Cite as:
Wei Xin Chan, Limsoon Wong. Accounting for treatment during the development or validation of prediction models. Journal of Bioinformatics and Computational Biology, 20(6):2271001, December 2022.

Journal of Bioinformatics and Computational Biology
© Imperial College Press

ACCOUNTING FOR TREATMENT DURING THE DEVELOPMENT OR VALIDATION OF PREDICTION MODELS

WEI XIN CHAN
School of Computing, National University of Singapore
13 Computing Drive, Singapore 117417
weixin@u.nus.edu

LIMSOON WONG
School of Computing, National University of Singapore
13 Computing Drive, Singapore 117417
wongls@comp.nus.edu.sg

Received (Day Month Year)
Revised (Day Month Year)
Accepted (Day Month Year)

Clinical prediction models are widely used to predict adverse outcomes in patients, and are often employed to guide clinical decision making. Clinical data typically consist of patients who received different treatments. Many prediction modelling studies fail to account for differences in patient treatment appropriately, which results in the development of prediction models that show poor accuracy and generalizability. In this manuscript, we list the most common methods used to handle patient treatments and discuss certain caveats associated with each method. We believe that proper handling of differences in patient treatment is crucial for the development of accurate and generalizable models. As different treatment strategies are employed for different diseases, the best approach to properly handle differences in patient treatment is specific to each individual situation. We use the Ma-Spore acute lymphoblastic leukemia data set as a case study to demonstrate the complexities associated with differences in patient treatment, and offer suggestions on incorporating treatment information during evaluation of prediction models. In clinical data, patients are typically treated on a case by case basis, with unique cases occurring more frequently than expected. Hence, there are many subtleties to consider during the analysis and evaluation of clinical prediction models.

Keywords: Clinical prediction model; Treatment; Risk prediction.

1. Introduction

Clinical prediction models are used for estimating the absolute risk of clinically important outcomes in patients. Prediction models can be employed to support clinical decision making, with estimated risks of clinically poor outcomes being used to guide treatment initiation in individuals. Examples of real-world applications include the Framingham risk score in cardiovascular disease (CVD) and Apgar score for the prognosis of newborns.[1,2]

Clinical data used in the development and validation of prediction models often consist of patients that underwent different treatment interventions. If treatment effects are not handled properly during the development of prediction models, the accuracy and generalizability of the models would be negatively affected.[3,4] This can be illustrated by a famous example presented by Caruana *et al.*,[5] where a machine learning model predicts that patients with a history of asthma have a lower risk of dying from pneumonia. Failure to account for the fact that asthmatic patients were admitted to the intensive care unit (ICU) and received treatment that lowered their mortality risk resulted in the model learning that asthma is associated with lower risk. Deploying the model in a population where asthmatics are not admitted to the ICU would prove disastrous, as these patients would be predicted as low risk when they in fact require special treatment.

Despite the need to properly account for patient treatments, few prediction modelling studies report patient treatments in sufficient detail. Surveys on prediction modelling studies in CVD highlighted that a significant portion of studies do not mention treatment use, and that most studies fail to mention treatment initiated after the time of prediction.[6,7]

Treatment strategies developed for diseases are highly specific and usually involve a combination of different treatments that may include both medical and non-medical interventions. In addition to treatment strategies differing between diseases, treatment strategies for the same disease evolve over time as well. In general, treatments differ in whether they are initiated prior to prediction (baseline treatment) or after the time of prediction (treatment drop-in). In complex cases, intensity levels of treatment are altered throughout the course of treatment.

As the treatment for each disease varies greatly, there is no one-size-fits-all approach in handling treatment effects during the development or validation of prediction models; the appropriate approach often depends on each individual situation. As a result, there are many subtleties to consider when accounting for treatment effects during the development or validation of prediction models in different settings.

In this manuscript, we introduce common methods that are used to handle treatments during the development of prediction models, and discuss issues associated with these methods. When treatment strategies are the same in the development and deployment cohorts, ignoring treatment may be a viable approach. However, when treatment strategies are different between cohorts, patient treatments should be accounted for. We illustrate the benefits of incorporating treatment information during the validation of prediction models through a case study of the Ma-Spore ALL data set. We suggest a scoring scheme that incorporates patient treatment information to assess whether treatment intensity predictions are correct, and present a way to visualize results from the scoring scheme that facilitates analysis of model predictions. We encourage examining individual cases in detail during the analysis and evaluation of clinical prediction models in order to properly account for the heterogeneities in clinical data.

