Cite this as:
Wei Xin Chan, Limsoon Wong. Obstacles to effective model deployment in healthcare.
Journal of Bioinformatics and Computational Biology, 21(2):2371001, April 2023.

Journal of Bioinformatics and Computational Biology
© Imperial College Press

OBSTACLES TO EFFECTIVE MODEL DEPLOYMENT IN HEALTHCARE

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)

Despite an exponential increase in publications on clinical prediction models over recent years, the number of models deployed in clinical practice remains fairly limited. In this paper, we identify common obstacles that impede effective deployment of prediction models in healthcare, and investigate their underlying causes. We observe a key underlying cause behind most obstacles - the improper development and evaluation of prediction models. Inherent heterogeneities in clinical data complicate the development and evaluation of clinical prediction models. Many of these heterogeneities in clinical data are unreported because they are deemed to be irrelevant, or due to privacy concerns. We provide real-life examples where failure to handle heterogeneities in clinical data, or sources of biases, led to the development of erroneous models.

The purpose of this paper is to familiarize modelling practitioners with common sources of biases, and heterogeneities in clinical data, both of which have to be dealt with to ensure proper development and evaluation of clinical prediction models. Proper model development and evaluation, together with complete and thorough reporting, are important pre-requisites for a prediction model to be effectively deployed in healthcare.

Keywords: Clinical prediction models. Deployment. Machine learning.

1. Introduction

Advances in high-throughput technologies and the rapid digitization of healthcare have led to an explosion in the amount of medical data freely available for research. The increased availability of medical data, coupled with recent advancements in machine learning, have led to a resurgence of interest in the use of predictive modelling in healthcare. This has resulted in an exponential increase in the number of publications on prediction models in healthcare.[1,2,3] However, despite the copious publications on prediction models in healthcare, the number of models deployed in clinical practice still remains fairly limited.[4,5]

In this paper, we identify common obstacles facing effective model deployment in healthcare, and investigate their underlying causes. One of the most common obstacles is the lack of reproducibility and replicability.[6] Failure to replicate the performance of a prediction model on independent data sets casts doubts on the model's ability to perform effectively when deployed. Another common obstacle facing effective model deployment is the high risk of unfairness displayed by models. We provide examples of unfairness exhibited by prediction models deployed in real-life in this paper.

We observe that a key underlying cause behind the various obstacles is the improper development and evaluation of prediction models. Healthcare/clinical prediction models are especially susceptible to improper development and evaluation due to the inherent heterogeneities in clinical data. We highlight common heterogeneities in clinical data and sources of biases that should be dealt with in order to avoid improper development and evaluation of models. Improper development of models often leads to inaccurate models that show poor generalizability, while improper evaluation of models may lead to overly optimistic results that cannot be replicated in independent data sets.

The purpose of this paper is to familiarize modelling practitioners with the most common obstacles facing effective model deployment, and their underlying causes. By being aware of the potential problems that may occur during development and evaluation of prediction models, practitioners would be better able to identify and deal with these issues. This would increase their chances of developing a model that can be effectively deployed.

2. Poor reproducibility and replicability

A core tenet of the scientific discovery process lies in the ability of the scientific community to either confirm or refute previous discoveries through independent studies. Scientific results and inferences that are replicable through independent studies have a higher likelihood of being true, and confidence in their reliability is built through repeated independent validation.[7] To avoid the many inconsistent definitions of reproducibility and replicability in scientific literature, we follow the definitions set out by National Academies of Sciences, Engineering, and Medicine[7] in this paper:

> Reproducibility is obtaining consistent results using the same input data; computational steps, methods, and code; and conditions of analysis. Replicability is obtaining consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data.

