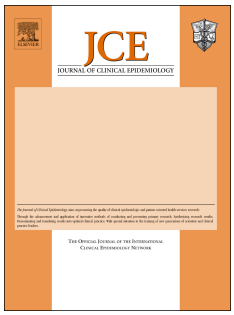
Accepted Manuscript



Explicit inclusion of treatment in prognostic modelling was recommended in observational and randomised settings

R.H.H. Groenwold, K.G.M. Moons, R. Pajouheshnia, D.G. Altman, G.S. Collins, T.P.A. Debray, J.B. Reitsma, R.D. Riley, L.M. Peelen

PII: S0895-4356(16)30030-0

DOI: 10.1016/j.jclinepi.2016.03.017

Reference: JCE 9145

To appear in: *Journal of Clinical Epidemiology*

Received Date: 15 September 2015

Revised Date: 1 March 2016

Accepted Date: 23 March 2016

Please cite this article as: Groenwold RHH, Moons KGM, Pajouheshnia R, Altman DG, Collins GS, Debray TPA, Reitsma JB, Riley RD, Peelen LM, Explicit inclusion of treatment in prognostic modelling was recommended in observational and randomised settings, *Journal of Clinical Epidemiology* (2016), doi: 10.1016/j.jclinepi.2016.03.017.

Explicit inclusion of treatment in prognostic modelling was recommended in observational and randomised settings

R.H.H. Groenwold[1], K.G.M. Moons[1,2], R. Pajouheshnia[1], D.G. Altman[3], G. S. Collins[3], T.P.A. Debray[1,2], J.B. Reitsma[1,2], R.D. Riley[4], L.M. Peelen[1]

1. Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, the Netherlands
2. Dutch Cochrane Center, University Medical Center Utrecht, the Netherlands
3. Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, OX3 7LD, United Kingdom.
4. Research Institute for Primary Care and Health Sciences, Keele University, Staffordshire, UK, ST5 5BG

Corresponding author:

R.H.H. Groenwold, MD, PhD

Julius Center for Health Sciences and Primary Care

University Medical Center Utrecht

PO Box 85500, 3508 GA Utrecht, the Netherlands

T: +31 88 755 9365; F: +31 88 756 8099; E: r.h.h.groenwold@umcutrecht.nl

Funding

Rolf H.H. Groenwold was funded by the Netherlands Organization for Scientific Research (NWO-Veni project 916.13.028). Karel G.M. Moons receives funding from the Netherlands Organisation for Scientific Research (project 9120.8004 and 918.10.615). Richard Riley and Doug Altman receive funding from an MRC Partnership Grant for the PROGnosis RESearch Strategy (PROGRESS) group (grant reference number: G0902393). Gary S. Collins receives funding from the Medical Research Council (grant number G1100513). The study sponsors had no role in the design of the study, the collection, analysis or interpretation of data, the writing of the report or the decision to submit the article for publication.

What is new

- When developing a prognostic model (that aims to produce accurate probabilities of the outcome in case the patient is not treated) using data from a randomised trial in which individuals from one arm do not receive treatment, restricting the analysis to untreated individuals may be a suitable strategy. However, removing all patients in the treatment group will reduce the sample size, leading to greater uncertainty around predictions and also to prognostic models that are more prone to overfitting.
- When developing a prognostic model using data from observational studies with treated patients, restricting the analysis to untreated individuals is not appropriate if treatment status depends on patient characteristics, including the predictors of the developed model.
- For either randomised or observational studies, it is preferable to explicitly model treatment when developing a prognostic model.

ABSTRACT

Objective: To compare different methods to handle treatment when developing a prognostic model that aims to produce accurate probabilities of the outcome of individuals if left untreated.

Study Design and Setting: Simulations were performed based on two normally distributed predictors, a binary outcome, and a binary treatment, mimicking a randomised trial or an observational study. Comparison was made between simply ignoring treatment (SIT), restricting the analytical dataset to untreated individuals (AUT), inverse probability weighting (IPW), and explicit modelling of treatment (MT). Methods were compared in terms of predictive performance of the model and the proportion of incorrect treatment decisions.

Results: Omitting a genuine predictor of the outcome from the prognostic model decreased model performance, in both an observational study and a randomised trial. In randomised trials, the proportion of incorrect treatment decisions was smaller when applying AUT or MT, compared to SIT and IPW. In observational studies, MT was superior to all other methods regarding the proportion of incorrect treatment decisions.

Conclusion: If a prognostic model aims to produce correct probabilities of the outcome in the absence of treatment, ignoring treatments that affect that outcome can lead to suboptimal model performance and incorrect treatment decisions. Explicitly modeling treatment is recommended.

Key words: Prognosis; Models, statistical; Computer simulation; Decision support techniques

Running title: Handling treatment in prognostic modelling

1. Introduction

Prognostic models (or risk scores) are increasingly important for clinical decision making.[1,2] For example, the predicted probability of an outcome, obtained through a prognostic model, may serve as the starting point for considerations of treatment initiation: high risks may lead to starting treatment, whereas in the case of low risks treatments may be withheld or delayed. For example, in the guideline of the European Society of Cardiology,[3] it is mentioned that "at risk levels >10%, drug treatment is more frequently required", although the authors caution that "no threshold is universally applicable". To guide individual treatment decisions, prognostic outcome predictions should ideally reflect the predicted course or outcome risk of disease if a patient were to remain untreated.[2,4]

Prognostic models are often developed using data from a randomised trial or an observational study, in which (at least part of the) individuals are treated.[5] If treatments are effective in reducing the risk of the predicted outcomes, simply ignoring those treatments in the development of a prognostic model may result in incorrect predictor-outcome associations and hence incorrect risk predictions of the natural history when used in new individuals.[6] Even though predictions are correct for those among whom the model was developed (the 'derivation set'), they may not generalize to future individuals who may be treated differently. In other words there is a danger of risk predictions being *confounded* by treatment: risk predictions appear low because of treatment, but in future patients the true risk might be substantially higher if they remain untreated. Further complications arise when treatment decisions in the data available were already being based on the values of the predictors in the model. For example, in patients with hypertension the observed predictive effect of blood pressure for cardiovascular outcomes is likely to be diluted, as those with high blood pressure will receive anti-hypertensive treatment, based on the observed high blood

pressure, in turn lowering their predicted risk. Thus if a prognostic model is developed using these data, the effect of blood pressure is likely to be downwardly biased and therefore risk predictions may be too low in future untreated individuals.

Methods to account for treatments in the development of a prognostic model to be used for predicting the health course of individuals in the absence of treatment include simply ignoring treatment,[5] restricting the development set to untreated individuals,[6] censoring observations after treatment has started,[7] and explicit modeling of the treatment.[8] Also, in the TRIPOD statement, there is an item on the reporting of treatment received among participants of a study developing or validating a multivariable prediction model for diagnosis or prognosis.[9]

In this article, we evaluate these different methods in situations that aim to develop a prognostic model generating predictions in case individuals were to remain untreated, which serve as input for treatment decisions. In particular, we examine how the methods impact upon the predictive performance and proportion of correct indications of treatment of a prognostic model being developed using data from a randomised or observational study.

2. Consequences of ignoring treatment in different phases of model development

The development and introduction of a new prognostic model comprises four distinct phases: derivation, validation, impact assessment, and implementation of the model.[1] As indicated above, for a model to be used to guide treatment decisions, the predictions made by the model should be the outcome risks of individuals if no treatment were to be given. This implies that such models should be developed in untreated populations. Nevertheless, in all phases of prognostic modelling research, some portion of the study population may actually be treated by an effective treatment.

When deriving a model in a treatment naïve population, the model will indeed provide risk predictions that reflect what will happen if a future but similar individual remains untreated. However, when part of the population is treated and treatment is ignored in the model derivation phase, the risk predictions from the model will be too low when validated or applied in individuals who are yet untreated. To what extent the predictions will be too low likely depends on the proportion of treated individuals in the derivation set and the magnitude of the treatment effect. Figure 1 illustrates this impact of ignoring treatment in the development of a prognostic model.

The impact of ignoring treatment when validating the developed model in new individuals obviously depends on what cohort of patients have been used in the derivation phase. If the model was derived in a treatment naïve population, the model will provide correct predictions if the individuals in the validation set are all untreated too; the predicted risks will correspond reasonably well with the observed risks. However, if such a developed model is validated in a (partly) treated population, the predicted risks will appear to be too high, if treatment is simply ignored in the validation phase.

When a model is derived in a (partly) treated population and this treatment is ignored in the development, the predicted risk will be too low, when validating the model in a treatment naïve population. If the proportion of treated individuals and the reasons for treatment are the same in the derivation and the validation phase, however, the predicted risks will appear to be correct in the validation phase, while in fact both are too low in those who are treated, if these risks are considered the outcome risks if no treatment were to be given.

How treatment was handled in the derivation and validation phase of model development directly impacts the usefulness of a prognostic model in daily clinical care, because incorrect predicted risks may lead to incorrect treatment decisions (e.g., in the presence of a risk threshold above which treatment is administered). When predicted risks are too high, too many individuals may receive treatment, while predicted risks that are too low risk may lead to undertreatment.

