npj | digital medicine

Perspective

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-024-01037-4

Why do probabilistic clinical models fail to transport between sites

Check for updates

Thomas A. Lasko, Eric V. Strobl & William W. Stead

The rising popularity of artificial intelligence in healthcare is highlighting the problem that a computational model achieving super-human clinical performance at its training sites may perform substantially worse at new sites. In this perspective, we argue that we should typically expect this *failure to transport*, and we present common sources for it, divided into those under the control of the experimenter and those inherent to the clinical data-generating process. Of the inherent sources we look a little deeper into site-specific clinical practices that can affect the data distribution, and propose a potential solution intended to isolate the imprint of those practices on the data from the patterns of disease cause and effect that are the usual target of probabilistic clinical models.

Those of us who build prediction models from Electronic Health Record (EHR) data commonly find that a model that works well at its original site doesn't work nearly as well at some other site[1–4]. (By *site* we mean a location in time and space, so the failure could be at the same institution, but a later time). When the new site's data is included in training and test sets, performance improves to match that of the original[3], but then performance at a third site can be back to nearly random. While this *failure to transport* is quite frustrating, we argue that we should expect this sort of behavior from all probabilistic clinical models, whether they are supervised predictive models or unsupervised discovery models, whether they use probabilities explicitly or implicitly, and regardless of how meticulous we are in their training.

Experimental errors, improper analysis, and failure to document experimental details can lead to a *failure to replicate*, in which an intended identical experiment performed at a new site produces conflicting results[5–8]. These issues can be difficult to avoid[9–11], and the problem is exacerbated by a weak culture of replication[12]. But while failure to transport could be considered an instance of failure to replicate[13–16], we see it as a sufficiently distinct phenomenon to warrant specific attention.

Failure to transport was recognized as a problem by the earliest medical AI pioneers. In 1961, Homer Warner implemented the first probabilistic model of symptoms and disease[17] using a Naïve Bayes method[18] with local conditional probabilities, which then failed on external data[19]. A follow-up by Bruce and Yarnall[19] using data from three sites noted similar failure to transport between sites, which they attributed to differences in conditional probabilities. A decade later, Alvan Feinstein argued that the very idea of probabilistic diagnosis was fatally flawed, on the grounds that the observational accuracy, the prevalence, and even the definitions of collected clinical observations varied across sites[20]. A decade after that, in the inaugural issue of *Medical Decision Making*, Tim de Dombal doubted ever being able to design a probabilistic diagnosis engine with data from one site that worked at others, because large-scale surveys demonstrated that disease prevalence and presentation vary dramatically across locations[21].

All probabilistic transport failure can be attributed to differences in the multivariate distribution of the training dataset vs. the application dataset, because the dataset is the sole means by which site-specific phenomena communicate with the model. (We define *application dataset* to mean the data with which the model will be used in practice, as opposed to the *training* and *test sets* that are used during development.) The difference could include slight mismatches in the univariate distribution of a variable, or more radical differences in the dependencies between many variables.

In this perspective, we consider various sources of these differences, dividing them into *Experimental Sources* that can be minimized by the experimental configuration, and *Inherent Sources* that, because they are internal to the way the clinical data are generated, are not so easy to avoid (Fig. 1). We use the term *unstable* for distributions that vary between sites due to either experimental or inherent sources[22–24]. Of the inherent sources, we will discuss in some depth the problem of site-specific clinical workflow processes that are difficult to account for.

Experimental sources

Some distributional differences between training and application datasets can be minimized by sufficient attention to the training pipeline. We call the sources of these differences *experimental* in the broad sense that training any model is a computational experiment, regardless of whether the model is supervised or unsupervised, whether it uses observational or interventional data, or whether it is addressing a hypothesis-driven or discovery-based

Vanderbilt University Medical Center, Nashville, TN, USA.
e-mail: tom.lasko@vanderbilt.edu

Experimental Sources

- Model Overfitting
- Information Leaks During Training
- Different Variable Definitions in Application Data
- Application to the Wrong Question

Inherent Sources

- Application Data with Different Causal Prevalence
- Presence of Site-Specific Processes

Fig. 1 | Potential Sources of Model Transport Failure. These are categorized as *Experimental Sources* that are under the direct control of the researcher, and *Inherent Sources* that are more difficult to address.

question. Experimental sources of instability are quite common, and they can be difficult to recognize.

Model overfitting

An overfit model performs well on training data but poorly on test data, even though the two datasets are drawn from the *same* underlying distribution[25]. Despite the fact that there should be no distributional difference between the two datasets, an overfit model has come to rely in part on patterns that are present *only by chance* in the training set, and which are necessarily different in application data. Getting a model to perform well while avoiding overfitting is the central task of machine learning[25], and we continue to learn surprising things about it[26–29]. A noticeable subset of failure-to-transport results, including those getting recent attention[13], is really just unrecognized overfitting[30].

Information leaks during training

The unintended presence of information in a training dataset that would not be present at application time is an information leak[31,32], a machine-learning equivalent to inadvertent unblinding in a clinical trial[33]. For example, mostly-positive and mostly-negative cases may be collected from separate sources, then patient identifiers assigned sequentially by source. Given this data for training, a model can easily learn that earlier patient identifiers are more likely to be positive[31].

A common leak is the use of positive instances constructed from, say, the time of hospital admission to the predicted event of interest (perhaps the onset of sepsis), and negative instances constructed from the full length of the admission (because there was no sepsis). In this case, variables such as time since admission or the presence of typically near-discharge events (such as weaning from a ventilator) leak information about the probability of a positive label.

Any information generated after the intended moment of application can leak information about the label. Often, this information leaks from times after the label was known, but more subtle leaks can occur from earlier times[32]. For example, information about treating the condition (perhaps starting antibiotics for sepsis) can make its way into training instances, leaking information about the condition's presence. But other indicators such as fever or tachycardia can originate from the period before sepsis is formally labeled, but clinically obvious, and a model predicting sepsis at that point would be less clinically useful[32].

