracted cohort using the methods outlined in Sections 2 and 3 with the aim of estimating the marginal and joint probabilities of AKI and total ICU LOS $\geq$ 5 days. All the models included identical predictors (Table S2), and therefore differed only in their estimation processes. To emphasize, the purpose was not to derive a new CPM for use in clinical practice, but rather to illustrate and apply the proposed analytical methods using real-world clinical data. We evaluated each of the models in terms of their respective calibration and discrimination metrics, using the same techniques as described in Section 4.5. The large sample size means that overfitting was not a concern[28]; nonetheless, we report predictive performance in a random hold-out sample (30%), simply to have some unused data to check the models in, rather than relying on development data alone (noting that split-sample method is not preferred in practice[1-3]).

Missing data in any predictor variables was imputed using multiple imputation, where we generated 20 imputed datasets.[69] The imputation models included all of the covariates, plus the two outcomes.[70] Within each imputed dataset, the CPMs were developed using each analytical method, which were then applied to the hold-out test samples to estimate the calibration and discrimination of each model. Performance metrics upon validation were then pooled across the imputations using Rubin's rules.[69]

All data extraction was performed using an SQL script written by the lead author (which is available on Github: https://github.com/GlenMartin31/Multivariate-Binary-CPMs). Analysis was performed using R version 4.0.2,[56] along with the packages stated in Section 4.6.

5.2 | Empirical study results

A total of 24 459 ICU admissions were included in the analysis; Table 2 presents an overview of a baseline summary of the whole cohort extracted from the MIMIC-III database. The correlation between the outcomes was 0.08, with an observed joint probability of both outcomes being 4.46%. The marginal probability of long LOS was 20.3% and for AKI was 16.1%.

Table 3 shows the calibration and discrimination (in the hold-out sample) for each method in terms of estimating the joint outcome risks. As with the simulation study, methods that account for dependence in the outcomes (ie, probabilistic classification chains, multinomial logistic regression, multivariate logistic regression and multivariate Bayesian probit regression) were well calibrated for all outcomes, with calibration-in-the-large and calibration slope close to 0 and 1, respectively. The models that do not account for outcome dependency significantly under-predicted the joint outcome risk, with a calibration-in-the-large over 0, although the 95% CI for the calibration slope spanned 1. A flexible nominal calibration plot[52,53] for the joint outcome performance is given in Figure 5, which shows the overall underestimation of the joint outcome proportion (ie, $P(\text{ICU} > 5_i = 1, \text{AKI}_i = 1)$) for the univariate and stacked regression models; in contrast the calibration plot for multinomial logistic regression and multivariate Bayesian probit regression show good calibration, which is supported by the calibration-in-the-large and calibration slope estimates (Table 3). All models had similar discrimination, both in terms of joint-outcome-combination specific PDI (Table 3) and overall PDI (Figure S6). These results align with those from the simulation study; here, we have an observed correlation in the outcomes of 0.08, which approximately corresponds to the simulation scenario where $\rho = 0.25$ (see Table 1).

6 | DISCUSSION

This study presents four methods for developing CPMs that respect the dependence between multiple clinical outcomes. As expected, only the methods that condition on each outcome (probabilistic classification chains and multinomial logistic regression) or model the correlation explicitly (multivariate logistic regression and multivariate Bayesian probit regression) provide reliable estimates of joint risks. All methods had similar predictive performance in terms of predicting the marginal risks of each outcome. Our results suggest that probabilistic classification chains, multinomial logistic regression or the multivariate probit model might be the most appropriate choice for developing multivariate CPMs for multiple binary outcomes.

There has been little previous research published on developing CPMs that aim to predict multiple outcomes simultaneously. Most CPMs have been developed for individual or composite outcomes. However, many medical contexts demand a multivariate approach to prediction. An example of such a context is multi-morbidity, which is becoming an increasing priority for health services around the world. Traditional approaches to do this in multi-morbidity (eg, Charlson Comorbidity Index[71]) involve rudimentary metrics, assigning crude weights to different conditions that cannot predict the co-occurrence of outcomes. There are many other clinical examples where a multivariate approach to prediction would be warranted. This study shows that accurate prediction of joint outcome risks is only achieved by developing CPMs that account for dependence in the outcomes. Models that do not account for outcome dependence underestimate the joint outcome risks.

While the findings of this study are intuitive from a statistical perspective, CPMs are usually not developed in a multivariate manner.[4,14,17,18] As such, the findings from this study have implications for multi-outcome risk prediction, which is becoming an increasing priority. Advantageously, all the methods proposed can be implemented using standard statistical software (although multivariate logistic regression and multivariate Bayesian probit regression do require user coding), meaning they could be readily applied to develop real-world CPMs for multi-outcome prediction. Of note is that as the number of outcomes under consideration increases, the computational demand of fitting the models also increases. Indeed, as the number of outcomes increases, the size of the parameter space for probabilistic classification chains, multinomial logistic regression and the multivariate probit model increases rapidly, albeit to a less extent for multinomial logistic regression (Table S3).

