Supplement to “The Clinician and Dataset Shift in Artificial Intelligence”
by Samuel G Finlayson, Adarsh Subbaswamy, Karandeep Singh, John Bowers, Annabel Kupke, Jonathan Zittrain, Isaac S Kohane, and Suchi Saria

https://www.nejm.org/doi/full/10.1056/NEJMc2104626

Table of contents:

Table S1: Expanded version of main text Table 1 with additional details and references.

<table>
<tr><td colspan="2"><b>Overview of identification and mitigation approach</b>[1,2]</td><td colspan="3">● 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 ongoing monitoring process [3] that periodically evaluates for AI malfunction risk from dataset shift related to the categories listed below (e.g., new data acquisition devices, new information technology (IT) practices and so on).
● Implement a process for frontline staff to flag scenarios where there may be concern for a dataset shift in order to facilitate a more formal review process by the governance committee.</td></tr>
<tr><th>Checklist Considerations</th><th>Example</th><th>Comments and References</th><th>Recognition strategies</th><th>Mitigation Strategies</th></tr>
<tr><td colspan="5"><i>Dataset Shift Category: Changes in Technology</i></td></tr>
<tr><td>Are there <b>new data acquisition device types</b> upstream from the model?</td><td>A computer-aided diagnostic (CAD) model developed to predict hip fractures was shown to rely on specific x-ray scanner models and technicians.</td><td>Badgely et al [4] investigated a diagnostic model for hip fractures built on deep learning. Even when fed only raw radiographs as input, the algorithm had learned to detect and depend heavily upon clinical confounding factors such as scanner model, scanner brand, and radiograph order date. When patient images were matched on these variables, diagnostic performance dropped significantly. This demonstrated that sudden changes in scanner model (or other healthcare process variables) could result in the sudden malfunction of such an AI system.</td><td><i>Governance Committee:</i>
For new implementations, check for differences in input device types between what the model expects versus what is being used in the current care environment.
For ongoing monitoring,</td><td>When new input devices are added, model outputs are checked for validity and models are retrained or tuned if needed.</td></tr>
</table>

			proactively identify when data acquisition devices or acquisition protocols change. *Frontline clinicians:* Flag when there are changes in data acquisition protocols.	
	The adoption of high-sensitivity troponin assays changes clinical interpretation of detectable troponin levels.	Vaz et al [5] describe the challenges faced by *human* clinicians in adapting to new high-sensitivity troponin assays. Previous versions of the laboratory test were positive almost exclusively in the setting of acute myocardial infarction, whereas newer tests can be positive in a variety of settings. AI models need to be updated to account for this and other new lab assays in the same way.	*Governance Committee:* For new implementations, check for differences in input device types between what the model expects versus what is being used in the current care environment. For ongoing monitoring, proactively identify when data acquisition devices or acquisition	When new input devices are added, model outputs are checked for validity and models are retrained or tuned if needed.

			protocols change. *Frontline clinicians:* Flag when there are changes in data acquisition protocols.	
Are there **new IT practices** (e.g., terminologies used to store data) upstream from the model?	A model developed with diagnoses defined using ICD-9 codes may not be accurate in hospitals that have adopted ICD-10 because of differences in definitions.	Ellis et al [6] analyzed insurance claims data from more than 18 million adults around the time of transition from ICD-9 to ICD-10, and found that 1 in 6 of diagnostic categories experienced an instantaneous shift in prevalence of more than 20%. As the authors conclude, “diagnostic classification systems developed with ICD-9-CM data may need to be refined for use with ICD-10-CM data for disease surveillance, performance assessment, or risk-adjusted payment.”	*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 whose data cannot be directly mapped from previous format.
Are there **new IT software/infrastructure** (e.g., EHR	Adopting a new EHR platform (or module) or even	Nestor et al [7] describe an ML algorithm that was successfully trained to predict mortality in ICU patients using laboratory	*Governance Committee:*	When model behavior changes after a major IT update,

Systems) on which the model relies?	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.	tests and vital signs, but experienced a sudden deterioration in performance. The researchers revealed that the drop in accuracy occurred when the data management software was switched from CareVue to MetaVision, subtly altering how even identical clinical measurements would be recorded in the electronic health record (EHR) and destroying precise relationships on which the model relied. A modification to the data input then allowed the model to generalize across vendors.	Prior to 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 EHR updates. Flag models that appear to have changed in behavior for one or more patient populations after EHR update.	multidisciplinary root cause analysis may identify updates for variable mappings, and/or require model retraining.