In addition to information leaking from the future, information can also leak between the training, test, and validation sets[31]. These leaks can be easy to miss, as in the well-known CheXNet paper[34], where researchers originally split the dataset by image, rather than by patient, leaking patient-level information between training and test sets[35]. (Although even after the leak was corrected, this impressive model still failed to transport to other sites[36–38]).

Different variable definitions in application data

If an application dataset defines a prediction target differently from the training data, model performance will suffer. This may seem obvious, but its role in failure to transport is easily missed, as happened[39] when the performance of EHR vendor Epic's internal sepsis prediction model dropped at least in part due to a more careful definition of the sepsis label in the application dataset[40].

Similarly, if semantically equivalent data variables are encoded with different identifiers in a new dataset, then the performance of a model that relies on those identifiers will suffer[41,42]. This may also seem obvious, but it can be a subtle problem, because, for example, different institutions that both use LOINC codes[43,44] to identify laboratory test results may actually use *different* LOINC codes for clinically equivalent tests[45]. And, of course, test results can be reported in different units (and possibly mislabeled) even within the same dataset[46].

A definitional mismatch can also cause the model to see what appear to be distribution differences in the affected variables, when in reality the model is seeing different variables with the same name.

Application to the wrong question

A model trained to answer question A is not likely to be as accurate if we try to use it to answer question B. This is yet another obvious statement, but again the problem can be subtle and easy to misdiagnose, especially if the questions are related.

This has happened multiple times with a 1997 model that predicts the mortality of pneumonia in hospitalized patients[47]. Researchers trained the model to answer question A, for which they had data: *How likely is this patient to survive when given usual inpatient care?* Then they applied it to answering question B, which is what they really wanted to know: How likely is this patient to survive if sent home *without* inpatient care? The original researchers explicitly stated that they were assuming that the answer to the two questions would be similar in patients with low probability of inpatient mortality.

Unfortunately, subsequent interpretations of the work appear to have missed the explicit assumption and misdiagnosed the source of resulting problems. A 2015 evaluation[48] of an interpretable learning method identified cases where the assumption was violated, such as for patients with asthma. Pneumonia patients with asthma were more likely than other pneumonia patients to survive with usual inpatient care because they were more likely to be immediately admitted to the ICU. But they are of course much *less* likely to survive if sent home *without* inpatient care. The 2015 work shows the clear value of interpretable models, because it highlighted this problem. But rather than pointing out serious assumption violations and the application to the wrong question, the researchers' next step was to patch the model by manually changing the weights for the *asthma* term (and any others that seemed counterintuitive). Later authors, citing only the 2015 work, claimed that the error was an unintended consequence of machine learning because the model had learned subtle patterns and was missing vital context[49]. Others recognized that the data distribution appeared distorted[50,51] (which it was, relative to question B), but that wasn't the core problem. The core problem was that the model was answering a different question than its application users were asking[52]. Alternatively, we could say that the users were asking a causal, counterfactual question that the non-causal, predictive model was not designed to answer: "What would happen if we intervened by sending the patient home instead of the usual practice of admitting to the hospital?"

Inherent sources

Application data distributions can differ from training data distributions because some real phenomenon of interest affects them differently (as opposed to overfitting, in which the difference is only due to random variation). Using a model under these conditions is an Out-of-Distribution (OOD) application[53–56]. Optimizing for OOD performance is an interesting and growing research direction[56–66].

The need to consider OOD performance in clinical prediction models arises from the fact that, as the AI pioneers observed, clinical data distributions actually do change between sites for reasons inherent to the data-generating mechanism, a phenomenon now known as *dataset shift*[54] or

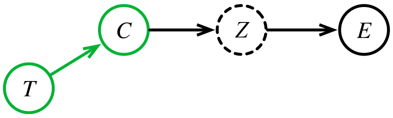

Fig. 2 | A naïve conceptual causal model of disease. Green: Variables or processes that can change between sites. Dotted circle: Unobserved variables. Variable names as in Table 1.

distribution shift[57]. As models have grown in their power to use complex distributional patterns, so has their need to adapt to this shift.

Application data with different causal prevalence

The most obvious distributional difference between sites is the underlying disease prevalence, which can vary drastically. If the prevalence of the predicted outcome differs between the training data and the application data, this can affect the performance of a model, especially its calibration[67,68]. Fortunately, differences in prevalence are easy to accommodate, at least in simple linear models[67,69,70].

However, while we usually speak conceptually about prevalence of *disease*, what actually varies is the prevalence of its *causes*. Consider a naïve probabilistic causal model of disease (Fig. 2), where Z is the core conceptual problem, with its causes C, and effects E. When the prevalence of causes varies with the site T, that will affect the downstream prevalence of the disease and its effects.

The path $C \rightarrow Z \rightarrow E$ is an abstraction representing a large network of causes and effects; we choose a node Z as our condition of interest, which defines upstream nodes as causes, and downstream nodes as effects.

For example, the core conceptual problem of Acute Myocardial Infarction (AMI, or heart attack) is the sudden death of heart muscle. There are many causes of AMI, of which the most immediate is an abrupt reduction in blood flow to the heart muscle. Causes further upstream in the network may include arteries narrowed by plaque, a sudden rupture of that plaque, a blood clot abruptly blocking an artery, or vasospasm stimulated by mediators released from platelets.

So far, none of the causes should inherently vary between sites. But even further upstream are causes related to genetics, diet, exercise, and environmental or medication exposures, all of which do vary with geographic location, local patient populations, local practice patterns, or time. That variation will eventually affect the downstream prevalence of AMI in the population, as well as that of any downstream effects, such as chest pain, shortness of breath, release into the bloodstream of intracellular cardiac enzymes, unconsciousness, or death.

Sometimes, a condition of interest is really the union of a set of subtypes, within which some are easier to predict than others[38,71,72]. In this case, a change in the distribution of subtypes will affect the performance of the model. Performance can even improve on an external dataset if it contains a larger proportion of an easy subtype. This fact can be used nefariously to construct test sets that dramatically increase the apparent performance of a model by including a high prevalence of an easily predicted subtype (such as an easy negative subtype).

