Systematic Review of Approaches to Preserve Machine Learning Performance in the Presence of Temporal Dataset Shift in Clinical Medicine

Lin Lawrence Guo[1] Stephen R. Pfohl[2] Jason Fries[2] Jose Posada[2] Scott Lanyon Fleming[2] Catherine Aftandilian[4] Nigam Shah[2] Lillian Sung[1,3]

[1] Program in Child Health Evaluative Sciences, The Hospital for Sick Children, Toronto, Canada
[2] Biomedical Informatics Research, Stanford University, Palo Alto, California, United States
[3] Division of Haematology/Oncology, The Hospital for Sick Children, Toronto, Canada
[4] Division of Pediatric Hematology/Oncology, Stanford University, Palo Alto, United States

Address for correspondence Lillian Sung, MD, PhD, Division of Haematology/Oncology, The Hospital for Sick Children, 555 University Avenue, Toronto, Ontario M5G1 × 8, Canada (e-mail: lillian.sung@sickkids.ca).

Appl Clin Inform 2021;12:808–815.

Abstract

Objective The change in performance of machine learning models over time as a result of temporal dataset shift is a barrier to machine learning-derived models facilitating decision-making in clinical practice. Our aim was to describe technical procedures used to preserve the performance of machine learning models in the presence of temporal dataset shifts.

Methods Studies were included if they were fully published articles that used machine learning and implemented a procedure to mitigate the effects of temporal dataset shift in a clinical setting. We described how dataset shift was measured, the procedures used to preserve model performance, and their effects.

Results Of 4,457 potentially relevant publications identified, 15 were included. The impact of temporal dataset shift was primarily quantified using changes, usually deterioration, in calibration or discrimination. Calibration deterioration was more common ($n = 11$) than discrimination deterioration ($n = 3$). Mitigation strategies were categorized as model level or feature level. Model-level approaches ($n = 15$) were more common than feature-level approaches ($n = 2$), with the most common approaches being model refitting ($n = 12$), probability calibration ($n = 7$), model updating ($n = 6$), and model selection ($n = 6$). In general, all mitigation strategies were successful at preserving calibration but not uniformly successful in preserving discrimination.

Conclusion There was limited research in preserving the performance of machine learning models in the presence of temporal dataset shift in clinical medicine. Future research could focus on the impact of dataset shift on clinical decision making, benchmark the mitigation strategies on a wider range of datasets and tasks, and identify optimal strategies for specific settings.

Keywords
- dataset shift
- machine learning
- clinical data
- systematic review

received
April 28, 2021
accepted after revision
July 12, 2021

DOI https://doi.org/10.1055/s-0041-1735184.
ISSN 1869-0327.

Background and Significance

Over 250,000 risk stratification model-related papers have been published, primarily over the last two decades.[1] The substantial increase in the ability to create predictive models in health care systems is largely due to the widespread adoption of electronic health records (EHRs) and the dramatic increase in the capacity to store and perform computations with large amounts of data. Many machine learning models developed using EHRs have demonstrated excellent performance with regards to discrimination and calibration.[2,3] For these models to be effectively adopted in health care systems, they need to sustain a high level of performance to outweigh the estimated $200,000 cost of integrating each model into clinical workflows[4] as well as to gain and maintain the trust of health care professionals that incorporate them into their clinical decision-making processes.[5] However, maintenance of model performance may be difficult because of changes in the health care environment over time.

Changes in health care over time can occur at the level of patients, practice, or administration. Variation in patients can occur based on changes in demographic characteristics of a catchment area, referral patterns, or emergence of novel diseases, as examples. Variation in practice can occur based on the results of major trials or guidelines; evolving practice patterns of health care professionals[6]; and changes in personnel, drug or test availability, and reimbursement policies at an institution. Variation in administration reflects those affecting the EHR such as EHR modifications, change in EHR vendor,[7] choice of coding system/version,[8] and coding practices. Together, these changes introduce a dataset shift due to mismatch between the distribution of the data used for model development and deployment.[9] Dataset shifts over time can be abrupt, gradual, incremental, or recurring (►Supplementary Fig. S1 [available in the online version]) and can have varying degrees of impact.