Dataset Shift Category: 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.	Adamson and Smith[8] provide an excellent and clinically-accessible primer on the risk of health disparities that emerges when specific populations are excluded from an ML algorithm's training data. They point to several real-world examples of this phenomenon, which amount to a special case of dataset shift.	Demographics for 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 on which clinical models are deployed -- including demographic and phenotypic breakdowns. Flag patient populations (demographic and/or comorbidity-based) for whom predictive models exhibit poorer	Retrain and/or redesign models using more inclusive datasets with careful attention to accuracy across subgroups. Specialized algorithms can detect and adapt when data from new populations arise.

			accuracy [3].	
Is the model being deployed in a new **clinical practice setting**?	Models developed in academic or specialty settings may not generalize to community use.	Many papers have generalized this phenomenon. One prominent example is the CC-Cruiser AI system for cataract diagnosis, which demonstrated high accuracy (>98%) in an initial pilot study, but showed worse performance (<89%) when tested in a multi-hospital trial that included various clinical practice settings across China[9]. Similarly, Norgeot et al [10] developed an algorithm for predicting rheumatoid arthritis outcomes using data from a university hospital, and saw AUROC decrease from 0.91 to 0.74 when tested in a public safety-net hospital.	*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 versus inpatient settings.	Model retraining/tuning with additional data from new deployment contexts. Shift-stable learning algorithms can often be adopted that are insensitive to site-specific biases [11,12].
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.	Pate et al [13] analyzed cardiovascular disease risk predictions generated by a predictive model, and found significant miscalibration develop over time due to shifting trends in cardiovascular disease that could be best attributed to the adoption of statins. Other examples may occur primarily on the consumer-side: for example, Hajek et	*Governance Committee:* Monitor model accuracy and calibration.	Retrain models with data from after the adoption of new therapies.

		al [14] shows improvements in smoking cessation among e-cigarrete users, which could have great impact on clinical predictive models for smoking cessation such as that of Luo et al.	*Frontline clinicians:* Flag models that begin to systematically overpredict or underpredict risk due to shifting standard of care.	
Have there been **changes in disease incidence** in patients for whom the model is applied?	A CAD model for chest-xray interpretation exhibited a poor ability to generalize across hospitals with different underlying rates of pneumonia.	Zech et al [15] conducted a thorough analysis of convolutional neural networks trained to analyze chest radiographs for detection of pneumonia. Even when trained on pooled data from multiple sites with different rates of pneumonia, they found that their models failed to generalize to external data from hospitals with different incidences.	*Governance Committee:* Monitor distribution of diagnoses over time, as well as model accuracy and calibration. Employ monitoring solutions that automatically flag shifts leading to deterioration in model performance.[3,16] *Frontline clinicians:* Flag models that begin to systematically overpredict or underpredict risk for specific clinical populations.	Recalibrate models in light of shifting incidence. Re-train models if necessary. Shift-stable learning algorithms can often be adopted that are insensitive to site-specific biases [1,12].

Is the model's clinical application affected by **seasonality**	Influenza cases spike in winter so incidence varies by month.	The Google Flu Trends product illustrates several phenomena, described by Lazer et al [17] as the "Parable of Google Flu." Google Flu Trends, a product developed to predict U.S. flu burden, exhibited an over-reliance on simple seasonal trends and an under-utilization on additional available public data. This caused the model to steadily worsen over time and grow to markedly overestimate flu prevalence. The model was especially unprepared to detect the non seasonal influenza A–H1N1 pandemic, which it completely missed.	*Governance Committee:* Monitor model performance, establish open channels for clinician reports. *Frontline clinicians:* Flag models that may be affected by recent unexpected events.	Retrain models to account for seasonality, or deploy distinct models at different times of year.
Has the model's clinical application been affected by **new diseases or other unexpected "black swan" events**?	Google Flu showed high accuracy in monitoring influenza, but failed completely to capture the influenza A–H1N1 pandemic			Mitigation measures (temporary model deactivation, model retraining) will depend on the specific etiology of the problem.
Dataset Shift Category: 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.	Gohil et al[18] investigated the pattern of sepsis diagnoses before and after the Centers of Medicare and Medicaid Services (CMS) introduced new sepsis codes and and medical severity diagnosis-related group (MS-DRG) systems. They found that these policies resulted effectively overnight in a 2-fold increase in non-severe sepsis, a 2.8-fold increase in severe sepsis, a 3.8-fold increase in sepsis on admission. Concomitantly, the reported mortality in this population decreased. Such policy-induced changes in clinical practice and reporting have clear ramifications for sepsis predictive	*Governance Committee:* Monitor model accuracy and calibration. Solicit feedback on major forthcoming changes in coding practices from clinical and administrative groups.	Use of high-quality clinical phenotypes independent of billing practices can ensure models that are stable to coding-related shifts.[19,20] Retrain or tune models, as needed.