Nonetheless, we note that all CPMs are developed with a particular prediction task in mind, and as such not all CPMs will aim—or indeed need—to accurately estimate joint risks of multiple outcomes. This study shows that all approaches accurately predicted the risks of each outcome individually, meaning the models in Section 3 are still useful at predicting marginal risk. It is important to emphasize that the methods that utilize data from multiple outcomes can also leverage the information contained across outcomes, with the associated advantages that this brings. For example, such advantages

TABLE 2 Baseline summary of the patient demographics, characteristics and lab/vital results over the first 24 hours of an intensive care unit (ICU) admission, for the whole cohort

Characteristic	Summary	Missing data, *n* (% of cohort)
N	24 459	
Demographics		
Age, mean (min, max)	60.92 (18.02, 99.28)	0 (0%)
Age group, n (%)		0 (0%)
<30	1406 (5.75%)	
30-40	1534 (6.27%)	
40-50	3091 (12.6%)	
50-60	4958 (20.3%)	
60-70	5522 (22.6%)	
70-80	4616 (18.9%)	
>80	3332 (13.6%)	
Male, n (%)	14 438 (59.0%)	0 (0%)
Admission type, n (%)		0 (0%)
Elective	4627 (18.9%)	
Urgent	609 (2.49%)	
Emergency	19 223 (78.6%)	
Ethnicity, *n* (%)		2665 (10.9%)
White	17 637 (72.1%)	
Asian	593 (2.42%)	
Black	1998 (8.17%)	
Hispanic	867 (3.54%)	
Other	709 (2.90%)	
Lab tests – Summary over first 24 hours on ICU		
Mean bicarbonate, mean (min, max)	24.4 (8.00, 51.5)	123 (0.50%)
Mean creatinine, mean (min, max)	0.82 (0.10, 4.73)	0 (0%)
Mean chloride, mean (min, max)	105.2 (64.5, 142.0)	85 (0.35%)
Mean hemoglobin, mean (min, max)	11.1 (3.33, 19.9)	59 (0.24%)
Mean platelet count, mean (min, max)	224.6 (7.50, 1646.2)	84 (0.34%)
Mean potassium, mean (min, max)	4.09 (2.30, 8.70)	2 (0.01%)
Mean partial thromboplastin time, mean (min, max)	35.3 (14.4, 150.0)	2596 (10.6%)
Mean international normalized ratio, mean (min, max)	1.37 (0.50, 18.2)	2537 (10.4%)
Mean prothrombin time, mean (min, max)	14.9 (8.00, 131.1	2542 (10.4%)
Mean white blood cell count, mean (min, max)	11.8 (0.10, 247.9)	137 (0.56%)
Vital signs—summary over first 24 hours on ICU		
Mean heart rate, mean (min, max)	86.2 (31.2, 155.0)	203 (0.83%)
Mean systolic blood pressure, mean (min, max)	119.0 (74.1, 203.0)	220 (0.90%)
Mean diastolic blood pressure, mean (min, max)	61.9 (27.4, 127.0)	220 (0.90%)
Mean respiration rate, mean (min, max)	18.5 (8.00, 41.8)	225 (0.92%)
Mean temperature (Celsius), mean (min, max)	36.9 (32.6, 39.8)	703 (2.87%)

(Continues)

TABLE 2 (Continued)

Characteristic	Summary	Missing data, *n* (% of cohort)
Mean oxygen saturation, mean (min, max)	97.5 (73.5, 100.0)	211 (0.86%)
Mean glucose, mean (min, max)	135.8 (52.0, 661.8)	299 (1.22%)
Outcomes		
Total ICU Length of Stay ≥ 5 days, n (%)	4957 (20.3%)	0 (0%)
Acute Kidney Injury by day 3 on ICU, n (%)	3930 (16.1%)	0 (0%)

TABLE 3 Internal validation (hold-out sample) calibration-in-the-large, calibration slope and polytomous discrimination index (PDI) performance results for each model in the MIMIC-III dataset. For the outcome column, $P11$ denotes $P(\text{ICU} > 5_i = 1, \text{AKI}_i = 1)$, $P10$ denotes $P(\text{ICU} > 5_i = 1, \text{AKI}_i = 0)$, and $P01$ denotes $P(\text{ICU} > 5_i = 0, \text{AKI}_i = 1)$

