Biostatistics (2020) **21**, 2, *pp.* 345–352
doi:10.1093/biostatistics/kxz041
Advance Access publication on November 19, 2019

From development to deployment: dataset shift, causality, and shift-stable models in health AI

ADARSH SUBBASWAMY

Department of Computer Science, Johns Hopkins University, 160 Malone Hall, 3400 N. Charles Street, Baltimore, MD, USA

SUCHI SARIA*

Department of Computer Science; Department of Applied Math & Statistics, and Department of Health Policy & Management, Johns Hopkins University, 160 Malone Hall, 3400 N. Charles Street, Baltimore, MD, USA

ssaria@cs.jhu.edu

Keywords: Causal inference; Dataset shift; Generalizability; Machine learning.

The deployment of machine learning (ML) and statistical models is beginning to transform the practice of healthcare, with models now able to help clinicians diagnose conditions like pneumonia and skin cancer, and to predict which hospital patients are at risk of adverse events such as septic shock. A major concern, however, is that model performance is heavily tied to details particular to the dataset the model was developed on—even slight deviations from the training conditions can result in wildly different performance. For example, when researchers trained a model to diagnose pneumonia from chest X-rays using data from one health system, but evaluated on data from an external health system, they found the model performed significantly worse than it did internally (Zech *and others*, 2018). The model failed to *generalize* (i.e., predict accurately) due to the *shifts* between the training conditions (health system one) and the deployment/testing conditions (health system two). These shifts are very common when moving a model from the training phase to deployment and can take a variety of forms, including changes in patient demographics, disease prevalence, measurement timing, equipment, treatment patterns, and more. Beyond contributing to poor performance, failing to account for shifts can also lead to dangerous decisions in practice: the system can fail to diagnose severely ill patients or recommend harmful treatments. This problem of shifting conditions which prevent generalization is referred to as *dataset shift* (Quiñonero-Candela *and others*, 2009), and in this article, we explain what it is, why it occurs, give an overview of the types of existing solutions, and discuss open challenges that remain.

Generalization is crucial for successfully deploying models since we want model predictions to be accurate when applied to new situations that were not in the training dataset. In order to ensure that a

*To whom correspondence should be addressed.

model will generalize, a core requirement is to check that the model or learning procedure satisfies the *stability* property: is model performance robust when the data are perturbed in various ways? In particular, when addressing dataset shift, we want models to be stable to perturbations in (i.e., shifts in) how the data were generated—such as changes to patient demographics or clinician treatment patterns—as opposed to leave-one-out sampling perturbations (Yu, 2013; Giordano *and others*, 2019; Yu and Kumbier, 2019) or adversarial perturbations of model inputs (Madry *and others*, 2018; Hendrycks and Dietterich, 2019). Dataset shift is of particular importance because it frequently occurs when deploying models, is difficult to test stability against (as we will discuss later), and is highly relevant to the ongoing discussion about the challenges of regulating ML-driven medical devices (U.S. Food and Drug Administration, 2019) (see accompanying Stern and Price (2020) for more on this).

Consider the mortality risk prediction model trained by Caruana *and others* (2015) on a dataset of hospitalized pneumonia patients, using information such as lab measurements, vital signs, and comorbidities. While the model had high predictive accuracy on one dataset, it was unstable to shifts in the choices driving which patients get admitted to the ICU versus the floor. As a result, when they evaluated it for triaging pneumonia patients upon ED presentation, they found that their model incorrectly predicted lower risk for patients with pneumonia and asthma versus those with only pneumonia. This shift in ICU admission policy, while subtle, had big implications: had the model been deployed for triage, it would have greatly endangered asthmatic pneumonia patients by suggesting they should be sent home. In machine learning, a *policy* is defined as a distribution over the possible actions that can be taken in any given scenario. Even subtle shifts in when or whether or not a lab was ordered can impact predictions—these forms of shifts are called *policy shifts* (Schulam and Saria, 2017). More generally, dataset shifts can come in various forms (from shifts in patient demographics to policy shifts in measurement frequency), which makes dataset shift challenging to address.

