Journal of Machine Learning Research 23 (2022) 1-61 Submitted 11/20; Revised 12/21; Published 08/22

Underspecification Presents Challenges for Credibility in Modern Machine Learning

Alexander D'Amour* ALEXDAMOUR@GOOGLE.COM
Katherine Heller* KHELLER@GOOGLE.COM
Dan Moldovan* MDAN@GOOGLE.COM
Ben Adlam ADLAM@GOOGLE.COM
Babak Alipanahi BABAKA@GOOGLE.COM
Alex Beutel ALEXBEUTEL@GOOGLE.COM
Christina Chen CHRISTINIUM@GOOGLE.COM
Jonathan Deaton JDEATON@GOOGLE.COM
Jacob Eisenstein JEISENSTEIN@GOOGLE.COM
Matthew D. Hoffman MHOFFMAN@GOOGLE.COM
Farhad Hormozdiari FHORMOZ@GOOGLE.COM
Neil Houlsby NEILHOULSBY@GOOGLE.COM
Shaobo Hou SHAOBOHOU@GOOGLE.COM
Ghassen Jerfel GHASSEN@GOOGLE.COM
Alan Karthikesalingam ALANKARTHI@GOOGLE.COM
Mario Lucic LUCIC@GOOGLE.COM
Yian Ma YIANMA@UCSD.EDU
Cory McLean CYM@GOOGLE.COM
Diana Mincu DMINCU@GOOGLE.COM
Akinori Mitani AMITANI@GOOGLE.COM
Andrea Montanari MONTANARI@STANFORD.EDU
Zachary Nado ZNADO@GOOGLE.COM
Vivek Natarajan NATVIV@GOOGLE.COM
Christopher Nielson† CHRISTOPHER.NIELSON@VA.GOV
Thomas F. Osborne† THOMAS.OSBORNE@VA.GOV
Rajiv Raman DRRRN@SNMAIL.ORG
Kim Ramasamy KIM@ARAVIND.ORG
Rory Sayres SAYRES@GOOGLE.COM
Jessica Schrouff SCHROUFF@GOOGLE.COM
Martin Seneviratne MARTSEN@GOOGLE.COM
Shannon Sequeira SHNNN@GOOGLE.COM
Harini Suresh HSURESH@MIT.EDU
Victor Veitch VICTORVEITCH@GOOGLE.COM
Max Vladymyrov MXV@GOOGLE.COM
Xuezhi Wang XUEZHIW@GOOGLE.COM
Kellie Webster WEBSTERK@GOOGLE.COM
Steve Yadlowsky YADLOWSKY@GOOGLE.COM
Taedong Yun TEDYUN@GOOGLE.COM
Xiaohua Zhai XZHAI@GOOGLE.COM
D. Sculley DSCULLEY@GOOGLE.COM

Editor: David Jensen

∗. These authors contributed equally to this work.
†. This paper represents the views of the authors, and not of the VA.

Abstract

Machine learning (ML) systems often exhibit unexpectedly poor behavior when they are deployed in real-world domains. We identify underspecification in ML pipelines as a key reason for these failures. An ML pipeline is the full procedure followed to train and validate a predictor. Such a pipeline is underspecified when it can return many distinct predictors with equivalently strong test performance. Underspecification is common in modern ML pipelines that primarily validate predictors on held-out data that follow the same distribution as the training data. Predictors returned by underspecified pipelines are often treated as equivalent based on their training domain performance, but we show here that such predictors can behave very differently in deployment domains. This ambiguity can lead to instability and poor model behavior in practice, and is a distinct failure mode from previously identified issues arising from structural mismatch between training and deployment domains. We provide evidence that underspecfication has substantive implications for practical ML pipelines, using examples from computer vision, medical imaging, natural language processing, clinical risk prediction based on electronic health records, and medical genomics. Our results show the need to explicitly account for underspecification in modeling pipelines that are intended for real-world deployment in any domain.

Keywords: distribution shift, spurious correlation, fairness, identifiability, computer vision, natural language processing, medical imaging, electronic health records, genomics

1. Introduction

In many applications of machine learning (ML), a trained model is not only required to predict well in the training domain; it is also expected to satisfy additional behavioral requirements in deployment. For example, in domains such as medical diagnostics, it is often desirable for the model to be sensitive to physiological signals, while being invariant to environmental or operational signals that can change between deployment contexts. In other domains, such as natural language processing, the behavioral requirements are determined by the details of the application; for example, the requirements in question answering, where sensitivity to world knowledge is important, may be different from those in translation, where isolating semantic knowledge is desirable. These requirements are often presented as criteria for determining whether an ML-based predictor can be trusted in practice.

Unfortunately, many ML pipelines—that is, the procedure that is followed to train and validate a predictor—are poorly set up to credibly satisfy these requirements. Here, we study a workflow that we call the standard ML pipeline. The standard ML pipeline includes a model specification, a training data source, and, importantly, an independent and identically distributed (iid) evaluation procedure, which validates a predictor's expected predictive performance on data drawn from the training distribution, such as a randomly held-out set. This standard paradigm has enabled transformational progress in a number of problem areas, but its blind spots are now becoming more salient. In particular, the evaluations in this pipeline are agnostic to the particular signal used by the trained model to produce predictions. As a result, concerns regarding "spurious correlations" and "shortcut learning" in trained models are now widespread (e.g., Geirhos et al., 2020; Arjovsky et al., 2019).

The purpose of this paper is to explore this gap in the standard ML pipeline. A common explanation is simply that, in many situations, the ML pipeline has a structural design flaw, such that there is a fundamental conflict between iid performance and desirable behavior in deployment. For example, this can occur when the main signal at training time comes from selection bias or features that will not be available at deployment time. In these cases, models optimized for iid performance will necessarily incorporate inappropriate signal (Caruana et al., 2015; Arjovsky et al., 2019; Ilyas et al., 2019).

However, even when these clear structural flaws are avoided, ML pipelines can still produce predictors that behave unexpectedly in deployment. One common observation from practitioners is that predictors optimized on the same data to achieve the same level of iid generalization will often show widely divergent behavior when applied to real-world settings. This observation is not

adequately explained by the standard "structural flaw" narrative, from which we would expect predictors with similar iid performance to show similar defects in deployment.

In this paper, we identify *underspecification* in ML pipelines as a distinct failure mode that explains this behavior. In general, the solution to a problem is underspecified if there are many distinct solutions that solve the problem equivalently. For example, the solution to an underdetermined system of linear equations (i.e., more unknowns than linearly independent equations) is underspecified, with an equivalence class of solutions given by a linear subspace. In the context of ML, we say an ML pipeline is underspecified if there are many distinct ways for the pipeline to produce a predictor that satisfies the pipeline's validation criterion equivalently, even if the specification of the pipeline (e.g., model specification, training data) is held constant. Various notions of underspecification are well-documented in the ML literature; for example, it is a core idea in deep ensembles, double descent, Bayesian deep learning, and loss landscape analysis (Lakshminarayanan et al., 2017; Fort et al., 2019; Belkin et al., 2018; Nakkiran et al., 2020). However, the substantive implications of underspecification in practical ML pipelines have been under-studied.

Here, we make two main claims about the role of underspecification in modern machine learning. The first claim is that underspecification in ML pipelines, and the standard ML pipeline in particular, is a key obstacle to reliably training models that behave as expected in deployment. Specifically, when an ML pipeline can produce many predictors that satisfy its validation criterion equivalently, the particular predictor that is returned (and thus the real-world behavior of the deployed model), will be determined by opaque, often arbitrary choices made in the training pipeline. Thus, in the standard ML pipeline, even if there exists a predictor with the strongest achievable iid performance that also behaves appropriately in deployment, we cannot guarantee that such a model will be returned when the pipeline is underspecified. In Section 3, we demonstrate this issue in several examples that incorporate simple models: one simulated, one theoretical, and one a real empirical example from medical genomics. In these examples, we show how underspecification manifests as sensitivity to arbitrary choices that keep iid performance fixed, but can have substantial effects on performance in model deployment.

The second claim is that underspecification is ubiquitous in modern applications of ML, and has substantial practical implications. We support this claim with an empirical study, in which we apply a simple experimental protocol across plausibly deployable deep learning pipelines in computer vision, medical imaging, natural language processing (NLP), and electronic health record (EHR) based prediction. The protocol is designed to detect underspecification by showing that a predictor's performance on *stress tests*—empirical evaluations that probe the model's behavior along practically important dimensions—is sensitive to iid-performance-preserving perturbations to the training pipeline, such as the choice of random seed. We show that these perturbations induce substantial variation in stress test performance, indicating that these behavioral characteristics of the model are poorly constrained by the training and validation pipeline. This variation distinguishes underspecification-induced failure from the more familiar case of structural mismatch, which would predict uniformly poor performance on stress tests. We find evidence of underspecification in all applications, with downstream effects on robustness, fairness, and causal grounding.

Together, our findings indicate that underspecification can, and does, degrade the credibility of ML systems in applications, even in settings where the pipeline specification is well-aligned with the goals of an application. The direct implication of our findings is that substantive real-world behavior of ML predictors can be determined in unpredictable ways by choices that are made for convenience, such as initialization schemes or step size schedules chosen for trainability—even when these choices do not affect iid performance. More broadly, our results suggest a need to explicitly test models for required behaviors in all cases where these requirements are not directly guaranteed by iid evaluations. Finally, these results suggest a need for training and evaluation techniques tailored to address underspecification, such as flexible methods to constrain ML pipelines to produce predictors that satisfy the requirements of specific applications. Interestingly, our findings suggest that enforcing

these constraints need not introduce hard tradeoffs with iid performance, a hypothesis that has been explored in subsequent work (see discussion in Section 9).