Model	Outcome	Calibration-in-the-large (95% CI)	Calibration Slope (95% CI)	Outcome-specific PDI (min, max)*
Univariate	P11	0.23(0.11, 0.35)	0.94(0.8, 1.08)	0.42(0.42, 0.43)
Univariate	P10	−0.15(−0.21, −0.08)	0.98(0.84, 1.13)	0.38(0.38, 0.39)
Univariate	P01	−0.09(−0.17, −0.01)	0.98(0.88, 1.09)	0.42(0.42, 0.42)
SR	P11	0.23(0.11, 0.35)	0.98(0.83, 1.13)	0.42(0.42, 0.43)
SR	P10	−0.15(−0.21, −0.08)	1.03(0.87, 1.18)	0.38(0.38, 0.39)
SR	P01	−0.09(−0.17, −0.01)	1.02(0.91, 1.13)	0.42(0.42, 0.42)
PCC	P11	−0.07(−0.19, 0.05)	1(0.85, 1.16)	0.42(0.42, 0.43)
PCC	P10	−0.05(−0.12, 0.01)	0.98(0.83, 1.12)	0.39(0.39, 0.4)
PCC	P01	0.03(−0.04, 0.11)	0.96(0.86, 1.07)	0.43(0.42, 0.43)
MLR	P11	−0.06(−0.19, 0.06)	0.91(0.77, 1.04)	0.44(0.44, 0.44)
MLR	P10	−0.05(−0.12, 0.02)	1(0.86, 1.13)	0.42(0.41, 0.42)
MLR	P01	0.03(−0.05, 0.1)	1(0.9, 1.11)	0.45(0.45, 0.45)
MLM	P11	−0.04(−0.16, 0.08)	1.08(0.91, 1.24)	0.42(0.42, 0.43)
MLM	P10	−0.07(−0.13, 0)	0.96(0.82, 1.11)	0.39(0.38, 0.39)
MLM	P01	0.01(−0.07, 0.09)	0.95(0.85, 1.06)	0.42(0.42, 0.43)
MPM	P11	−0.07(−0.2, 0.07)	1.06(0.89, 1.22)	0.42(0.42, 0.43)
MPM	P10	−0.06(−0.13, 0.02)	1(0.85, 1.15)	0.39(0.39, 0.39)
MPM	P01	0.03(−0.05, 0.11)	0.98(0.87, 1.09)	0.43(0.42, 0.43)

Abbreviations: MLM, multivariate logistic model; MPM, multivariate Bayesian probit model; MLR, multinomial logistic regression; PCC, probabilistic classification chains; SR, stacked regression; Univariate, two independent clinical prediction models.

[a]A depiction of the overall PDI is given in Appendix S1; min/max values for the PDI are taken across the 20 multiple imputed datasets.

have been widely shown in the multivariate IPD meta-analysis literature,[72] the literature on joint modeling,[73,74] and also in cross-sectional data of correlated binary outcomes.[14,17,18]

Here we focus on predicting binary outcomes, but continuous, ordinal and time-to-event outcomes are also commonly required from CPMs.[3] While most of the methods described in this paper generalize naturally to predicting continuous outcomes, further consideration will be required for ordinal and time-to-event data. Explicitly, for time-to-event, one should account for competing risks (eg, death), especially in cases where outcomes might occur over several years. Multi-state survival models present a way of doing this,[19] and have been used to develop CPMs to predict risk of moving between disease/condition/pathway states (eg, [20]). However, the use of multistate models to develop CPMs for multimorbidity