3. Methods

3.1 Outline of simulations

In this section, we focus on different methods to handle treatment in the derivation of a prognostic model and their possible impact in terms of incorrect treatment decisions. The simulated scenarios are outlined in Figure 2. In all scenarios, there are two (possibly correlated) variables (denoted X_1 and X_2), which are associated with the outcome of interest (Y). Each of these variables can be considered as a single predictor, or as a summary of multiple predictors (i.e., X_1 and X_2 are variables or vectors of variables). In addition, there is a single binary treatment (T). Half of the chosen scenarios mimic data from a randomised trial, in which treatment was a random process (i.e., independent of the variables X_1 and X_2), and the remaining scenarios are chosen to mimic observational data, in which treatment decisions were based on the values of the variables X_1 and X_2. We assume that this treatment is effective in reducing the outcome. We also assume that there are no other sources of bias, apart from the potential confounding effect by X_1 and X_2.

3.2 Simulation setup

All simulations and analyses were performed in R for windows, version 3.1.3.[10] The simulation code is available upon request.

Simulated datasets of 1000 individuals were generated. Each dataset consisted of four variables: two continuous, standard normally distributed variables (indicated by X_1 and X_2), a binary treatment (indicated by T), and a binary outcome (indicated by Y). Data were generated by first sampling X_1 and X_2 from a multivariate normal distribution with a correlation of 0 or 0.3 between the two variables, which is a realistic range of correlations typically observed in behavioural, socioeconomic, and physiological factors in biomedical

research.[11] Next, the binary treatment status and outcome were sequentially generated by sampling from a Bernoulli distribution with individual-specific probabilities of treatment and outcome status, $\pi = p_{i,t}$ and $\pi = p_{i,y}$, respectively. The true individual-specific probabilities of treatment status ($p_{i,t}$) (i.e., the probability of receiving treatment) were generated using the logistic model:

$$logit(p_{i,t}) = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} , \qquad (1)$$

which implies that treatment decisions are based on the variables X_1 and X_2. The true individual-specific probabilities of outcome status ($p_{i,y}$) were generated using the logistic model:

$$logit(p_{i,y}) = \beta_0 + \beta_1 T_i + \beta_2 X_{1i} + \beta_3 X_{2i} , \qquad (2)$$

which implies that the probability of the outcome depends on the variables X_1 and X_2 as well as treatment status (T). The values of the parameters α_0, α_1, α_2, β_0, β_1, and β_2 (Figure 2) differed between scenarios (see Section 3.3). Notably, in those scenarios that mimicked a randomised trial, the parameters α_1 and α_2 were set to zero, in order to make treatment allocation a random process. A step-by-step guide to the simulation study is now outlined.

3.3 Step 1: Choose one of 10 simulated scenarios

To assess the impact of different methods to handle treatment in the development of a prognostic model, 10 different scenarios were considered and, in each of these, data were generated using the set-up described in Section 3.2. The different scenarios and parameter settings are summarized in Table A1 (in the Appendix).

Scenarios 1-5 represent the development of a prognostic model using data from a randomised trial. In scenario 1 (default scenario), the variables X_1 and X_2 were considered independent ($\rho=0$). Both variables were equally associated with the outcome: both increased the log(odds) of the outcome by 1 per unit increase (i.e., $\beta_1=\beta_2= 1$). Treatment assignment was a random process ($\alpha_1=\alpha_2=0$) and treatment was present in approximately 50% of the individuals. Treatment was considered to be effective in preventing the outcome (OR=0.5). Approximately 10% of the individuals experienced the outcome of interest. In scenarios 2-5 one of the simulation parameters from scenario 1 was changed. In scenario 2, the variables X_1 and X_2 were correlated ($\rho=0.3$). In scenario 3, the association between X_1 and the outcome was doubled ($\beta_1=2$), while the association between X_2 and the outcome remained unchanged ($\beta_2=1$). This resembles a situation in which one of two (sets of) variables has a larger contribution in predicting the outcome. In a similar way, in scenario 4, the association between X_2 and the outcome was doubled (and thus twice as large as the association between X_1 and the outcome). In scenario 5, the treatment was less effective (OR=0.9).

Scenarios 6-10 were the same as scenarios 1-5, respectively, except that now treatment assignment was not a random process, but depended on X_1 and X_2, thus mimicking the development of a prognostic model using data from an observational study in which some patients received treatment. Both X_1 and X_2 increased the log(odds) of the treatment by 1 per unit increase (i.e., $\alpha_1=\alpha_2= 1$).

3.4 Step 2: Implement the different methods to develop prognostic models in the presence of treatments

In each simulated dataset, eight different approaches were applied to develop a prognostic model to predict outcome Y. The eight different methods are summarized in Table A2 (in the Appendix) and described hereafter.

For all methods, the model relating the outcome to the predictors was a logistic regression model (in line with the data generating model). For half of the methods both predictors (i.e. X_1 and X_2) were considered observed, whereas for the other half the predictor X_2 was considered unobserved. The latter may also correspond to a situation in which a possible predictor of the outcome is intentionally omitted from the model, for example because the measurement of the predictor is very costly, or invasive. As previously indicated, each of the variables X_1 and X_2 can be considered as a combination (or reflection) of multiple predictors, thus even the models including only X_1 could be considered as prognostic models including multiple predictors. Each of the methods differed in the way treatment was accounted for.

For methods 1 and 2, which are the simply ignore treatment (SIT) methods, treatment was simply ignored: method 1 was a model regressing Y on X_1, ignoring X_2 and T; method 2 was a model regressing Y on X_1 and X_2, ignoring T.

For methods 3 and 4, analysis of untreated individuals (AUT), analysis was performed using information on untreated individuals only (i.e., restriction to those for whom T=0); the outcome models were the same as in methods 1 and 2.

Methods 5 and 6 were based on inverse probability weighting (IPW). First, a logistic regression model was fitted regressing treatment (T) on the predictor X_1 (method 5), or regressing T on the predictors X_1 and X_2 (method 6). This yielded individual probabilities of

being treated for all individuals in the dataset. Next, treated individuals were weighted by the inverse of the probability of being treated, while untreated individuals were weighted by the inverse of the probability of not being treated. Weighting thus created a pseudo-population in which treatment status was independent of the predictors X_1, or X_1 and X_2. A weighted regression model was then fitted regressing Y on X_1 (method 5), or regressing Y on X_1 and X_2 (method 6).

In methods 7 and 8, treatment was explicitly modelled as a separate predictor (MT): method 7 was a model regressing Y on X_1 and T, ignoring X_2; method 8 was a model regressing Y on X_1, X_2, and T. None of the methods included corrections for optimism, as no selection procedure was used to select predictors for inclusion in the final model.[12]

3.5 Step 3: Calculate and compare parameters of apparent performance of the developed models

For each scenario, 1000 datasets of 1000 individuals each were created, and the prognostic models were developed and evaluated for each dataset. The values of the performance measures were then averaged across all 1000 datasets. For each scenario separately, the apparent performance of the different methods to handle treatment in the development of a prognostic model was compared using the Brier score,[13] Harrell's c-statistic,[14] the observed by expected risk prediction ratio, the standard errors of the association between the predictor X_1 and the outcome, and the proportion of incorrect treatment decisions. These performance measures were also compared against their optimal value, which was calculated based on the data generating model (model 2 in Section 3.2). For each individual, the untreated probability of the outcome could be calculated based on their values of the variables X_1, X_2, while setting

treatment status to untreated (T=0). This model-based untreated probability of the outcome was then used to estimate the optimal values of the different performance measures.

The impact on treatment decisions was assessed by calculating the proportion of false-positive treatment decisions (i.e., the proportion of individuals who are treated, while in fact they should not be treated) and the proportion of false-negative treatment decisions (i.e., the proportion of individuals who are not treated, while in fact they should be treated). For each individual, the true probability of the occurrence of the outcome if the individual remained untreated was calculated based on the true outcome model. Based on this true untreated probability, a correct treatment decision could be made: if the untreated probability of an outcome event exceeds an (a-priori chosen) treatment threshold, an individual should be treated, while the individual should not be treated if the probability does not exceed the threshold. This was compared to the actual treatment decision, which was based on the prediction model that was developed with one of the considered methods (described in Section 3.4). For the scenarios 1 (randomised trial) and 6 (observational data), the proportions of incorrect treatment decisions were estimated for a range of treatment threshold between 0.025 and 0.5. For all other scenarios, the treatment threshold was set at 10%, i.e., individuals receive treatment if their predicted risk exceeds 10%, while they remain untreated if their predicted risk is lower than the threshold.

4. Results

4.1 Development of a prognostic model using data from a randomised trial

Figure 3 shows the impact different methods have on treatment decisions for scenario 1, which mimics development of a prognostic model in a randomised study. Since the results for the models SIT_1, SIT_2, MT_1, and MT_2 were equal to IPW_1, IPW_2, AUT_1, and AUT_2, respectively, only the results for the first four are plotted. Both panels show that as the treatment threshold increases, the probability of a false-positive treatment decision decreases and the probability of a false-negative treatment decision increases. The models in which only one of the two predictors is included (SIT_1 and MT_1) are clearly inferior to the models in which both predictors are included (SIT_2 and MT_2). Although the SIT_2 model results in fewer false-positive treatment decisions than the MT_2 model, the latter model results in fewer false-negative treatment decisions.