Organization The paper is organized as follows. We present some core concepts and review relevant literature in Section 2. We present a set of examples of underspecification in simple, analytically tractable models as a warm-up in Section 3. We then present a set of four deep learning case studies in Sections 5–8. We close with a discussion in Section 9.

Overall, our strategy in this paper is to provide a broad range of examples of underspecification in a variety of modeling pipelines. Readers may not find it necessary to peruse every example to appreciate our argument, but different readers may find different domains to be more familiar. As such, the paper is organized such that readers can take away most of the argument from understanding one example from Section 3 and one case study from Sections 5–8. However, we believe there is benefit to presenting all of these examples under the single banner of underspecification, so we include them all in the main text.

2. Preliminaries and Related Work

2.1 Machine Learning Pipelines

In this paper, the object we focus on is an ML pipeline, which is the procedure that is followed to train and validate a predictor. We consider a supervised learning setting, where the goal is to obtain a predictor $f : \mathcal{X} \mapsto \mathcal{Y}$ that maps inputs x (e.g., images, text) to labels y. The specification of an ML pipeline includes the training data $\mathcal{D}$ drawn from a training distribution P; a specification of the model class $\mathcal{F}$ from which a predictor $f(x)$ will be chosen; a procedure for choosing predictors from that class; and a procedure for validating the candidate predictor. Usually, the pipeline selects $f \in \mathcal{F}$ by approximately minimizing the predictive risk on the training distribution $\mathcal{R}_{\mathsf{P}}(f) := \mathbb{E}_{(X,Y)\sim\mathsf{P}}[\ell(f(X), Y)]$ for some loss function ℓ. Here, we focus on the *standard ML pipeline*, which validates the predictor f by evaluating its predictions on an independent and identically distributed test set D' also drawn from P, e.g., a hold-out set selected completely at random. Here, we assume that the pipeline is tuned to return a predictor that exhibits the best achievable validation performance given the constraints of its specification.

In practice, even when core components of the ML pipeline are fixed—such as the training distribution model, specification, and validation procedure—there are still many degrees of freedom that often remain unspecified. These may include the specific optimization algorithm, any number of hyperparameters such as learning rate, weight initialization, data ordering, etc. These details of the ML pipeline are often left to automatic search procedures, or simply randomized, as in the case of weight initialization and data ordering. With advances in automated machine learning (AutoML) techniques such as neural architecture search (see Elsken et al., 2019, for a review), more aspects of ML pipelines are being left as degrees of freedom rather than explicit design decisions.

2.2 Underspecification

We say that an ML pipeline is *underspecified* if there are many predictors f that a pipeline could return that satisfy the pipeline's validation criteria equivalently. This often occurs if there are many degrees of freedom in the pipeline specification, such as its random seed, that have little effect on validation performance. We denote this set of achievable validation-equivalent predictors $\mathcal{F}^* \subset \mathcal{F}$. In the standard ML pipeline, $\mathcal{F}^*$ corresponds to a set of predictors with the highest practically achievable iid performance given the pipeline specification. Previously, Breiman (2001) referred to similar sets of equivalently performant models as "Rashomon sets".

Underspecification creates difficulties when the predictors in $\mathcal{F}^*$ process inputs in systematically different ways, resulting in different generalization behavior on distributions that differ from P. When

this is true, even when $\mathcal{F}^*$ contains a predictor that would behave appropriately in deployment, a pipeline may return a different predictor because it cannot distinguish between them.

The ML literature has studied various notions of underspecification in more theoretical contexts. In the deep learning literature specifically, much of the discussion has focused on the shape of the loss landscape $\mathbb{E}_{(X,Y)\sim\mathsf{P}}[\ell(f(X), Y)]$, and of the geometry of non-unique risk minimizers, including discussions of wide or narrow optima (see, e.g. Chaudhari et al., 2019), and connectivity between global modes in the context of model averaging (Izmailov et al., 2018; Fort et al., 2019; Wilson and Izmailov, 2020) and network pruning (Frankle et al., 2020). Underspecification also plays a role in recent analyses of overparametrization in theoretical and real deep learning models (Belkin et al., 2018; Mei and Montanari, 2019; Nakkiran et al., 2020). Here, underspecification is a direct consequence of having more degrees of freedom than datapoints.

Our work here complements these efforts in two ways: first, our goal is to understand how underspecification relates to model behaviors beyond the training distribution P; and secondly, the primary object that we study is practical ML *pipelines* rather than the loss landscape itself. This latter distinction is particularly important for this work, as there is often a gap between analysis of a learning problem and analysis of an ML pipeline built around that problem. For example, ML pipelines built around well-conditioned learning problems can exhibit underspecification if the choice of convex regularizer is left as a degree of freedom. Likewise, ML pipelines built around poorly conditioned problems may show little sign of underspecification, for example, if the optimization algorithm is fixed and induces a strong implicit regularization. In fact, the pipelines we study incorporate a number of standard tricks, such as early stopping, which are ubiquitous in ML as it is applied to real problems, but can widen the gap between learning theory and practice. This being said, these analyses are clearly connected, as we discuss in Section 3.

Our treatment of underspecification is more closely related to work on "Rashomon sets" (Fisher et al., 2019), "predictive multiplicity" (Marx et al., 2019), and methods that search for validation-equivalent predictors that are "right for the right reasons" (Ross et al., 2017). These lines of work similarly note that a single learning problem specification can admit many near-optimal solutions, and that these solutions may have very different properties along axes such as interpretability or fairness. Here, we draw out additional implications of this phenomenon in practical ML pipelines.

2.3 Structural and Underspecified Failure Modes

Underspecification differs from more commonly-studied structural failure modes that arise in applications of machine learning. In structural failure modes, there is an explicit tension between iid generalization and desirable behavior at deployment time. In these scenarios, a predictor that behaves as required will have inferior iid performance compared to a predictor that uses so-called "spurious" associations that are strongly predictive of the label in the training data, but do not appear in plausible deployment settings.

For example, structural failure modes have been widely reported in medical applications of ML, where the training inputs often include markers of a doctor's diagnostic judgment (Oakden-Rayner et al., 2020). As an illustration, Winkler et al. (2019) report on a CNN model used to diagnose skin lesions that exhibited strong reliance on surgical ink markings around skin lesions that doctors had deemed to be cancerous. Because the judgment that went into the ink markings may have used information not available in the image itself, a predictor that incorporated this feature could achieve better iid predictive performance than one that did not. However, these markings would not be expected to be present in deployment, where the predictor would be used pre-diagnosis, using only unmarked images.

Structural failure modes often indicate a design flaw in the ML pipeline: the learning problem that the ML pipeline was designed to solve is not well-aligned with the real-world problem encountered in deployment. As such, they can often be addressed by changing core parts of the ML pipeine, such as more careful selection of training data. Several algorithmic approaches have also been proposed to

overcome structural issues, including Peters et al. (2016); Heinze-Deml et al. (2018); Arjovsky et al. (2019); Magliacane et al. (2018). These approaches often use data collected in multiple environments to identify causal invariances.

Underspecification is distinct and complementary to structural failure modes. This is evident in settings where structural issues are not present. For example, in many perception problems, the label can be recovered with high certainty using only desired signal (e.g., the shape of the foreground object in an image recognition task), but ML models will often exhibit sensitivity to inappropriate signal (e.g., background features that are correlated with the foreground object at training time), even though this is not necessary to achieve competitive iid performance. In this work, we argue that underspecification factors into these failures.

Geirhos et al. (2020) connects this idea to the notion of "shortcut learning". They point out that there may be many predictors that generalize well in iid settings, but only some that align with the intended solution to the prediction problem. In addition, they also note (as we do) that opaque aspects of ML pipelines, such as the optimization procedure, can make certain features easier for a pipeline to represent, and note the need for future investigation in this area. Our work offers additional empirical support to this argument. Furthermore, we show that even pipelines that are identical up to their random seed can produce predictors that rely differently on distinct shortcuts, emphasizing the relevance of underspecification.

Finally, we note that structural and underspecified failure modes can coexist in the same problem. In problems where there is a fundamental tradeoff between iid performance and required deployment behavior, $\mathcal{F}^*$ may not contain any predictors with ideal behavior in deployment; however, the particular strength of the tradeoff and is often underspecified by the standard ML pipeline. In many of the examples that we consider here, particularly the experiments with deep learning pipelines, both types of failures are likely to be present.

2.4 Stress Tests and Credibility

Our core claims revolve around how underspecification creates ambiguity in real-world predictor behavior, which can undermine the credibility of an ML system. In particular, we are interested in behavior that is *not* tested by iid evaluations, but has observable implications in practically important situations. To assess these behavioral requirements, we use *stress tests*, or evaluations that probe the properties of a predictor by observing its outputs on specifically designed inputs. Strategies of this type are one way to warrant trust in an ML system (Jacovi et al., 2020).