Presence of site-specific processes

With some thought, we can identify additional inherent sources of instability that are not represented in Fig. 2. To understand these, note that in general, the variables in the $C \rightarrow Z \rightarrow E$ *cause-effect network* are not directly recorded in clinical practice[73,74], largely because (as with heart muscle death) they are not directly observable under typical clinical conditions. Instead, the related *observational* variables O (such as laboratory test results, clinical images, medication records, billing codes, and narrative clinical text) are observed and recorded using site-specific processes (Fig. 3). While these are intended to reliably reflect the latent variables in the cause-effect network, in practice they cover only a subset of those variables, and the observations depend on case mix, practice patterns, specific instruments, reagents, and personnel, as well as financial incentives[71,75,76]. We might expect this dependence to be standardized across sites, but it turns out to be an important source of distribution shift, even for observations as straightforward clinical laboratory tests, which are affected by spectrum bias[72], test ordering patterns[77], variation in measurement[78], and variation in reference values[79].

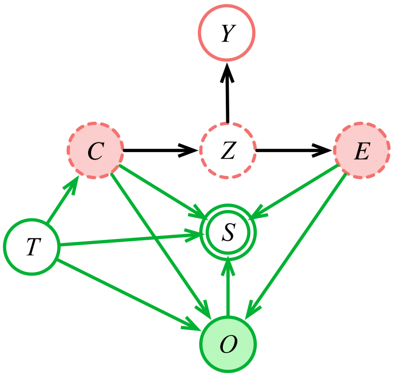

Fig. 3 | A more complete model of disease, including a predicted target *Y*. Conditioning on observations O (shaded green) to produce $P(Y|O)$ allows the prediction to be affected by site T and selection variables S. In contrast, conditioning on the latent but true causes and effects (shaded red) to produce $P(Y|C,E)$, does not. As in other figures, arrows represent the direction of causality, not the direction of inference. Variable names as in Table 1.

Instability arises not only from *how* a variable is observed but also *which* variables and *when*[66,77]. The *which* and *when* decisions are generally made by expert recognition of a developing clinical picture, and can depend on any other variable or process at the time of the decision. These include pathophysiologic phenomena, but also practical phenomena such as the patient deciding to present for care, being discovered unresponsive in public, brought to the hospital by ambulance, assigned to a particular clinical team, or other factors that may or may not be recorded in the EHR.

Decisions to observe are represented in our model by a set of selection variables S (Fig. 3)[80]. The variable $S_o \in S$ represents the decision of *when* to record (record when $S_o = 1$, not when $S_o = 0$), and each $S_{O_i} \subseteq S$ is a set of variables that collectively represent the decision of *which* O_i to record (record O_i when all elements of $S_{O_i} = 1$). Any decision in S may depend on the site T, causes C, effects E, or observations O. The *when* and *which* decisions together produce what has been described as "selection bias on selection bias"[80]. (For simplicity, we omit a discrete time index t in the graph and the notation. But the notation can be extended for all variables so that, for example, S_o becomes $S_o[t]$, meaning S_o at the time t.)

In clinical practice, many measurements are prompted by the suspicion that a relevant observational variable O_i may be outside of its healthy range, which means that the distribution of observed variables $P(O_i|S_{O_i} = 1)$ is different from the unobserved $P(O_i|S_{O_i} = 0)$, also known as *informative missingness*[81] or *informative presence*[82]. The decision to observe is subject to disagreement about what is relevant (even between clinicians of the same specialty trying to answer the same clinical question about the same patient[83–86]), and can become subject to feedback loops if the decisions themselves are used as predictors in a model[81,87]. Given this variation in *how*, *when*, and *which* observations are made, it is difficult to see how *any* of the observational variables could be stable across sites.

Typically, we want to estimate $P(Y|O)$. But conditioning on observations O does not block the influence of the site T, which is the root cause of all instability (Fig. 3), and a sufficiently powerful model can estimate that influence, given enough data. This can happen even if the observations in O are only the pixels of a radiographic image[88]. A powerful model can learn what it means clinically that a chest X-ray was done on a particular machine, that a laboratory test was done on a weekend, or that it was ordered by a given physician. In a multi-site dataset, the identity of the originating site can easily be inferred from site-specific dependencies and exploited for prediction[37].

Perhaps because the model is exploiting unanticipated dependencies in the data that don't match causal pathophysiologic pathways, these dependencies have been called *shortcuts*[89], and considered cheating. But blaming the model isn't a productive research direction. The algorithm doesn't know what we intend for it to learn. It can't tell the difference between pathophysiologic and process-related patterns. All it knows is that the task is to take $O \in R^n$ and predict $Y \in R$ or $Y \in \{0, 1\}$, and it uses all available patterns to do that.

Moreover, human experts also exploit process-related information. A radiologist doesn't just look at the film to identify signs of disease, she also wants to know why the film was ordered, who ordered it, and what was going on with the patient at the time. A pathologist doesn't only look at the slide under the microscope, he also wants to know the location on the body where the biopsy was taken, and what was the clinical scenario that led to ordering it. If a computational model can infer such things from the input data, there is no reason why it shouldn't use them to improve its performance.

Potential solutions

Experimental sources of instability have known solutions that can take effort to implement, but these are at least under the control of the experimenter[14,16,90]. Minimizing the effects of inherent sources is a harder problem, because they are actually part of the data-generating mechanism[54,89,91–93]. We are unaware of any experiments to quantify the relative prevalence and magnitude of experimental vs. inherent sources of instability, although we argue above that inherent sources are likely to drive many transport failures. If this turns out to be true, how could we minimize their impact?