In Table 1, the impact different methods have on treatment decisions is shown for all scenarios, using a treatment threshold of 10%. Explicitly modelling treatment (MT) and analysis of untreated individuals (AUT) led to similar proportions of incorrect treatment decisions. Irrespective of the method used, proportions of incorrect treatment decisions (either false-positive or false-negative treatment decisions) increase when omitting predictor X_2 from the model and can be as large as 0.444 (MT_1, scenario 4). When modelling both predictors X_1 and X_2, yet ignoring treatment (SIT_2 or IPW_2), the probability of a false-negative treatment decision (i.e., not treating an individual when in fact they should be treated) is still considerably large. For example, in scenario 1, the probability of a false-negative treatment decision is 0.208 for methods SIT_2 and IPW_2, whereas it is 0.066 and 0.058 for AUT_2 and MT_2, respectively. The reason is that the predicted probabilities of the outcome in the absence

of treatment are too low: a probability below the treatment threshold might actually be the result of treatment and thus in the absence of treatment, this probability should be higher.

Irrespective of the method used, adding a genuine predictor to the model improves the Brier score (smaller values indicate better performance) as well as the c-statistic (larger values indicate better performance) (Table 2). Methods in which predictor X_2 is considered unobserved performed better under scenario 3 than under scenario 4, because in scenario 4 the predictor X_2 is the most influential predictor, whereas in scenario 3 the most influential predictor is X_1. Although performance improves when explicitly modelling treatment (MT_1 and MT_2), this improvement is small and ignoring treatment (SIT_1 and SIT_2) appears to have relatively little impact. The observed-to-expected ratio was 1.000 for all methods, except for the analysis of untreated individuals (AUT), because the treatment was effective and the expected probability of the outcome among the untreated is higher than the overall probability of the outcome. Restriction to just untreated individuals reduces the sample size, which results in larger standard errors compared to the other methods.

4.2 Development of a prognostic model using data from an observational study

Figure 4 shows the impact different methods have on treatment decisions for scenario 6, which mimics development of a prognostic model in observational data. Similar patterns are observed as in Figure 3. However, the probability of false-positive treatment decisions is less affected by excluding the second predictor from the model. Although analysis of untreated individuals (AUT) results in lowest probabilities of false-positive treatment decisions, this model is clearly inferior to a model in which treatment is explicitly modelled (MT) when comparing probabilities of false-negative treatment decisions.

Table 3 shows the impact different methods have on treatment decisions if the prognostic model is developed using observational data, using a treatment threshold of 10%. Again, including both predictors X_1 and X_2 improves treatment decisions. Compared to the results of simulations mimicking a randomised trial (Table 1), in the simulations of observational data the analysis of untreated individuals more often leads to false-negative treatment decisions (and less often to false-positive treatment decisions). The reason is that in the simulated scenarios on average the untreated individuals have a relatively low probability of the outcome, leading to an underestimation of the actual probability of the outcome and, hence, an increased probability of a false-negative treatment decision.

The methods SIT, IPW, and MT showed similar performance (Brier score and c-statistic) (Table 4). The only exception is the analysis of untreated individuals (AUT), which yields Brier scores that are smaller, e.g., the Brier score of the model including treatment: e.g., 0.040 for AUT_2 vs 0.064 for MT_2 (scenario 3). The observed-to-expected ratio was larger than 1 for the analysis of untreated individuals (AUT), because the simulated scenarios were such that particularly high-risk individuals were treated (thus selecting individuals with a relatively low probability of the outcome for the analysis of untreated individuals). Again, restriction to untreated individuals (AUT) reduces sample size, which results in larger standard errors compared to the other methods.

5. Discussion

This simulation study shows that when developing a prognostic model, ignoring an effective treatment results in incorrect predictions of the outcome if an individual were to remain untreated. To resolve this, in the case of randomised trials one can either restrict analyses to untreated individuals, or include all treated and untreated individuals with treatment included as a predictor in the model. The latter approach is recommended as the sample size stays larger and is thus far more efficient to identify genuine predictor-outcome associations. When prognostic models are developed using data from observational studies, analysis of untreated individuals only is not appropriate because in observational data, those who are untreated may have a relatively high (or low) probability of the outcome, leading to an overestimation (or underestimation) of the outcome risk. Including treated as well as untreated individuals and including treatment as a predictor in the model will overcome this problem.

Typically, the development of a prognostic model starts with derivation of the model in one cohort, followed by validation in another cohort, and finally implementation in clinical practice.[1,9] Here, we focused on the first step, i.e., derivation of the prognostic model and its apparent (internal) performance in the same data used to develop the model. If the model is derived in a treatment-naïve population, yet validated in a non-treatment-naïve population (or vice versa), the performance of the model may be poor if treatment is ignored at either of the two phases. This may partly explain poor performance when applying a prognostic model outside the population in which the model was derived.[15] Also, importantly, our simulations show that developing or validating the model in a subset of untreated individuals may not yield optimal performance, if treatment assignment is not a random process (scenarios 6-10). In that case, treatment should explicitly be taken in to account when modelling the outcome. Notably, when developing a prognostic model using data from an observational study, some

confounders of the treatment-outcome relation may be unobserved. The models in which the predictor X_2 is considered unobserved mimic this situation (i.e., the models SIT_1, AUT_1, IPW_1, and MT_1). When comparing explicit modelling of treatment (MT_1) with ignoring treatment (SIT_1), in the scenarios considered the former approach is superior in terms of false-positive treatment decision, Brier score, and c-statistic, while inferior in terms of standard error and false-negative treatment decision. The decision to model treatment explicitly should therefore take into account which of the performance measures is considered most important.

For each scenario and for each method considered, we assessed the impact of omitting one (set of) predictor(s) for the outcome from the prognostic model (specifically, the predictor X_2 was considered unobserved and thus omitted from the model). Obviously, omitting a relatively weak predictor from the model has less impact on the performance of the prognostic model than omitting a relatively strong predictor. Likewise, when the treatment has a relatively small effect on the outcome compared to the predictors included in the model (scenarios 5 and 10), ignoring it probably will have less impact compared to ignoring a treatment that has a large effect on the outcome.

A clear advantage of simulation studies, in contrast to using empirical data, is that methods can be compared to a reference (in this case the 'true' probability of the outcome if an individual remains untreated). An obvious downside of simulation studies is that simulated scenarios may be deemed unrealistic. For example, we simulated only two continuous predictors of the outcome, whereas in prognostic research multiple predictors are likely to be considered (including non-continuous ones). However, these two predictors can of course also be considered as combinations of multiple predictors, including dichotomous, categorical and continuous predictors, and the results of our simulations likely also apply to such settings.

Furthermore, only binary outcomes were considered, and time to occurrence of the outcome was ignored. In addition, we focused on the development of prognostic models in a setting in which treatment was initiated at the start of follow-up for each individual and remained constant during follow-up. Interactions between predictors and treatment were not considered in the simulated scenarios (i.e., no treatment effect modification). In prognostic studies, in which the strength of a prediction changes in case treatment is given, such interactions may be required to model the data appropriately.

In the method that applied inverse probability weighting, the treatment was not included as a predictor in the (weighted) model regressing the outcome on the predictors. Consequently, the method in which treatment was simply ignored (SIT) and the IPW method yielded the same results in case of developing the model using data from a randomised trial. The IPW method could be improved upon by including the treatment in the weighted outcome model.

Future research could address the possible impact of time-varying treatments in this setting. In randomised trials, information on allocated treatment (i.e., intention-to-treat) may be insufficient and detailed information on actual use may be required. Also other ‘treatments’ such as lifestyle changes (including dietary habits) and non-pharmacological interventions such as surgical interventions should be considered. Furthermore, the consequences of the choice of method to handle treatments in diagnostic prediction research might be different from prognostic research, because for example treatments already may have been started based on symptoms of the target condition before the measurements of the diagnostic test(s) under evaluation are made. Likewise, in prognostic studies, the treatment may have been started before the measurement of the predictor too, and subsequently affect the predictor value. These situations were not addressed in this study.

Based on the results, several recommendations can be made, which are summarized in the Text Box.

We conclude that ignoring treatments that affect the outcome in the development of a prognostic model can result in incorrect predicted probabilities for individuals if they were to remain untreated, which in turn may lead to incorrect treatment decisions. A solution is to explicitly model such treatments in the development of a prognostic model, although this may be challenging particularly when treatment status changes over time or when treatment effect is modified by patient-level covariates. Regardless, researchers who develop a prognostic model must be explicit in how treatment was handled, as recommended in the TRIPOD Statement for reporting prediction models,[9] and be clear how absolute risk prediction derived from a prognostic model should be viewed in the context of current treatment strategies.