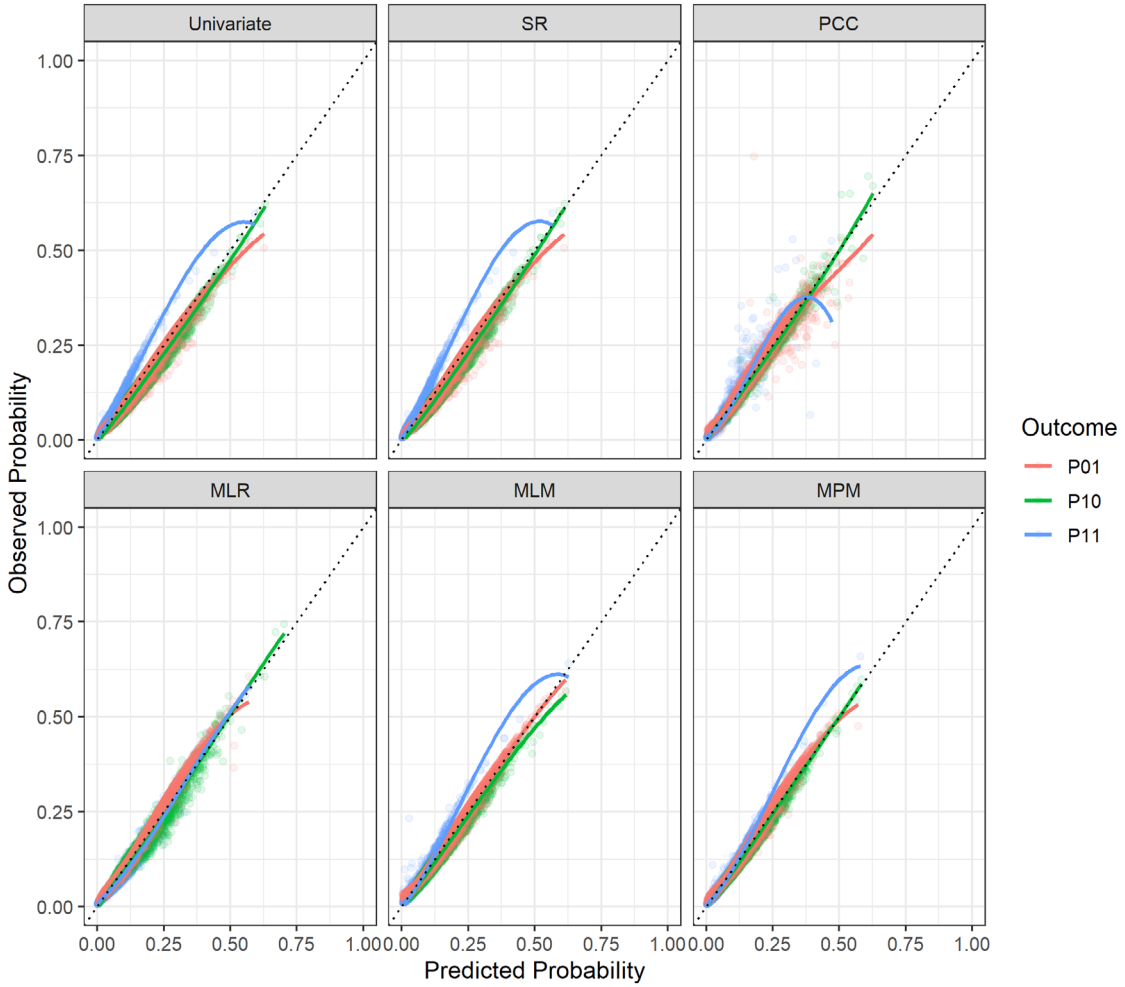

FIGURE 5 Internal validation (hold-out sample) flexible (nonparametric) nominal calibration plot fitted with vector splines on the linear predictors for each model in the MIMIC-III dataset. For illustration, this is taken from the 20th imputed dataset (similar results in the other imputed datasets). The scatter of points in each plot arises due to the multivariate nature of the plots; the lines are smoothing splines to show average shape of the calibration plot. $P11$ denotes $P(\text{ICU} > 5_i = 1, \text{AKI}_i = 1)$, $P10$ denotes $P(\text{ICU} > 5_i = 1, \text{AKI}_i = 0)$, and $P01$ denotes $P(\text{ICU} > 5_i = 0, \text{AKI}_i = 1)$. The models are as follows and as described in the methods section: Univariate, two independent clinical prediction models; SR, stacked regression; PCC, probabilistic classification chains; MLR, multinomial logistic regression; MLM, multivariate logistic model; MPM, multivariate Bayesian probit model [Colour figure can be viewed at wileyonlinelibrary.com]

prediction is not commonplace and would require methodological developments due to combinatorial complexities arising by the number of states (ie, outcome combinations). As such, further research is required to extend the approaches in this paper to consider time-to-event CPMs through a multistate and competing risk framework; some of the methods outlined here might provide a foundation to doing this.

Similarly, a patient's care pathway comprises a mixture of outcome "types"; thus, it would be advantageous if the methods to develop CPMs for multiple outcomes could handle this heterogeneity. For example, one might need to predict continuous (eg, blood test), binary (eg, procedural complication), and time-to-event (eg, time-to-readmission) outcomes simultaneously. While stacked regression provides a natural way of doing this by operating on the linear predictor scale (Section 2.3), this method relies on conditional independence so cannot estimate joint outcome risks. In contrast, the multivariate probit model has been used to model continuous and binary outcomes simultaneously by correlating the error terms of a linear model with the latent error term for the binary outcome.[18,75] The use of copula methods within the multivariate probit model might be one way to generalize this approach to work for any pair of outcome types.[76] Further research is warranted to explore this, and to consider the extension of the other methods to handle different outcomes.

