Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

N Engl J Med. Author manuscript; available in PMC 2022 January 15.

Published in final edited form as:

N Engl J Med. 2021 July 15; 385(3): 283–286. doi:10.1056/NEJMc2104626.

The Clinician and Dataset Shift in Artificial Intelligence

Samuel G. Finlayson, Ph.D.,
Harvard Medical School, Boston, MA

Adarsh Subbaswamy, B.S.,
Johns Hopkins University, Baltimore, MD

Karandeep Singh, M.D., M.M.Sc.,
University of Michigan Medical School, Ann Arbor, MI

John Bowers, B.A.,
Yale Law School, New Haven, CT

Annabel Kupke, B.A., B.S.,
Boston University School of Law, Boston, MA

Jonathan Zittrain, J.D., M.P.A.,
Harvard Law School, Cambridge, MA

Isaac S. Kohane, M.D., Ph.D.,
Harvard Medical School, Boston, MA

Suchi Saria, Ph.D.
Bayesian Health, New York, NY

TO THE EDITOR:

Artificial intelligence (AI) systems are now regularly being used in medical settings,[1] although regulatory oversight is inconsistent and undeveloped.[2,3] Safe deployment of clinical AI requires informed clinician-users, who are generally responsible for identifying and reporting emerging problems. Clinicians may also serve as administrators in governing the use of clinical AI. A natural question follows: are clinicians adequately prepared to identify circumstances in which AI systems fail to perform their intended function reliably?

A major driver of AI system malfunction is known as "dataset shift."[4,5] Most clinical AI systems today use machine learning, algorithms that leverage statistical methods to learn key patterns from clinical data. Dataset shift occurs when a machine-learning system underperforms because of a mismatch between the data set with which it was developed and the data on which it is deployed.[4] For example, the University of Michigan Hospital implemented the widely used sepsis-alerting model developed by Epic Systems; in April 2020, the model had to be deactivated because of spurious alerting owing to changes

ssaria@bayesianhealth.com .
Dr. Finlayson and Mr. Subbaswamy contributed equally to this letter.
Disclosure forms provided by the authors are available with the full text of this letter at NEJM.org.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

in patients' demographic characteristics associated with the coronavirus disease 2019 pandemic. This was a case in which dataset shift fundamentally altered the relationship between fevers and bacterial sepsis, leading the hospital's clinical AI governing committee (which one of the authors of this letter chairs) to decommission its use. This is an extreme example; many causes of dataset shift are more subtle. In Table 1, we present common causes of dataset shift, which we group into changes in technology (e.g., software vendors), changes in population and setting (e.g., new demographics), and changes in behavior (e.g., new reimbursement incentives); the list is not meant to be exhaustive.

Successful recognition and mitigation of dataset shift require both vigilant clinicians and sound technical oversight through AI governance teams.[4,5] When using an AI system, clinicians should note misalignment between the predictions of the model and their own clinical judgment, as in the sepsis example above. Clinicians who use AI systems must frequently consider whether relevant aspects of their own clinical practice are atypical or have recently changed. For their part, AI governance teams must be sure that it is easy for clinicians to report concerns about the function of AI systems and provide feedback so that the clinician who is reporting will understand that the registered concern has been noted and, if appropriate, actions to mitigate the concern have been taken. Teams must also establish AI monitoring and updating protocols that integrate technical solutions and clinical voices into an AI safety checklist, as shown in Table 1.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Supported by grants from the Food and Drug Administration (5 U01 FD005942–05), the Sloan Foundation (FG-2018–10877), the National Science Foundation (1840088), and the National Institute of General Medical Sciences (T32GM007753).

References

1. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med 2019; 380: 1347–58. [PubMed: 30943338]
2. Ross C Hospitals are using AI to predict the decline of Covid-19 patients — before knowing it works. Stat 4 24, 2020 (https://www.statnews.com/2020/04/24/coronavirus-hospitals-use-ai-to-predict-patient-decline-before-knowing-it-works/).
3. Artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD) action plan Silver Spring, MD: Food and Drug Administration, January 2021 (https://www.fdagov/media/145022/download).
4. Subbaswamy A, Saria S. From development to deployment: dataset shift, causality, and shift-stable models in health AI. Biostatistics 2020; 21: 345–52. [PubMed: 31742354]
5. Saria S, Subbaswamy A. Tutorial: safe and reliable machine learning. In: Proceedings of the conference on fairness, accountability, and transparency, January 29–31, 2019. New York: Association for Computing Machinery, 2019. DOI: 10.1056/NEJMc2104626

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 1.