REFERENCES

1. Moons KG, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. BMJ. 2009;338:b606.
2. Campbell W, Ganna A, Ingelsson E, Janssens AC. Prediction impact curve is a new measure integrating intervention effects in the evaluation of risk models. J Clin Epidemiol. 2015 Jun 25. pii: S0895-4356(15)00318-2. doi: 10.1016/j.jclinepi.2015.06.011. [Epub ahead of print].
3. Perk J, De Backer G, Gohlke H, Graham I, Reiner Z, Verschuren M, Albus C, Benlian P, Boysen G, Cifkova R, Deaton C, Ebrahim S, Fisher M, Germano G, Hobbs R, Hoes A, Karadeniz S, Mezzani A, Prescott E, Ryden L, Scherer M, Syvänne M, Scholte op Reimer WJ, Vrints C, Wood D, Zamorano JL, Zannad F; European Association for Cardiovascular Prevention & Rehabilitation (EACPR); ESC Committee for Practice Guidelines (CPG). European Guidelines on cardiovascular disease prevention in clinical practice (version 2012). The Fifth Joint Task Force of the European Society of Cardiology and Other Societies on Cardiovascular Disease Prevention in Clinical Practice (constituted by representatives of nine societies and by invited experts). Eur Heart J. 2012;33(13):1635-701.
4. Hemingway H, Croft P, Perel P, Hayden JA, Abrams K, Timmis A, Briggs A, Udumyan R, Moons KG, Steyerberg EW, Roberts I, Schroter S, Altman DG, Riley RD; PROGRESS Group. Prognosis research strategy (PROGRESS) 1: a framework for researching clinical outcomes. BMJ. 2013;346:e5595.
5. Liew SM, Doust J, Glasziou P. Cardiovascular risk scores do not account for the effect of treatment: a review. Heart. 2011;97(9):689-97.
6. Frankel MR, Morgenstern LB, Kwiatkowski T, Lu M, Tilley BC, Broderick JP, Libman R, Levine SR, Brott T. Predicting prognosis after stroke: a placebo group analysis from

the National Institute of Neurological Disorders and Stroke rt-PA Stroke Trial. Neurology. 2000;55(7):952-9.

7. Lawton M, Tilling K, Robertson N, Tremlett H, Zhu F, Harding K, Oger J, Ben-Shlomo Y. A longitudinal model for disease progression was developed and applied to multiple sclerosis. J Clin Epidemiol. 2015;68(11):1355-65.
8. van Leeuwen N, Lingsma HF, Perel P, Lecky F, Roozenbeek B, Lu J, Shakur H, Weir J, Steyerberg EW, Maas AI; International Mission on Prognosis and Clinical Trial Design in TBI Study Group; Corticosteroid Randomization After Significant Head Injury Trial Collaborators; Trauma Audit and Research Network. Prognostic value of major extracranial injury in traumatic brain injury: an individual patient data meta-analysis in 39,274 patients. Neurosurgery. 2012;70(4):811-8.
9. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. J Clin Epidemiol. 2015;68(2):134-43.
10. R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
11. Smith GD, Lawlor DA, Harbord R, Timpson N, Day I, Ebrahim S. Clustered environments and randomized genes: a fundamental distinction between conventional and genetic epidemiology. PLoS Med. 2007;4(12):e352.
12. Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. New York: Springer; 2009.
13. Brier GW. Verification of Forecasts Expressed in Terms of Probability". Monthly Weather Review 1950;78:1-3.

14. Harrell FE Jr, Lee KL, Califf RM, Pryor DB, Rosati RA. Regression modelling strategies for improved prognostic prediction. Stat Med. 1984;3(2):143-52.

15. Cook NR, Ridker PM. Further insight into the cardiovascular risk calculator: the roles of statins, revascularizations, and underascertainment in the Women's Health Study. JAMA Intern Med. 2014;174(12):1964-71.

FIGURES

Figure 1. Impact of ignoring treatment in prognostic modelling research.

Legend Figure 1

The top panel shows risk predictions of the outcome made by a prognostic model derived in a treatment-naïve population or a population in whom everyone is treated. Treatment is assumed to be equally effective on a relative scale in all individuals (constant risk ratio for treatment), yet ignored in the development of the prediction model. Hence, the predictions based on the model developed in the treated population underestimate the true untreated risk.

The middle panel shows a hypothetical distribution of baseline risk in a population in which treatment decision are to be made. In the presence of a treatment threshold, above which treatment is initiated, the predicted untreated risks based on the model derived in the treatment-naïve population yield correct treatment decisions (shaded grey area).

The bottom panel shows the same hypothetical distribution of baseline risk in a population in which again treatment decision are to be made. The predicted untreated risks based on the model derived in the treated population are too low and thus for some subjects their predicted untreated risk drops below the threshold, leading to incorrectly withholding treatment (i.e., undertreatment, false-negative treatment decision). These are indicated by the striped grey area. The shaded grey area indicates correct treatment initiation irrespective of the incorrect predicted risks.

Figure 3. Incorrect treatment decisions based on prognostic models developed using data from a randomised trial.

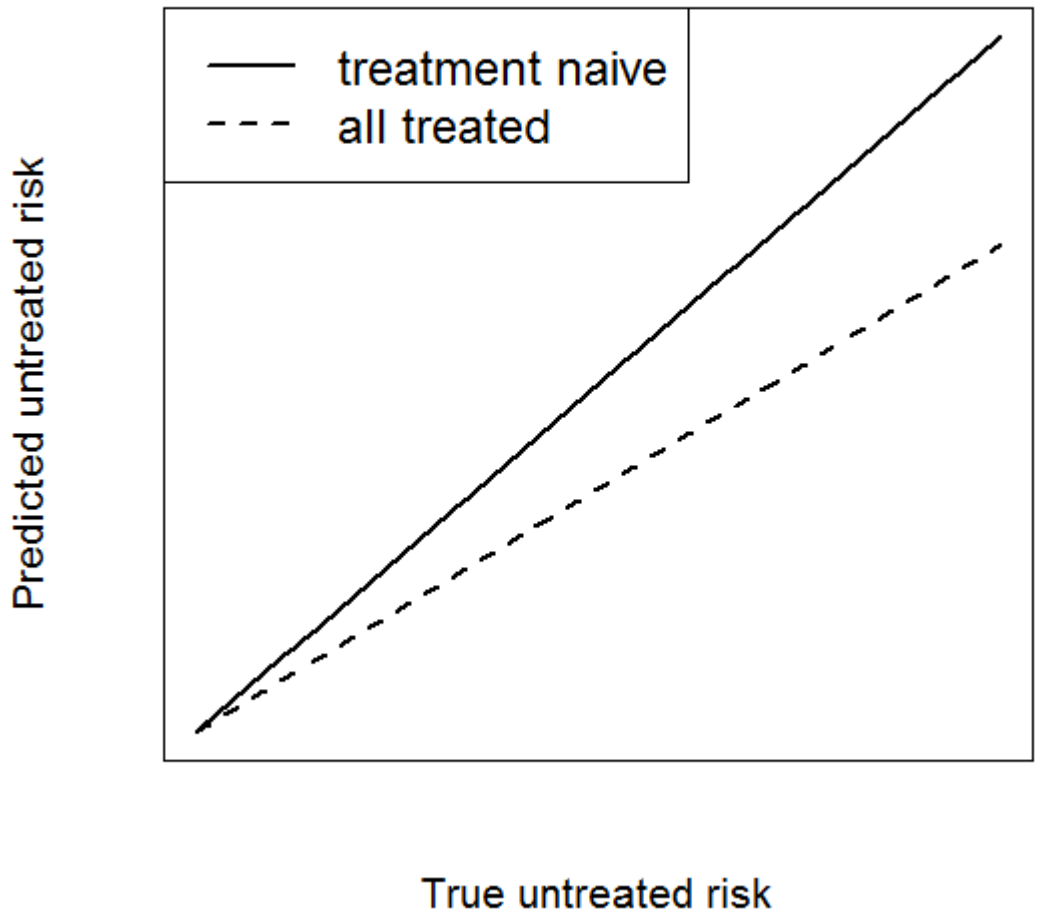
Legend Figure 3

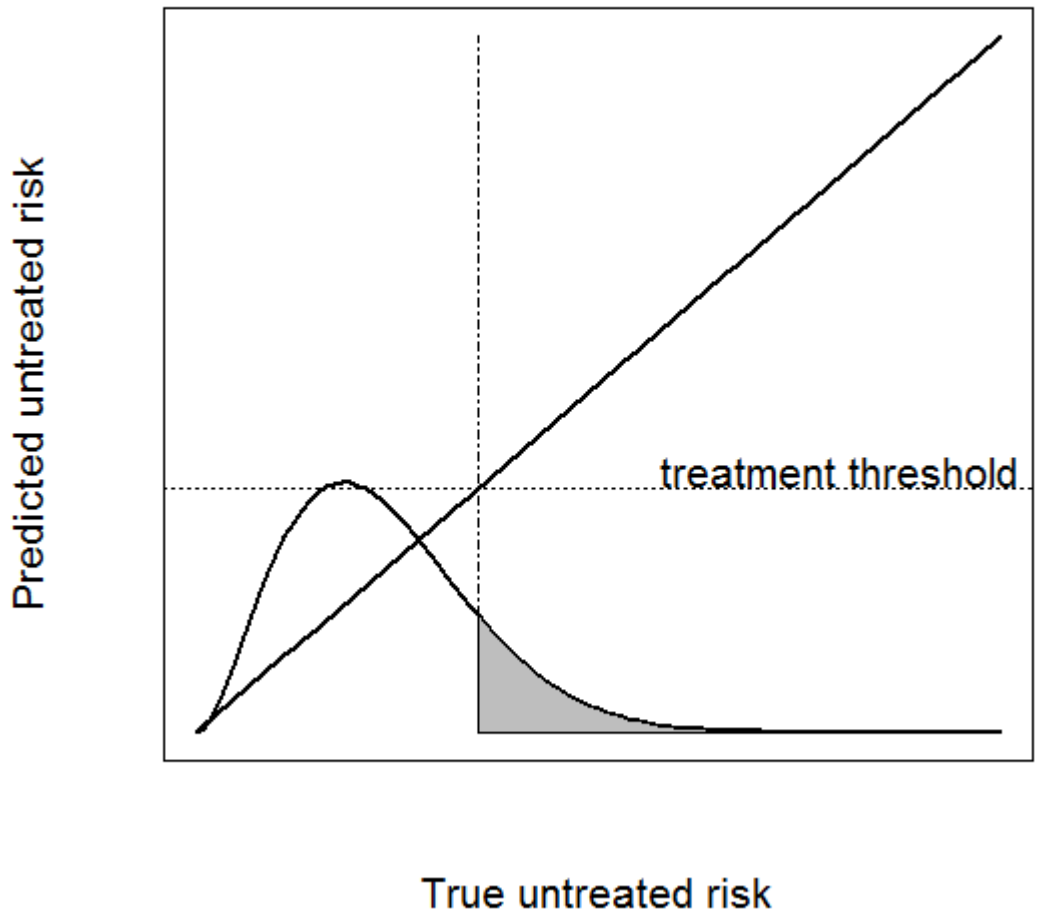
Graphs show probability of false positive (left panel) and false negative (right panel) treatment decisions when developing a model in the presence of an effective treatment, which is differently handled by the different methods. SIT: simply ignore treatment (in SIT_1 predictor X_2 is considered unobserved; in SIT_2 both predictors X_1 and X_2 are considered observed); MT: model includes treatment (in MT_1 predictor X_2 is considered unobserved; in MT_2 both predictors X_1 and X_2 are considered observed). See text for details.