An obvious solution is to somehow identify and exclude the inherent unstable patterns from the model. There are powerful and sophisticated methods to do this, whether the unstable information resides in nodes, edges, or a combination[51,56,60,65,66,94–96]. A related approach is to use data from multiple sites during training, under leave-one-site-out cross validation, so the model identifies only stable features to begin with[2]. These all can be effective methods in specific circumstances. However, we might expect that removing *any* non-redundant information will decrease a model's performance, inevitably trading performance for stability[56]. If unstable patterns are as ubiquitous as expected in clinical data, removing them could have a disastrous effect.

Is it possible to keep the information contained in unstable patterns, but minimize their impact on transportability? One potential direction is to notice that the task of training a stable model reduces to blocking the influence of the site T on the estimate of disease variables Z and Y. There are a host of ways of doing this, including carefully collecting prospectively randomized data, or specifying an appropriate set of conditioning variables using prior knowledge to block the effects in a particular dataset[66]. There are also measurement error models that can correct for measurement differences between sites[97,98]. These methods can be effective, but properly specifying the appropriate conditioning set or measurement error distribution remains challenging in this setting. Instead, we would like an approach that could be more easily applied to untamed EHR data.

In general, we can block the influence of the site T on the estimate of disease variables Z and Y by conditioning on the latent cause and effect variables C and E, because $Z,Y \perp\!\!\!\perp \mathrm{T}|\mathrm{C,E}$ (Fig. 3). However, conditioning on latent variables is difficult; to address this difficulty, we consider separating our model into a *Process Model* that estimates the latent variables, and a *Disease Model* that uses them for prediction (Fig. 4).

We expect that both models would need to be learned from data. The Process Model must infer the site-specific relationships $P(C, E|O, T, S)$, where S is given implicitly by the missing elements of O. It represents the imprint on the data of local implementations of care processes, and must be trained for each site. The Disease Model would infer the stable relationships $P(Z|C, E)$, and $P(Y|Z)$ if a specific label or target is required (Fig. 4). It represents how disease behaves; it generalizes across all sites, and could even be trained using estimates $\{C, Z, E, Y\}$ pooled from multiple sites.

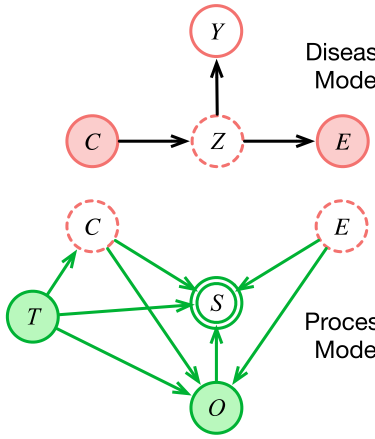

Fig. 4 | Separating the site-specific Process Model $P(C,E|O,T,S)$ from the stable Disease Model $P(Z,Y|C,E)$. The Process Model produces point estimates for C and E, which are used by the Disease Model to infer values for Z and Y. Variable names as in Table 1.

Table 1 | Variable Definitions

Variable	Definition
T	Data Collection Site (in time and space)
C	(Latent) Disease causes or risk factors.
Z	(Latent) Core disease characteristics.
E	(Latent) Downstream effects.
Y	Binary label or continuous target
O	Observed variables
S	Selection variables

To achieve stability under this arrangement, the Process Model must infer sufficiently accurate point estimates of the latent variables, or else information about T can leak through the conditioning. Given enough information in O, such as by a sufficiently large number of variables in O, acceptable accuracy should be achievable[99,100]. We know of no implementations of this strategy, but some promising initial steps have been made.

First, de Fauw and colleagues[101] found that a model of retinal disease from 3-dimensional optical coherence tomography images ported poorly between different types of scanners, with an error rate of 46% on external data for predicting specialty referral (vs. 5.5% on the internal test set). To address the instability, they created a Process Model that produced a tissue segmentation $P(C, E|O, T)$, from the raw image pixels and a Disease Model $P(Z, Y|C, E)$ that recognized retinal pathology from the segmentation and made a referral recommendation. Retraining only the Process Model on the new scanner's data dramatically reduced the external error rate to 3%. This worked because the investigators knew what the latent variables were (the physical structure of tissue layers), and were able to label them for training the Process Model. But the results demonstrate the promise of concentrating the instability into the Process Model.

Next, Lasko and Mesa[102] used probabilistic independence to infer unsupervised data signatures that represent 2000 latent variables in the $C \rightarrow Z \rightarrow E$ network from a large EHR dataset. Then they separately trained a supervised Disease Model using the estimated values of those variables to predict liver transplant 10 years in the future. The top predictors in the Disease Model corresponded *in correct rank order* to the leading causes of hepatocellular carcinoma, suggesting that the method had accurately estimated causal latent variables. Strobl and Lasko[103] later proved theoretically that the latent variables learned by this approach do, in fact, correspond to root causes of disease (meaning the furthest upstream variables in C or Z with unique causal effects on Y), and can identify those causes specific to each patient case. These results demonstrate the promise of estimating and then conditioning on latent disease variables, even when they are not known in advance. The approach has since been extended to accommodate heteroscedasticity[104] and latent confounders[105].

But is it actually worth the effort to separate one model into two? If we must retrain a Process Model at each site anyway, why not just retrain the whole thing? For example, why not start with a dataset from just a few sites, using a site identifier as an input variable? Additional sites could be added to the model by simply continuing the training using the new site's data, with its new site identifier. The question recalls de Dombal's note that a model using global probabilities performed much better at each data-contributing site than cross-site probabilities did, though not quite as well as same-site probabilities[21]. A lighter variation on this could be to develop a single foundational model, which each site would fine-tune or update separately with local data[3,69,106–108]. These approaches can be quite practical and effective.

The largest benefit we see to our proposal of isolating the two types of patterns is that while the updated or cumulative model solutions may address the transportability problem, they miss out on what could be substantial advances that exploit the natural division between process relationships and disease relationships.