2. Handling treatments: Common methods

Clinical prediction models are frequently employed to facilitate treatment initiation decisions, which requires estimating the risk of adverse outcomes in the absence of treatment.[3,8] As clinical data often consists of a mix of treated and untreated patients, it is essential to properly account for patient treatments when developing a model to support treatment initiation decisions. The most common methods used to handle treatments include ignoring treatment, restricting analysis to untreated patients, including treatment together with the adverse outcome as a composite outcome, modelling treatment, and hypothetical prediction.[3,4,9] However, modelling treatments will only work if knowledge of treatment use is known at the time of prediction (i.e. will not work for treatment drop-ins).

2.1. *Ignoring treatment*

Many prediction modelling studies choose to ignore the fact that treated patients are present in the data being using to develop the models.[7] If the treatment was effective in reducing risk of adverse outcome in patients, treated patients would have a lower chance of suffering from the adverse outcome. Not accounting for treated patients in the data used to train the model would result in a model that under-estimates risk when deployed on untreated patients. The effectiveness of treatment and proportion of treated patients in the data would determine the extent to which the developed models would under-estimate risk.[3]

Ignoring treatment also leads to a “treatment paradox”.[10] Suppose a model predicts patients as high risk due to a specific predictor, and treatment is given to patients predicted as high risk. If the treatment is effective and lowers the patients’ probability of having an adverse outcome, the predictor-outcome association would be attenuated. If the data is used to develop a model without accounting for treatment, the model would learn that the predictor is now associated with low risk instead, contradicting the original prediction.

2.2. *Restricting analysis*

Another commonly used method is to restrict analysis to untreated patients only. This method is susceptible to selection bias, as often patients deemed to be low risk are excluded from treatments.[4] Developing models using data consisting only of untreated patients with a lower probability of the outcome might lead to an under-estimation of risk when the model is validated on the full range of patients with different levels of risk.[3]

2.3. *Composite outcome*

A method that can be used to handle treatment drop-ins is to combine the treatment together with the original outcome as a composite outcome. Essentially, we are estimating the risk of occurrence of the outcome or administration of treatment in a patient. One of the cases better suited for this method is when treatment most likely prevented the occurrence of the outcome (e.g. patients would most probably suffer from myocardial infarction if they did not undergo surgery).

2.4. *Modelling treatment*

Baseline treatments received by patients can be accounted for by explicitly including treatment as a predictor in the prediction model. However, as information about treatment drop-ins are unavailable at the time of prediction, this method does not apply to them.

2.5. *Hypothetical prediction*

The hypothetical prediction method accounts for treatments by performing a counterfactual prediction - predicting risk in a hypothetical world where treated patients are not given treatment instead. Causal inference methods, such as marginal structural models and g-formulas, are used in the estimation of hypothetical untreated risk.[4,11] However, the validity of these estimates are conditioned on three key assumptions: exchangeability of treated and untreated patients given measured confounders, positivity (having a non-zero number of treated and untreated patients for all covariate patterns) and consistency, where a patient's hypothetical risk is equal to her observed risk in the real world.[9,11] These assumptions are often unverifiable empirically in observational studies, where treatment is not randomized. In addition, it is challenging to account for differences between interventional and conditional distributions, to identify all potential confounders and colliders, and to avoid model misspecification.[12]

3. Estimands

Accounting for treatment using different methods during the development of prediction models results in different models with subtly different estimands.[9] Ignoring treatment effectively results in a model that estimates a patient's risk of the adverse outcome, given that patients are treated according to the current treatment strategy. Deployment of the model in a population with the same treatment strategy would yield legitimate predictions, as similar associations between predictors, treatment and outcome exist. If the deployed model predicted a good outcome for a patient, the patient should continue on the same treatment strategy; however, if the model predicted a poor outcome, the patient would do better not to follow the same treatment strategy. In cases where the deployment population does not have the same treatment strategy, it is best to account for treatment in order to avoid inaccurate predictions.