Dataset shift is a major barrier to the generalizability of machine learning models across health care institutions and over time.[10] Although model generalizability across both geography and time are desirable, temporal generalization places more emphasis on producing deployable models aimed at preserving performance in a specific healthcare system.[11] Because in actual deployment dataset shifts are often difficult to anticipate and only identified when changes in calibration or discrimination are examined, approaches that make machine learning models robust to these changes are an important step toward the reliable application of machine learning in healthcare. Despite the existence of hundreds of publications on methods of dataset shift detection and mitigation,[12] it was unclear how many had been applied in clinical medicine. This calls for a systematic review of mitigation strategies aimed at reducing the impact of dataset shift on clinical prediction models to identify promising solutions and determine future directions.

Objectives

The aim was to describe technical procedures used to preserve the performance of machine learning models in the presence of temporal dataset shift in clinical medicine.

Methods

We followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) recommendations for reporting.[13]

Data Sources and Searches

The literature search was conducted by a library scientist in the following databases: Medline, Medline in-process, Medline epubs ahead of print, Embase, APA PsycInfo, arXiv and web of science. ►Supplementary Table S1 (available in the online version) describes the full search strategy; it included publications from database inception to January 21, 2021. For the search, we included the Medical Subject Heading terms and text words that identified machine learning (including text words for machine learning package names and algorithms) and dataset shift (including dataset, distribution, domain, covariate, and concept shift or drift). We further included text words that identified consequences of dataset shift (including performance and calibration shift or drift). The set was limited to English publications and studies involving humans.

Study Selection

Eligibility criteria were defined a priori. Studies were included if they were fully published studies that used machine learning and implemented a technical procedure to address temporal dataset shift. We defined machine learning as methods that learn a model using a dataset (training data) by automatically determining a function that maps a set of inputs (features) to their corresponding outputs (labels) with the goal of predicting an outcome using the trained model in a new dataset not yet seen (test data).

Studies were excluded if they did not implement a mitigation strategy for temporal dataset shift, if they did not address a clinical problem such as predicting a patient outcome or if they were duplicate publications. We also excluded studies focused on sensor data (i.e., physiological signals) evaluated within single patients, as our intent was to address temporal dataset shift occurring across different patients over a time span of months to years.

Two reviewers (L.L.G. and L.S.) independently evaluated the titles and abstracts of studies identified using the search strategy, and potentially relevant publications were retrieved in full. Both reviewers then applied the eligibility criteria to the full text articles and made decisions independently. Discrepancies were resolved by consensus or arbitration by a third author (S.R.P.) if required.

Data Abstraction and Methodological Approach

Two reviewers (L.L.G. and L.S.) abstracted all data in duplicate; discrepancies were resolved by consensus. The primary outcome was the technical procedure used where the goal was to preserve the performance of machine learning models in the presence of temporal dataset shift. These were classified as model-level or feature-level mitigation strategies. Model-level mitigation strategies were further categorized into fixed methods, characterized by models with static parameters

upon training; online learning, characterized by models with dynamic parameters upon training; and model selection. Fixed methods included probability calibration (methods that adjust the predicted probabilities of a base model using logistic regression while retaining the base model's parameters) and model refitting (re-estimating model parameters using updating data, and once the parameters are re-estimated, they become fixed until the arrival of new data). Online learning included model updating (methods that incrementally update [instead of entirely refit] the model parameters as new data become available), and ensemble models (methods that combine the predictions of a set of models and weigh their contributions). Model selection involved statistical tests to select the best mitigation strategy among a set of strategies. Feature-level mitigation strategies process features prior to model fitting and were categorized into learning-based (data driven) and expert knowledge-based (domain expertise driven) methods.

We recorded if the mitigation strategy was successful at preserving the performance of the machine learning model over time. In addition, we described factors reported to be associated with temporal dataset shift by the authors, and how the impact of temporal dataset shift was quantified.