First, we see the Disease Model as representing what we actually want to know about health and disease. It is described using the same relationships that clinicians learn in the classroom phase of medical school to guide clinical thinking. It is stable, transferrable knowledge that can be directly shared between institutions, forming the core of a learning healthcare system. It is the representation that abstracts away all site-specific information.

Second, we see the Process Model as representing the site-specific processes that clinicians learn in their clerkships and residency, by which patients are diagnosed, monitored, and treated. It gives a window onto what we want to know about healthcare delivery: What are our care processes? How do they differ between institutions? How do they evolve over time? What are their inefficiencies and conflicts? Those questions are probably best answered by direct observation, but the processes involved do leave an imprint on the data record, which when isolated could provide clues or signals about the processes. This idea recalls Hripcsak and Albers[76] from a decade ago, who insightfully described the EHR as an artifact of the recording process, rather than a direct record of disease. They called for studying the EHR as a phenomenon of its own, "deconvolving" the actual patient state from what is recorded. Their call maps directly onto learning and analyzing Process Models .

Summary

It should not surprise us when high-performance probabilistic clinical models fail to transport to new sites. The failure is directly caused by differences in the multivariate distribution between the training data and the application data. Some of these differences are due to controllable experimental factors, such as overfitting, information leaks, data definitions, and misuse, but others are inherent in site-specific data-generating mechanisms, and much more difficult to avoid. How best to minimize the inherent factors is an open question, as is the relative impact of experimental vs. inherent sources in transport failure. However, we argue that simply removing unstable patterns and variables from clinical datasets is unlikely to succeed, because nearly all recorded observations and their relationships with disease variables are rendered unstable by site-specific clinical practices. Instead, we propose embracing the instability and exploiting all information in a record to maximize performance, by training first a Process Model that uses the site-specific information to infer latent cause and effect variables, and then a Disease Model that uses those latent variables to make stable inferences about a patient's clinical state. There are early but promising indicators of potential merit in the approach.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Received: 9 June 2023; Accepted: 14 February 2024;
Published online: 01 March 2024

References

1. Van Calster, B., Steyerberg, E. W., Wynants, L. & van Smeden, M. There is no such thing as a validated prediction model. *BMC Med.* **21**, 70 (2023).
2. de Jong, V. M. T., Moons, K. G. M., Eijkemans, M. J. C., Riley, R. D. & Debray, T. P. A. Developing more generalizable prediction models from pooled studies and large clustered data sets. *Stat. Med.* **40**, 3533–3559 (2021).
3. Debray, T. P. A. et al. Meta-analysis and aggregation of multiple published prediction models. *Stat. Med.* **33**, 2341–2362 (2014).
4. Siontis, G. C. M., Tzoulaki, I., Castaldi, P. J. & Ioannidis, J. P. A. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *J. Clin. Epidemiol.* **68**, 25–34 (2015).
5. Begley, C. G. & Ioannidis, J. P. A. Reproducibility in science. *Circ. Res.* **116**, 116–126 (2015).
6. Motulsky, H. J. Common misconceptions about data analysis and statistics. *Naunyn. Schmiedebergs Arch. Pharmacol.* **387**, 1017–1023 (2014).
7. Goodman, S. N., Fanelli, D. & Ioannidis, J. P. A. What does research reproducibility mean? *Sci. Transl. Med.* **8**, 341ps12–341ps12 (2016).
8. Baker, M. 1,500 scientists lift the lid on reproducibility. *Nature* **533**, 452–454 (2016).
9. Ostropolets, A. et al. Reproducible variability: assessing investigator discordance across 9 research teams attempting to reproduce the same observational study. *J. Am. Med. Inform. Assoc.* **30**, 859–868 (2023).
10. Botvinik-Nezer, R. et al. Variability in the analysis of a single neuroimaging dataset by many teams. *Nature* **582**, 84–88 (2020).
11. Errington, T. M., Denis, A., Perfito, N., Iorns, E. & Nosek, B. A. Challenges for assessing replicability in preclinical cancer biology. *eLife* **10**, e67995 (2021).
12. Coiera, E. & Tong, H. L. Replication studies in the clinical decision support literature–frequency, fidelity, and impact. *J. Am. Med. Inform. Assoc.* **28**, 1815–1825 (2021).
13. Sohn, E. The reproducibility issues that haunt health-care AI. *Nature* **613**, 402–403 (2023).
14. McDermott, M. B. A. et al. Reproducibility in machine learning for health research: Still a ways to go. *Sci. Transl. Med.* **13**, eabb1655 (2021).
15. Van Calster, B., Wynants, L., Timmerman, D., Steyerberg, E. W. & Collins, G. S. Predictive analytics in health care: how can we know it works? *J. Am. Med. Inform. Assoc.* **26**, 1651–1654 (2019).
16. Heil, B. J. et al. Reproducibility standards for machine learning in the life sciences. *Nat. Methods* **18**, 1132–1135 (2021).
17. Warner, H. R., Toronto, A. F., Veasey, L. G. & Stephenson, R. A Mathematical approach to medical diagnosis: application to congenital heart disease. *JAMA* **177**, 177–183 (1961).
18. Ledley, R. S. & Lusted, L. B. Reasoning foundations of medical diagnosis. *Science* **130**, 9–21 (1959).
19. Bruce, R. A. & Yarnall, S. R. Computer-aided diagnosis of cardiovascular disorders. *J. Chronic Dis.* **19**, 473–484 (1966).
20. Feinstein, A. R. An analysis of diagnostic reasoning. II. The strategy of intermediate decisions. *Yale J. Biol. Med.* **46**, 264–283 (1973).
21. de Dombal, F. T., Staniland, J. R. & Clamp, S. E. Geographical variation in disease presentation: does it constitute a problem and can information science help? *Med. Decis. Mak.* **1**, 59–69 (1981).
22. Bao, Y. et al. Association of nut consumption with total and cause-specific mortality. *N. Engl. J. Med.* **369**, 2001–2011 (2013).
23. Yu, B. Stability. *Bernoulli* **19**, 1484–1500 (2013).
24. Yu, B. & Kumbier, K. Veridical data science. *Proc. Natl. Acad. Sci.* **117**, 3920–3929 (2020).