Given that dataset shift is highly varied and can lead to dangerous failures if left unaccounted for, what should we do to address it? Typically, learning methods are developed on a specific dataset. A common practice for moving the model to a different setting is to adapt or re-learn using a dataset obtained from this new setting. The challenge with this approach is that it is unrealistic to assume that data from all possible deployment settings are available upfront. Within a *failure prevention* paradigm, developers anticipate and guard against likely sources of shifts between environments during model learning without the need for data from each setting (Saria and Subbaswamy, 2019). This involves three main components: (i) determining likely shifts we want models to be robust to, (ii) testing the stability of a model to those shifts, and (iii) employing learning methods that come with stability guarantees regarding how the model will perform under those shifts. In order to successfully perform any component, we need a technical language for identifying and expressing exactly what conditions are shifting across settings.

1. REPRESENTING AND REASONING ABOUT DATASET SHIFT USING GRAPHS

To identify the likely shifts we want to protect against, we need to understand how the variables in the data were generated and what aspects of this process are vulnerable to shifts across environments. Consider Figure 1, which shows how the variables in a clinical mortality prediction model are related. An edge, such as the one from Asthma to Pneumonia, denotes that one variable generates or causes another (e.g., asthma is a risk factor for pneumonia). By placing edges as needed between the variables, we can express the various physiologic and decision factors and how these are related: the values of lab measurements and vital signs like blood pressure are affected by illnesses such as pneumonia, and all of these determine a patient's risk of mortality. The resulting causal directed acyclic graph represents the *data generating process* and encodes how all variables are related.

Using graphs, complex processes can be broken down into individual components (variables and factors driving them in the graph) that are inspected for vulnerability to shifts. In Figure 1, we identify

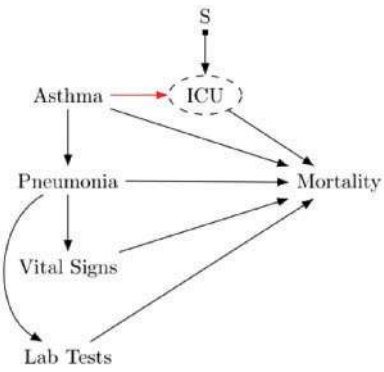

Fig. 1. Graph for the pneumonia mortality prediction example. Red denotes an unstable edge. Dashed node is unobserved. Color figures are available at *Biostatistics* online.

the component corresponding to ICU admission criteria as likely to shift between datasets. This can be visually distinguished with a colored edge or through the addition of auxiliary *selection variables* (Pearl and Bareinboim, 2011; Subbaswamy *and others*, 2019b) (denoted by square S nodes) which point to variables whose generative process can shift across datasets (the $S \rightarrow ICU$ edge). The visual representation also naturally provides a statistical characterization of the shift, which we will see later is important for deriving methodological solutions. Each component in the graph corresponds to a conditional distribution (of the variable given its parents in the graph), so when we identify that the ICU admission policy is likely to shift, we mean that we expect $P(ICU \mid Asthma)$ to differ across datasets. Importantly, using a graphical language, we can express richer extensions of commonly considered instances of dataset shift, such as *covariate shift* (Shimodaira, 2000) and *target/label shift* (Storkey, 2009; Zhang *and others*, 2013).