6.1 | Limitations

Several limitations should be considered for this study. First, we have only considered binary outcomes where the occurrence of one outcome does not prevent the occurrence of the others. In practice, competing risks will need to be accounted for and the biases incurred by failing to consider this were not explored here. Second, we only generated two predictors in the simulation study; we acknowledge that most CPMs include more than two covariates, but one could regard the two simulated predictors as summaries of several variables. Similarly, neither the simulation nor the critical care example considered variable selection. Variable selection might be especially important where different predictors are associated with different outcomes, or where the direction of the association of a given predictor differs across outcomes. Considering variable selection and heterogeneity in associations should be considered for future work. Third, we only considered case studies and simulations with large sample sizes; therefore, further research is needed to explore the concepts of this paper in settings where overfitting might be a concern (eg, penalization).[28] Fourth, the simulations and real-world example only considered two outcomes, and computational cost of all methods might increase if aiming to predict more outcomes simultaneously; future work should explore this directly. Finally, we only considered methods embedded within a regression framework, and we acknowledge that we have not considered machine learning, classification-based approaches, such as multilabel neural networks.[26] Nonetheless, interpretable, machine learning multivariate CPMs are a grand challenge.

6.2 | Conclusion

This paper reports four approaches that can advance CPMs beyond the current disconnected prediction of single conditions to combinatorial approaches that reflect the real-world challenge of multiple-outcome health care. Any CPM that aims to predict joint risk of multiple outcomes should only be based on methods that explicitly model the correlation structure. In such a situation, our results suggest that probabilistic classification chains, multinomial logistic regression or the multivariate probit model might be the most appropriate choice. Approaches that model outcome dependency more accurately reflect real-world health care and benefit from the well-known advantages to inference that analyzing multiple outcomes simultaneously offers.

ACKNOWLEDGEMENTS

This publication was funded by an MRC-NIHR Methodology Research Programme grant (grant number: MR/T025085/1). Kym Snell is funded by the National Institute for Health Research School for Primary Care Research (NIHR SPCR). This publication presents independent research partially funded by the National Institute for Health Research (NIHR). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

CONFLICT OF INTEREST

The authors declare no potential conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in the lead author's Github page, at https://github.com/GlenMartin31/Multivariate-Binary-CPMs

ORCID

Glen P. Martin https://orcid.org/0000-0002-3410-9472
Matthew Sperrin https://orcid.org/0000-0002-5351-9960
Richard D. Riley https://orcid.org/0000-0001-8699-0735

REFERENCES

1. Steyerberg EW. *Clinical Prediction Models*. New York, NY: Springer; 2009.
2. Riley RD, Windt D, Croft P, Moons K. *Prognosis Research in Healthcare: Concepts, Methods, and Impact*. Oxford: Oxford University Press; 2019.
3. Harrell FE. *Regression Modeling Strategies*. 2nd ed. New York, NY: Springer; 2015.