25. Abu-Mostafa, Y. S., Magdon-Ismail, M. & Lin, H.-T. Overfitting. in *Learning from data: A short course* (AMLbook, 2012).
26. Advani, M. S., Saxe, A. M. & Sompolinsky, H. High-dimensional dynamics of generalization error in neural networks. *Neural Netw.* **132**, 428–446 (2020).
27. Belkin, M., Hsu, D., Ma, S. & Mandal, S. Reconciling modern machine-learning practice and the classical bias-variance trade-off. *Proc. Natl. Acad. Sci. USA* **116**, 15849–15854 (2019).
28. Belkin, M. Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation. *Acta Numer.* **30**, 203–248 (2021).
29. d'Ascoli, S., Sagun, L. & Biroli, G. Triple descent and the two kinds of overfitting: where and why do they appear? In *Advances in neural information processing systems* (eds. Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F. & Lin, H.) vol. 33 3058–3069 (Curran Associates, Inc., 2020).
30. Yu, K.-H. et al. Reproducible machine learning methods for lung cancer detection using computed tomography images: algorithm development and validation. *J. Med. Internet Res.* **22**, e16709 (2020).
31. Kaufman, S., Rosset, S. & Perlich, C. Leakage. in data mining: formulation, detection, and avoidance. In *Proce. 17th ACM SIGKDD international conference on Knowledge discovery and data mining* 556–563 (Association for Computing Machinery, New York, NY, USA, 2011). https://doi.org/10.1145/2020408.2020496.
32. Davis, S. E., Matheny, M. E., Balu, S. & Sendak, M. P. A framework for understanding label leakage in machine learning for health care. *J. Am. Med. Inform. Assoc.* https://doi.org/10.1093/jamia/ocad178 (2023).
33. Rosset, S., Perlich, C., Świrszcz, G., Melville, P. & Liu, Y. Medical data mining: insights from winning two competitions. *Data Min. Knowl. Discov.* **20**, 439–468 (2010).
34. Rajpurkar, P. et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. Preprint at https://doi.org/10.48550/arXiv.1711.05225 (2017).
35. Guts, Y. Target Leakage in Machine Learning. https://www.youtube.com/watch?v=dWhdWxgt5SU (2018).
36. Perry, T. Andrew Ng X-Rays the AI Hype. *IEEE Spectrum*. https://spectrum.ieee.org/andrew-ng-xrays-the-ai-hype (2021).
37. Zech, J. R. et al. Confounding variables can degrade generalization performance of radiological deep learning models. *PLOS Med.* **15**, e1002683 (2018).
38. Oakden-Rayner, L., Dunnmon, J., Carneiro, G. & Ré, C. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. *Proc. ACM Conf. Health Inference Learn.* **2020**, 151–159 (2020).
39. Habib, A. R., Lin, A. L. & Grant, R. W. The epic sepsis model falls short—the importance of external validation. *JAMA Intern. Med.* **181**, 1040–1041 (2021).
40. Wong, A. et al. External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients. *JAMA Intern. Med.* **181**, 1065–1070 (2021).
41. Nestor, B. et al. Feature Robustness in Non-stationary Health Records: Caveats to Deployable Model Performance in Common Clinical Machine Learning Tasks. *Proceedings of the 4th Machine Learning for Healthcare Conference, in Proceedings of Machine Learning Research* **106**, 381–405 (2019). Available from https://proceedings.mlr.press/v106/nestor19a.html.
42. Gong, J. J., Naumann, T., Szolovits, P. & Guttag, J. V. Predicting Clinical Outcomes Across Changing Electronic Health Record Systems. In *Proc. 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1497–1505 (Association for Computing Machinery, New York, NY, USA, 2017); https://doi.org/10.1145/3097983.3098064.
43. McDonald, C. J. et al. LOINC, a universal standard for identifying laboratory observations: A 5-Year Update. *Clin. Chem.* **49**, 624–633 (2003).
44. Stram, M. et al. Logical observation identifiers names and codes for laboratorians: potential solutions and challenges for interoperability. *Arch. Pathol. Lab. Med.* **144**, 229–239 (2019).
45. Parr, S. K., Shotwell, M. S., Jeffery, A. D., Lasko, T. A. & Matheny, M. E. Automated mapping of laboratory tests to LOINC codes using noisy labels in a national electronic health record system database. *J. Am. Med. Inform. Assoc.* **25**, 1292–1300 (2018).
46. Abhyankar, S., Demner-Fushman, D. & McDonald, C. J. Standardizing clinical laboratory data for secondary use. *J. Biomed. Inform.* **45**, 642–650 (2012).
47. Cooper, G. F. et al. An evaluation of machine-learning methods for predicting pneumonia mortality. *Artif. Intell. Med.* **9**, 107–138 (1997).
48. Caruana, R. et al. Intelligible models for HealthCare: predicting pneumonia risk and hospital 30-day readmission. In *Proc. 21th ACM SIGKDD international conference on knowledge discovery and data mining (KDD'15)* (2015). https://doi.org/10.1145/2783258.2788613.
49. Cabitza, F., Rasoini, R. & Gensini, G. F. Unintended consequences of machine learning in medicine. *JAMA J. Am. Med. Assoc.* **318**, 517–518 (2017).
50. Subbaswamy, A. & Saria, S. From development to deployment: dataset shift, causality, and shift-stable models in health AI. *Biostatistics* **21**, 345–352 (2020).
51. Subbaswamy, A. & Saria, S. I-SPEC: An End-to-End Framework for Learning Transportable, Shift-Stable Models. Preprint at https://doi.org/10.48550/arXiv.2002.08948 (2020).
52. Lasko, T. A., Walsh, C. G. & Malin, B. Benefits and risks of machine learning decision support systems. *JAMA J. Am. Med. Assoc.* **318**, 2355 (2017).
53. Shen, Z. et al. Towards Out-Of-Distribution Generalization: A Survey. Preprint at https://doi.org/10.48550/arXiv.2108.13624 (2021).
54. *Dataset Shift in Machine Learning*. (The MIT Press, Cambridge, Mass, 2008).
55. Moreno-Torres, J. G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N. V. & Herrera, F. A unifying view on dataset shift in classification. *Pattern Recognit.* **45**, 521–530 (2012).
56. Subbaswamy, A., Chen, B. & Saria, S. A unifying causal framework for analyzing dataset shift-stable learning algorithms. *J. Causal Inference* **10**, 64–89 (2022).
57. Koh, P. W. et al. WILDS: A Benchmark of in-the-Wild Distribution Shifts. *Proceedings of the 38th International Conference on Machine Learning, in Proceedings of Machine Learning Research* **139**, 5637–5664 (2021). Available from https://proceedings.mlr.press/v139/koh21a.html.
58. Zhou, K., Liu, Z., Qiao, Y., Xiang, T. & Loy, C. C. Domain Generalization: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell*. 1–20 https://doi.org/10.1109/TPAMI.2022.3195549 (2022)
59. Wang, J., Lan, C., Liu, C., Ouyang, Y. & Qin, T. Generalizing to Unseen Domains: A Survey on Domain Generalization. In *Proc. Thirtieth International Joint Conference on Artificial Intelligence* 4627–4635 (International Joint Conferences on Artificial Intelligence Organization, Montreal, Canada, 2021); https://doi.org/10.24963/ijcai.2021/628.
60. Pearl, J. & Bareinboim, E. Transportability of causal and statistical relations: a formal approach. *Proc. AAAI Conf. Artif. Intell.* **25**, 247–254 (2011).
61. Arjovsky, M., Bottou, L., Gulrajani, I. & Lopez-Paz, D. Invariant Risk Minimization. Preprint at https://doi.org/10.48550/arXiv.1907.02893 (2020).
62. Bellot, A. & van der Schaar, M. Accounting for Unobserved Confounding in Domain Generalization. Preprint at https://doi.org/10.48550/arXiv.2007.10653 (2022).