Study Demographics and Risk of Bias

Demographic information included year published, pediatric versus adult cohort, population studied, data source, machine learning algorithm(s) implemented, and the number of time periods (i.e., discrete time windows) in which temporal dataset shift was examined. We also abstracted the label, whether models were developed by using data from a single center or multiple centers, the number of mitigation strategies implemented and whether calibration and discrimination deterioration were reported to be present, absent, or not reported.

Data related to the risk of bias was based upon an approach suggested by Luo et al.[14] We abstracted whether descriptions of inclusion and exclusion criteria, sample, response variable, information leakage prevention, data preprocessing (including handling of missing data), data splitting, and validation metrics were reported. We also determined if the code used to train and validate models was made publicly available.

Statistical Methods

Based upon the nature of the outcomes, synthesis was not performed. Statistical analysis involved describing proportions for the categorical outcomes.

Results

►Supplementary Fig. S2 (available in the online version) illustrates the flow diagram of study identification and selection. A total of 4,457 potentially relevant references were identified; 75 manuscripts were retrieved for full-text evaluation. After the exclusion of 61 manuscripts, 14 were retained in the systematic review. The most common reason for exclusion was the focus on a nonclinical problem ($n = 46$). One additional publication was identified from an author's personal reference list. Thus, 15 manuscripts were included in the systematic review.

►Table 1 and ►Supplementary Table S2 (available in the online version) describe the demographic characteristics of the 15 studies. Eight studies were published in or after 2018. The most common machine learning algorithm was logistic regression ($n = 13$), and five studies employed more than one

Table 1 Characteristics of studies addressing dataset shift in clinical medicine ($n = 15$)

Characteristic	*n* (%)
Published in 2018 or later	8 (53)
Pediatric study population	1 (7)
Population	
Intensive care or neonatal intensive care	5 (33)
Surgical	5 (33)
Inpatients	4 (27)
Prostate biopsy	1 (7)
Data source	
Administrative	4 (27)
Registry	4 (27)
Electronic health record	4 (27)
Trial	2 (13)
Combination	1 (7)
Machine learning algorithm[a]	
Logistic regression	13 (87)
Random forest	4 (27)
Gradient boosting	1 (7)
Artificial neural network	2 (13)
Multiple	5 (33)
Number of time periods examined	
1–4	5 (33)
5–10	6 (40)
>10	4 (27)
Factors reported to be associated with temporal dataset shift[a]	
Change in outcome rate	8 (53)
Change in case mix	3 (20)
Change in predictor-outcome association	2 (13)
Change of record-keeping system	2 (13)
Not reported	5 (33)
Quantification of the impact of temporal dataset shift[a]	
Change in calibration	11 (73)
Change in discrimination	12 (80)
Change in both calibration and discrimination	9 (60)

[a]As a study could belong to multiple categories, the total number does not equal the number of studies.

algorithm. The number of time periods examined ranged from 1 to 36. Temporal dataset shift was not formally defined but were reported in some studies to be associated with changes in outcome rate ($n = 8$), case mix ($n = 3$), predictor-outcome association ($n = 2$), and record-keeping system ($n = 2$). All but one study quantified temporal dataset shift using change in calibration (for example, Cox recalibration intercepts, or slopes[15]) or discrimination (typically the area-under-receiver-operating-characteristic curve). While all 11 studies evaluating calibration reported temporal calibration deterioration, only three of 12 studies evaluating discrimination reported temporal discrimination deterioration. Furthermore, there was no consensus in delineating a difference threshold to define model deterioration. ►Supplementary Table S2 (available in the online version) also illustrates that the three studies reporting discrimination deterioration were single-center studies by using data from Beth Israel Deaconess Medical Center, while the nine studies reporting that discrimination deterioration was absent or uncertain were multicenter studies. ►Supplementary Table S3 (available in the online version) summarizes the risk of bias assessment across the 15 studies. The most poorly reported domain was code availability, which was present in only two manuscripts.

►Table 2 summarizes the mitigation strategies used to preserve machine learning performance in the presence of temporal dataset shift. The most common approach was model refitting ($n = 12$), followed by probability calibration ($n = 7$), model updating ($n = 6$), and model selection ($n = 6$). Below, we separately describe each category and its success in mitigating the impact of temporal dataset shift.