2. Checking a model's susceptibility to shifts

Graphical representations of shifts help us check susceptibility to shifts, which is important because existing methods for testing stability are limited. Current approaches are primarily empirically based: Zech *and others* (2018) trained and validated a pneumonia diagnosis model on one dataset and compared the performance when applied to datasets from new medical centers, and Nestor *and others* (2019) trained a mortality prediction model on data collected from one hospital in 2001 and measured its performance on data collected in subsequent years at that hospital. Ideally, under the failure prevention paradigm, we want to train models that are robust to prespecified types of shifts (more on this in the next section), or at least be able to test robustness to the prespecified shifts. Current empirical evaluations, however, make it difficult to consider specific shifts. Suppose we want evidence that a mortality prediction model will generalize to a different hospital, so we apply it to a new dataset and see that the performance deteriorated. It is difficult to determine exactly why the deterioration occurred, which means we cannot determine if we are robust to a particular shift. Is it because there was a shift in patient population demographics? Or perhaps there was a shift in antibiotic prescription habits or lab ordering patterns? Unless test data come from the intended deployment environment, empirical evaluations give little insight into exactly what shifts the model is (or is not) robust to, making it difficult to draw meaningful conclusions about whether the model will generalize.

A key reason graphical representations of shifts are useful is because they allow us to bypass some of these limitations. Using common graph analysis tools such as *d-separation* (Pearl, 1988), we can determine if the model we are fitting is stable (Subbaswamy and Saria, 2018). For example, Figure 2

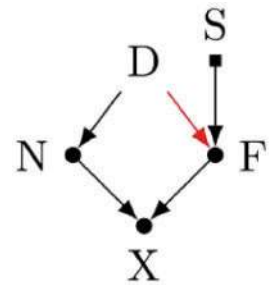

Fig. 2. Graph for the pneumonia diagnosis from chest X-rays example. Red denotes an unstable edge. Color figures are available at *Biostatistics* online.

depicts a simplified graph for the problem of diagnosing pneumonia (N) from chest X-rays (X) which contains style features (F). The style features (e.g., marks on the image which convey if the X-ray is front-to-back or back-to-front) are chosen depending on the equipment and clinician preferences in the particular hospital department (D). The selection variable (S) tells us that style feature preferences are likely to shift from dataset to dataset. To determine if a particular model (e.g., a model of $P(N|X,F)$ which learns to diagnose pneumonia from the X-ray and style features) is stable to a shift, we need to check if the target variable N is d-separated (i.e., conditionally independent) from the selection variable S given the features we are using. In this example, $P(N|X,F)$ is unstable to shifts in style preferences, while if we include the department as a feature, the model of $P(N|X,F,D)$ will be stable. Thus, graphs augment what we can currently do with empirical approaches.

3. ALGORITHMS FOR LEARNING MODELS THAT GUARANTEE STABILITY AGAINST SHIFTS

Even more desirable than providing evidence that a model will generalize, is using stable learning algorithms which allow us to train models with stability guarantees. Broadly, there are two types of stable training methods: *reactive* and *proactive* approaches. Reactive approaches use data from the intended deployment environment to correct for shifts. Common examples include many *domain adaptation* algorithms which use *importance sampling* techniques to reweight training data (Shimodaira, 2000; Sugiyama *and others*, 2006; Huang *and others*, 2007; Gretton *and others*, 2009). Unfortunately, deployment data are often difficult to collect prior to model training in modern ML applications in which there are many possible deployment environments (e.g., for cloud-based ML models) or when the deployment environment is left unspecified (in which case no additional data is available). In these settings, it is important to use proactive learning approaches which learn models that are stable to any anticipated shifts (Subbaswamy and Saria, 2018), including stable algorithms that are robust to policy shifts (Schulam and Saria, 2017).

While proactive approaches do not require data from the deployment environment(s), this brings with it new challenges. In general, learning algorithms learn everything they can from a dataset. However, when shifts occur between development and deployment, we only want to learn stable information that will generalize. This means we need a way to constrain what the model learns, and that the goal of a stable learning algorithm is to learn all stable information without learning any unstable information. Fortunately, by defining operators on a graph, we have a way to retain stable information while getting rid of unstable information. For example, in Figure 2, if we had some way to delete the red edge (denoting the unstable style feature preferences) and only this edge, and learn the rest of the information in the graph (all of which is stable), then such a model would be the optimal stable model. Thus, graphs give us a framework for comparing different stable models in terms of the *stability-accuracy tradeoff*: as we

constrain what the model can learn, its training accuracy decreases while the model's ability to generalize improves (Subbaswamy *and others*, 2019a).