Stress tests are becoming a key part of standards of evidence in a number of applied domains, including medicine (Collins et al., 2015; Liu et al., 2020a; Rivera et al., 2020), economics (Mullainathan and Spiess, 2017; Athey, 2017), public policy (Kleinberg et al., 2015), and epidemiology (Hoffmann et al., 2019). In many settings where stress tests have been proposed in the ML literature, they have often uncovered cases where models fail to generalize as required for direct real-world application. Our aim is to show that underspecification can play a role in these failures.

Here, we review three types of stress tests that we consider in this paper, and make connections to existing literature where they have been applied.

Stratified Performance Evaluations Stratified evaluations (i.e., subgroup analyses) test whether a predictor f behaves differently on inputs from different strata of a dataset. We choose a particular feature A and stratify a standard test dataset $\mathcal{D}'$ into strata $\mathcal{D}'_a = \{(x_i, y_i) : A_i = a\}$. A performance metric can then be calculated and compared across different values of a. Often, A is chosen to be a feature that is not mechanistically related to the label, but is salient in deployment.

Stratified evaluations have been presented in the literature on fairness in machine learning, where examples are stratified by socially salient characteristics like skin type and gender (Buolamwini and Gebru, 2018); the ML for healthcare literature (Obermeyer et al., 2019; Oakden-Rayner et al., 2020), where examples are stratified by subpopulations; and the natural language processing and computer

vision literatures where examples are stratified by topic or notions of difficulty (Hendrycks et al., 2019; Zellers et al., 2018).

Shifted Performance Evaluations Shifted performance evaluations test whether the average performance of a predictor f generalizes when the test distribution differs in a specific way from the training distribution. These tests define a new data distribution $\mathsf{P}' \neq \mathsf{P}$ from which to draw the test dataset $\mathcal{D}'$, then evaluate a performance metric with respect to this shifted dataset.

There are several strategies for generating P', which test different properties of f. For example, to test whether f exhibits invariance to a particular transformation $T(x)$ of the input, one can define P' to be the distribution of the variables $(T(x), y)$, when (x, y) are drawn from the training distribution P (e.g., noising of images in ImageNet-C in Hendrycks and Dietterich (2019)). One can also define P' less formally, for example by changing the data scraping protocol used to collect the test dataset (e.g., ObjectNet in Barbu et al. (2019)), or changing the instrument used to collect data.

Shifted performance evaluations form the backbone of empirical evaluations in the literature on robust machine learning and task adaptation (e.g., Hendrycks and Dietterich, 2019; Wang et al., 2019; Djolonga et al., 2020; Taori et al., 2020). Shifted evaluations are also required in some reporting standards, including those for medical applications of AI (Collins et al., 2015; Liu et al., 2020a; Rivera et al., 2020).

Contrastive Evaluations Shifted evaluations that measure aggregate performance can be useful for surfacing potential deployment failures, but the aggregation involved can obscure more fine-grained patterns. Contrastive evaluations can support localized analysis of particular model behaviors. Contrastive evaluations are performed on the example, rather than distribution level, and check whether a particular modification of the input x causes the output of the model to change in unexpected ways. Formally, a contrastive evaluation makes use of a dataset of matched sets $\mathcal{C} = \{z_i\}_{i=1}^{|\mathcal{C}|}$, where each matched set z_i consists of a base input x_i that is modified by a set of transformations $\mathcal{T}$, $z_i = (T_j(x_i))_{T_j \in \mathcal{T}}$. In contrastive evaluations, metrics are computed with respect to matched sets, and can include, for example, measures of similarity or ordering among the examples in the matched set. For instance, if it is assumed that each transformation in $\mathcal{T}$ should be label-preserving, then a measurement of disagreement within the matched sets reveals unexpected behavior.

Contrastive evaluations are common in the ML fairness literature, e.g., to assess counterfactual notions of fairness (Garg et al., 2019; Kusner et al., 2017). They are also increasingly common as robustness or debugging checks in the natural language processing literature (Ribeiro et al., 2020; Kaushik et al., 2020).

3. Warm-Up: Underspecification in Simple Models

To build intuition for how underspecification manifests in practice, we demonstrate its consequences in three relatively simple models before moving on to study practical deep neural networks. Here, we examine examples of underspecification in three different settings: (1) a simple parametric model for an epidemic in a simulated setting; (2) a linear model in a real-world medical genomics setting, where such models are currently state-of-the-art; and (3) a shallow random feature model in the theoretical infinitely wide limit. In each case, we show how modeling pipelines that have known degeneracies return substantively different predictors when the pipeline is perturbed. This previews our strategy for perturbing deep learning pipelines with unknown degeneracies, where we observe similar results.

3.1 Underspecification in a Simple Epidemiological Model

One core task in infectious disease epidemiology is forecasting the trajectory of an epidemic. Dynamical models are often used for this task. Here, we consider a simple simulated setting where the data is generated exactly from this model; thus, unlike a real setting where model misspecification is a primary concern, the only challenge here is to recover the true parameters of the generating

process, which would enable an accurate forecast. We show that even in this simplified setting, underspecification is a key challenge in the forecasting task.

Specifically, we consider the simple Susceptible-Infected-Recovered (SIR) model that is often used as the basis of epidemic forecasting models in infectious disease epidemiology. This model is specified in terms of the rates at which the number of susceptible (S), infected (I), and recovered (R) individuals in a population of size N, change over time:

$$\frac{dS}{dt} = -\beta \left(\frac{I}{N}\right) S, \qquad \frac{dI}{dt} = -\frac{I}{D} + \beta \left(\frac{I}{N}\right) S, \qquad \frac{dR}{dt} = \frac{I}{D}.$$

In this model, the parameter β represents the transmission rate of the disease from the infected to susceptible populations, and the parameter D represents the average duration that an infected individual remains infectious.

To simulate the forecasting task, we generate a full trajectory from this model for a full time-course T. We specify a forecasting pipeline that takes in data of observed infections up to some time $T_{\text{obs}} < T$, and estimates the parameters (β, D) by minimizing squared-error loss on predicted infections at each timepoint using gradient descent (susceptible and recovered are usually not observed). [1] Importantly, during the early stages of an epidemic, when T_{obs} is small, the parameters of the model are poorly identified by this training task. This is because, at this stage, the number of susceptible is approximately constant at the total population size (N), and the number of infections grows approximately exponentially at rate $\beta - 1/D$. The data only determine this rate. Thus, there are many pairs of parameter values (β, D) that describe the exponentially growing timeseries of infections equivalently well.

However, when used to forecast the trajectory of the epidemic past T_{obs}, these parameters yield very different predictions. In Figure 1(a), we show two predicted trajectories of infections corresponding to two parameter sets (β, D). Despite fitting the observed data identically, these models predict peak infection numbers, for example, that are orders of magnitude apart.

Because the training objective cannot distinguish between parameter sets (β, D) that yield equivalent growth rates $\beta - 1/D$, arbitrary choices in the learning pipeline determine which set of observation-equivalent parameters are returned by the learning algorithm. In Figure 1(c), we show that by changing the point D_0 at which the parameter D is initialized in the least-squares minimization procedure, we obtain a wide variety of predicted trajectories from the model. In addition, the particular distribution used to draw D_0 (Figure 1(b)) has a substantial influence on the distribution of predicted trajectories.

In realistic epidemiological models that have been used to inform policy, underspecification is dealt with by testing models in forecasting scenarios (i.e., stress testing), and constraining the problem with domain knowledge and external data, for example about viral dynamics in patients (informing D) and contact patterns in the population (informing β) (see, e.g. Flaxman et al., 2020).

3.2 Underspecification in a Linear Polygenic Risk Score Model

For a more realistic example of underspecification, we turn to polygenic risk scores. Polygenic risk scores (PRS) in medical genomics leverage patient genetic information (genotype) to predict clinically relevant characteristics (phenotype). Typically, they are linear models built on categorical features that represent genetic variants. PRS have shown great success in some settings (Khera et al., 2018), but face difficulties when applied to new patient populations (Martin et al., 2017; Duncan et al., 2019; Berg et al., 2019).

We show that underspecification plays a role in this difficulty with generalization. Specifically, we show that there is a non-trivial set of predictors $\mathcal{F}^*$ that show strong performance in iid validations, but transfer very differently to a new population. Thus, a modeling pipeline based on iid performance alone cannot reliably return a predictor that transfers well.

1. Here, we omit a separate validation procedure because in-sample fit is sufficient validation in such a simple model.

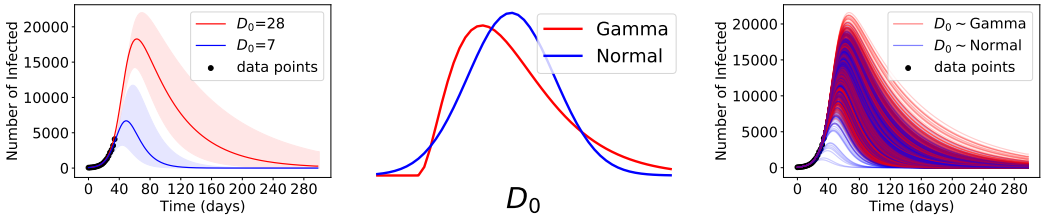

Figure 1: **Underspecification in a simple epidemiological model.** A training pipeline that only minimizes predictive risk on early stages of the epidemic leaves key parameters underspecified, making key behaviors of the model sensitive to arbitrary training choices. Because many parameter values are equivalently compatible with fitting data from early in the epidemic, the trajectory returned by a given training run depends on where it was initialized, and different initialization distributions result in different distributions of predicted trajectories.