Model-Level Mitigation Strategies

All 15 studies employed mitigation strategies at the model level, with or without additional feature processing. Among the 12 studies that used a fixed method, models were trained by using data from a specific time window[16–24] or data across all past time windows.[22,25–27] Seven studies applied probability calibration in the form of mean correction[16–18,20,23,24,27] (adding an intercept), proportional change[16–18,20,23,24,27] (adding a slope), or nonlinear mappings between baseline predictions and outcomes.[16,24] Along with adjusting the model predictions, Su et al[23] and Janssen et al[20] added individual predictor variables to the logistic model, thus allowing additional parameters to be estimated. All probability calibration methods were reported to be successful in mitigating the impact of temporal dataset shift on calibration across several scenarios. However, these methods often did not improve discrimination.[17,18,23] Furthermore, there was no single best approach among the probability calibration methods. The best approach depended on the size of the updating data, the complexity of the base model and the factor associated with the shift.[16–18,24]

Model refitting was used in 12 studies.[16–27] In four studies, it served as a comparator against other mitigation strategies.[16,24–26] In five other studies, it improved calibration, but often not more than probability calibration methods.[19–21,23,27] Nestor et al found that model refitting using data from the previous year protected against discrimination deterioration related to a change in the record keeping system.[25,26] Adam et al later used simulations to show that refitting using all available data better protected the model from biases to do with feedback loops in which imperfect

Table 2 Mitigation strategies to address dataset shift in clinical medicine

	Model level					Feature level	
	Fixed		Online learning		Model selection		
Study (Year)	Probability calibration	Model refitting	Model updating	Ensemble model		Learned	Expert knowledge
Feng (2020)[28]				●			
Adam (2020)[22]		●	●				
Davis (2019)[16]	●	●			●		
Davis (2019)[24]	●	●			●		
Nestor et al (2019)[25]		●				●	●
Siregar (2019)[17]	●	●	●		●		
Nestor et al (2018)[26]		●					●
Su (2018)[23]	●	●	●	●			
Davis (2017)[30]					●		
Davis (2017)[29]					●		
Siregar et al (2016)[18]	●	●	●		●		
Strobl et al (2015)[27]	●	●	●				
Hickey et al (2013)[19]		●	●				
Janssen (2008)[20]	●	●					
Parry (2003)[21]		●					

Note: The black dot indicates that the mitigation strategy was used in the corresponding study.

model predictions (such as false positives) can influence future labels.[22]

Online learning was used in 11 studies and consisted of model updating ($n = 6$) and ensemble models ($n = 2$). The most common model updating approach was Bayesian model updating ($n = 7$).[17–19,23,27] Although this method was reported to be successful at mitigating the impact of dataset shift on model calibration,[19] it did not outperform probability calibration methods.[17,18,23,27] The other model updating approach was a single step gradient descent that used the updating data to update model parameters. This updating method performed using all historical data worked as well as refitting the model using the same data.[22]

An ensemble model approach was used by two studies. Feng et al developed an ensemble method that used updating data to learn how to approve modifications to an existing random forest model. The method produced predictions using the weighted average of a family of strategies that differed in their optimism for the modifications.[28] This method safely and autonomously approved new modifications while adapting to temporal dataset shift. Su et al submitted the output of two dynamic linear models to a logistic regression to be used as predictors, also known as model stacking. Although this approach achieved success in reducing the impact of temporal dataset shift on model calibration, this method was no more effective than each individual dynamic linear model.[23]

Model selection was used in six studies.[16–18,24,29,30] One approach by Davis et al was a data-driven selection procedure that balanced performance against the simplicity of the mitigation strategy. Using the updating data, the procedure nonparametrically compared the performance of several methods that varied in complexity with respect to data requirements and analytical resource demands including no updating, probability calibration, and model refitting. The procedure selected the simplest method that had statistically indistinguishable performance compared with the more complex methods. Complexity of the mitigation strategy recommended by this selection procedure increased with the severity of calibration deterioration,[16] size of the updating data,[16] and model complexity.[16,24]

Feature-Level Mitigation Strategies

Two studies used learned and expert knowledge-based methods to address temporal dataset shift caused by a change in the record-keeping system at a single center.[25,26]