4. Case study: Ma-Spore acute lymphoblastic leukemia data set

Different diseases require different treatment strategies, each with their own set of issues and complexities. Treatment strategies should be reported in detail so that necessary actions can be taken to account for them during both the development and validation of prediction models. We use data from the Ma-Spore acute lymphoblastic leukemia (ALL) 2003 and 2010 studies[13,14] as a case study to demonstrate some of the issues and complexities related to treatment, and offer suggestions on how to deal with them in a nuanced manner.

Patients in the Ma-Spore ALL data set underwent risk-adapted chemotherapy and were treated at different intensities at different times throughout the course of treatment. In addition, eligible patients in the high risk subtype BCR-ABL underwent a bone marrow transplant (BMT). As patient treatment information could not be found in the public Ma-Spore ALL data set, we requested for available treatment information from the authors and obtained treatment information regarding the final treatment intensity patients were treated at and whether they underwent BMT.[14] However, details regarding a patient's treatment intensities at other points in time during the therapy were not available. We inferred each patient's treatment intensities throughout the course of treatment by using information from patient metadata and the set of decision rules in the risk-adapted treatment protocol. We concluded that the varying treatment intensities that patients received at different times throughout the course of treatment could be reduced down to four possible treatment routes. Patients will undergo one out of four possible treatment routes, which we denote as standard risk (SR), intermediate risk (IR), high risk one (HR1) and high risk two (HR2).

All patients start off on IR treatment intensity, and may be re-assigned to other treatment intensity levels at only two specific time points, Day 8 and post-induction (i.e. after the induction phase of chemotherapy). Patients who are escalated to HR treatment intensity will remain at HR treatment intensity for the remaining course of treatment. In the HR1 treatment route, patients are elevated to HR treatment intensity at Day 8. Patients who are assigned to the HR1 treatment route are primarily from high risk subtypes, namely the BCR-ABL, MLL and hypodiploid subtypes. In the HR2 treatment route, patients are elevated to HR treatment intensity post-induction. Patients that undergo the HR2 treatment route mostly have poor treatment response which is evidenced by their high Day 33 or Week 12 minimal residual disease (MRD) values. Patients that meet a set of stringent criteria, including having low Day 33 and Week 12 MRD values, will be assigned to the SR treatment route where patients will be treated on IR treatment intensity before being de-escalated to SR treatment intensity post-induction. In the IR treatment route, patients are treated on IR treatment intensity throughout the course of chemotherapy.

Table 1: Scoring scheme

Outcome	Treatment	Prediction	Score
Remission	SR	SR	1
Remission	SR	IR	0
Remission	SR	HR	0
Remission	IR	SR	1
Remission	IR	IR	1
Remission	IR	HR	0
Remission	HR	SR	0
Remission	HR	IR	1
Remission	HR	HR	1
Relapse	SR	SR	0
Relapse	SR	IR	1
Relapse	SR	HR	1
Relapse	IR	SR	0
Relapse	IR	IR	0
Relapse	IR	HR	1
Relapse	HR	SR	0
Relapse	HR	IR	0
Relapse	HR	HR	1

5. Some nuances

The treatment strategy for ALL is more complicated than the simple case where patients were either treated or untreated. Not only are ALL patients treated at different intensities at different points in time throughout chemotherapy, eligible high risk patients also receive BMT. In this section, we propose a scoring scheme that incorporates the use of patient treatment information. We demonstrate why it is beneficial to utilize treatment information during the validation of prediction models.

Consider a prediction model that predicts the risk of relapse given that patients are treated according to the risk-adapted treatment protocol. The predicted risk of relapse is used to recommend the treatment intensity (i.e. treatment route) that patients should undergo (SR, IR or HR). For simplicity, we consider both HR1 and HR2 treatment routes to be equivalent and term them both as high risk (HR) treatment routes. The correct treatment intensity has to achieve a balance between maximizing a patient's probability of achieving continuous complete remission (CCR) and reducing the cytotoxic effects of chemotherapy. As achieving CCR is significantly more important than eliminating cytotoxic side-effects, the correct treatment intensity is essentially the lowest treatment intensity level at which a patient still achieves CCR.