To construct distinct, validation-equivalent predictors, we exploit a core ambiguity in PRS, namely, that many genetic variants that are used as features are nearly collinear. This collinearity makes it difficult to distinguish causal and correlated-but-noncausal variants (Slatkin, 2008). A common approach to this problem is to partition variants into clusters of highly-correlated variants and to only include one representative of each cluster in the PRS (e.g., International Schizophrenia Consortium et al., 2009; CARDIoGRAMplusC4D Consortium et al., 2013). Usually, standard heuristics are applied to choose clusters and cluster representatives as a pre-processing step (e.g., "LD clumping", Purcell et al., 2007).

Importantly, because of the high correlation of features within clusters, the choice of cluster representative leaves the iid risk of the predictor largely unchanged. Thus, distinct PRS predictors that incorporate different cluster representatives can be treated as members of the validation-equivalent set $\mathcal{F}^*$. However, this choice has strong consequences for model generalization.

To demonstrate this effect, we examine how feature selection influences behavior in a stress test that simulates transfer of PRS across populations. Using data from the UK Biobank (Sudlow et al., 2015), we examine how a PRS predicting a particular continuous phenotype called the *intraocular pressure* (IOP) transfers from a predominantly British training population to "non-British" test population (see Appendix D for definitions). We construct an ensemble of 1000 PRS predictors that sample different representatives from each feature cluster, including one that applies a standard heuristic from the popular tool PLINK (Purcell et al., 2007).

The three plots on the left side of Figure 2 confirm that each predictor with distinct features attains comparable performance in the training set (left panel) and iid test set (middle panel), with the standard heuristic (red dots) slightly outperforming random representative selection. However, on the shifted "non-British" test data (right panel), we see far wider variation in performance, and the standard heuristic fares no better than the rest of the ensemble. More generally, within this set of predictors, performance on the British test set is only weakly associated with performance on the "non-British" set (Spearman $\rho = 0.135; 95\%$ CI 0.070-0.20; Figure 2, right).

Thus, because the PRS training pipeline is underspecified, it cannot reliably return a predictor that transfers as required between populations, despite some predictors in $\mathcal{F}^*$ having acceptable transfer performance. For full details of this experiment and additional background information, see Appendix D.

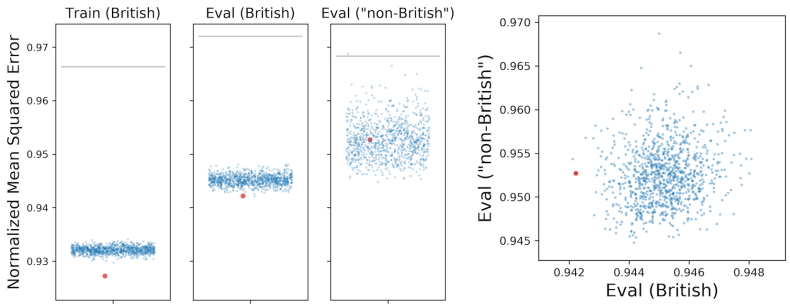

Figure 2: **Underspecification in linear models in medical genomics. (Left)** Performance of a PRS model using genetic features in the British training set, the British evaluation set, and the "non-British" evaluation set, as measured by the normalized mean squared error (MSE divided by the true variance, lower is better). Each dot represents a PRS predictor (using both genomic and demographic features). Large red dots are PRS predictors using the "index" variants of the clusters of correlated features selected by PLINK. Gray lines represent the baseline models using only demographic information. The x-axis is random jitter added to visualize the points more easily. The increased error from the left panel to the middle panel is the standard generalization gap from train to test on iid data; meanwhile, the large dispersion of errors in the right panel relative to the middle panel is evidence of underspecification of transfer accuracy. **(Right)** Comparison of model performance (NMSE) in British and "non-British" eval sets (middle and right panels on the Left side of the figure), given the same set of genomic features (Spearman $\rho = 0.135; 95\%$ CI 0.070-0.20). Transfer performance is difficult to predict from iid performance.

3.3 Theoretical Analysis of Underspecification in a Random Feature Model

We close this set of examples with a theoretical analysis of the population risk of random feature models in the infinitely-wide limit. The content of this section is substantially more mathematical than the rest of the paper, and can be safely skipped by readers more interested in empirical results.

In this problem, underpsecification results from overparameterization: when there are more parameters than datapoints, the learning problem is inherently underspecified. Overparameterization is a key property of many modern neural network models. Much recent work has shown that this underspecification has interesting regularizing effects on iid generalization, but there has been little focus on its impact on how models behave on other distributions. Here, we show that we can recover the effect of underspecification on out-of-distribution generalization in an asymptotic analysis of a simple random feature model, which is often used as a model system for neural networks in the infinitely wide regime.

We consider for simplicity a regression problem: we are given data $\{(\boldsymbol{x}_i, y_i)\}_{i \leq n}$, with $\boldsymbol{x}_i \in \mathbb{R}^d$ vector of covariates and $y_i \in \mathbb{R}$ a response. As a tractable and yet mathematically rich setting, we use the random features model of Neal (1996) and Rahimi and Recht (2008). This is a one-hidden-layer neural network with random first layer weights $\boldsymbol{W}$ and learned second layer weights $\boldsymbol{\theta}$. We learn a predictor $f_{\boldsymbol{W}} : \mathbb{R}^d \to \mathbb{R}$ of the form

$$f_{\boldsymbol{W}}(\boldsymbol{x}) = \boldsymbol{\theta}^{\mathsf{T}} \sigma(\boldsymbol{W}\boldsymbol{x}).$$

Here, $\boldsymbol{W} \in \mathbb{R}^{N \times d}$ is a random matrix with rows $\boldsymbol{w}_i \in \mathbb{R}^d$, $1 \leq i \leq N$ that are not optimized and define the featurization map $\boldsymbol{x} \mapsto \sigma(\boldsymbol{W}\boldsymbol{x})$. We take $(\boldsymbol{w}_i)_{i \leq N}$ to be iid and uniformly random with $\|\boldsymbol{w}_i\|_2 = 1$. We consider data $(\boldsymbol{x}_i, y_i)$, where $\boldsymbol{x}_i$ are uniformly random with $\|\boldsymbol{x}_i\|_2 = \sqrt{d}$ and a linear target $y_i = f_*(\boldsymbol{x}_i) = \boldsymbol{\beta}_0^{\mathsf{T}} \boldsymbol{x}_i$.

We analyze this model in a setting where both the number of datapoints n and the neurons N both tend toward infinity with a fixed overparameterization ratio N/n. For $N/n < 1$, we learn the

second layer weights using least squares. For $N/n \geq 1$ there exist choices of the parameters $\boldsymbol{\theta}$ that perfectly interpolate the data $f_{\boldsymbol{W}}(\boldsymbol{x}_i) = y_i$ for all $i \leq n$. We choose the minimum ℓ_2-norm interpolant (which is the model selected by GD when $\boldsymbol{\theta}$ is initialized at 0):

$$\begin{aligned} &\text{minimize } \|\boldsymbol{\theta}\| \\ &\text{subject to } f_{\boldsymbol{W}}(\boldsymbol{x}_i) = y_i \text{ for all } i. \end{aligned}$$

We analyze the predictive risk of the predictor $f_{\boldsymbol{W}}$ on two test distributions, P, which matches the training distribution, and P_Δ, which is perturbed in a specific way that we describe below. For a given distribution Q, we define the prediction risk as the mean squared error for the random feature model derived from $\boldsymbol{W}$ and for a test point sampled from Q:

$$R(\boldsymbol{W}, \mathsf{Q}) = \mathbb{E}_{(X,Y)\sim\mathsf{Q}}(Y - \hat{\boldsymbol{\theta}}(\boldsymbol{W})^\mathsf{T}\sigma(\boldsymbol{W}X))^2.$$

This risk depends implicitly on the training data through $\hat{\boldsymbol{\theta}}$, but we suppress this dependence.

Building on the work of Mei and Montanari (2019) we can determine the precise asymptotics of the risk under certain distribution shifts in the limit $n, N, d \to \infty$ with fixed ratios n/d, N/n. We provide detailed derivations in Appendix E, as well as characterizations of other quantities such as the sensitivity of the prediction function $f_{\boldsymbol{W}}$ to the choice of $\boldsymbol{W}$.

In this limit, any two independent random choices $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ induce trained predictors $f_{\boldsymbol{W}_1}$ and $f_{\boldsymbol{W}_2}$ that have indistinguishable in-distribution error $R(\boldsymbol{W}_i, \mathsf{P})$. However, given this value of the risk, the prediction functions $f_{\boldsymbol{W}_1}(\boldsymbol{x})$ and $f_{\boldsymbol{W}_2}(\boldsymbol{x})$ are nearly as orthogonal as they can be, and this leads to very different test errors on certain shifted distributions P_Δ.