4. Biesheuvel CJ, Vergouwe Y, Steyerberg EW, Grobbee DE, Moons KGM. Polytomous logistic regression analysis could be applied more often in diagnostic research. *J Clin Epidemiol*. 2008;61(2):125-134. https://doi.org/10.1016/j.jclinepi.2007.03.002.
5. Jong VMT, Eijkemans MJC, Calster B, et al. Sample size considerations and predictive performance of multinomial logistic prediction models. *Stat Med*. 2019;38(9):1601-1619. https://doi.org/10.1002/sim.8063.
6. Chong C-F, Li Y-C, Wang T-L, Chang H. Stratification of adverse outcomes by preoperative risk factors in coronary artery bypass graft patients: an artificial neural network prediction model. *AMIA. Annu Symp proceedings AMIA Symp*. 2003;2003:160-164.
7. Prins C, de Villiers Jonker I, Botes L, Smit FE. Cardiac surgery risk-stratification models. *Cardiovasc J Afr.* 2012;23(3):160-164. https://doi.org/10.5830/CVJA-2011-047.
8. Schuit E, Kwee A, Westerhuis M, et al. A clinical prediction model to assess the risk of operative delivery. *BJOG An Int J Obstet Gynaecol.* 2012;119(8):915-923. https://doi.org/10.1111/j.1471-0528.2012.03334.x.
9. Damen JAAG, Hooft L, Schuit E, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. *BMJ*. 2016;353:i2416. https://doi.org/10.1136/bmj.i2416.
10. Vickers AJ. Prediction models in cancer care. *CA Cancer J Clin*. 2011;61(5):315-326. https://doi.org/10.3322/caac.20118.
11. Echouffo-Tcheugui JB, Kengne AP. Risk models to predict chronic kidney disease and its progression: a systematic review. *PLoS Med*. 2012;9(11):e1001344. https://doi.org/10.1371/journal.pmed.1001344.
12. Bayliss EA, Bayliss MS, Ware JE, Steiner JF. Predicting declines in physical function in persons with multiple chronic medical conditions: what we can learn from the medical problem list. *Health Qual Life Outcomes*. 2004;2:47. https://doi.org/10.1186/1477-7525-2-47.
13. Fortin M, Lapointe L, Hudon C, Vanasse A, Ntetu AL, Maltais D. Multimorbidity and quality of life in primary care: a systematic review. *Health Qual Life Outcomes*. 2004;2:51. https://doi.org/10.1186/1477-7525-2-51.
14. Carey V, Zeger SL, Diggle P. Modelling multivariate binary data with alternating logistic regressions. *Biometrika.* 1993;80(3):517-526. https://doi.org/10.1093/biomet/80.3.517.
15. Breiman L, Friedman J. Predicting multivariate responses in multiple linear regression. *J R Stat Soc B*. 1997;59(1):3-54.
16. Kip KE, Hollabaugh K, Marroquin OC, Williams DO. The problem with composite end points in cardiovascular studies. The story of major adverse cardiac events and percutaneous coronary intervention. *J Am Coll Cardiol*. 2008;51(7):701-707. https://doi.org/10.1016/j.jacc.2007.10.034.
17. Chib S, Greenberg E. Analysis of multivariate probit models. *Biometrika*. 1998;85(2):347-361. https://doi.org/10.1093/biomet/85.2.347.
18. Teixeira-Pinto A, S-LT N. Correlated bivariate continuous and binary outcomes: issues and applications. *Stat Med*. 2009;28(13):1753-1773. https://doi.org/10.1002/sim.3588.
19. Putter H, Fiocco M, Gekus RB. Tutorial in biostatistics: competing risk and multi-state models. *Stat Med*. 2007;26(11):2389-2430. https://doi.org/10.1002/sim.2712.
20. Upshaw JN, Konstam MA, Van Klaveren D, Noubary F, Huggins GS, Kent DM. Multistate model to predict heart failure hospitalizations and all-cause mortality in outpatients with heart failure with reduced ejection fraction. *Circ Hear Fail.* 2016;9(8):e003146. https://doi.org/10.1161/CIRCHEARTFAILURE.116.003146.
21. Freisling H, Viallon V, Lennon H, et al. Lifestyle factors and risk of multimorbidity of cancer and cardiometabolic diseases: a multinational cohort study. *BMC Med*. 2020;18(1):5. https://doi.org/10.1186/s12916-019-1474-7.
22. Zhang ML, Zhou ZH. A review on multi-label learning algorithms. *IEEE Trans Knowl Data Eng.* 2014;26(8):1819-1837. https://doi.org/10.1109/TKDE.2013.39.
23. Madjarov G, Kocev D, Gjorgjevikj D, Džeroski S. An extensive experimental comparison of methods for multi-label learning. *Pattern Recognition*. Vol 45. 2012;9:3084–3104. https://doi.org/10.1016/j.patcog.2012.03.004.
24. Zhang ML, Li YK, Liu XY, Geng X. Binary relevance for multi-label learning: an overview. *Front Comput Sci.* 2018;12(2):191-202. https://doi.org/10.1007/s11704-017-7031-7.
25. Read J, Pfahringer B, Holmes G, Frank E, Classifier Chains for Multi-label Classification. Buntine W, Grobelnik M, Mladenić D, Shawe-Taylor J. *Machine Learning and Knowledge Discovery in Databases*. In: Springer, Berlin, Germany; 2009;5782:254–269.
26. Zhang ML, Zhou ZH. Multilabel neural networks with applications to functional genomics and text categorization. *IEEE Trans Knowl Data Eng*. 2006;18(10):1338-1351. https://doi.org/10.1109/TKDE.2006.162.
27. Dudbridge F. Criteria for evaluating risk prediction of multiple outcomes. *Stat Methods Med Res*. 2020;29:3492-3510. https://doi.org/10.1177/0962280220929039.
28. Riley RD, Snell KI, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. *Stat Med*. 2019;38(7):1276-1296. https://doi.org/10.1002/sim.7992.
29. Pavlou M, Ambler G, Seaman SR, et al. How to develop a more accurate risk prediction model when there are few events. *BMJ*. 2015;351:h3868. https://doi.org/10.1136/BMJ.H3868.
30. Park T, Casella G. The Bayesian lasso. *J Am Stat Assoc*. 2008;103(482):681-686. https://doi.org/10.1198/016214508000000337.
31. Debray TPAA, Koffijberg H, Vergouwe Y, Moons KGMM, Steyerberg EW, Steyerberg EW. Aggregating published prediction models with individual participant data: a comparison of different approaches. *Stat Med*. 2012;31(23):2697-2712. https://doi.org/10.1002/sim.5412.
32. Martin GP, Mamas MA, Peek N, Buchan I, Sperrin M. A multiple-model generalisation of updating clinical prediction models. *Stat Med*. 2018;37(8):1343-1358. https://doi.org/10.1002/sim.7586.
33. Xing L, Lesperance M, Zhang X. Simultaneous prediction of multiple outcomes using revised stacking algorithms. *Hancock J, Ed. Bioinformatics.* 2019;36:65-72. https://doi.org/10.1093/bioinformatics/btz531.
34. Debray TPA, Koffijberg H, Nieboer D, Vergouwe Y, Steyerberg EW, Moons KGM. Meta-analysis and aggregation of multiple published prediction models. *Stat Med*. 2014;33(14):2341-2362. https://doi.org/10.1002/sim.6080.