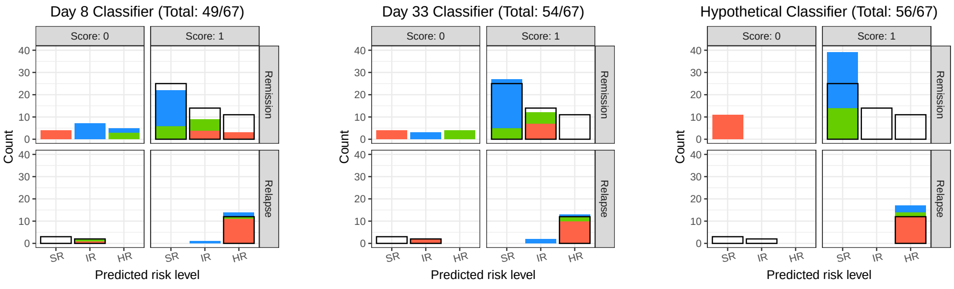
Determining the correct treatment intensity patients should receive requires knowledge of individual treatment outcome under counterfactual treatments (i.e. the effect SR/IR/HR treatment intensities have on patient's risk of relapse and side-effects). However, there is no ground truth of counterfactual events as they do not occur in the real world. For example, if a patient was predicted to require

SR treatment, but was in real life put on IR treatment, we would not be able to know how the patient would respond to SR treatment. We overcome the absence of ground truth by logically deducing whether a treatment recommendation is correct or not based on each patient's treatment information and treatment outcome. We logically deduce whether treatment intensity recommendations are correct or not for all combinations of a patient's treatment outcome (relapse or CCR), treatment route (SR, IR or HR) and the predicted treatment intensity recommendation (SR, IR or HR). We present a scoring scheme encompassing all cases in Table 1. All conceivably correct recommendations are awarded a score of one, while incorrect recommendations are given a score of zero.

We elaborate on a plausible framework used to logically deduce whether treatment intensity recommendations are correct or incorrect. For patients who relapsed, treatment intensity recommendations that are above the treatment intensities patients were treated at are deemed to be correct, while recommendations that are below or equal to the treatment intensities patients received are incorrect. This stems from the reasoning that patients who relapsed were treated at an insufficient treatment intensity level, and would benefit from a higher treatment intensity level. In the case that the patient relapses even when treated on the highest intensity level (HR), we take HR to be the correct treatment intensity. For patients that achieved CCR, recommendations that are equal to or less intense (by one risk level) than the actual treatment intensity received are deemed to be correct, while a recommendation at a higher level than the treatment level received would be incorrect. This penalizes cases of over-treatment, while awarding treatment de-escalation that will help reduce toxic side-effects of chemotherapy.

However, one subtlety regarding the above framework is that it relies on the implicit assumption that patients are able to complete the treatment intensity they were prescribed. In reality, some patients may discontinue treatment for various reasons, such as being unable to tolerate the side effects of the prescribed treatment intensity. Consequently, these patients suffer from relapse or treatment failure. For example, patients that receive HR treatment but are unable to tolerate the high dosage typically discontinue treatment, thereby ending up suffering from relapse or treatment failure. There has been sporadic evidence that such patients could potentially benefit from a lower dosage.[15] These patients are different from patients that are able to complete treatment at the prescribed intensity but still relapse. In these cases (patients who relapsed previously but who might not be able to tolerate high intensity treatments), it is not straightforward to decide whether these recommendations of increased intensity are correct or incorrect.

We demonstrate in Figure 1 a way to visualize results from our scoring scheme that facilitates analysis of the predictions made by a model. In particular, our approach allows users to easily analyze whether incorrect treatment recommendations are primarily made up of over-treatments or under-treatments. We demonstrate the use of our scoring scheme in evaluating the performance of three prediction models on the Ma-Spore ALL data set.



(a) **(b)** **(c)**

Fig. 1: Frequency of scores awarded by the proposed scoring scheme when used to evaluate treatment intensity predictions from three separate prediction models. Correct and incorrect recommendations are awarded a score of one and zero, respectively. Scores are aggregated for each prediction model, and the total score for each model expressed out of the maximum score attainable is presented in each chart title. Stacked bars represent the cumulative frequency of scores awarded to recommendations. The bars are coloured according to the actual treatment intensity received, with blue, green and red denoting standard risk (SR), intermediate risk (IR) and high risk (HR), respectively. Bar outlines represent the frequency of scores if the actual treatment intensities that patients received were used as treatment intensity recommendations. Each subfigure is divided into four panels, with patients in each panel having the same score and treatment outcome. Performance of prediction models that are variations of a subtype-specific prediction model, based on different sets of overlapping features that are available at various time-points throughout the course of treatment, namely a) Day 8 and b) Day 33. c) Performance of a hypothetical prediction model that recommends HR treatment for patients who relapse, and SR treatment for patients that achieve continuous complete remission.