Specifically, we define P_Δ in terms of an adversarial mean shift. We consider test inputs $\boldsymbol{x}_{\text{test}} = \boldsymbol{x}_0 + \boldsymbol{x}$, where $\boldsymbol{x}$ is an independent sample from the training distribution, but $\boldsymbol{x}_0$ is a constant mean-shift defined with respect to a fixed set of random feature weights W_0. We denote this shifted distribution with $\mathsf{P}_{\Delta,\boldsymbol{W}_0}$. For a given $\boldsymbol{W}_0$, a shift $\boldsymbol{x}_0$ can be chosen such that (1) it has small norm ($||\boldsymbol{x}_0|| < \Delta \ll ||\boldsymbol{x}||$), (2) it leaves the risk of an independently sampled $\boldsymbol{W}$ mostly unchanged ($R(\boldsymbol{W}, \mathsf{P}_{\Delta,\boldsymbol{W}_0}) \approx R(\boldsymbol{W}, \mathsf{P}_{\text{train}})$), but (3) it drastically increases the risk of $\boldsymbol{W}_0$ ($R(\boldsymbol{W}_0, \mathsf{P}_{\Delta,\boldsymbol{W}_0}) > R(\boldsymbol{W}_0, \mathsf{P}_{\text{train}})$). In Figure 3 we plot the risks $R(\boldsymbol{W}, \mathsf{P}_{\Delta,\boldsymbol{W}_0})$ and $R(\boldsymbol{W}_0, \mathsf{P}_{\Delta,\boldsymbol{W}_0})$ normalized by the iid test risk $R(\boldsymbol{W}, \mathsf{P}_{\text{train}})$ as a function of the overparameterization ratio for two different data dimensionalities. The upper curves correspond to the risk for the model against which the shift was chosen adversarially, producing a 3-fold increase in risk. Lower curves correspond to the risk for the same distributional shift for the independent model, resulting in very little risk inflation.

These results show that any predictor selected by min-norm interpolation is vulnerable to shifts along a certain direction, while many other models with equivalent risk are not vulnerable to the same shift. The particular shift itself depends on a random set of choices made during model training. Here, we argue that similar dynamics are at play in many modern ML pipelines, under distribution shfits that reveal practically important model properties.

4. Empirical Strategy for Probing Underspecification in Deep Learning Pipelines

Having motivated the notion of underspecification in ML pipelines, we now turn to an empirical study of underspecification in modern deep learning pipelines. We show that underspecification affects the real-world behavior of predictors trained using the standard ML pipeline in three domains: computer vision (including both basic research and medical imaging), natural language processing, and clinical risk prediction using electronic health records. In each case, we use a simple experimental protocol to show that the standard pipeline admits a non-trivial set $\mathcal{F}^*$ of high-performing validation-equivalent predictors, but that different predictors in $\mathcal{F}^*$ exhibit systematically different real-world behavior.

Similarly to our approach in Section 3, our protocol approaches underspecification constructively by instantiating a set of predictors from the set $\mathcal{F}^*$ and probing their behavior. However, for deep

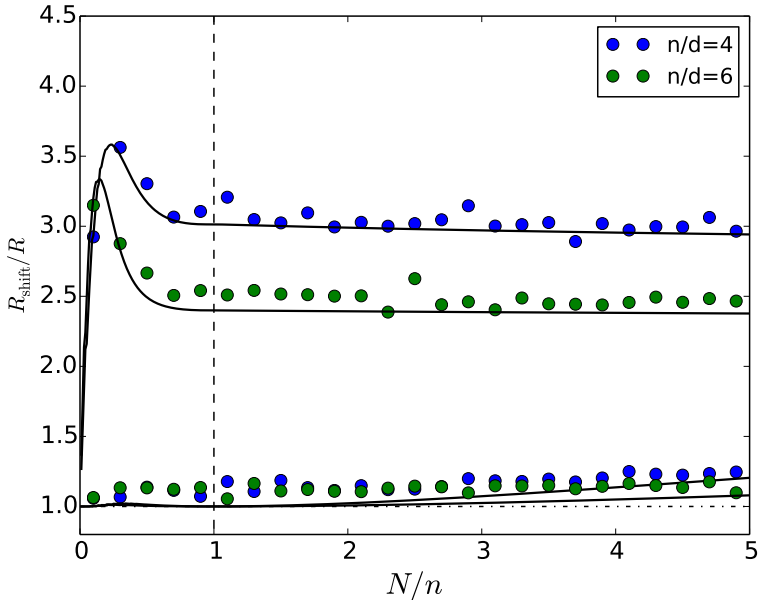

Figure 3: **Random feature models with identical in-distribution risk show distinct risks under mean shift.** Expected risk (averaging over random features $\boldsymbol{W}_0$, $\boldsymbol{W}$) of predictors $f_{\boldsymbol{W}_0}$, $f_{\boldsymbol{W}}$ under a $\boldsymbol{W}_0$-adversarial mean-shift at different levels of overparameterization (N/n) and sample size-to-parameter ratio (n/d). Upper curves: Normalized risk $\mathsf{E}_{\boldsymbol{W}_0} R(\boldsymbol{W}_0; \mathsf{P}_{\boldsymbol{W}_0,\Delta})/\mathsf{E}_{\boldsymbol{W}} R(\boldsymbol{W}; \mathsf{P})$ of the adversarially targeted predictor $f_{\boldsymbol{W}_0}$. Lower curves: Normalized risk $\mathsf{E}_{\boldsymbol{W},\boldsymbol{W}_0} R(\boldsymbol{W}; \mathsf{P}_{\boldsymbol{W}_0,\Delta})/\mathsf{E}_{\boldsymbol{W}} R(\boldsymbol{W}; \mathsf{P})$ of a predictor $f_{\boldsymbol{W}}$ defined with independently drawn random weights $\boldsymbol{W}$. Under this particular shift, the normalized risk of the predictor $f_{\boldsymbol{W}_0}$ is several times larger than the normalized risk of $f_{\boldsymbol{W}}$ across a range of overparameteriation settings, despite having indistinguishable risk under the iid test distribution P'. Here the input dimension is $d = 80$, N is the number of neurons, and n the number of samples. We use ReLU activations; the ground truth is linear with $\|\boldsymbol{\beta}_0\|_2 = 1$. Circles are empirical results obtained by averaging over 50 realizations. Continuous lines correspond to the analytical predictions detailed in the supplement.

models, it is difficult to specify predictors in this set analytically. Instead, we construct an ensemble of predictors from a given model by perturbing small parts of the ML pipeline (e.g., the random seed used in training), and retraining the model several times. When there is a non-trivial set $\mathcal{F}^*$, such small perturbations are often enough to push the pipeline to return a different choice $f \in \mathcal{F}^*$. This strategy does not yield an exhaustive exploration of $\mathcal{F}^*$; rather, it is a conservative indicator of which predictor properties are well-constrained and which are underspecified by the modeling pipeline.

Once we obtain an ensemble, we make several measurements. First, we empirically confirm that the models in the ensemble have near-equivalent iid performance, and can thus be considered to be members of $\mathcal{F}^*$. Secondly, we evaluate the ensemble on one or more application-specific stress tests that probe an aspect of model behavior that is important in the deployment context (see Section 2.4). Variability in stress test performance—beyond what can be explained by differences in iid performance—provides evidence that the modeling pipeline is underspecified along a practically important dimension.

The experimental protocol we use to probe underspecification is closely related to uncertainty quantification approaches based on deep ensembles (e.g., Lakshminarayanan et al., 2017; Dusenberry

et al., 2020). In particular, by averaging across many randomly perturbed predictors from a single modeling pipeline, deep ensembles have been shown to be effective tools for detecting out-of-distribution inputs, and correspondingly for tamping down the confidence of predictions for such inputs (Snoek et al., 2019). These approaches probe a particular notion of algorithmic stability, where the training pipeline for a model is perturbed even when the data are held constant (Yu et al., 2013).

To establish that observed variability in stress test performance is a genuine indicator of underspecification, we evaluate three properties.

- First, we consider the *magnitude* of the variation, either relative to iid performance (when they are on the same scale), or relative to external benchmarks, such as comparisons between ML pipelines featuring different model architectures.
- Secondly, when sample size permits, we consider *unpredictability* of the variation from iid performance. Even if the observed difference in iid performance in our ensemble is small, if stress test performance tracks closely with iid performance, this would suggest that our characterization of $\mathcal{F}^*$ is too permissive. We assess this by computing correlation coefficients between the iid validation metric and the stress test metric.
- Finally, we establish that the variation in stress tests indicates *systematic differences* between the predictors in the ensemble. Often, the magnitude of variation in stress test performance alone will be enough to establish systematicness. However, in some cases we supplement with a mixture of quantitative and qualitative analyses of stress test outputs to illustrate that the differences between predictors does align with important dimensions of the application.

In all cases that we consider, we find evidence that practically important predictor behaviors are underspecified by the standard ML pipeline. In some cases the evidence is obvious, while in others it is more subtle, owing in part to the conservative nature of our exploration of $\mathcal{F}^*$. Our results interact with a number of research areas in each of the fields that we consider, so we close each case study with a short application-specific discussion.

5. Case Studies in Computer Vision

Computer vision is one of the flagship application areas in which deep learning on large-scale training sets has advanced the state of the art. Here, we focus on an image classification task, specifically on the ImageNet validation set (Deng et al., 2009). We examine two pipeline specifications: one that trains a ResNet-50 model (He et al., 2016) on ImageNet, and another that pretrains a ResNet-101x3 Big Transfer (BiT) model (Kolesnikov et al., 2019) on the JFT-300M dataset (Sun et al., 2017) then fine-tunes it on ImageNet. The former is a standard baseline in image classification. The latter is scaled-up ResNet designed for transfer learning, which attains state-of-the-art, or near state-of-the-art, on many image classification benchmarks, including ImageNet.