35. Martin GP, Mamas MA, Peek N, Buchan I, Sperrin M. Clinical prediction in defined populations: a simulation study investigating when and how to aggregate existing models. *BMC Med Res Methodol*. 2017;17(1):1. https://doi.org/10.1186/s12874-016-0277-1.
36. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Ser B*. 1996;58(1):267-288.
37. Dembczynski K, Cheng W, Hullermeier E. Bayes optimal multilabel classification via probabilistic classifier chains. *Proceedings of the 27th International Conference on International Conference on Machine Learning*. Haifa, Israel: Omnipress; 2010:279-286.
38. Venables WN, Ripley BD. *Modern Applied Statistics with S*. 4th ed. New York, NY: Springer; 2002.
39. Gauvreau K, Pagano M. The analysis of correlated binary outcomes using multivariate logistic regression. *Biometrical J*. 1997;39(3):309-325. https://doi.org/10.1002/bimj.4710390306.
40. Gumbel EJ. Bivariate logistic distributions. *J Am Stat Assoc*. 1961;56(294):335-349. https://doi.org/10.1080/01621459.1961.10482117.
41. Nikoloulopoulos AK. *Copula-Based Models for Multivariate Discrete Response Data*. Berlin, Germany: Springer; 2013:231-249.
42. Genest C, Nikoloulopoulos AK, Rivest LP, Fortin M. Predicting dependent binary outcomes through logistic regressions and meta-elliptical copulas. *Brazilian J Probab Stat*. 2013;27(3):265-284. https://doi.org/10.1214/11-BJPS165.
43. Stefanescu C, Turnbull BW. On the multivariate Probit model for exchangeable binary data with covariates. *Biometrical J*. 2005;47(2):206-218. https://doi.org/10.1002/bimj.200410101.
44. Edwards YD, Allenby GM. Multivariate analysis of multiple response data. *J Mark Res*. 2003;40(3):321-334.
45. Burke DL, Bujkiewicz S, Riley RD. Bayesian bivariate meta-analysis of correlated effects: impact of the prior distributions on the between-study correlation, borrowing of strength, and joint inferences. *Stat Methods Med Res*. 2018;27(2):428-450. https://doi.org/10.1177/0962280216631361.
46. Plummer M. rjags: Bayesian Graphical Models using MCMC. 2018. https://cran.r-project.org/package=rjags.
47. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Stat Med*. 2019;38(11):2074-2102. https://doi.org/10.1002/sim.8086.
48. Emrich LJ, Piedmonte MR. A method for generating high-dimensional multivariaten binary variates. *Am Stat*. 1991;45(4):302-304. https://doi.org/10.1080/00031305.1991.10475828.
49. Touloumis A. Simulating correlated binary and multinomial responses under marginal model specification: the SimCorMultRes package. *R J*. 2016;8(2):79-91. journal.r-project.org/archive/2016/RJ-2016-034/RJ-2016-034.pdf.
50. Ted Li S, Hammond JL. Generation of pseudorandom numbers with specified univariate distributions and correlation coefficients. *IEEE Trans Syst Man Cybern*. 1975;SMC-5(5):557-561. https://doi.org/10.1109/TSMC.1975.5408380.
51. Cox D. Two further applications of a model for binary regression. *Biometrika*. 1958;45(3):562-565.
52. Van Hoorde K, Vergouwe Y, Timmerman D, Van Huffel S, Steyerberg EW, Van Calster B. Assessing calibration of multinomial risk prediction models. *Stat Med*. 2014;33(15):2585-2596. https://doi.org/10.1002/sim.6114.
53. Van Hoorde K, Van Huffel S, Timmerman D, Bourne T, Van Calster B. A spline-based tool to assess and visualize the calibration of multiclass risk predictions. *J Biomed Inform*. 2015;54:283-293. https://doi.org/10.1016/j.jbi.2014.12.016.
54. Van Calster B, Vergouwe Y, Looman CWN, Van Belle V, Timmerman D, Steyerberg EW. Assessing the discriminative ability of risk models for more than two outcome categories. *Eur J Epidemiol*. 2012;27(10):761-770. https://doi.org/10.1007/s10654-012-9733-3.
55. Van Calster B, Van Belle V, Vergouwe Y, Timmerman D, Van Huffel S, Steyerberg EW. Extending the c-statistic to nominal polytomous outcomes: the polytomous discrimination index. *Stat Med*. 2012;31(23):2610-2626. https://doi.org/10.1002/sim.5321.
56. R Core Team. *R: A Language and Environment for Statistical Computing*. Team RDC, ed. Vienna, Austria: R Foundation for Statistical Computing; 2020. https://www.R-project.org/.
57. Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen T, Miller E, Bache S, Müller K, Ooms J, Robinson D, Seidel D, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H. Welcome to the Tidyverse. *Journal of Open Source Software*. 2019;4(43):1686. http://dx.doi.org/10.21105/joss.01686.
58. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinform*. 2011;12:77.
59. Plummer M, Best N, Cowles K, Vines K. CODA: convergence diagnosis and output analysis for MCMC. *R News*. 2006;6(1):7-11.
60. CRAN. Fortran code by Genz A, R code by Kenkel B. pbivnorm: Vectorized Bivariate Normal CDF. 2015. https://cran.r-project.org/package=pbivnorm.
61. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*. 2010;33(1):1-22. https://doi.org/10.18637/jss.v033.i01.
62. Yee TW. *Vector Generalized Linear and Additive Models: With an Implementation in R*. New York, NY: Springer; 2015.
63. Yee TW, Wild CJ. Vector generalized additive models. *J R Stat Soc Ser B*. 1996;58(3):481-493.
64. Yee TW. The VGAM package for categorical data analysis. *J Stat Softw*. 2010;32(10):1-34. http://www.jstatsoft.org/v32/i10/.
65. Johnson AEW, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. *Sci Data*. 2016;3(1):160035. https://doi.org/10.1038/sdata.2016.35.
66. Gentimis T, Alnaser AJ, Durante A, Cook K, Steele R. Predicting hospital length of stay using neural networks on MIMIC III data. Paper presented at: Proceedings of 2017 IEEE 15th International Conference on Dependable, Autonomic and Secure Computing, 2017 IEEE 15th International Conference on Pervasive Intelligence and Computing, 2017 IEEE 3rd International Conference on Big Data Intelligence and Computing. Vol 2018; IEEE; 2018:1194-1201. doi:https://doi.org/10.1109/DASC-PICom-DataCom-CyberSciTec.2017.191
67. National Institute for Health and Care Excellence. *Acute Kidney Injury: Prevention, detection and management up to the point of renal replacement therapy*. London, UK: Royal College of Physicians; 2013.