The two prediction models in Figures 1a and 1b are different variations of a subtype-specific prediction model we developed using the Ma-Spore ALL data set. These models estimate a patient's probability of achieving CCR based on different sets of overlapping features that are available at different time-points in the patient's treatment plan, namely Day 8 and Day 33. The estimates are subsequently stratified according to a set of thresholds into three different treatment intensities (SR, IR and HR). These treatment intensities are the predictions or recommendations given by the models, and are not the actual treatment intensities that the patients were treated on. These two sets of treatment intensity recommendations are evaluated using our scoring scheme, and the results are visualized in Figures 1a and 1b.

Both sets of treatment intensity recommendations performed fairly well for patients, with an average of approximately 77% of patients receiving correct treatment recommendations. Correct recommendations can either be patients who achieved CCR being recommended the same treatment intensities that they underwent, or patients who relapsed being recommended higher treatment intensities than the intensity they were treated on. Majority of incorrect recommendations made by the Day 8 and Day 33 prediction models were recommendations that would result in over-treatment (cases where patients achieved CCR after undergoing actual treatment intensities that are lower than the recommended treatment). Although these

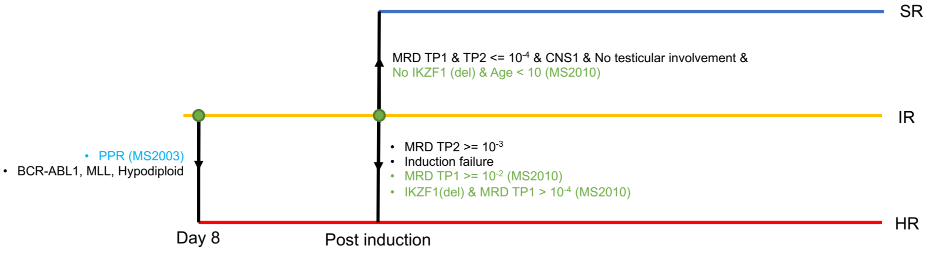

SR
MRD TP1 & TP2 <= 10^-4 & CNS1 & No testicular involvement &
No IKZF1 (del) & Age < 10 (MS2010)
IR
• PPR (MS2003)
• BCR-ABL1, MLL, Hypodiploid
• MRD TP2 >= 10^-3
• Induction failure
• MRD TP1 >= 10^-2 (MS2010)
• IKZF1(del) & MRD TP1 > 10^-4 (MS2010)
HR
Day 8
Post induction

Fig. 2: Risk-adapted treatment plan employed in Ma-Spore acute lymphoblastic leukemia 2003 and 2010 studies (MS2003 and MS2010). All patients start off on the intermediate risk (IR) treatment arm, and treatment intensity may be altered at two decision time-points, namely Day 8 of chemotherapy and post induction. The criteria for escalation to high risk (HR) or de-escalation to standard risk (SR) is stated next to the arrowheads. Bullet points represent an "OR" relationship between criteria. Criteria specific to MS2003, MS2010 are highlighted in light blue and light green text, respectively. PPR: Poor prednisolone response, MRD: Minimal residual disease, TP1: Time-point one, TP2: Time-point two.

recommendations are incorrect, they have a less detrimental impact than incorrect recommendations that result in under-treatment. In addition, most of the incorrect over-treatment recommendations only exceeded by a single risk level (e.g. patients who underwent IR treatment and achieved CCR were recommended HR treatment).

6. Benefits of incorporating treatment information

To illustrate the benefits of incorporating treatment information during validation of prediction models, we consider a classifier C that is able to predict treatment outcome labels perfectly. Originally, patients in the Ma-Spore ALL data set are treated using a risk-adapted treatment strategy, where decisions are made whether to alter patient treatment intensity at two time-points, based on individual patient features such as ALL subtype and MRD (see Figure 2). We look to adopt a new treatment strategy, where instead of deciding whether to alter patient intensity according to the original strategy, we use predictions from classifier C to decide whether to escalate or de-escalate patient treatment intensity. Patients that are predicted to relapse would have their treatment escalated to HR, while patients predicted to achieve CCR would be de-escalated to SR treatment.