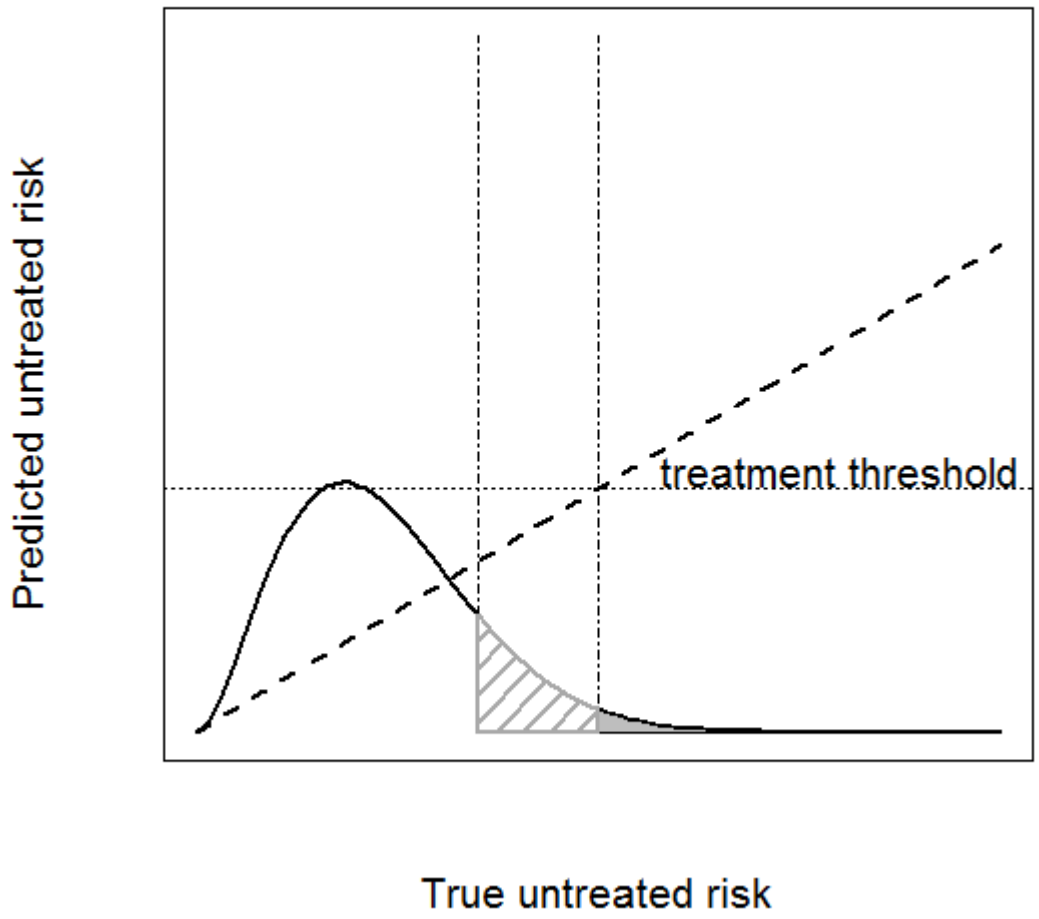
Figure 4. Incorrect treatment decisions based on prognostic models developed using observational data.

Legend Figure 4

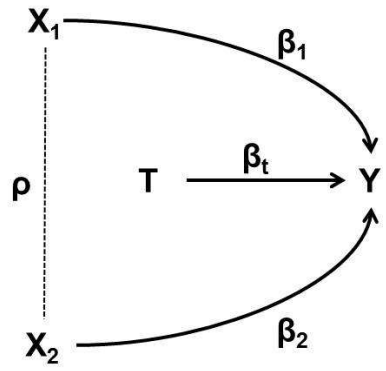
Graphs show probability of false positive (left panels) and false negative (right panels) treatment decisions when developing a model in the presence of an effective treatment, which is differently handled by the different methods. SIT: simply ignore treatment (in SIT_1 predictor X_2 is considered unobserved; in SIT_2 both predictors X_1 and X_2 are considered observed); AUT: analysis of untreated individuals (in AUT_1 predictor X_2 is considered unobserved; in AUT_2 both predictors X_1 and X_2 are considered observed); IPW: inverse probability weighting (in IPW_1 predictor X_2 is considered unobserved; in IPW_2 both predictors X_1 and X_2 are considered observed); MT: model includes treatment (in MT_1 predictor X_2 is considered unobserved; in MT_2 both predictors X_1 and X_2 are considered observed). See text for details.