68. Levey AS, Coresh J, Greene T, et al. Using standardized serum creatinine values in the modification of diet in renal disease study equation for estimating glomerular filtration rate. *Ann Intern Med*. 2006;145(4):247-254. https://doi.org/10.7326/0003-4819-145-4-200608150-00004.
69. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons; 1987.
70. Sterne JAC, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*. 2009;338(1):b2393-b2393. https://doi.org/10.1136/bmj.b2393.
71. Charlson ME, Pompei P, Ales KL, CR MK. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J Chronic Dis*. 1987;40(5):373-383.
72. Riley RD, Price MJ, Jackson D, et al. Multivariate meta-analysis using individual participant data. *Res Synth Methods*. 2015;6(2):157-174. https://doi.org/10.1002/jrsm.1129.
73. Hickey GL, Philipson P, Jorgensen A, Kolamunnage-Dona R. Joint modelling of time-to-event and multivariate longitudinal outcomes: recent developments and issues. *BMC Med Res Methodol*. 2016;16(1):1-15. https://doi.org/10.1186/s12874-016-0212-5.
74. Rizopoulos D, Molenberghs G, Lesaffre EMEH. Dynamic predictions with time-dependent covariates in survival analysis using joint modeling and landmarking. *Biometrical J*. 2017;59(6):1261-1276. https://doi.org/10.1002/bimj.201600238.
75. Dunson DB. Bayesian latent variable models for clustered mixed outcomes. *J R Stat Soc Ser B Stat Methodol*. 2000;62(2):355-366. https://doi.org/10.1111/1467-9868.00236.
76. De Leon AR, Wu B. Copula-based regression models for a bivariate mixed discrete and continuous outcome. *Stat Med*. 2011;30(2):175-185. https://doi.org/10.1002/sim.4087.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Martin GP, Sperrin M, Snell KIE, Buchan I, Riley RD. Clinical prediction models to predict the risk of multiple binary outcomes: a comparison of approaches. *Statistics in Medicine*. 2021;40:498–517. https://doi.org/10.1002/sim.8787