Evaluation of the new treatment strategy using traditional metrics such as accuracy may give misleading results. If patients who relapsed are assumed to require HR treatment and patients who achieved CCR are assumed to be suitable for de-escalation to SR treatment, classifier C would be deemed to have an accuracy of 100%. However, some patients who achieved CCR were treated on HR treatment intensity, and treatment recommendations of SR may not be sufficient for these patients, who may end up relapsing instead. Evaluation using our scoring scheme, which incorporates treatment information, would identify these recommendations as incorrect (see Figure 1c).

It is tricky to evaluate the above group of patients, as there are no ground truths available for these counterfactual claims. Patients that were treated at HR intensity and achieved CCR may either require HR treatment intensity to achieve CCR, or on the other hand may benefit from de-escalated treatment. A way to infer which of the two possibilities patients belong to is by examining a feature that is indicative of treatment response, such as the Day 33 MRD. Patients with good treatment response will be more likely to benefit from de-escalation of treatment.

7. Closing remarks

We believe that treatment information of patients should be reported in detail for clinical data used in the development and validation of clinical prediction models. Firstly, this allows for proper handling of patient treatments during development of prediction models, hence ensuring that these models are able to generalize to other cohorts where treatment strategies are not the same. In addition, incorporation of treatment information during validation of the prediction model allows for a more detailed evaluation of the prediction model, and helps in the handling of heterogeneities in clinical data.

We suggest a simple approach to handle complex risk-adapted treatment plans, for use in situations where the treatment plans remain the same in both the development and deployment cohort. Typically, these treatment plans have multiple decision time points where patient treatment intensities may be altered. This approach involves establishing that the aim of the model is to predict whether patients should continue to be treated on the original treatment plan, or be excluded from it. We can adapt a prediction model that predicts treatment outcome in ALL patients to recommend whether patients should or should not continue with the original treatment plan that they are set on. If patients are predicted to achieve CCR, they should continue with the original treatment plan, and if patients are predicted to relapse, they should be excluded from the original treatment plan. For the latter group of patients, if they are able to tolerate higher treatment intensity, they should be recommended to receive escalated treatment. If not, they may benefit from de-escalated treatment.[15] In the worst case where they do not respond to treatment, taking patients off treatment would eliminate painful side effects of treatment and help in saving costs.

Patients are typically treated on a case by case basis, with clinicians prescribing treatment based on a patient's individual condition. Hence, the analysis and evaluation of clinical prediction models benefits greatly from the examination of details of individual patients and explicit consideration of individual cases in the data set. As illustrated earlier, a scoring scheme which incorporates patient treatment information during validation of clinical prediction models allows for a more nuanced evaluation of correct and incorrect treatment intensity recommendations by considering different specific cases.

8. Acknowledgements

We would like to thank Allen Yeoh, Zhenhua Li and Kean Hui Chiew for supplying us with data from the Ma-Spore ALL 2003 and 2010 studies, and for their help in clarifying details regarding patient treatment strategies.