One study used a learned method to apply principal component analysis to reduce the dimensionality of the feature space. This method was not successful in reducing discrimination deterioration.[25] The two studies that used expert knowledge-based mitigation strategies evaluated code mapping[25] and feature grouping.[25,26] Code mapping is an automatic procedure that maps the identifier of each feature to its associated Concept Unique Identifier using the Unified Medical Language System.[31] Code mapping was not effective in reducing discrimination deterioration. In contrast, manual grouping of features into their underlying concept by clinical experts was the only feature-level mitigation strategy that was successful in reducing temporal discrimination deterioration.

Discussion

This systematic review described the technical procedures used in clinical medicine to preserve the performance of machine learning models in the presence of temporal dataset shift. We identified 15 publications that quantified the impact of temporal dataset shift on clinical prediction models and examined technical procedures to address the impact. We found that temporal calibration deterioration was more common than temporal discrimination deterioration. Model-level mitigation strategies to address temporal dataset shift were more common than feature-level mitigation strategies, with the most common approaches being model refitting, probability calibration, model updating, and model selection. In general, all mitigation strategies were successful at preserving calibration but not uniformly successful at preserving discrimination.

The number of identified publications examining mitigation strategies to address temporal dataset shift in clinical medicine was small, and even smaller if only unique approaches were considered. This stood in contrast to the large body of literature evaluating mitigation strategies outside of clinical medicine. Because our search strategy and screening of titles and abstracts would have omitted some nonclinical publications, the 46 articles excluded at full text screening because the setting was nonclinical is a subset of the total nonclinical literature. This estimation is supported by a review describing 130 publications on temporal concept shift.[32] Our finding suggests that methodological research addressing this important topic has lagged in clinical medicine, a result that is important since mitigation strategies successful in nonclinical settings may not be successful when applied to clinical data.[33]

The identified studies suggested that the best choice of mitigation strategy depended on the type and severity of dataset shift.[16] Currently, there is no standard approach that maps a type of dataset shift within a specific setting to a specific mitigation strategy. Moreover, there is often variability in how the term dataset shift and its subcategories are defined.[9,32] To begin to address these issues, we first recommend the standardization of terminology and common assumptions related to dataset shift. We suggest basing temporal dataset shift terminology upon previously used terms and definitions.[9,34] Typical categories of dataset shift are expressed in terms of assumptions as to which statistical relationships are likely to be stable or change across time on the basis of the assumed directionality and stability of the causal relationships between the features, the outcome, and any unobserved confounders. In this framing, the general problem of dataset shift is one where joint distributions of the training and the test data are different, that is, $P_{train}(y, x) \neq P_{test}(y, x)$, where y represents the outcome variable and x represents a set of features or covariates.

If it is assumed that the outcome causally depends on the features X (i.e., an X → Y assumption consistent with the prediction of future outcomes), then plausible settings may include covariate shift, where $P_{train}(x) \neq P_{test}(x)$ and $P_{train}(y|x) = P_{test}(y|x)$. This corresponds to a change in the distribution of features without an accompanying change in the relationship between the features and the outcome. In contrast, a change in the relationship between the features and the outcome is termed concept shift, that is, $P_{train}(y|x) \neq P_{test}(y|x)$ and $P_{train}(x) = P_{test}(x)$. Note that covariate shift and concept shift can coexist. Conversely, under the Y → X assumption (consistent with image classification where the disease Y causes the change in pixels X),[35] prior probability shift occurs if the probability of the outcome changes without a corresponding change in the relationship between the outcome and the features, that is, $P_{train}(y) \neq P_{test}(y)$ and $P_{train}(x|y) = P_{test}(x|y)$. ►**Supplementary Fig. S3** (available in the online version) diagrammatically describes each category of shift and provides illustrative clinical examples that align with an X → Y assumption.