Overview of Our Recommended Approach to Recognizing and Mitigating Dataset Shift.*

Dataset Shift Category and Checklist Considerations	Examples of Dataset Shift	Recognition Strategies	Mitigation Strategies
Changes in technology			
Are there new types of data-acquisition devices upstream from the model?	A CAD model developed to predict hip fractures was shown to rely on specific radiographic scanner models and technicians. The adoption of high-sensitivity troponin assays changes clinical interpretation of detectable troponin levels.	Governance committee: For new implementations, check for differences in input-device types between what the model expects and what is being used in the current care environment. For ongoing monitoring, proactively identify when data-acquisition devices or protocols change. Frontline clinicians: Flag when there are changes in data-acquisition protocols.	When new input devices are added, model outputs are checked for validity and models are retrained or tuned if needed.
Are there new IT practices (e.g., terminologies used to store data) upstream from the model?	A model developed with diagnoses defined with the use of ICD-9 codes may not be accurate in hospitals that have adopted ICD-10 because of differences in definitions.	Governance committee: Routine IT protocols should flag all institution-wide IT changes that are upstream from clinical predictive models. Frontline clinicians: Flag changes in IT and electronic documentation practices (e.g., new templates) that may be missed by IT.	Retrain models in which data cannot be directly mapped from the previous format.
Is there new IT software or infrastructure (e.g., EHR systems) on which the model relies?	Adopting a new EHR platform (or module) or even routine updates to an existing platform can cause models to malfunction. For example, routine EHR updates may result in internal changes in variable definitions that may inadvertently change definitions of predictors that lead to incorrect model predictions.	Governance committee: Before deployment of new EHR platforms, carefully review variable mapping for predictive models (similar to the process followed for clinical decision support alerts). After deployment of new EHR platforms, rigorously monitor for statistical changes in the inputs to or outputs of predictive models. Frontline clinicians: Flag inadvertent errors in variable mappings introduced during HER updates. Flag models that appear to have changed in behavior for one or more patient populations after EHR update.	When model behavior changes after a major IT update, multidisciplinary rootcause analysis may identify updates for variable mappings, require model retraining, or both.
Changes in population and setting			
Is the model being applied to new clinical demographics?	Models trained in predominantly White populations may underperform on patients from underrepresented racial or ethnic groups. Patient populations may change within a given health system through mergers. For example, an urban hospital that acquires primary care practices in a rural area may have changes in hospitalized population demographics.	Demographic characteristics of the population in which the model was developed are typically available in a peer-reviewed publication or model information sheet. Model vendors will commonly provide updated local performance measures. Governance committee: Carefully monitor baseline characteristics of populations in which clinical models are deployed, including demographic and phenotypic breakdowns. Flag patient populations (on the basis of demographic characteristics, coexisting conditions, or both) for whom predictive models have poorer accuracy. Frontline clinicians: Report to the AI governance committee patient demographics that differ from those commonly seen by their service (e.g., visitors from another country) to request verification that the algorithm has been evaluated on this population.	Retrain or redesign models with the use of more inclusive data sets and with careful attention to accuracy across subgroups. Specialized algorithms can detect and adapt when data from new populations arise.
Is the model being deployed in a new clinical practice setting?	Models developed in academic or specialty settings may not generalize to community use.	Governance committee: Consider “locally validating” models by running them silently first (without showing the output to clinicians) when rolling out to new clinical contexts. Frontline clinicians: Flag models whose outputs appear to be less sensible when applied — for example, in outpatient as compared with inpatient settings.	Model retraining or tuning with additional data from new deployment contexts. Shift-stable learning algorithms can often be adopted that are insensitive to site-specific biases.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Dataset Shift Category and Checklist Considerations	Examples of Dataset Shift	Recognition Strategies	Mitigation Strategies
Have new treatments or standard of care been implemented for patients and diseases for whom the model is applied?	Statin therapies result in miscalibration of cardiovascular predictive models.	Governance committee: Monitor model accuracy and calibration. Frontline clinicians: Flag models that begin to systematically overpredict or underpredict risk owing to a shifting standard of care.	Retrain models with data obtained after the adoption of new therapies.
Have there been changes in disease incidence among patients for whom the model is applied?	A CAD model for chest-radiographic interpretation had a poor ability to generalize across hospitals with different underlying rates and types of pneumonia.	Governance committee: Monitor distribution of diagnoses over time, as well as model accuracy and calibration. Use monitoring solutions that automatically flag shifts that lead to deterioration in model performance. Frontline clinicians: Flag models that begin to systematically overpredict or underpredict risk for specific clinical populations.	Recalibrate models in light of shifting incidence. Retrain models if necessary.
Is the clinical application of the model affected by seasonality?	Over- or underreliance on seasonal trends for diseases such as influenza can result in model errors.	Governance committee: Monitor for seasonal patterns in model performance. Frontline clinicians: Flag models that appear to overpredict or underpredict during specific seasons.	Retrain models to account for seasonality, or deploy distinct models at different times of year.
Has the clinical application of the model been affected by new diseases or other unexpected “black swan” events?	The Google Flu Trends product failed to capture the swine flu epidemic.	Governance committee: Monitor model performance and establish open channels for clinician reports. Frontline clinicians: Flag models that may be affected by recent unexpected events.	Mitigation measures (temporary model deactivation, model retraining) will depend on the specific cause of the problem.
Changes in behavior			
Have new clinical behavioral incentives arisen that influence the data on which the model is applied?	Differential reimbursement of sepsis relative to other causes of death has resulted in a measurable rise in documented diagnosis of sepsis.	Governance committee: Monitor model accuracy and calibration. Solicit feedback on major forthcoming changes in coding practices from clinical and administrative groups. Frontline clinicians: Flag models that depend on diagnostic codes, because the choice of a specific code for a condition may have changed since model training.	Retrain or tune models, as needed.
Have changes in patient behavior arisen that influence the data on which the model is applied?	After the diagnosis of a highprofile celebrity, patients may seek diagnostic evaluation with fewer or no symptoms.	Governance committee: Review and assess implicit underlying behavioral assumptions of any AI model. (Models that predict health behavior may issue predictions with disproportionate effects on vulnerable populations even in the absence of dataset shift.) Frontline clinicians: Flag models that may be affected by patient behavioral trends noted in the clinic or in the literature.	Retrain or redesign models as necessary to account for dynamic patient behavior.
Have changes in clinical practice arisen that influence the data on which the model is applied?	Adoption of new order sets, or changes in their timing, can heavily affect predictive model output. Surgical skin markings affect the accuracy of dermatology classifiers, a practice that varies according to clinical setting.	Governance committee: Coordinate with health system leadership (e.g., chief medical officer), clinical departments or groups (e.g., internal medicine), or health system committees (e.g., cardiopulmonary resuscitation committee) to flag major institutional changes in practice patterns. Use monitoring solutions that automatically flag high-risk scenarios. Frontline clinicians: Flag subtle changes in practice patterns that may be relevant to clinical predictive models.	Retrain or redesign (e.g., predictor redefinition) in light of new practices. Shiftstable learning algorithms can often correct for biases related to practice patterns.
Have changes in clinical nomenclature arisen that influence the data on which the model is applied?	Formal reclassification of disorders, such as the creation of autism spectrum disorders under the DSM-5, requires updating of models operating on clinical text or diagnostic codes. Competing guidelines for sepsis phenotyping result in variance across hospitals and over time.	Governance committee: Coordinate with clinical committees (e.g., hospital sepsis committee) to recheck model performance when clinical criteria meaningfully change for a condition being predicted by a model. Frontline clinicians: Flag relevant models for reassessment when clinical societies or new literature results in new nomenclature.	Retraining or redesign will probably be necessary to account for new nomenclature.
Has the AI system induced behavioral	Overreliance on a CAD system for mammography worsened the	Governance committee: Support ongoing clinical education for clinicians and clinical	Recalibrate or retrain models over time to