References

1. V Apgar. A proposal for a new method of evaluation of the newborn. *Classic Papers in Critical Care*, 32(449):97, 1952.
2. Peter WF Wilson, Ralph B D'Agostino, Daniel Levy, Albert M Belanger, Halit Silbershatz, and William B Kannel. Prediction of coronary heart disease using risk factor categories. *Circulation*, 97(18):1837–1847, 1998.
3. Rolf HH Groenwold, Karel GM Moons, Romin Pajouheshnia, Doug G Altman, Gary S Collins, Thomas PA Debray, Johannes B Reitsma, Richard D Riley, and Linda M Peelen. Explicit inclusion of treatment in prognostic modeling was recommended in observational and randomized settings. *Journal of Clinical Epidemiology*, 78:90–100, 2016.
4. Matthew Sperrin, Glen P Martin, Alexander Pate, Tjeerd Van Staa, Niels Peek, and Iain Buchan. Using marginal structural models to adjust for treatment drop-in when developing clinical prediction models. *Statistics in Medicine*, 37(28):4142–4154, 2018.
5. Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1721–1730, 2015.
6. SM Liew, J Doust, and P Glasziou. Cardiovascular risk scores do not account for the effect of treatment: a review. *Heart*, 97(9):689–697, 2011.
7. Romin Pajouheshnia, Johanna AAG Damen, Rolf HH Groenwold, Karel GM Moons, and Linda M Peelen. Treatment use in prognostic model research: a systematic review of cardiovascular prognostic studies. *Diagnostic and Prognostic Research*, 1(1):1–10, 2017.
8. Harry Hemingway, Peter Croft, Pablo Perel, Jill A Hayden, Keith Abrams, Adam Timmis, Andrew Briggs, Ruzan Udumyan, Karel GM Moons, Ewout W Steyerberg, et al. Prognosis research strategy (PROGRESS) 1: a framework for researching clinical outcomes. *BMJ*, 346, 2013.
9. Nan van Geloven, Sonja A Swanson, Chava L Ramspek, Kim Luijken, Merel van Diepen, Tim P Morris, Rolf HH Groenwold, Hans C van Houwelingen, Hein Putter, and Saskia le Cessie. Prediction meets causal inference: the role of treatment in clinical prediction models. *European Journal of Epidemiology*, 35(7):619–630, 2020.
10. N Peek, M Sperrin, M Mamas, T Van Staa, and I Buchan. Hari Seldon, QRISK3, and the prediction paradox. *BMJ*, 357:j2099, 2017.
11. Barbra A Dickerman, Issa J Dahabreh, Krystal V Cantos, Roger W Logan, Sara Lodi, Christopher T Rentsch, Amy C Justice, and Miguel A Hernán. Predicting counterfactual risks under hypothetical treatment strategies: an application to HIV. *European Journal of Epidemiology*, pages 1–10, 2022.
12. Mattia Prosperi, Yi Guo, Matt Sperrin, James S Koopman, Jae S Min, Xing He, Shannan Rich, Mo Wang, Iain E Buchan, and Jiang Bian. Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nature Machine Intelligence*, 2(7):369–375, 2020.
13. Allen Eng Juh Yeoh, Hany Ariffin, Elaine Li Leng Chai, Cecilia Sze Nga Kwok,

Yiong Huak Chan, Kuperan Ponnudurai, Dario Campana, Poh Lin Tan, Mei Yoke Chan, Shirley Kow Yin Kham, et al. Minimal residual disease-guided treatment deintensification for children with Acute Lymphoblastic Leukemia: Results from the Malaysia-Singapore Acute Lymphoblastic Leukemia 2003 Study. *Journal of Clinical Oncology*, 30(19), 2012.

14. Allen Eng Juh Yeoh, Yi Lu, Winnie Hui Ni Chin, Edwynn Kean Hui Chiew, Evelyn Huizi Lim, Zhenhua Li, Shirley Kow Yin Kham, Yiong Huak Chan, Wan Ariffin Abdullah, Hai Peng Lin, et al. Intensifying treatment of childhood B-lymphoblastic leukemia with IKZF1 deletion reduces relapse and improves overall survival: results of Malaysia-Singapore ALL 2010 study. *Journal of Clinical Oncology*, 36(26):2726–2735, 2018.
15. Benjamin Kye Jyn Tan, Chong Boon Teo, Xavier Tadeo, Siyu Peng, Hazel Pei Lin Soh, Sherry De Xuan Du, Vilianty Wen Ya Luo, Aishwarya Bandla, Raghav Sundar, Dean Ho, et al. Personalised, rational, efficacy-driven cancer drug dosing via an artificial intelligence SystEm (PRECISE): a protocol for the PRECISE CURATE.AI pilot clinical trial. *Frontiers in Digital Health*, 3:635524, 2021.

Wei Xin Chan is a Ph.D. student in the School of Computing at the National University of Singapore. He received his B.Sc. in Biochemistry from University College London in 2014. He is currently working on predictive modelling in acute lymphoblastic leukemia. His research interests include predictive modelling, batch effects and gene expression analysis.

Limsoon Wong is a Professor in the School of Computing at the National University of Singapore. He was also a professor (now honorary) of pathology in the Yong Loo Lin School of Medicine at NUS. He currently works mostly on knowledge discovery technologies and their application to biomedicine. He is a Fellow of the ACM, named in 2013 for his contributions to database theory and computational biology.