The lack of reproducibility is one of the most common and problematic issues found in clinical prediction models, and stems from incomplete and unclear reporting.[6,8] To ensure that researchers are able to reproduce results from the original data, clear and complete reporting of all aspects concerning the prediction model should be performed. Firstly, the prediction task and clinical setting that a model is to be deployed in has to be clearly defined. This includes defining the prediction target and deployment population clearly. Secondly, population characteristics of patient data used to train and validate the prediction model should be described in detail, so that researchers have the necessary context to interpret the results. Thirdly, data pre-processing steps that were taken to transform the data should be clearly described, such as feature selection methodology. In addition, predictors in the model should be clearly described, including details regarding how and when they were measured (if applicable). Fourthly, model details such as its type, parameters and architecture should be reported, along with design choices made regarding the model. Lastly, details regarding evaluation methodology should be clearly stated. The above-mentioned details are not exhaustive, and any other detail necessary to reproduce the original results should also be clearly reported. Modelling practitioners are encouraged to standardize reporting according to suggested guidelines in order to avoid incomplete and unclear reporting.[1,6,8,9]

Another major obstacle facing effective model deployment is the poor replicability of clinical prediction models. Models that do not show consistent results on independent data sets are deemed to be less reliable and often fail to gain the confidence of the scientific community. Even so, majority of publications on clinical prediction models do not perform external validation.[10,11] This is compounded by the fact that publications that do perform external validation often report worse model performance during external validation.[11] For instance, Yeoh *et al.*[12] performed external validation of three treatment outcome prediction models[13,14,15] that were developed using their own respective cohort of acute lymphoblastic leukemia (ALL) patients. All of the models were unable to discriminate between treatment outcomes when evaluated on an independent data set of ALL patients.

There are two possible reasons why the prediction models mentioned above showed good internal validation while having poor external validation. Firstly, the data sets used to develop these models were small in size, and had low proportions of patients with the event of interest (i.e. relapse). Performing internal validation involves partitioning the small data sets into even smaller train and test sets, which may result in unstable results during evaluation.[16] Secondly, the development data set and the data set used for external validation may consist of patients from different sub-populations (e.g. from different geographic regions). If that were the case, it would be more accurate to describe the models as having poor generalizability instead of poor replicability.

External validation of prediction models should be performed whenever possible to demonstrate its replicability.[10] Demonstrating the replicability of a model helps to build confidence in its reliability and improves its chances of clinical deployment. In cases where external validation is not possible, the next best alternative would be temporal validation, where the development data set can be split into training and test sets according to time.[17]

3. Unfairness in prediction models

Another key reason behind the limited number of prediction models that are deployed in healthcare is due to the high risk of unfairness in prediction models. More recently, researchers have become aware of biases exhibited by prediction models that have been deployed in real-world applications. Most prediction models derive their function by recognizing implicit patterns in the data that they are trained on. These models tend to learn the hidden biases that exist in the training data. As as result, these models make biased predictions that result in unfairness against certain groups or individuals. On the other hand, some prediction algorithms make biased predictions even when trained on data that is devoid of biases.

There are many different sources of biases in both data and algorithm. The introduction of these biases is mostly unintentional; often we are only made aware of their presence upon the discovery of errors or unfairness in model predictions. Procedures or design choices that seem innocuous are often responsible for introducing biases to data or algorithm. It is extremely hard to ensure that data is free of bias, as doing so would require pre-empting every possible source of bias, which requires an inordinate amount of care. A more reasonable goal would be to minimize the number of biases by learning the most frequent sources of biases that have recurred in previous literature. We enumerate below a few of the most prevalent sources of biases in prediction models deployed specifically in healthcare. For a more comprehensive summary of the sources of biases in machine learning, please refer to Mehrabi *et al.*[18].