		models, which are one of the more common use-cases for clinical predictive modeling.[19,20]		
Have new **changes in patient behavior** arisen that influence the data on which the model is applied?	Following the diagnosis of a high-profile celebrity, patients self-refer for diagnostic evaluation with fewer or no symptoms.	The "Angelina Jolie effect," in which a high profile celebrity diagnosis led to increased patient self-referral for genetic testing for breast cancer, is an example where public opinion and patient behavior changed quickly about a specific health behavior.[21]	*Governance Committee:* Review and assess implicit underlying behavioral assumptions of any AI model. (Note: Models predicting health behavior may issue predictions with disproportionate impacts 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 new **changes in clinical practice** arisen that influence the data	Adoption of new order sets, or changes in their timing, can heavily impact predictive	Agniel et al [22], found that in up to 86% of lab tests, the timing of the laboratory test order was more important to a clinical predictive model than the resultant value of the test. Thus, changes in lab timing	*Governance Committee:* Coordinate with health system	Retrain or redesign (e.g. predictor redefinition) in light of new practices.

on which the model is applied?	model output.	could heavily alter model output. More broadly, many clinical predictive models rely on specific lab orders. If those lab orders are placed routinely as part of an order set, they may artificially inflate patient risk scores.	leadership (e.g., chief medical officer), clinical departments/groups (e.g., internal medicine), or health system committees (e.g., cardiopulmonary resuscitation committee) to flag major institutional changes in practice patterns. Employ monitoring solutions that automatically flag high risk scenarios.[3,16] *Frontline clinicians:* Flag subtle changes in practice patterns that may be relevant to clinical predictive models.	Shift-stable learning algorithms can often correct for biases related to practice patterns [11,12].
	Surgical skin markings impact the accuracy of dermatology classifiers, a practice which varies by clinical setting.	Winkler et al [23] investigated a commercially-available deep neural network for the detection of melanoma, and analyzed the diagnostic performance on the same lesions before and after they were marked with a surgical skin marker. They found that the specificity of the algorithm differed by up to 40% depending on the skin marker. This implies that such models could behave very differently when used by dermatologists versus primary care physicians or patients themselves, who may mark the lesions at different rates or with different techniques.		
Have new **changes in clinical nomenclature** arisen that influence the data on which the model is applied?	Guidelines for sepsis phenotyping changed over the last decade to incorporate more granular clinical criteria.	Saria and Henry [20] describe the large variance in meaning of a “sepsis” diagnosis as a result of many competing clinical criteria. As hospitals shift their diagnostic criteria, clinical predictive models will need to be adjusted accordingly.	*Governance Committee:* Coordinate with clinical committees (e.g., hospital sepsis committee) to recheck model performance when	Retraining or redesign will likely be necessary to account for new phenotypes. Retraining or redesign will likely be necessary to account for new nomenclature.

	Formal reclassification of disorders, such as the creation of autism spectrum disorders under the DSM-5, require updating of models operating on clinical text or diagnostic codes.	Nalfon and Kuo [24] and others have described the significant impact of the redefinition of “autism” by the DSM-5, which reclassification by the DSM-5 of Autism and associated conditions such as Asperger syndrome into a single Autism Spectrum Disorder. Models operating on clinical notes or diagnostic codes need to be adjusted whenever terminology changes so substantially.	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 result in new nomenclature.	
Has the **AI-system induced behavioral changes** that affect how it is used?	Automation bias: Overreliance on a CAD system for mammography worsened the sensitivity of human radiologists to disease.	Lyell and Coiera [25] provide an overview of the phenomenon of automation bias, one of many ways in which the users of an automated system can become over reliant on decision support. CAD systems for mammography provide a case study in automation bias, because they appeared to improve human clinician performance based on preliminary studies, but some subsequent analyses -- such as Lehman et al[26] and Alberdi et al.[27]. -- demonstrated decreased human performance when using CAD.	*Governance Committee:* Support ongoing clinical education for clinicians and clinical departments using any AI model, to ensure that they understand how to correctly use any such model, and specifically how not to use it. Employ automated monitoring solutions to check for under- and over-reliance on AI.[3,16]	Recalibrated or retrained models over time to account for behavioral changes.

			Frontline clinicians: Understand the intended use of any AI system, and strive to remain vigilant for cognitive biases.	

References

1. Subbaswamy A, Saria S. From development to deployment: dataset shift, causality, and shift-stable models in health AI. Biostatistics 2019;21(2):345–52.

2. Saria S, Subbaswamy A. Tutorial: Safe and Reliable Machine Learning. In: FAT* ’19: Proceedings of the Conference on Fairness, Accountability, and Transparency. New York: Association for Computing Machinery (ACM); 2019. https://arxiv.org/abs/1904.07204.

3. Subbaswamy A, Adams R, Saria S. Evaluating Model Robustness and Stability to Dataset Shift. In: Proceedings of the 24th International Conference on Artificial Intelligence and Statistics. PMLR; 2021. p. 2611–9.