More broadly, there is a hierarchy of shift-stable solutions (Subbaswamy *and others*, 2019a) (or more accurately, distributions) which correspond to three classes of operators on a graph (Pearl, 2000; Shpitser and Tchetgen Tchetgen, 2016): (i) conditioning, (ii) intervening, and (iii) computing counterfactuals. At the lowest level of the hierarchy, conditioning, we remove unstable parts of the graph by essentially deleting variables (and all paths from them). Compare this to the highest level operator, computing counterfactuals, which precisely deletes individual edges in the graph. Intervening falls between the two, and deletes all edges into a variable. Due to the increased precision of the highest level operators (levels 2 and 3), these operators are more efficient at removing unstable components from the graph while retaining as much stable information as possible. Consider Figure 2 again, and suppose that the department an X-ray was taken in is not recorded in the data (i.e., the variable D is unobserved). Using conditioning alone, there is no stable solution (we can retain no stable information about X or F without including the unstable edge). However, there are levels 2 and 3 solutions which only delete the unstable edge, and thus they are optimal in this case (Subbaswamy *and others*, 2019a).

Solutions at the highest level do have limitations, however. First, while only one model needs to be fitted for conditional solutions, levels 2 and 3 solutions require fitting multiple submodels which increases chances for model misspecification. While there has been work on *doubly robust* estimation to counteract model misspecification (Bang and Robins, 2005; Funk *and others*, 2011), thus far this work has only considered specific types of graph structures. Second, level 2 distributions are not always *identifiable* (i.e., not able to be estimated from the training data alone), and level 3 distributions require knowledge beyond the structure of the graph. Regarding the identifiability limitation, Subbaswamy *and others* (2019b) provide an algorithm which determines if an identifiable level 2 (or level 1) solution exists, and if so, returns the optimal one.

To end this section, we want to note that there are also existing stable learning algorithms which do not require an explicit graph (though they can be cast in terms of the graphical hierarchy; see Subbaswamy *and others*, 2019a for a discussion). These include dataset-driven approaches (e.g., Rojas-Carulla *and others*, 2018; Magliacane *and others*, 2018) which learn stable models using data from many environments, and *bounded magnitude distributionally robust* methods (see, e.g., Rothenhäusler *and others*, 2018) which assume shifts take a particular form and have a known magnitude. Perhaps the most important implication of the hierarchy is that it provides a common framework for comparing the optimality of different solutions and for developing new algorithms.

4. Open challenges and looking forward

We have now seen the many ways that explicit graphical representations of shifts are beneficial: graphs allow us to identify and proactively declare the shifts we want to guard against, they improve our ability to check if a model will be susceptible to shifts, and they provide a framework for developing shift-stable learning algorithms. One difficulty is that determining the graph structure requires working closely with domain experts to elicit prior knowledge in a way that can be used to place and orient edges. To make this process easier, a promising direction is to combine prior knowledge with *structure learning* algorithms to partially learn the graph from data. Further, methods for performing sensitivity analysis to changes or uncertainty in the graph structure would help to reduce concerns about misspecification of the graph. Regarding testing susceptibility to shifts, the inability of current empirical evaluations to test robustness to particular shifts and their reliance on the existence of representative datasets remain fundamental limitations. New approaches which could semi-synthetically generate data under a particular shift would allow for targeted evaluation of stability to particular shifts, and would serve as useful empirical evidence that a trained model will generalize to new environments.

Finally, we want to mention that addressing dataset shift during and prior to learning is only one part of building reliable ML systems, with two other critical pieces being *monitoring* and *maintenance* (Saria and Subbaswamy, 2019). Inevitably, unanticipated shifts will occur, and deployed models become stale resulting in performance decay. Thus, an important part of handling dataset shift is to develop methods that can detect when and what type of shifts have occurred. Further, once unanticipated shifts are detected, we need maintenance strategies for updating or replacing the model. Through the combined development of failure prevention learning techniques, tests for model stability, methods for monitoring, and maintenance protocols, a comprehensive framework for handling dataset shift is on the horizon.