Omitted variable bias occurs when variables which have an impact on the dependent variable (i.e. prediction target) are omitted from the model. This may cause the model to attribute the effect of the omitted variable to variables that are included in the model. Usually, the omitted variable is excluded or in some cases not recorded due to oversight. In some cases, variables may be omitted due to privacy concerns. An example of omitted variable bias was observed when a rule-based learning algorithm was trained to predict a patient's risk of dying from pneumonia.[19] One of the rules the model learned was that patients with a history of asthma had a lower risk of mortality from pneumonia. The rule was counter-intuitive, and on closer inspection it was discovered that patients with a history of asthma who contracted pneumonia were sent to the intensive care unit (ICU). Patients admitted into the ICU received intensive care which greatly lowered their risk of mortality. This resulted in the model erroneously associating lowered risk of mortality with asthma, instead of the omitted variable ICU admission.

Representation bias occurs when the sampled data is not representative of the underlying population. When prediction models are trained on sampled data lacking in representation of certain sub-populations, they may exhibit poor prediction performance when deployed on these sub-populations. An example was when the Framingham risk score for cardiovascular disease, which was developed on a data set that was predominantly white and male, was found to be inaccurate when deployed on black populations.[20] To maximize the prediction performance of a model on the target population or group, the model should be trained on data that contains a sizeable number of patients representative of the target group.

Measurement bias is the systematic error that arises during improper measurement of data. An example of measurement bias can be observed in the case of the algorithm used to facilitate COVID-19 relief funding allocation.[21] COVID-19 infection rates may be subject to measurement bias as it may be affected by differing diagnostic testing coverage between poorer and wealthier counties. Wealthier counties may receive higher diagnostic testing coverage, leading to a larger number of cases detected and thus higher COVID-19 infection rate.

4. Underlying causes

There are countless possible causes that may impede effective model deployment. We mention several of the causes above, but they are by no means exhaustive. Out of the many causes, we highlight an underlying cause that recurs frequently: improper development and evaluation of models. We elaborate on how improper development and improper evaluation of prediction models impedes model deployment through the use of real-life examples in the sections below.

4.1. *Improper development of models*

Throughout the development process of prediction models, there are many actions that constitute improper development. These actions often result in models that do not function well or exhibit unfairness when deployed on the target population. One such action is when an incorrect proxy for the prediction target is used to train a prediction model; we show why this leads to a biased model using a real-life example.

This example concerns a commercial prediction model that was developed to predict the health risk of primary care patients (i.e. risk of onset of common chronic illnesses). The model was used to identify patients that would benefit from high-risk care management programs. Developing the model using healthcare cost as a proxy to health risk resulted in a biased model, as healthcare cost was not an accurate proxy for health care. Obermeyer *et al.*[22] highlighted that for the same risk score predicted by the model, Blacks had a higher number of chronic illnesses than Whites. This reflected an inherent bias in the development data - for the same number of chronic illnesses, healthcare costs of Blacks were lower than that of Whites.

Other examples of improper model development include the sources of biases mentioned above, such as developing models on samples that are not representative of the deployment population, or not ensuring the uniformity of measurements across different sub-populations when collecting development data.

4.1.1. *Heterogeneities in clinical data*

Clinical prediction models are especially susceptible to improper development and evaluation due to the many possible heterogeneities in clinical data. It is important to deal with these heterogeneities when they arise, in order to avoid improper model development and evaluation.

The most important heterogeneity in clinical data that has to be accounted for is the differences in patient treatment. This is because different treatments have different magnitudes of effect on patient outcome. Failing to account for treatment differences when developing models that predict patient outcome results in models that produce biased risk estimates. Chan & Wong[23] provide a detailed discussion of differences in patient treatment, and how to account for them during model development and evaluation.

We use the MIMIC-III electronic health records (EHR) data set[24] to demonstrate how failure to account for differences in treatment results in the development of models that show sub-optimal performance. This example also emphasizes the importance of complete reporting, especially of factors that have a causal effect on the prediction target. MIMIC-III is a real-world EHR data set that comprises of data on patients that stayed in the ICU of the Beth Israel Deaconess Medical Center (in Boston) between 2001 and 2012. Patient year of care is randomized in the data set for privacy reasons. However, this also prevents modelling practitioners from accounting for differences in treatment year of care when developing prediction models.