treatment naive
all treated
Predicted untreated risk
True untreated risk


Predicted untreated risk
treatment threshold
True untreated risk

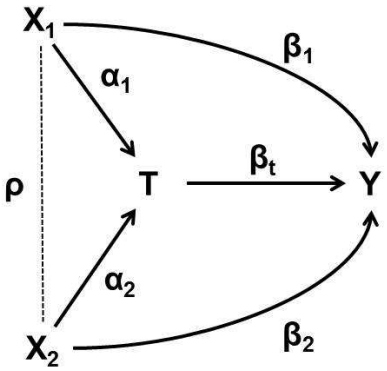

Predicted untreated risk
treatment threshold
True untreated risk

Randomized trial

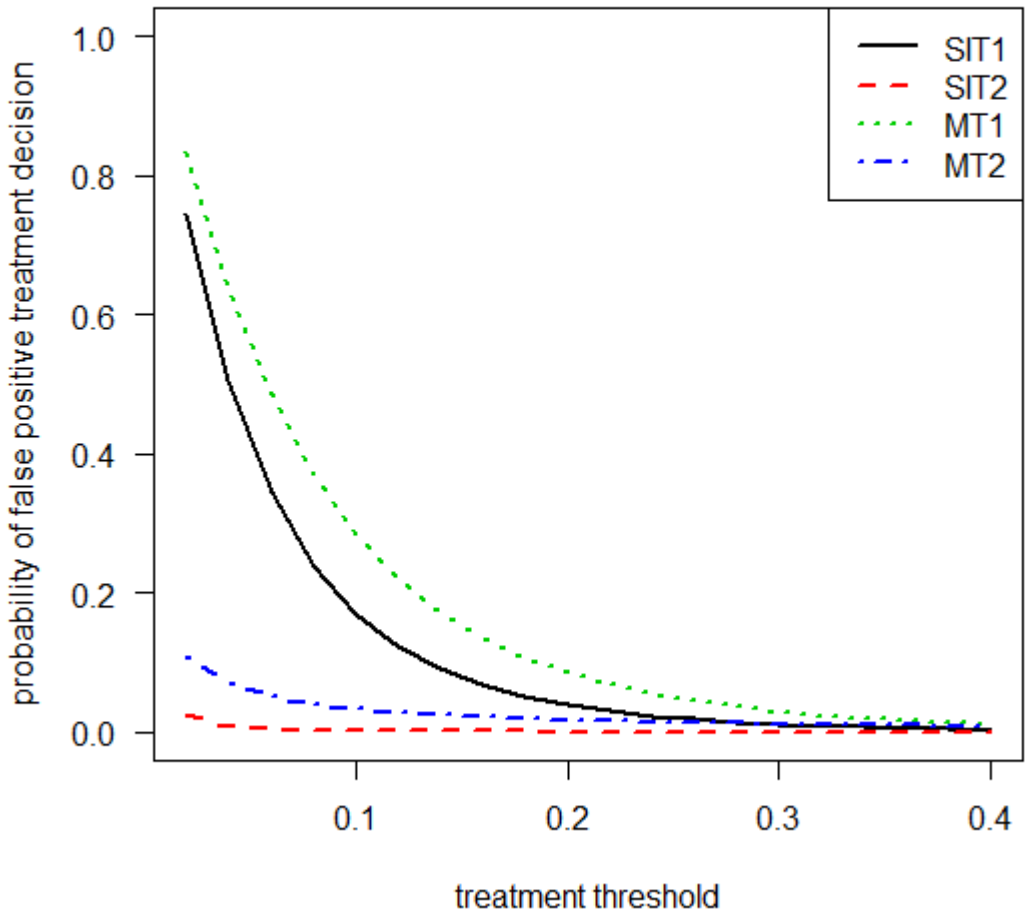


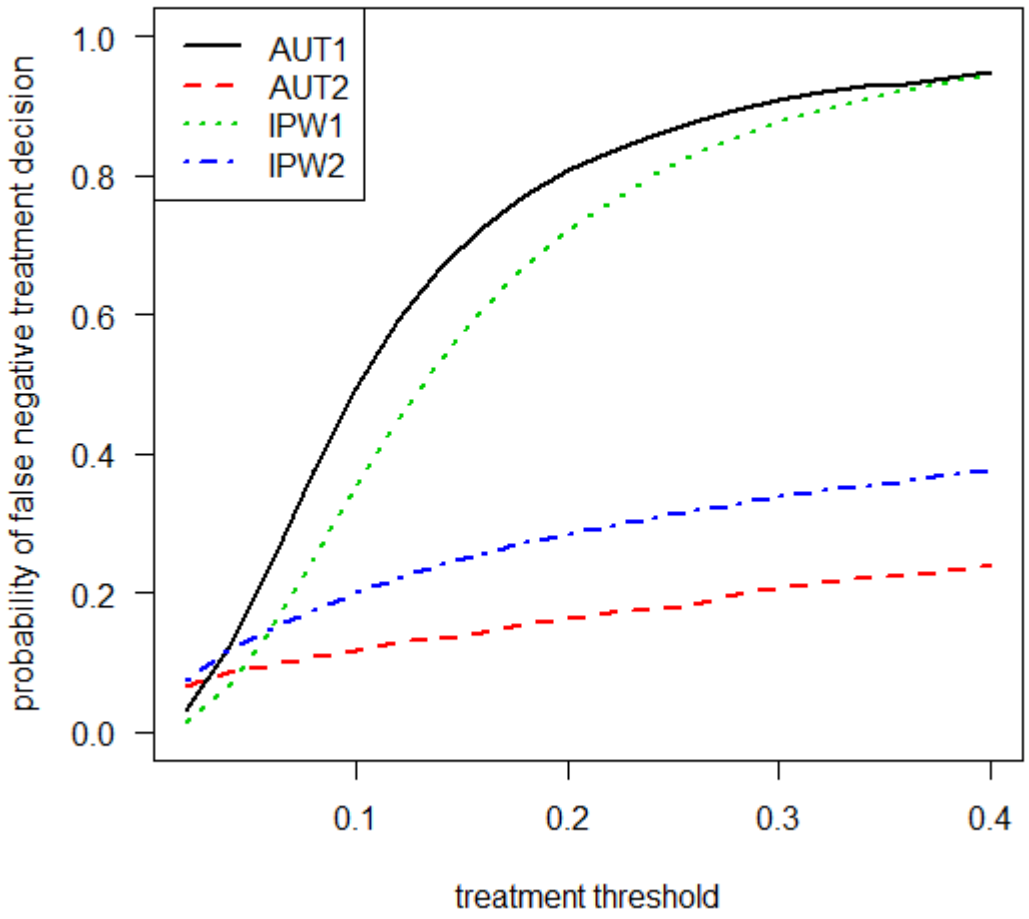
Simulation parameter	Default setting	Modified setting
ρ	0	0.3
β_1	1	2
β_2	1	2
β_t	log(0.5)	log(0.9)
P(T=1)	0.5	
P(Y=1)	0.2	