Beyond standardization of terminologies, we encourage benchmarking of established mitigation strategies from the machine learning literature in different datasets and in different patient populations to identify, if there are mitigation methods that are preferred depending on a specific type of shift or clinical setting. Several promising approaches to address differences between training and test distributions (not restricted to temporal dataset shift) have been developed in recent research on machine learning outside of clinical studies. These approaches aim to produce robust models, for instance, by incorporating more expressive domain knowledge as to which causal mechanisms are likely to be stable or change across time[36] or by estimating invariant relationships across different environments.[37,38]

One issue that has not been highlighted prominently is how temporal dataset shift affects clinical decision-making.[10] Regardless of the degree to which there is deterioration in calibration or discrimination, it is important to evaluate the impact of temporal dataset shift in the context of its impact on clinical decision-making and downstream outcomes.[39] We suggest that this element be explicitly examined in future studies.

The strength of this review is the focus on an issue highly relevant to the deployment of machine learning models in the clinical setting, namely temporal dataset shift. Another strength is the use of two reviewers for each step in the systematic review. However, there are several limitations. First, despite our attempt to be exhaustive in the search, some conference proceedings (e.g., proceedings of machine learning research) with potentially relevant papers were missed. Nonetheless, our search identified and evaluated many preprints of papers in these proceedings obtained from arXiv. Second, some deployed clinical prediction models may have built-in periodic recalibration, refitting, or incorporated other approaches that mitigate temporal dataset shift but were not published. These approaches would not have been identified by this review. Third, we focused on temporal dataset shift and did not also examine geographic dataset shift. While we recognize both areas are important, we chose to focus on temporal shift, as this would have greater relevance in a common deployment setting, where models are trained and evaluated using a single institution's data. Lastly, our search strategy excluded studies that delineated temporal dataset shift without applying a clinical prediction model. Such methods are complementary to the mitigation strategies reviewed in this study.[40]

Conclusion

In conclusion, the objective of this systematic review was to describe technical procedures used to preserve the performance of machine learning models in the presence of temporal dataset shift in clinical medicine. We identified 15 studies in total, and consequently there was limited research in this area. Future research could evaluate the impact of dataset shift on clinical decision making, benchmark mitigation strategies on a wider range of datasets and identify optimal approaches for specific settings.

Clinical Relevance Statement

Temporal dataset shift associated with changes in health care overtime is a barrier to deploying machine learning-based clinical decision support systems. This systematic review identified limited methodological research that aimed to mitigate the impact of temporal dataset shift on the discrimination performance of clinical prediction models. We recommend more benchmarking of mitigation strategies on a wider range of datasets and tasks to better characterize the impact of temporal dataset shift and identify suitable solutions for specific settings.

Multiple Choice Questions

1. Which of the following options is a feature-level mitigation strategy aimed to reduce the impact of temporal dataset shift on clinical prediction model performance?
   a. Periodic re-estimation of model parameters (i.e., model refitting).
   b. Ensemble methods that combine the predictions of a set of models and weight their contributions.
   c. Aggregation of features according to their underlying concept by clinical experts.
   d. Methods that adjust the predicted probabilities of a base model using, for example, logistic regression.

   Correct Answer: The correct answer is option c. Model refitting, ensemble methods, and probability calibration are model-level mitigation strategies. See ►**Table 2** for grouping of mitigation strategies.

2. Which of the following options fits the definition of dataset shift as when the joint distributions of the training and the test data are different? For all options, x represents a set of features or covariates and y represents the outcome variable.

a. $P_{train}(y, x) \neq P_{test}(y, x)$
b. $P_{train}(x) \neq P_{test}(x)$
c. $P_{train}(y|x) \neq P_{test}(y|x)$
d. $P_{train}(y) \neq P_{test}(y)$

Correct Answer: The correct answer is option a. Option b corresponds to a change in the distribution of features. Option c corresponds to a change in the association between features and outcome. Option d corresponds to a change in the distribution of outcome. Only option a corresponds to a change in the joint distribution of features and outcome.

Note

L.S. is the Canada Research Chair in Pediatric Oncology Supportive Care.

Author Contributions

L.L.G. and L.S. supported in data acquisition and data analysis. All authors helped in study concepts and design, and data interpretation; involved in drafting the manuscript or revising it critically for important intellectual content; carried out the final approval of version to be published; and granted agreement to be accountable for all aspects of the work.