In a study by Nestor *et al.*[25], the authors obtained a license to access the year of care of patients in the MIMIC-III data set; they highlighted two heterogeneities in the data set related to patient year of care. Firstly, clinical measurements were recorded in a different manner after 2008 due to a change in the data management system. Secondly, care practices and population demographics evolved through the years, resulting in temporal drift in the data. The authors developed models trained only on prior year data, by taking into account patient year of care. These models showed better discriminative performance than models developed without accounting for year of care.

Other heterogeneities that occur in clinical data include batch effects, which commonly arise due to the processing of patient data in batches because of limited patient availability at each point in time. The overwhelming amount of heterogeneous details in clinical data makes it hard to ascertain which essential details have to be dealt with when developing prediction models. As a result, these essential details are frequently unreported and unaccounted for, which impedes the proper development and evaluation of models. Great care has to be taken to identify and handle these details to achieve proper model development and evaluation.

4.2. *Improper evaluation of models*

Clinical prediction models are susceptible to improper evaluation, mainly due to issues with data availability. Improper evaluation of models often produces overly optimistic model assessments which are not reflective of model performance under actual deployment. Healthcare decision makers are less likely to trust the authenticity of evaluation results if they observe evidence of improper evaluation. Performing proper evaluation is a key foundation for successful model deployment. In this section, we provide examples of improper evaluation.

The most common flaw when evaluating clinical prediction models is not performing external validation.[10,11] External validation on an external cohort of patients (i.e. not from the development cohort) should be performed whenever possible. A model may perform well on a hold-out test set partitioned from the development cohort but perform badly on an external cohort (i.e. patients from another study). Performing well during external validation is essential to prove a model's replicability.[10]

Other flaws in evaluation are less discernible; we present a flaw in the evaluation of a deep learning model[26] that is used to predict gene mutation from small cell lung cancer histopathology images. In the original study, histopathology slide images are randomly split into training, validation and test sets. However, Oner *et al.*[27] discovered that some of the slides in the training and test sets were highly correlated as they were derived from the same patient, which led to overly optimistic results during evaluation. Proper evaluation involved splitting slide images at the patient level, so that slides from the same patient can only be present together in either the training, validation or test set. The model showed significantly worse performance when properly evaluated. This example highlights the importance of proper evaluation, in giving an accurate assessment of a model's ability to generalize to new patients during real-world deployment.

Evaluating the performance of prediction models usually entails a few established procedures and metrics, such as plotting the receiver operating characteristic (ROC) and precision-recall curves, and calculating the c-statistic (i.e. area under the ROC curve), calibration and net benefit. Nonetheless, there is no one-size-fits-all approach in evaluating prediction models, with the most appropriate metrics to use differing according to the intended use cases of the models.[1] Evaluating models using unsuitable metrics will give inconsequential results that offer little indication of how models would perform during real-world deployment. In addition, modelling practitioners should be aware that metrics such as accuracy and precision vary according to the proportion of positive and negative samples in the test set, even when model performance remains the same.

In cases where either the prediction task or data set is more complex than usual, we propose that evaluation should be customized accordingly. The aim is to evaluate the performance of the model on its prediction task as accurately as possible. Traditional metrics are often suited for standard tasks, but lack the finesse to accurately evaluate model performance on more specialized tasks. Chan & Wong[23] provide an example of a customized evaluation method that incorporates differences in patient treatment.

5. Key takeaways

Heterogeneities in data and sources of bias during model development differ according to each data set and situation. It is best for modelling practitioners to familiarize themselves with common heterogeneities in data and common sources of biases during model development. The best way to handle heterogeneities and biases is highly dependent on the characteristics of each individual situation. It is also crucial for practitioners to be aware of the intricacies in handling heterogeneities and biases.