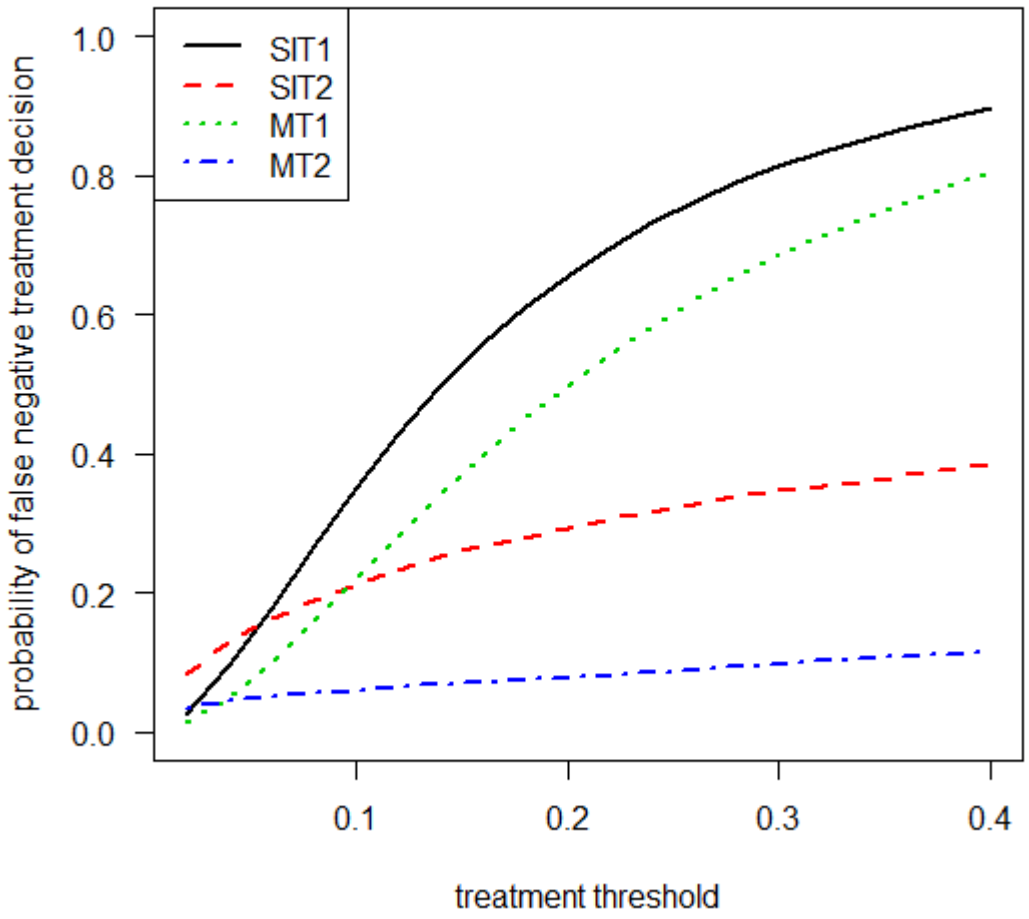
Observational study

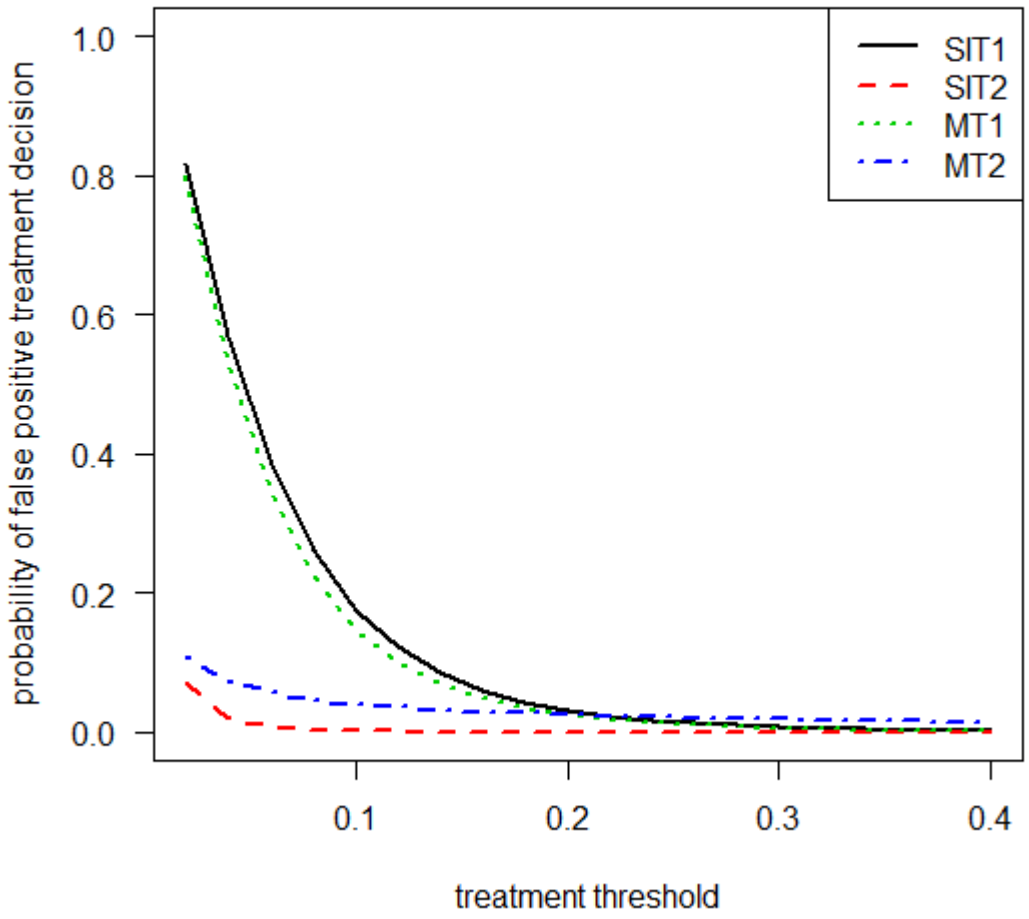


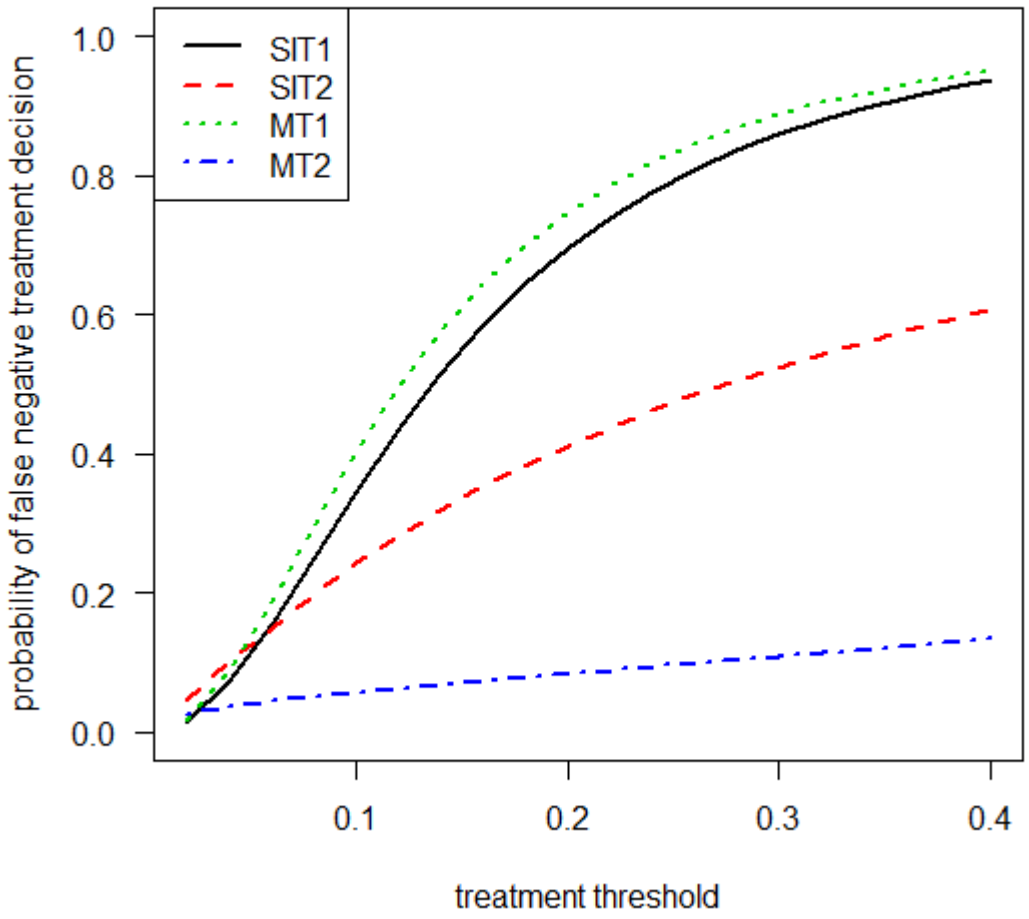
Simulation parameter	Default setting	Modified setting
ρ	0	0.3
α_1	1	
α_2	1	
β_1	1	2
β_2	1	2
β_t	log(0.5)	log(0.9)
P(T=1)	0.5	
P(Y=1)	0.2	

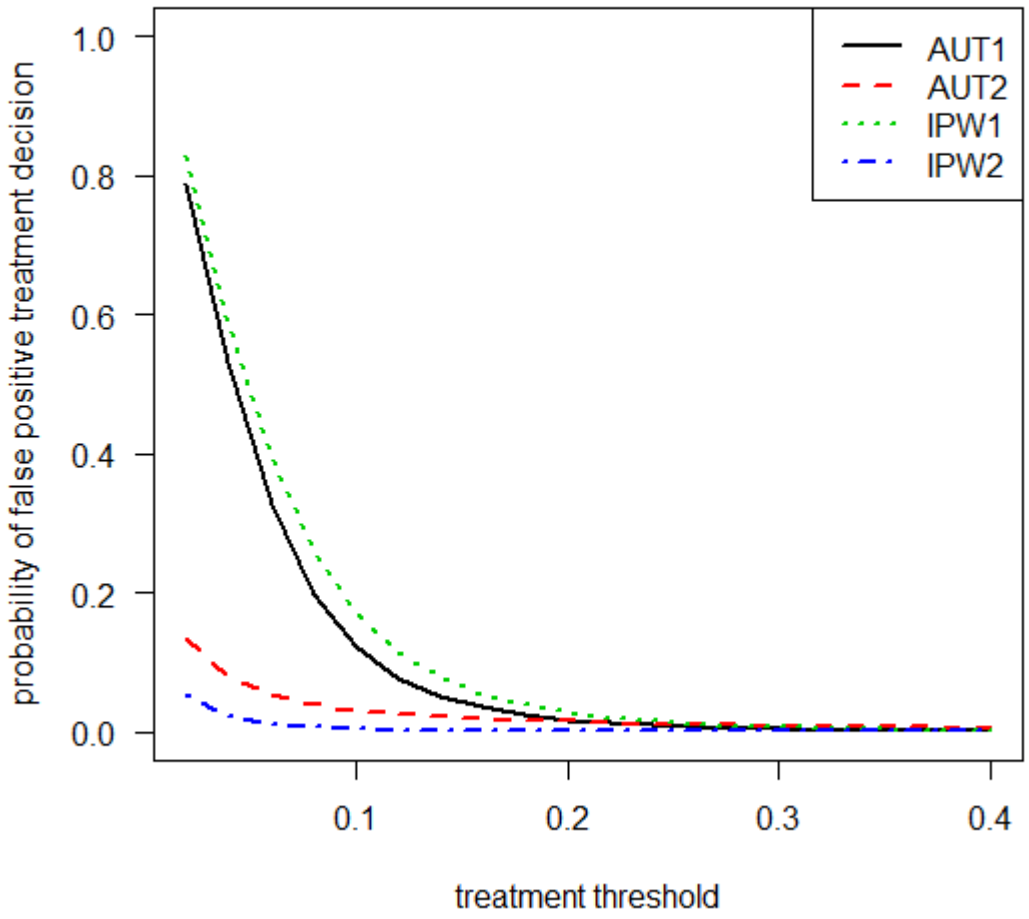

SIT1
SIT2
MT1
MT2
probability of false positive treatment decision
1.0
0.8
0.6
0.4
0.2
0.0
0.1
0.2
0.3
0.4
treatment threshold


AUT1
AUT2
IPW1
IPW2
probability of false negative treatment decision
1.0
0.8
0.6
0.4
0.2
0.0
0.1
0.2
0.3
0.4
treatment threshold


SIT1
SIT2
MT1
MT2
probability of false negative treatment decision
1.0
0.8
0.6
0.4
0.2
0.0
0.1
0.2
0.3
0.4
treatment threshold


SIT1
SIT2
MT1
MT2
probability of false positive treatment decision
1.0
0.8
0.6
0.4
0.2
0.0
0.1
0.2
0.3
0.4
treatment threshold


SIT1
SIT2
MT1
MT2
probability of false negative treatment decision
1.0
0.8
0.6
0.4
0.2
0.0
0.1
0.2
0.3
0.4
treatment threshold


AUT1
AUT2
IPW1
IPW2
probability of false positive treatment decision
1.0
0.8
0.6
0.4
0.2
0.0
0.1
0.2
0.3
0.4
treatment threshold

Text box. Recommendations on handling treatments in the development of prognostic models.

- If a prognostic model aims to produce accurate individual probabilities of the outcome in the absence of treatment, ignoring treatments that affect the outcome in the development of such a model can lead to suboptimal model performance, incorrect predicted probabilities, and thus suboptimal treatment decisions.
- Restricting the analysis to untreated individuals may only be a suitable strategy when developing a prognostic model using data from a randomised trial in which individuals from one treatment arm truly receive no treatment (or placebo), but not in the case of a randomised trial that compares two active treatments. Furthermore, restriction to untreated individuals reduces the sample size and thus the precision of estimated predictor-outcome associations.
- Restricting the analysis to untreated individuals is not appropriate when prognostic models are developed using data from observational studies in which treatment status depends on patient characteristics (including the predictors). Instead, it is preferred to explicitly model treatments that affect the outcome when developing a prognostic model.

TABLES

Table 1. Impact on treatment decisions of methods to develop prognostic models using randomised data.