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Dataset Shift Category and Checklist Considerations	Examples of Dataset Shift	Recognition Strategies	Mitigation Strategies
changes that affect how it is used?	sensitivity of human radiologists to disease (automation bias).	departments using any AI model to ensure that they understand how to correctly use any such model and specifically how not to use it. Use automated monitoring solutions to check for under- and overreliance on AI. Frontline clinicians: Understand the intended use of any AI system and strive to remain vigilant for cognitive biases.	account for behavioral changes.

* With a principled approach to the various causes of dataset shift, informed clinicians and artificial intelligence (AI) governance committees can partner with system developers to implement best practices. General recommendations include the following: establish a governance committee with multidisciplinary expertise in the AI system and how it will be used clinically, partner with solution developers in implementing a checklist and an ongoing monitoring process that evaluates for AI malfunction risk from dataset shift, and implement a process for frontline staff to flag scenarios in which there may be concern for a dataset shift in order to facilitate a more formal review process by the governance committee.4,5 Additional discussion and references for all examples are provided in Table S1 in the Supplementary Appendix, available with the full text of this letter at NEJM.org. CAD denotes computer-aided diagnostic, DSM-5 *Diagnostic and Statistical Manual of Mental Disorders*, fifth edition, EHR electronic health record, ICD-9 *International Classification of Diseases, 9th Revision*, ICD-10 *International Classification of Diseases, 10th Revision*, and IT information technology.