63. Amodei, D. et al. Concrete Problems in AI Safety. Preprint at https://doi.org/10.48550/arXiv.1606.06565 (2016).
64. Degtiar, I. & Rose, S. A Review of Generalizability and Transportability. *Annu. Rev. Stat. Appl.* **10**, 501–524 (2023).
65. Correa, J. D., Lee, S. & Bareinboim, E. Counterfactual Transportability: A Formal Approach. *Proceedings of the 39th International Conference on Machine Learning, in Proceedings of Machine Learning Research* **162**, 4370–4390 (2022). Available from https://proceedings.mlr.press/v162/correa22a.html.
66. Bareinboim, E., Tian, J. & Pearl, J. Recovering from selection bias in causal and statistical inference. *Proceedings of the AAAI Conference on Artificial Intelligence* **28** (2014). https://doi.org/10.1609/aaai.v28i1.9074.
67. Morise, A. P., Diamond, G. A., Detrano, R., Bobbio, M. & Gunel, E. The effect of disease-prevalence adjustments on the accuracy of a logistic prediction model. *Med. Decis. Mak.* **16**, 133–142 (1996).
68. Davis, S. E., Lasko, T. A., Chen, G., Siew, E. D. & Matheny, M. E. Calibration drift in regression and machine learning models for acute kidney injury. *J. Am. Med. Inform. Assoc. JAMIA* **24**, 1052–1061 (2017).
69. Davis, S. E. et al. A nonparametric updating method to correct clinical prediction model drift. *J. Am. Med. Inform. Assoc.* **26**, 1448–1457 (2019).
70. Poses, R. M., Cebul, R. D., Collins, M. & Fager, S. S. The importance of disease prevalence in transporting clinical prediction rules. *Ann. Intern. Med.* **105**, 586–591 (1986).
71. Riley, R. D. et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. *BMJ* **353**, i3140 (2016).
72. Mulherin, S. A. & Miller, W. C. Spectrum bias or spectrum effect? Subgroup variation in diagnostic test evaluation. *Ann. Intern. Med.* **137**, 598–602 (2002).
73. Botsis, T., Hartvigsen, G., Chen, F. & Weng, C. Secondary use of EHR: data quality issues and informatics opportunities. *Summits Transl. Bioinform.* **2010**, 1–5 (2010).
74. Sarwar, T. et al. The secondary use of electronic health records for data mining: data characteristics and challenges. *ACM Comput. Surv.* **55**, 33:1–33:40 (2022).
75. Tellez, D. et al. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Med. Image Anal.* **58**, 101544 (2019).
76. Hripcsak, G. & Albers, D. J. Next-generation phenotyping of electronic health records. *J. Am. Med. Inform. Assoc.* **20**, 117–121 (2013).
77. Agniel, D., Kohane, I. S. & Weber, G. M. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *BMJ* **361**, k1479 (2018).
78. Joffe, M. et al. Variability of creatinine measurements in clinical laboratories: results from the CRIC study. *Am. J. Nephrol.* **31**, 426–434 (2010).
79. Siest, G. et al. The theory of reference values: an unfinished symphony. *Clin. Chem. Lab. Med.* **51**, 47–64 (2013).
80. Strobl, E. V., Visweswaran, S. & Spirtes, P. L. Fast causal inference with non-random missingness by test-wise deletion. *Int. J. Data Sci. Anal.* **6**, 47–62 (2018).
81. Groenwold, R. H. H. Informative missingness in electronic health record systems: the curse of knowing. *Diagn. Progn. Res.* **4**, 8 (2020).
82. Sisk, R. et al. Informative presence and observation in routine health data: a review of methodology for clinical risk prediction. *J. Am. Med. Inform. Assoc.* **28**, 155–166 (2021).
83. Herasevich, V., Ellsworth, M. A., Hebl, J. R., Brown, M. J. & Pickering, B. W. Information needs for the OR and PACU electronic medical record. *Appl. Clin. Inform.* **5**, 630–641 (2014).
84. Zeng, Q., Cimino, J. J. & Zou, K. H. Providing concept-oriented views for clinical data using a knowledge-based system: An Evaluation. *J. Am. Med. Inform. Assoc. JAMIA* **9**, 294–305 (2002).
85. Van Vleck, T. T., Stein, D. M., Stetson, P. D. & Johnson, S. B. Assessing data relevance for automated generation of a clinical summary. *Annu. Symp. Proc. AMIA Symp.* **2007**, 761–765 (2007).
86. Lasko, T. A. et al. User-centered clinical display design issues for inpatient providers. *Appl. Clin. Inform.* **11**, 700–709 (2020).
87. van Smeden, M., Groenwold, R. H. H. & Moons, K. G. M. A cautionary note on the use of the missing indicator method for handling missing data in prediction research. *J. Clin. Epidemiol.* **125**, 188–190 (2020).
88. Badgeley, M. A. et al. Deep learning predicts hip fracture using confounding patient and healthcare variables. *Npj Digit. Med.* **2**, 1–10 (2019).
89. Geirhos, R. et al. Shortcut learning in deep neural networks. *Nat. Mach. Intell.* **2**, 665–673 (2020).
90. Van Calster, B. et al. A calibration hierarchy for risk models was defined: from utopia to empirical data. *J. Clin. Epidemiol.* **74**, 167–176 (2016).
91. Pan, S. J. & Yang, Q. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **22**, 1345–1359 (2010).
92. D'Amour, A. et al. Underspecification presents challenges for credibility in modern machine learning. *J. Mach. Learn. Res.* **23**, 1–61 (2022).
93. Delétang, G. et al. Neural Networks and the Chomsky Hierarchy. *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=WbxHAzkeQcn (2023).
94. Saranrittichai, P., Mummadi, C. K., Blaiotta, C., Munoz, M. & Fischer, V. Overcoming Shortcut Learning in a Target Domain by Generalizing Basic Visual Factors from a Source Domain. In *Computer Vision – ECCV 2022. ECCV 2022. Lecture Notes in Computer Science*, Vol. 13685 (eds Avidan, S., Brostow, G., Cissé, M., Farinella, G. M. & Hassner, T.) (Springer, Cham., 2022). https://doi.org/10.1007/978-3-031-19806-9_17.
95. Magliacane, S. et al. Domain adaptation by using causal inference to predict invariant conditional distributions. In *Advances in Neural Information Processing Systems* vol. 31 (Curran Associates, Inc., 2018).
96. Atzmon, Y., Kreuk, F., Shalit, U. & Chechik, G. A causal view of compositional zero-shot recognition. In *Advances in Neural Information Processing Systems* vol. 33 1462–1473 (Curran Associates, Inc., 2020).
97. Stefanski, L. A. & Cook, J. R. Simulation-extrapolation: the measurement error Jackknife. *J. Am. Stat. Assoc.* **90**, 1247–1256 (1995).
98. Carroll, R. J., Roeder, K. & Wasserman, L. Flexible parametric measurement error models. *Biometrics* **55**, 44–54 (1999).
99. Wang, Y. & Blei, D. M. The blessings of multiple causes. *J. Am. Stat. Assoc.* **114**, 1574–1596 (2019).
100. Ogburn, E. L., Shpitser, I. & Tchetgen, E. J. T. Counterexamples to 'The Blessings of Multiple Causes' by Wang and Blei. Preprint at https://doi.org/10.48550/arXiv.2001.06555 (2020).
101. De Fauw, J. et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat. Med.* **24**, 1342–1350 (2018).
102. Lasko, T. A. & Mesa, D. A. Computational phenotype discovery via probabilistic independence. In *Proc KDD workshop on appl data sci for healthcare (DSHealth)* (2019). Available from https://doi.org/10.48550/arXiv.1907.11051.
103. Strobl, E. V. & Lasko, T. A. Identifying patient-specific root causes of disease. In *Proc. 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics* 1–10 (Association for Computing Machinery, New York, NY, USA, 2022); https://doi.org/10.1145/3535508.3545553.
104. Strobl, E. V. & Lasko, T. A. Identifying patient-Specific root causes heteroscedastic noise model. *J. Comput. Sci.* **72**, 102099 (2023).