		Method							
	Scenario	**SIT_1**	**SIT_2**	**AUT_1**	**AUT_2**	**IPW_1**	**IPW_2**	**MT_1**	**MT_2**
False positive treatment decision	1.	0.172	0.002	0.290	0.038	0.172	0.002	0.286	0.035
	2.	0.132	0.001	0.219	0.033	0.132	0.001	0.217	0.028
	3.	0.082	0.001	0.127	0.021	0.082	0.001	0.126	0.018
	4.	0.318	0.001	0.451	0.023	0.318	0.001	0.444	0.018
	5.	0.214	0.017	0.235	0.035	0.214	0.017	0.234	0.031
False negative treatment decision	1.	0.346	0.208	0.217	0.066	0.347	0.208	0.220	0.058
	2.	0.293	0.191	0.188	0.061	0.292	0.191	0.189	0.054
	3.	0.240	0.158	0.167	0.059	0.240	0.158	0.167	0.052
	4.	0.384	0.152	0.262	0.052	0.385	0.152	0.266	0.045
	5.	0.286	0.072	0.268	0.074	0.286	0.072	0.268	0.062

Legend Table 1

Abbreviations: SIT: simply ignore treatment (in SIT_1 predictor X_2 is considered unobserved; in SIT_2 both predictors X_1 and X_2 are considered observed); AUT: analysis of untreated individuals (in AUT_1 predictor X_2 is considered unobserved; in AUT_2 both predictors X_1 and X_2 are considered observed); IPW: inverse probability weighting (in IPW_1 predictor X_2 is considered unobserved; in IPW_2 both predictors X_1 and X_2 are considered observed); MT: model includes treatment (in MT_1 predictor X_2 is considered unobserved; in MT_2 both predictors X_1 and X_2 are considered observed).

Table 2. Performance of methods to develop prognostic models using randomised data.

		Method								
	Scenario	**SIT_1**	**SIT_2**	**AUT_1**	**AUT_2**	**IPW_1**	**IPW_2**	**MT_1**	**MT_2**	**Reference**
Brier score	1.	0.082	0.074	0.099	0.087	0.082	0.074	0.081	0.072	0.073
	2.	0.081	0.072	0.096	0.085	0.081	0.072	0.080	0.071	0.072
	3.	0.070	0.061	0.080	0.069	0.070	0.061	0.069	0.060	0.061
	4.	0.088	0.062	0.102	0.070	0.088	0.062	0.087	0.061	0.061
	5.	0.085	0.076	0.088	0.078	0.085	0.076	0.085	0.076	0.076
c-statistic	1.	0.728	0.817	0.726	0.818	0.728	0.817	0.742	0.826	0.824
	2.	0.780	0.843	0.778	0.842	0.780	0.843	0.790	0.851	0.849
	3.	0.860	0.899	0.858	0.898	0.860	0.899	0.866	0.903	0.902
	4.	0.685	0.899	0.682	0.898	0.685	0.899	0.698	0.903	0.902
	5.	0.729	0.818	0.728	0.819	0.729	0.818	0.731	0.820	0.818
O:E ratio	1.	1.000	1.000	0.801	0.801	1.000	1.000	1.000	1.000	1.000
	2.	1.000	1.000	0.807	0.807	1.000	1.000	1.000	1.000	1.000
	3.	1.000	1.000	0.838	0.838	1.000	1.000	1.000	1.000	1.000
	4.	1.000	1.000	0.835	0.835	1.000	1.000	1.000	1.000	1.000
	5.	1.000	1.000	0.970	0.970	1.000	1.000	1.000	1.000	1.000
Standard error	1.	0.121	0.132	0.158	0.175	0.120	0.130	0.122	0.133	
	2.	0.130	0.138	0.172	0.184	0.128	0.136	0.131	0.140	
	3.	0.164	0.187	0.223	0.257	0.157	0.179	0.167	0.191	
	4.	0.113	0.145	0.149	0.196	0.112	0.141	0.114	0.147	
	5.	0.119	0.130	0.167	0.184	0.119	0.129	0.120	0.131	

Legend Table 2

Abbreviations: SIT: simply ignore treatment (in SIT_1 predictor X_2 is considered unobserved; in SIT_2 both predictors X_1 and X_2 are considered observed); AUT: analysis of untreated individuals (in AUT_1 predictor X_2 is considered unobserved; in AUT_2 both predictors X_1 and X_2 are considered observed); IPW: inverse probability weighting (in IPW_1 predictor X_2 is considered unobserved; in IPW_2 both predictors X_1 and X_2 are considered observed); MT: model includes treatment (in MT_1 predictor X_2 is considered unobserved; in MT_2 both predictors X_1 and X_2 are considered observed). The O:E ratio is the ratio of the mean observed risk of the outcome and the mean predicted risk of the outcome. Ratios are average over 1000 simulations. Standard errors

are the average standard error of the association between the predictor X_1 and the outcome, averaged over 1000 simulations.

Note that the method in which treatment as well as the predictors X_1 and X_2 are explicitly modelled performs even better than the reference (which is based on the data generating model), because chance processes are also accounted for in the analytical method.

Table 3. Impact on treatment decisions of methods to develop prognostic models using data from an observational study.

		Method							
	Scenario	**SIT_1**	**SIT_2**	**AUT_1**	**AUT_2**	**IPW_1**	**IPW_2**	**MT_1**	**MT_2**
False positive treatment decision	6.	0.172	0.001	0.119	0.029	0.168	0.004	0.143	0.037
	7.	0.111	0.001	0.076	0.023	0.108	0.004	0.100	0.033
	8.	0.074	0.001	0.060	0.017	0.072	0.003	0.064	0.025
	9.	0.328	0.001	0.063	0.016	0.318	0.003	0.095	0.024
	10.	0.201	0.013	0.053	0.023	0.186	0.020	0.070	0.032
False negative treatment decision	6.	0.351	0.249	0.500	0.124	0.359	0.204	0.409	0.063
	7.	0.331	0.251	0.484	0.139	0.340	0.200	0.366	0.065
	8.	0.261	0.197	0.295	0.099	0.267	0.155	0.294	0.050
	9.	0.377	0.198	0.843	0.103	0.394	0.153	0.731	0.052
	10.	0.302	0.088	0.663	0.172	0.327	0.101	0.577	0.084

Legend Table 3

Abbreviations: SIT: simply ignore treatment (in SIT_1 predictor X_2 is considered unobserved; in SIT_2 both predictors X_1 and X_2 are considered observed); AUT: analysis of untreated individuals (in AUT_1 predictor X_2 is considered unobserved; in AUT_2 both predictors X_1 and X_2 are considered observed); IPW: inverse probability weighting (in IPW_1 predictor X_2 is considered unobserved; in IPW_2 both predictors X_1 and X_2 are considered observed); MT: model includes treatment (in MT_1 predictor X_2 is considered unobserved; in MT_2 both predictors X_1 and X_2 are considered observed).

Table 4. Performance of methods to develop prognostic models using observational data.

		Method								
	Scenario	**SIT_1**	**SIT_2**	**AUT_1**	**AUT_2**	**IPW_1**	**IPW_2**	**MT_1**	**MT_2**	**Reference**
Brier score	6.	0.085	0.078	0.063	0.058	0.085	0.079	0.085	0.077	0.078
	7.	0.079	0.072	0.051	0.048	0.079	0.073	0.078	0.071	0.072
	8.	0.072	0.065	0.045	0.041	0.072	0.066	0.072	0.064	0.064
	9.	0.089	0.065	0.052	0.041	0.090	0.066	0.088	0.064	0.064
	10.	0.083	0.074	0.046	0.043	0.083	0.075	0.082	0.074	0.075
c-statistic	6.	0.708	0.789	0.693	0.792	0.708	0.789	0.711	0.799	0.796
	7.	0.762	0.819	0.739	0.814	0.762	0.818	0.763	0.827	0.825
	8.	0.853	0.886	0.847	0.886	0.853	0.885	0.854	0.890	0.889
	9.	0.671	0.886	0.636	0.886	0.671	0.885	0.696	0.890	0.889
	10.	0.727	0.816	0.695	0.796	0.727	0.814	0.742	0.817	0.815
O:E ratio	6.	1.000	1.000	1.435	1.435	1.014	0.899	1.000	1.000	1.000
	7.	1.000	1.000	1.710	1.710	1.017	0.892	1.000	1.000	1.000
	8.	1.000	1.000	1.871	1.871	1.018	0.908	1.000	1.000	1.000
	9.	1.000	1.000	1.888	1.888	1.044	0.906	1.000	1.000	1.000
	10.	1.000	1.000	2.030	2.030	1.058	0.991	1.000	1.000	1.000
Standard error	6.	0.117	0.125	0.203	0.225	0.117	0.127	0.124	0.137	
	7.	0.128	0.135	0.244	0.258	0.129	0.139	0.139	0.147	
	8.	0.159	0.176	0.289	0.333	0.154	0.179	0.165	0.192	
	9.	0.111	0.138	0.218	0.275	0.111	0.140	0.118	0.152	
	10.	0.120	0.131	0.238	0.260	0.120	0.133	0.126	0.140	

Legend Table 4

Abbreviations: SIT: simply ignore treatment (in SIT_1 predictor X_2 is considered unobserved; in SIT_2 both predictors X_1 and X_2 are considered observed); AUT: analysis of untreated individuals (in AUT_1 predictor X_2 is considered unobserved; in AUT_2 both predictors X_1 and X_2 are considered observed); IPW: inverse probability weighting (in IPW_1 predictor X_2 is considered unobserved; in IPW_2 both predictors X_1 and X_2 are considered observed); MT: model includes treatment (in MT_1 predictor X_2 is considered unobserved; in MT_2 both predictors X_1 and X_2 are considered observed). The O:E ratio is the ratio of the mean observed risk of the outcome and the

mean predicted risk of the outcome. Ratios are average over 1000 simulations. Standard errors are the average standard error of the association between the predictor X_1 and the outcome, averaged over 1000 simulations.

Note that the method in which treatment as well as the predictors X_1 and X_2 are explicitly modelled performs even better than the reference (which is based on the data generating model), because chance processes are also accounted for in the analytical method.