4. Badgeley MA, Zech JR, Oakden-Rayner L, et al. Deep learning predicts hip fracture using confounding patient and healthcare variables. NPJ Digit Med 2019;2:31.

5. Vaz HA, Guimaraes RB, Dutra O. Challenges in high-sensitive troponin assay interpretation for intensive therapy. Rev Bras Ter Intensiva 2019;31(1):93–105.

6. Ellis RP, Hsu HE, Song C, et al. Diagnostic Category Prevalence in 3 Classification Systems Across the Transition to the International Classification of Diseases, Tenth Revision, Clinical Modification. JAMA Netw Open 2020;3(4):e202280.

7. 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. In: Proceedings of the Machine Learning for Healthcare Conference. PMLR; 2019. p. 381–405.

8. Adamson AS, Smith A. Machine Learning and Health Care Disparities in Dermatology. JAMA Dermatol 2018;154(11):1247–8.

9. Lin H, Li R, Liu Z, et al. Diagnostic Efficacy and Therapeutic Decision-making Capacity of an Artificial Intelligence Platform for Childhood Cataracts in Eye Clinics: A Multicentre Randomized Controlled Trial. EClinicalMedicine 2019;9:52–9.

10. Norgeot B, Glicksberg BS, Trupin L, et al. Assessment of a Deep Learning Model Based on Electronic Health Record Data to Forecast Clinical Outcomes in Patients With Rheumatoid Arthritis. JAMA Netw Open 2019;2(3):e190606.

11. Schulam P, Saria S. Reliable Decision Support using Counterfactual Models. In: Proceedings of the 31st Conference on Neural Information Processing System. Curran Associates Inc.; 2017. p. 1696–706.

12. Subbaswamy A, Schulam P, Saria S. Preventing Failures Due to Dataset Shift: Learning Predictive Models That Transport. In: Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics. PMLR; 2019. p. 3118–27.

13. Pate A, van Staa T, Emsley R. An assessment of the potential miscalibration of cardiovascular disease risk predictions caused by a secular trend in cardiovascular disease in England. BMC Med Res Methodol 2020;20(1):289.

14. Hajek P, Phillips-Waller A, Przulj D, et al. A Randomized Trial of E-Cigarettes versus Nicotine-Replacement Therapy. N Engl J Med 2019;380(7):629–37.

15. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLoS Med 2018;15(11):e1002683.

16. Nushi B, Kamar E, Horvitz E, Kossmann D. On Human Intellect and Machine Failures: Troubleshooting Integrative Machine Learning Systems. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. AAAI Press; 2017.

17. Lazer D, Kennedy R, King G, Vespignani A. Big data. The parable of Google Flu: traps in big data analysis. Science 2014;343(6176):1203–5.

18. Gohil SK, Cao C, Phelan M, et al. Impact of Policies on the Rise in Sepsis Incidence, 2000-2010. Clin Infect Dis 2016;62(6):695–703.

19. Henry KE, Hager DN, Osborn TM, Wu AW, Saria S. Comparison of Automated Sepsis Identification Methods and Electronic Health Record–based Sepsis Phenotyping: Improving Case Identification Accuracy by Accounting for Confounding Comorbid Conditions. Critical Care Explorations 2019;1(10). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7063888/. 10.1097/CCE.0000000000000053

20. Saria S, Henry KE. Too Many Definitions of Sepsis: Can Machine Learning Leverage the Electronic Health Record to Increase Accuracy and Bring Consensus? Crit Care Med 2020;48(2):137–41.

21. Evans DG, Wisely J, Clancy T, et al. Longer term effects of the Angelina Jolie effect: increased risk-reducing mastectomy rates in BRCA carriers and other high-risk women. Breast Cancer Res. 2015;17:143.

22. Agniel D, Kohane IS, Weber GM. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. BMJ 2018;361:k1479.

23. Winkler JK, Fink C, Toberer F, et al. Association Between Surgical Skin Markings in Dermoscopic Images and Diagnostic Performance of a Deep Learning Convolutional Neural Network for Melanoma Recognition. JAMA Dermatol 2019; http://dx.doi.org/10.1001/jamadermatol.2019.1735. 10.1001/jamadermatol.2019.1735

24. Halfon N, Kuo AA. What DSM-5 could mean to children with autism and their families. JAMA Pediatr 2013;167(7):608–13.

25. Lyell D, Coiera E. Automation bias and verification complexity: a systematic review. J Am Med Inform Assoc 2017;24(2):423–31.

26. Lehman CD, Wellman RD, Buist DSM, et al. Diagnostic Accuracy of Digital Screening Mammography With and Without Computer-Aided Detection. JAMA Intern Med 2015;175(11):1828–37.

27. Alberdi E, Povykalo A, Strigini L, Ayton P. Effects of incorrect computer-aided detection (CAD) output on human decision-making in mammography. Acad Radiol 2004;11(8):909–18.