Protection of Human and Animal Subjects

As this study is a systematic review of primary studies, human and/or animal subjects were not included in the project.

Funding

None.

Conflict of Interest

None declared.

References

1 Challener DW, Prokop LJ, Abu-Saleh O. The proliferation of reports on clinical scoring systems: issues about uptake and clinical utility. JAMA 2019;321(24):2405–2406
2 Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med 2018;1:18
3 Harutyunyan H, Khachatrian H, Kale DC, Ver Steeg G, Galstyan A. Multitask learning and benchmarking with clinical time series data. Sci Data 2019;6(01):96
4 Sendak MP, Balu S, Schulman KA. Barriers to Achieving Economies of Scale in Analysis of EHR Data. A Cautionary Tale. Appl Clin Inform 2017;8(03):826–831
5 Cutillo CM, Sharma KR, Foschini L, Kundu S, Mackintosh M, Mandl KDMI in Healthcare Workshop Working Group. Machine intelligence in healthcare-perspectives on trustworthiness, explainability, usability, and transparency. NPJ Digit Med 2020;3:47
6 Braithwaite J. Changing how we think about healthcare improvement. BMJ 2018;361:k2014
7 Johnson AE, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016;3:160035
8 National Center for Health Statistics. International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM). Centers for Disease Control and Prevention Accessed February 13, 2021 at: https://www.cdc.gov/nchs/icd/icd9cm.htm
9 Moreno-Torres JG, Raeder T, Alaiz-Rodríguez R, Chawla NV, Herrera F. A unifying view on dataset shift in classification. Pattern Recognit 2012;45(01):521–530
10 Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. Artificial intelligence, bias and clinical safety. BMJ Qual Saf 2019;28(03):231–237
11 Futoma J, Simons M, Panch T, Doshi-Velez F, Celi LA. The myth of generalisability in clinical research and machine learning in health care. Lancet Digit Health 2020;2(09):e489–e492
12 Gama J, Žliobaitė I, Bifet A, Pechenizkiy M, Bouchachia A. A survey on concept drift adaptation. ACM Comput Surv 2014;46(04):1–37
13 Moher D, Shamseer L, Clarke M, et al; PRISMA-P Group. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. Syst Rev 2015;4:1
14 Luo W, Phung D, Tran T, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. J Med Internet Res 2016;18(12):e323
15 Cox DR. Two further applications of a model for binary regression. Biometrika 1958;45(3–4):562–565
16 Davis SE, Greevy RA, Fonnesbeck C, Lasko TA, Walsh CG, Matheny ME. A nonparametric updating method to correct clinical prediction model drift. J Am Med Inform Assoc 2019;26(12):1448–1457
17 Siregar S, Nieboer D, Versteegh MIM, Steyerberg EW, Takkenberg JJM. Methods for updating a risk prediction model for cardiac surgery: a statistical primer. Review Interact Cardiovasc Thorac Surg 2019;28(03):333–338
18 Siregar S, Nieboer D, Vergouwe Y, et al. Improved prediction by dynamic modeling: an exploratory study in the adult cardiac surgery database of the netherlands association for cardio-thoracic surgery. Circ Cardiovasc Qual Outcomes 2016;9(02):171–181
19 Hickey GL, Grant SW, Caiado C, et al. Dynamic prediction modeling approaches for cardiac surgery. Circ Cardiovasc Qual Outcomes 2013;6(06):649–658
20 Janssen KJ, Moons KG, Kalkman CJ, Grobbee DE, Vergouwe Y. Updating methods improved the performance of a clinical prediction model in new patients. J Clin Epidemiol 2008;61(01):76–86
21 Parry G, Tucker J, Tarnow-Mordi WUK Neonatal Staffing Study Collaborative Group. CRIB II: an update of the clinical risk index for babies score. Lancet 2003;361(9371):1789–1791
22 Adam GA, Chang C-HK, Haibe-Kains B, Goldenberg A. Hidden risks of machine learning applied to healthcare: unintended feedback loops between models and future data causing model degradation. Presented at: Proceedings of the 5th Machine Learning for Healthcare Conference; Proceedings of Machine Learning Research. Accessed 2020 at: http://proceedings.mlr.press
23 Su TL, Jaki T, Hickey GL, Buchan I, Sperrin M. A review of statistical updating methods for clinical prediction models. Stat Methods Med Res 2018;27(01):185–197
24 Davis SE, Greevy RA, Lasko TA, Walsh CG, Matheny ME. Comparison of prediction model performance updating protocols: using a data-driven testing procedure to guide updating. AMIA Annu Symp Proc 2019 2019:1002–1010
25 Nestor B, McDermott MBA, Boag W, et al. Feature robustness in non-stationary health records: caveats to deployable model performance in common clinical machine learning tasks. Presented at: Proceedings of the 4th Machine Learning for Healthcare Conference. Accessed 2019 at: http://proceedings.mlr.press
26 Nestor B, McDermott MBA, Chauhan G, et al. Rethinking clinical prediction: why machine learning must consider year of care and feature aggregation. Available at: arXiv:181112583 [csLG] . Accessed 2018
27 Strobl AN, Vickers AJ, Van Calster B, et al. Improving patient prostate cancer risk assessment: Moving from static, globally-applied to dynamic, practice-specific risk calculators. J Biomed Inform 2015;56:87–93