A key challenge in computer vision is robustness under distribution shift. It has been well-documented that many deep computer vision models suffer from brittleness under distribution shifts that humans do not find challenging (Goodfellow et al., 2016; Hendrycks and Dietterich, 2019; Barbu et al., 2019). This brittleness has raised questions about deployments in open-world high-stakes application, and has given rise to an active literature on robustness in image classification (see, e.g., Taori et al., 2020; Djolonga et al., 2020). Recent work has connected lack of robustness to computer vision models' encoding counterintuitive features (Ilyas et al., 2019; Geirhos et al., 2019; Yin et al., 2019; Wang et al., 2020).

Here, we show concretely that the pipelines we study are underspecified in ways that align with distribution shifts. Going beyond previous work that show degradation in model performance across shifts, here we show that predictors trained by identical pipelines (up to random seed) can have very

different responses to corruptions and shifts, even when their iid performance is held nearly constant. We begin by constructing ensembles of trained ResNet-50 and BiT models: we train 50 ResNet-50 models on ImageNet using identical pipelines that differ only in their random seed, 30 BiT models that are initialized at the same JFT-300M-trained checkpoint, and differ only in their fine-tuning seed and initialization distributions (10 runs each of zero, uniform, and Gaussian initializations). On the ImageNet validation set, the ResNet-50 predictors achieve a $75.9\% \pm 0.11$ top-1 accuracy, while the BiT predictors achieve a $86.2\% \pm 0.09$ top-1 accuracy.

We evaluate these predictor ensembles on two stress tests that have been proposed in the image classification robustness literature: ImageNet-C (Hendrycks and Dietterich, 2019) and ObjectNet (Barbu et al., 2019). ImageNet-C is a benchmark dataset that replicates the ImageNet validation set, but applies synthetic but realistic corruptions to the images, such as pixelation or simulated snow, at varying levels of intensity. ObjectNet is a crowdsourced benchmark dataset that covers a set of classes included in the ImageNet validation set, but depicts them in a wider variety of settings and configurations. Both stress tests have been used as prime examples of the lack of human-like robustness in deep image classification models.

5.1 ImageNet-C

We show results from the evaluation on several ImageNet-C tasks in Figure 4. The tasks we show here incorporate corruptions at their highest intensity levels (level 5 in the benchmark). In the figure, we highlight variability in the accuracy across predictors in the ensemble, relative to the variability in accuracy on the standard iid test set. For both the ResNet-50 and BiT predictors, variation on some ImageNet-C tasks is an order of magnitude larger than variation in iid performance. Furthermore, within this ensemble, there is weak sample correlation between performance on the iid test set and performance on each benchmark stress test, and performance between tasks (all 95% CI's for Pearson correlation using $n = 50$ and $n = 30$ contain zero, see Figure 5). This indicates that the differences in stress test performance cannot be explained by some predictors simply being better-trained than others. For context, we report absolute model accuracies and ensemble standard deviations in Table 1.

5.2 ObjectNet

We also evaluate these ensembles along more "natural" shifts in the ObjectNet test set. Here, we compare the variability in model performance on the ObjectNet test set to a subset of the standard ImageNet test set with the 113 classes that appear in ObjectNet. The results of this evaluation are in Table 1. The relative variability in accuracy on the ObjectNet stress test is larger that the variability seen in the standard test set (standard deviation is 2x for ResNet-50 and 5x for BiT), although the difference in magnitude is not as striking as in the ImageNet-C case in Figure 4. There is also a slightly stronger relationship between standard test accuracy and test accuracy on ObjectNet (Spearman ρ 0.22 $(-0.06, 0.47)$ for ResNet-50, 0.47 $(0.13, 71)$ for BiT).

Nonetheless, the variability in accuracy suggests that some predictors in the ensembles are systematically better or worse at making predictions on the ObjectNet test set. We quantify this with p-values from a one-sided permutation test, which we interpret as descriptive statistics. Specifically, we compare the variability in model performance on the ObjectNet test set with variability that would be expected if prediction errors were randomly distributed between predictors. The variability of predictor accuracies on ObjectNet is large compared to this baseline ($p = 0.002$ for ResNet-50 and $p = 0.000$ for BiT). On the other hand, the variability between predictor accuracies on the standard ImageNet test set are more typical of what would be observed if errors were randomly distributed ($p = 0.203$ for ResNet-50 and $p = 0.474$ for BiT). In addition, the predictors in our ensembles disagree far more often on the ObjectNet test set than they do in the ImageNet test set, whether or not we

Dataset	ImageNet	pixelate	contrast	motion blur	brightness	ObjectNet
ResNet-50	0.759 (0.001)	0.197 (0.024)	0.091 (0.008)	0.100 (0.007)	0.607 (0.003)	0.259 (0.002)
BiT	0.862 (0.001)	0.555 (0.008)	0.462 (0.019)	0.515 (0.008)	0.723 (0.002)	0.520 (0.005)

Table 1: **Accuracies of ensemble members on stress tests.** Ensemble mean (standard deviations) of accuracy proportions on ResNet-50 and BiT models.

Dataset	ImageNet	ImageNet (subset)	ObjectNet
ResNet-50	0.160 (0.001)	0.245 (0.005)	0.509 (0.003)
BiT	0.064 (0.004)	0.094 (0.006)	0.253 (0.012)

Table 2: **Ensemble disagreement proportions for ImageNet vs ObjectNet models.** Average disagreement between pairs of predictors in the ResNet and BiT ensembles. The "subset" test set only includes classes that also appear in the ObjectNet test set. Models show substantially more disagreement on the ObjectNet test set.

consider the subset of the ImageNet test set examples that have classes that appear in ObjectNet (Table 2).

5.3 Conclusions

Our results show that practically important properties of image classifiers are poorly constrained by the standard ML pipeline. The fact that underspecification persists in the scaled-up BiT pipeline is particularly notable, because scaling up data and model size has become a popular strategy for improving performance across a wide range of robustness stress tests, aligning closely with how much this scaling improves performance on iid evaluations (Djolonga et al., 2020; Taori et al., 2020; Hendrycks et al., 2020). However, even as overall performance on stress tests improves, iid-performance-preserving perturbations to the pipeline still induce variability in stress test performance, indicating that important real-world behaviors are still not fully specified. Improving the composition of training data has the potential to mitigate underspecification, but these results suggest that underspecification requires more special attention as scaling strategies are explored.

6. Case Studies in Medical Imaging

Medical imaging is one of the primary high-stakes domains where deep image classification models are directly applicable. In this section, we examine underspecification in two medical imaging models designed for real-world deployment. The first classifies images of patient retinas, while the second classifies clinical images of patient skin. We show that, when trained using the standard ML pipeline, the behavior of these models is underspecified along dimensions that are practically important for deployment. These results confirm the need for explicitly testing and monitoring medical ML models in settings that accurately represent the deployment domain, as codified in recent best practices (Collins et al., 2015; Kelly et al., 2019; Rivera et al., 2020; Liu et al., 2020a).

6.1 Ophthalmological Imaging

Deep learning models have shown great promise in the ophthalmological domain (Gulshan et al., 2016; Ting et al., 2017). Here, we consider a pipeline that trains one such model to predict diabetic retinopathy (DR) and referable diabetic macular edema (DME) from retinal fundus images. The pipeline trains a model with an Inception-V4 backbone (Szegedy et al., 2017) by first pretraining on ImageNet, and then fine-tuning using de-identified retrospective fundus images from EyePACS in the

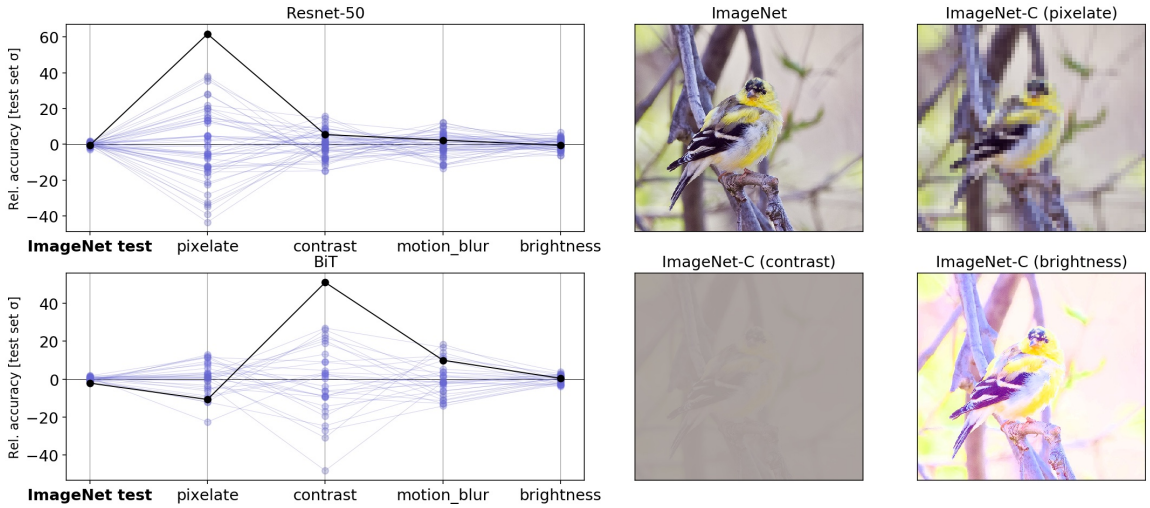

Figure 4: **Image classification model performance on stress tests is sensitive to random initialization in ways that are not apparent in iid evaluation. (Top Left)** Parallel axis plots showing the variation in accuracy between identically trained, randomly initialized ResNet-50 models on strongly corrupted ImageNet-C data. Lines represent the performance of each predictor in the ensemble on classification tasks using uncorrupted test data, as well as corrupted data (pixelation, contrast, motion blur, and brightness). Given values are the deviation in accuracy from the ensemble mean, scaled by the standard deviation of accuracies on the "clean" ImageNet test set. The solid black line highlights the performance of an arbitrarily selected model to show how performance on one test may not be a good indication of performance on others. **(Right)** Example image from the standard ImageNet validation set, with corrupted versions from the ImageNet-C benchmark. Right: Example images from the standard ImageNet test set, with corrupted versions from the ImageNet-C benchmark.