Certain heterogeneities in clinical data are often unreported in due to privacy reasons. Hence, it is important to anticipate, in particular, unreported heterogeneities that have a causal effect on the prediction target. A frequent example is treatment differences between patients. Permission to access patient treatment information for model development should be requested from data providers, as it is imperative to account for treatment differences during model development and evaluation. Failure to account for treatment differences would be equivalent to implicitly assuming that all patients receive the same treatment.

Proper evaluation of models helps to build confidence in their reproducibility and replicability. Firstly, external validation should be performed whenever possible, to ensure fair assessment of model performance. Secondly, care has to be taken to ensure that the evaluation methodology is free of errors (e.g. data leakage). Lastly, evaluation metrics used should be focused on assessing model performance in achieving the prediction objective.

Complete and clear reporting of all aspects of model development and evaluation is a simple yet often overlooked factor that improves the reproducibility of models. All details regarding the prediction model, such as the prediction objective, target deployment population, data characteristics and composition, data pre-processing, model specification and evaluation methodology, should be provided. As a rule of thumb, all details required to reproduce the original results should be reported when presenting the prediction model.

References

1. E Hope Weissler, Tristan Naumann, Tomas Andersson, Rajesh Ranganath, Olivier Elemento, Yuan Luo, Daniel F Freitag, James Benoit, Michael C Hughes, Faisal Khan, et al. The role of machine learning in clinical research: transforming the future of evidence generation. *Trials*, 22(1):1–15, 2021.
2. Adam Bohr and Kaveh Memarzadeh. The rise of artificial intelligence in healthcare applications. In *Artificial Intelligence in Healthcare*, pages 25–60. Elsevier, 2020.
3. Tao Guo, Yongzhen Fan, Ming Chen, Xiaoyan Wu, Lin Zhang, Tao He, Hairong Wang, Jing Wan, Xinghuan Wang, and Zhibing Lu. Cardiovascular implications of

fatal outcomes of patients with coronavirus disease 2019 (COVID-19). *JAMA Cardiology*, 5(7):811–818, 2020.

4. Pablo Perel, Phil Edwards, Reinhard Wentz, and Ian Roberts. Systematic review of prognostic models in traumatic brain injury. *BMC Medical Informatics and Decision Making*, 6(38), 2006.
5. Jeremy C Wyatt and Douglas G Altman. Commentary: Prognostic models: clinically useful or quickly forgotten? *BMJ*, 311(7019):1539–1541, 1995.
6. Alistair EW Johnson, Tom J Pollard, and Roger G Mark. Reproducibility in critical care: a mortality prediction case study. In *Machine Learning for Healthcare Conference*, pages 361–376. PMLR, 2017.
7. National Academies of Sciences, Engineering, and Medicine. *Reproducibility and Replicability in Science*. National Academies Press, 2019.
8. Laure Wynants, Ben Van Calster, Gary S Collins, Richard D Riley, Georg Heinze, Ewoud Schuit, Marc MJ Bonten, Darren L Dahly, Johanna A Damen, Thomas PA Debray, et al. Prediction models for diagnosis and prognosis of COVID-19: systematic review and critical appraisal. *BMJ*, 369:m1328, 2020.
9. Andrew L Beam, Arjun K Manrai, and Marzyeh Ghassemi. Challenges to the reproducibility of machine learning models in health care. *JAMA*, 323(4):305–306, 2020.
10. Sung Yang Ho, Kimberly Phua, Limsoon Wong, and Wilson Wen Bin Goh. Extensions of the external validation for checking learned model interpretability and generalizability. *Patterns*, 1(8):100129, 2020.
11. George CM Siontis, Ioanna Tzoulaki, Peter J Castaldi, and John PA Ioannidis. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *Journal of Clinical Epidemiology*, 68(1):25–34, 2015.
12. Allen E-J Yeoh, Zhenhua Li, Difeng Dong, Yi Lu, Nan Jiang, Jan Trka, Ah Moy Tan, Hai Peng Lin, Thuan Chong Quah, Hany Ariffin, et al. Effective Response Metric: a novel tool to predict relapse in childhood acute lymphoblastic leukaemia using time-series gene expression profiling. *British Journal of Haematology*, 181(5):653–663, 2018.
13. Deepa Bhojwani, Huining Kang, Renee X Menezes, Wenjian Yang, Harland Sather, Naomi P Moskowitz, Dong-Joon Min, Jeffrey W Potter, Richard Harvey, Stephen P Hunger, et al. Gene expression signatures predictive of early response and outcome in high-risk childhood acute lymphoblastic leukemia: a children's oncology group study. *Journal of Clinical Oncology*, 26(27):4376–4384, 2008.
14. Amy Holleman, Meyling H Cheok, Monique L den Boer, Wenjian Yang, Anjo JP Veerman, Karin M Kazemier, Deqing Pei, Cheng Cheng, Ching-Hon Pui, Mary V Relling, et al. Gene-expression patterns in drug-resistant acute lymphoblastic leukemia cells and response to treatment. *New England Journal of Medicine*, 351(6):533–542, 2004.
15. Lüder Hinrich Meyer, Sarah Mirjam Eckhoff, Manon Queudeville, Johann Michael Kraus, Marco Giordan, Jana Stursberg, Andrea Zangrando, Elena Vendramini, Anja Möricke, Martin Zimmermann, et al. Early relapse in ALL is identified by time to leukemia in NOD/SCID mice and is characterized by a gene signature involving survival pathways. *Cancer Cell*, 19(2):206–217, 2011.
16. Ewout W Steyerberg and Frank E Harrell. Prediction models need appropriate internal, internal–external, and external validation. *Journal of Clinical Epidemiology*, 69:245–247, 2016.
17. Chava L Ramspek, Kitty J Jager, Friedo W Dekker, Carmine Zoccali, and Merel van Diepen. External validation of prognostic models: what, why, how, when and where? *Clinical Kidney Journal*, 14(1):49–58, 2021.
18. Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram

Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(115):1–35, 2021.

19. 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.
20. Crystel M Gijsberts, Karlijn A Groenewegen, Imo E Hoefer, Marinus JC Eijkemans, Folkert W Asselbergs, Todd J Anderson, Annie R Britton, Jacqueline M Dekker, Gunnar Engström, Greg W Evans, et al. Race/ethnic differences in the associations of the Framingham risk factors with carotid IMT and cardiovascular events. *PLoS One*, 10(7):e0132321, 2015.
21. Pragya Kakani, Amitabh Chandra, Sendhil Mullainathan, and Ziad Obermeyer. Allocation of COVID-19 relief funding to disproportionately black counties. *JAMA*, 324(10):1000–1003, 2020.
22. Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.
23. Wei Xin Chan and Limsoon Wong. Accounting for treatment during the development or validation of prediction models. *Journal of Bioinformatics and Computational Biology*, 20(06):2271001, 2022. PMID: 36514873.
24. Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(160035), 2016.
25. Bret Nestor, Matthew BA McDermott, Willie Boag, Gabriela Berner, Tristan Naumann, Michael C Hughes, Anna Goldenberg, and Marzyeh Ghassemi. Feature robustness in non-stationary health records: caveats to deployable model performance in common clinical machine learning tasks. In *Machine Learning for Healthcare Conference*, pages 381–405. PMLR, 2019.
26. Nicolas Coudray, Paolo Santiago Ocampo, Theodore Sakellaropoulos, Navneet Narula, Matija Snuderl, David Fenyö, Andre L Moreira, Narges Razavian, and Aristotelis Tsirigos. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nature Medicine*, 24(10):1559–1567, 2018.
27. Mustafa Umit Oner, Yi-Chih Cheng, Hwee Kuan Lee, and Wing-Kin Sung. Training machine learning models on patient level data segregation is crucial in practical clinical applications. *medRxiv*, 2020. doi: https://doi.org/10.1101/2020.04.23.20076406.

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.