105. Strobl, E. & Lasko, T. A. Sample-Specific Root Causal Inference with Latent Variables. *Proceedings of the Second Conference on Causal Learning and Reasoning, in Proceedings of Machine Learning Research* **213**, 895–915 (2023). Available from https://proceedings.mlr.press/v213/strobl23b.html.

106. Vergouwe, Y. et al. A closed testing procedure to select an appropriate method for updating prediction models. *Stat. Med.* **36**, 4529–4539 (2017).

107. Janssen, K. J. M., Moons, K. G. M., Kalkman, C. J., Grobbee, D. E. & Vergouwe, Y. Updating methods improved the performance of a clinical prediction model in new patients. *J. Clin. Epidemiol.* **61**, 76–86 (2008).

108. Tanner, K., Keogh, R. H., Coupland, C. A. C., Hippisley-Cox, J. & Diaz-Ordaz, K. Dynamic updating of clinical survival prediction models in a rapidly changing environment. *Diagn. Progn. Res.* **7**, 24 (2023). https://doi.org/10.1186/s41512-023-00163-z.

Acknowledgements

This work was funded in part by grants from the National Library of Medicine (LM013807), the National Cancer Institute (CA253923), and the National Institute of Arthritis and Musculoskeletal and Skin Diseases (AR076516).

Author contributions

T.A.L. conceived the original direction, with substantial scientific refinement in discussion with E.V.S. and W.W.S. T.A.L. wrote the initial manuscript draft, with critical revision by E.V.S. and W.W.S. All authors accept responsibility for the final article.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-024-01037-4.

Correspondence and requests for materials should be addressed to Thomas A. Lasko.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.