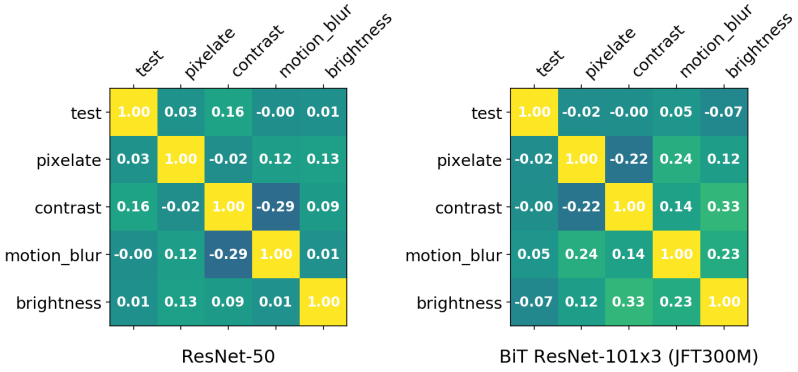

Figure 5: **Performance on ImageNet-C stress tests is unpredictable from standard test performance.** Spearman rank correlations of predictor performance, calculated from random initialization predictor ensembles. (Left) Correlations from 50 retrainings of a ResNet-50 model on ImageNet. (Right) Correlations from 30 ImageNet fine-tunings of a ResNet-101x3 model pre-trained on the JFT300M dataset.

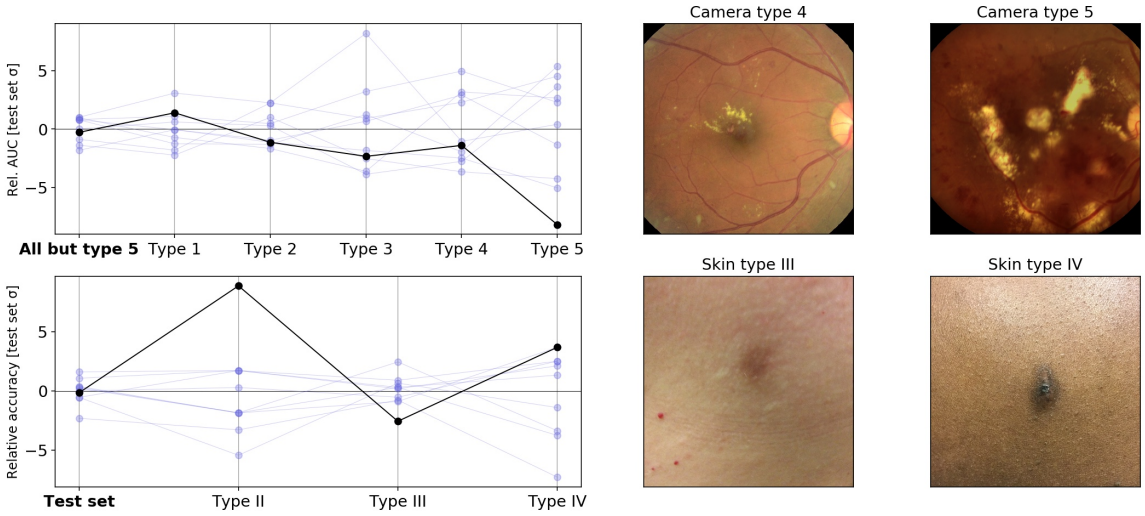

Figure 6: **Stress test performance varies across identically trained medical imaging models.** Points connected by lines represent metrics from the same model, evaluated on an iid test set (bold) and stress tests. Each axis shows deviations from the ensemble mean, divided by the standard deviation for that metric in the standard iid test set. These models differ only in random initialization at the fine-tuning stage. **(Top Left)** Variation in AUC between identical diabetic retinopathy classification models when evaluated on images from different camera types. Camera type 5 is a camera type that was not encountered during training. **(Bottom Left)** Variation in accuracy between identical skin condition classification models when evaluated on different skin types. **(Right)** Example images from the original test set (left) and the stress test set (right). Some images are cropped to match the aspect ratio.

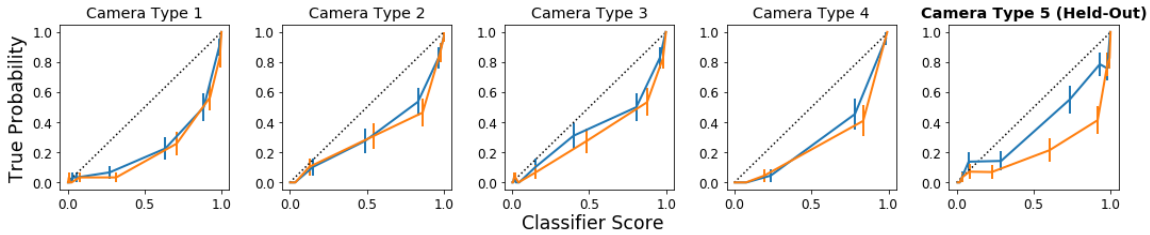

Figure 7: **Identically trained retinal imaging models show systematically different behavior on stress tests.** Calibration plots for two diabetic retinopathy classifiers (orange and blue) that differ only in random seed at fine-tuning. Calibration characteristics of the models are nearly identical for each in-distribution camera type 1–4, but are qualitatively different for the held-out camera type 5. Error bars are ±2 standard errors.

United States and from eye hospitals in India. Dataset and model architecture details are similar to those in (Krause et al., 2018). For our experiment, we restrict the pipeline to incorporate only standard iid validation.

A key use case for these models is to augment human clinical expertise in underserved settings, where doctor capacity may be stretched thin. As such, generalization to images taken by a range of cameras, including those deployed at different locations and clinical settings, is essential for system usability (Beede et al., 2020).

Here, we show that the performance of predictors produced by the standard ML pipeline is sensitive to underspecification. Specifically, we construct an ensemble of 10 models that differ only in random initialization at the fine-tuning stage. We evaluate these models on stress tests predicting DR using images taken by a camera type not encountered during training and validation.

The results are shown in Figure 6. Measuring accuracy in terms of AUC, variability in AUC on the held-out camera type is larger than that in the standard test set, both in aggregate, and compared to most strata of camera types in the training set. To establish that this larger variability is not easily explained away by differences in sample size, we conduct a two-sample z-test comparing the AUC standard deviation in the held-out camera test set ($n = 287$) against the AUC standard deviation in the standard test set ($n = 3712$) using jackknife standard errors, obtaining a z-value of 2.47 and a one-sided p-value of 0.007. In addition, models in the ensemble differ systematically in ways that are not revealed by performance in the standard test set. For example, in Figure 7, we show calibration plots of two predictors from the ensemble computed across camera types. The predictors have similar calibration curves for the cameras encountered during training, but have markedly different calibration curves for the held-out camera type. This suggests that these predictors process images in systematically different ways that only become apparent when evaluated on the held-out camera type.

6.2 Dermatological Imaging

Deep learning based image classification models have also been explored for applications in dermatology (Esteva et al., 2017). Here, we examine a model proposed in Liu et al. (2020b) that is trained to classify skin conditions from clinical skin images. This model has a similar Inception-V4 architecture to the one in Section 6.1. We also use a similar pipeline in this experiment: we pretrain on ImageNet, fine-tune on dermatological images, and restrict valdation to iid accuracy evaluations.

In this setting, one key concern is that the model may have variable performance across skin types, especially when these skin types are differently represented in the training data. Given the social salience of skin type, this concern is aligned with broader concerns about ensuring that machine learning does not amplify existing healthcare disparities (Adamson and Smith, 2018). In dermatology in particular, differences between the presentation of skin conditions across skin types has been linked to disparities in care (Adelekun et al., 2020).

Here, we show that model performance across skin types is sensitive to underspecification. Specifically, we construct an ensemble of 10 predictors with randomly initialized fine-tuning layer weights. We then evaluate the predictors on a stress test that stratifies the test set by skin type on the Fitzpatrick scale (Fitzpatrick, 1975) and measures Top-1 accuracy within each slice.