28 Feng J. Learning how to approve updates to machine learning algorithms in non-stationary settings. Available at: arXiv preprint arXiv:201207278. Accessed 2020
29 Davis SE, Lasko TA, Chen G, Matheny MECalibration drift among regression and machine learning models for hospital mortality AMIA Annual Symposium proceedings/AMIA Symposium. 2017; Annual Symposium proceedings. AMIA Symposium 625–634 Accessed 2017 at: https://pubmed.ncbi.nlm.nih.gov/29854127/
30 Davis SE, Lasko TA, Chen G, Siew ED, Matheny ME. Calibration drift in regression and machine learning models for acute kidney injury. J Am Med Inform Assoc 2017;24(06):1052–1061
31 Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004;32 (Database issue):D267–D270
32 Lu J, Liu A, Dong F, Gu F, Gama J, Zhang G. Learning under concept drift: a review. IEEE Trans Knowl Data Eng 2018;31(12):2346–2363
33 Zhang H, Dullerud N, Seyyed-Kalantari L, Morris Q, Joshi S, Ghassemi M. An empirical framework for domain generalization in clinical settings. Presented at: Proceedings of the Conference on Health, Inference, and Learning; Virtual Event, USA. Accessed 2021 at: https://doi.org/10.1145/3450439.3451878
34 Quiñonero-Candela J, Sugiyama M, Ben-David S, et al. Dataset Shift in Machine Learning. MIT Press; 2008
35 Schölkopf B, Janzing D, Peters J, Sgouritsa E, Zhang K, Mooij J. On causal and anticausal learning. Presented at: Proceedings of the 29th International Coference on International Conference on Machine Learning; Edinburgh, Scotland. Accessed 2012 at: https://icml.cc/2012/papers/625.pdf
36 Subbaswamy A, Schulam P, Saria S. Preventing failures due to dataset shift: learning predictive models that transport. Presented at: International Conference on Artificial Intelligence and Statistics (AISTATS); Naha, Japan. Accessed 2019 at: http://proceedings.mlr.press
37 Heinze-Deml C, Peters J, Meinshausen N. Invariant causal prediction for nonlinear models. J Causal Inference 2018;6(02):. Doi: 10.1515/jci-2017-0016
38 Arjovsky M, Bottou L, Gulrajani I, Lopez-Paz D. Invariant risk minimization. arXiv preprint arXiv:190702893. Accessed 2019 at: https://arxiv.org/abs/1907.02893
39 Liu VX, Bates DW, Wiens J, Shah NH. The number needed to benefit: estimating the value of predictive analytics in healthcare. J Am Med Inform Assoc 2019;26(12):1655–1659
40 Sáez C, Gutiérrez-Sacristán A, Kohane I, García-Gómez JM, Avillach P. EHRtemporalVariability: delineating temporal data-set shifts in electronic health records. Gigascience 2020;9(08): giaa079