ACKNOWLEDGMENTS

Conflict of Interest: None declared.

REFERENCES

BANG, H. AND ROBINS, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61**, 962–973.

CARUANA, R., LOU, Y., GEHRKE, J., KOCH, P., STURM, M. AND ELHADAD, N. (2015). Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission. In: Cao, L., Zhang, C. (editors), *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY: ACM, pp. 1721–1730.

FUNK, M. J., WESTREICH, D., WIESEN, C., STÜRMER, T., BROOKHART, M. A. AND DAVIDIAN, M. (2011). Doubly robust estimation of causal effects. *American Journal of Epidemiology* **173**, 761–767.

GIORDANO, R., STEPHENSON, W., LIU, R., JORDAN, M. AND BRODERICK, T. (2019). A Swiss army infinitesimal jackknife. In: Chaudhuri, K. and Sugiyama, M. (editors), *Proceedings of the The 22nd International Conference on Artificial Intelligence and Statistics*. Naha, Okinawa, Japan, PP. 1139–1147.

GRETTON, A., SMOLA, A., HUANG, J., SCHMITTFULL, M., BORGWARDT, K. AND SCHÖLKOPF, B. (2009). Covariate shift by kernel mean matching. In: Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A. and Lawrence, N. D. (editors), *Dataset Shift in Machine Learning*. Cambridge, MA: The MIT Press, pp. 131–160.

HENDRYCKS, D. AND DIETTERICH, T. (2019). Benchmarking neural network robustness to common corruptions and perturbations. In: Levine, S., Livescu, K. and Mohamed, S. (editors), *Proceedings of the 7th International Conference on Learning Representations*. New Orleans, LA, USA: OpenReview.net, https://openreview.net/forum?id=HJz6tiCqYm.

HUANG, J., GRETTON, A., BORGWARDT, K., SCHÖLKOPF, B. AND SMOLA, A. J. (2007). Correcting sample selection bias by unlabeled data. In: Schölkopf, B. Platt, J. C. and Hoffman, T. (editors), *Advances in Neural Information Processing Systems 19*. Cambridge, MA: MIT Press, pp. 601–608.

MADRY, A., MAKELOV, A., SCHMIDT, L., TSIPRAS, D. AND VLADU, A. (2018). Towards deep learning models resistant to adversarial attacks. In: Murray, I., Ranzato, M. A. and Vinyals, O. (editors), *Proceedings of the 6th International Conference on Learning Representations*. Vancouver, BC, Canada: OpenReview.net, https://openreview.net/forum?id=rJzIBfZAb.

MAGLIACANE, S., van OMMEN, T., CLAASSEN, T., BONGERS, S., VERSTEEG, P. AND MOOIJ, J. M. (2018). Domain adaptation by using causal inference to predict invariant conditional distributions. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N. and Garnett, R. (editors), *Proceedings of Advances in Neural Information Processing Systems 31 (NIPS 2018)*. Montreal, Canada, pp. 10846–10856.

NESTOR, B., MCDERMOTT, M. B. A., BOAG, W., BERNER, G., NAUMANN, T., HUGHES, M. C., Goldenberg, A. and Ghassemi, M. (2019). Feature robustness in non-stationary health records: caveats to deployable model performance in common clinical machine learning tasks. In: Doshi-Velez, F., Fackler, J., Kale, D. C., Wallace, B. C. and

Wiens, J. (editors), *Proceedings of the 4th Conference on Machine Learning in Healthcare*. Ann Arbor, Michigan, USA.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco, CA: Morgan Kaufmann.

Pearl, J. (2000). *Causality*. Cambridge, MA: MIT Press.

Pearl, J. and Bareinboim, E. (2011). Transportability of causal and statistical relations: a formal approach. In: Burgard, W. and Roth, D. (editors), *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*. San Francisco, California, USA: AAAI Press.

Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A. and Lawrence, N. D. (2009). *Dataset Shift in Machine Learning*. Cambridge, MA: The MIT Press.

Rojas-Carulla, M., Schölkopf, B., Turner, R. and Peters, J. (2018). Invariant Models for Causal Transfer Learning. *Journal of Machine Learning Research: JMLR* **19**, 1–34.

Rothenhäusler, D., Meinshausen, N., Bühlmann, P. and Peters, J. (2018). Anchor regression: heterogeneous data meets causality. Available at https://arxiv.org/pdf/1801.06229.pdf.

Saria, S. and Subbaswamy, A. (2019). Tutorial: safe and reliable machine learning. In: Morgenstern, J. and Boyd, D. (editors), *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019*. Atlanta, GA: ACM.

Schulam, P. and Saria, S. (2017). Reliable decision support using counterfactual models. In: Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N. and Garnett, R. (editors), *Proceedings of Advances in Neural Information Processing Systems 30 (NIPS 2017)*. Long Beach, CA, pp. 1697–1708.

Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference* **90**, 227–244.

Shpitser, I. and Tchetgen Tchetgen, E. (2016). Causal inference with a graphical hierarchy of interventions. *Annals of Statistics* **44**, 2433.

Stern, A. D. and Price, W. N. (2020). Regulatory oversight, causal inference, and safe and effective health care machine learning. *Biostatistics* **21**, 363–367.

Storkey, A. (2009) When training and test sets are different: characterizing learning transfer. In: Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A. and Lawrence, N. D. (editors), *Dataset Shift in Machine Learning*. Cambridge, MA, USA: The MIT Press, pp. 2–28.

Subbaswamy, A., Chen, B. and Saria, S. (2019a). A universal hierarchy of shift-stable distributions and the tradeoff between stability and performance. Available at https://arxiv.org/pdf/1905.11374.pdf.

Subbaswamy, A. and Saria, S. (2018). Counterfactual Normalization: proactively addressing dataset shift and improving reliability using causal mechanisms. In: Globerson, A. and Silva, R. (editors). *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence*. Monterey, CA, pp. 947–957.

Subbaswamy, A., Schulam, P. and Saria, S. (2019b). Preventing failures due to dataset shift: learning predictive models that transport. In: Chaudhuri, K. and Sugiyama, M. (editors), *Proceedings of The 22nd International Conference on Artificial Intelligence and Statistics*. Okinawa, Naha, Japan, pp. 3118–3127.

Sugiyama, M., Blankertz, B., Krauledat, M., Dornhege, G. and Müller, K.-R. (2006). Importance-weighted cross-validation for covariate shift. In: Franke K., Müller KR., Nickolay B. and Schäfer R. (editors), *Pattern Recognition. Lecture Notes in Computer Science*, vol 4174. Berlin, Heidelberg: Springer.

U.S. Food and Drug Administration. (2019). *Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD)*. Silver Spring, MD, USA: U.S. FDA.

Yu, B. (2013). Stability. *Bernoulli* **19**, 1484–1500.

YU, B. AND KUMBIER, K. (2019). Three principles of data science: predictability, computability, and stability (PCS). Available at https://arxiv.org/pdf/1901.08152.pdf.

ZECH, J. R., BADGELEY, M. A., LIU, M., COSTA, A. B., TITANO, J. J. AND OERMANN, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Medicine*, **15**, e1002683.

ZHANG, K., SCHÖLKOPF, B., MUANDET, K. AND WANG, Z. (2013). Domain adaptation under target and conditional shift. In: Dasgupta, S. and McAllester, D. (editors), *Proceedings of the 30th International Conference on Machine Learning*. Atlanta, GA, USA, pp. 819–827.

[*Received September 25, 2019; revised September 25, 2019; accepted for publication September 25, 2019*]