The results are shown at the bottom of Figure 6. Compared to overall test accuracy, there is larger variation in test accuracy within skin type strata across models, particularly in skin types II and IV, which form substantial portions ($n = 437$, or 10.7%, and $n = 798$, or 19.6%, respectively) of the test data. Based on this test set, some predictors in this ensemble would be judged to have higher discrepancies across skin types than others, even though they were all produced by an identical training pipeline.

Because the sample sizes in each skin type stratum differ substantially, we use a permutation test to explore the extent to which the larger variation in some subgroups can be accounted for by sampling size. This test shuffles the skin type indicators across examples in the test set, then calculates the variance of the accuracy across these random strata. We compute one-sided p-values with respect to this null distribution and interpret them as exploratory descriptive statistics. The key question is whether the larger variability in some strata, particularly skin types II and IV, can be explained away by sampling noise alone. (Our expectation is that skin type III is both large enough and similar enough to the iid test set that its accuracy variance should be similar to the overall variance, and the sample size for skin type V is so small that a reliable characterization would be difficult.) Here, we find that the variation in accuracy in skin types III and V are easily explained by

sampling noise, as expected ($p = 0.54, n = 2619$; $p = 0.42, n = 109$). Meanwhile the variation in skin type II is largely consistent with sampling noise ($p = 0.29, n = 437$), but the variation in skin type IV seems to be more systematic ($p = 0.03, n = 798$). These results are exploratory, but they suggest a need to pay special attention to this dimension of underspecification in ML models for dermatology.

6.3 Conclusions

Overall, the vignettes in this section demonstrate that underspecification can introduce complications for deploying ML, even in application areas where it has the potential to highly beneficial. In particular, these results suggest that one cannot expect ML models to automatically generalize to new clinical settings or populations, because this generalization behavior is underspecified. This confirms the need to tailor and test models for the clinical settings and populations in which they will be deployed. While current strategies exist to mitigate these concerns, addressing underspecification, and generalization issues more generally, could reduce a number of points of friction at the point of care (Beede et al., 2020).

7. Case Study in Natural Language Processing

Deep learning models play a major role in modern natural language processing (NLP). In particular, large-scale Transformer models (Vaswani et al., 2017) trained on massive unlabeled text corpora have become a core component of many NLP pipelines (Devlin et al., 2019). For many applications, a successful recipe is to pretrain by language modeling or denoising a large unlabeled corpus, and then fine-tune using labeled data from a task of interest, sometimes no more than a few hundred examples (e.g., Howard and Ruder, 2018; Peters et al., 2018). This workflow has yielded strong results across a wide range of tasks in natural language processing, including machine translation, question answering, summarization, sequence labeling, and more. As a result, a number of NLP products are built on top of publicly released pretrained checkpoints of language models such as BERT (Devlin et al., 2019).

However, recent work has shown that NLP systems built with this pattern often rely on "shortcuts" (Geirhos et al., 2020), which may be based on spurious phenomena in the training data (McCoy et al., 2019b). Shortcut learning presents a number of difficulties in natural language processing: for example, failure to satisfy intuitive invariances, such as invariance to typographical errors or seemingly irrelevant word substitutions (Ribeiro et al., 2020); ambiguity in measuring progress in language understanding (Zellers et al., 2019); and reliance on stereotypical associations with race and gender (Caliskan et al., 2017; Rudinger et al., 2018; Zhao et al., 2018; De-Arteaga et al., 2019).

In this section, we show that underspecification plays a role in shortcut learning in the pretrain/fine-tune approach to NLP, in both stages. In particular, we show that reliance on specific shortcuts can vary substantially between predictors that differ only in their random seed at fine-tuning or pretraining time. Following our experimental protocol, we perform this case study with an ensemble of predictors obtained from identical training pipelines that differ only in the specific random seed used at pretraining and/or fine-tuning time. Specifically, we train 5 instances of the BERT "large-cased" language model (Devlin et al., 2019), using the same Wikipedia and BookCorpus data that was used to train the public checkpoints. This model has 340 million parameters, and is the largest BERT model with publicly released pretraining checkpoints. For tasks that require fine-tuning, we fine-tune each of the five checkpoints 20 times using different random seeds.

In each case, we evaluate the ensemble of predictors on stress tests designed to probe for specific shortcuts, focusing on shortcuts based on stereotypical correlations, and find evidence of underspecification along this dimension in both pretraining and fine-tuning. As in the other cases we study here, these results suggest that shortcut learning is not enforced by model architectures, but can be a symptom of ambiguity in the ML pipeline.

Underspecification has a wider range of implications in NLP. We connect our results to previous work that reported instability in performance on stress tests designed to diagnose shortcut learning in Natural Language Inference tasks (McCoy et al., 2019b; Naik et al., 2018). Using the same protocol, we replicate the results (McCoy et al., 2019a; Dodge et al., 2020; Zhou et al., 2020), and extend them to show sensitivity to the pretraining random seed. We also explore how underspecification affects representations in static word embeddings.

7.1 Gendered Correlations in Downstream Tasks

We begin by examining gender-based shortcuts on two previously proposed benchmarks: a semantic textual similarity (STS) task and a pronoun resolution task.

7.1.1 Semantic textual similarity (STS)

In the STS task, a predictor takes in two sentences as input and scores their similarity. We obtain predictors for this task by fine-tuning BERT checkpoints on the STS-B benchmark (Cer et al., 2017), which is part of the GLUE suite of benchmarks for representation learning in NLP (Wang et al., 2018). Our ensemble of predictors achieves consistent iid accuracy, measured in terms of correlation with human-provided similarity scores, ranging from 0.87 to 0.90. This matches reported results from Devlin et al. (2019), although better correlations have subsequently been obtained by pretraining on larger datasets (Liu et al., 2019; Lan et al., 2019; Yang et al., 2019).

To measure reliance on gendered correlations in the STS task, we use a set of stress test templates proposed by Webster et al. (2020): we create a set of triples in which the noun phrase in a given sentence is replaced by a profession, "a man", or "a woman", e.g., "a doctor/woman/man is walking." The model's gender association for each profession is quantified by the *similarity delta* between pairs from this triple, e.g.,

$$\text{sim}(\text{“a woman is walking”}, \text{“a doctor is walking”}) - \text{sim}(\text{“a man is walking”}, \text{“a doctor is walking”}).$$

A model that does not learn a gendered correlation for a given profession will have an expected similarity delta of zero. We are particularly interested in the extent to which the similarity delta for each profession correlates with the percentage of women actually employed in that profession, as measured by U.S. Bureau of Labor Statistics (BLS; Rudinger et al., 2018).

7.1.2 Pronoun resolution

In the pronoun resolution task, the input is a sentence with a pronoun that could refer to one of two possible antecedents, and the predictor must determine which of the antecedents is the correct one. We obtain predictors for this task by fine-tuning BERT checkpoints on the OntoNotes dataset (Hovy et al., 2006). Our ensemble of predictors achieves accuracy ranging from 0.960 to 0.965.

To measure gendered correlations on the pronoun resolution task, we use the stress test templates proposed by Rudinger et al. (2018). In these templates, there is a gendered pronoun with two possible antecedents, one of which is a profession. The linguistic cues in the template are sufficient to indicate the correct antecedent, but predictors may instead learn to rely on the correlation between gender and profession. In this case, the similarity delta is the difference in predictive probability for the profession depending on the gender of the pronoun.

7.1.3 Gender correlations and underspecification

We find significant variation in the extent to which the predictors in our ensemble incorporate gendered correlations. For example, in Figure 8 (Left), we contrast the behavior of two predictors (which differ only in pretraining and fine-tuning seed) on the STS task. Here, the slope of the line is a proxy for the predictor's reliance on gender. One fine-tuning run shows strong correlation with

	F (p-value)	Spearman ρ (95% CI)
Semantic text similarity (STS)		
Test Accuracy	5.66 (4e-04)	—
Gender Correlation	9.66 (1e-06)	0.21 (-0.00, 0.40)
Pronoun resolution		
Test Accuracy	48.98 (3e-22)	—
Gender Correlation	7.91 (2e-05)	0.08 (-0.13, 0.28)

Table 3: **Summary statistics for structure of variation on gendered shortcut stress tests.** For each dataset, we measure the accuracy of 100 predictors, corresponding to 20 randomly initialized fine-tunings from 5 randomly initialized pretrained BERT checkpoints. Models are fine-tuned on the STS-B and OntoNotes training sets, respectively. The F-statistic quantifies how systematic differences are between pretrainings. Specifically, it is the ratio of between-pretraining variance to within-pretraining variance in the accuracy statistics. In the absense of sampling noise, F-statistic of 0 would indicate that pretraining seed explains no variance in performance. p-values are reported to give a sense of scale, but not for inferential purposes; it is unlikely that assumptions for a valid F-test are met. F-values of this magnitude are consistent with systematic between-group variation. The Spearman ρ statistic quantifies how ranked performance on the fine-tuning task correlates with the stress test metric of gender correlation.

BLS statistics about gender and occupations in the United States, while another shows a much weaker relationship. For an aggregate view, Figures 8 (Center) and (Right) show these correlations in the STS and coreference tasks across all predictors in our ensemble, with predictors produced from different pretrainings indicated by different markers. These plots show three important patterns:

1. There is a large spread in correlation with BLS statistics: on the STS task, correlations range from 0.3 to 0.7; on the pronoun resolution task, the range is 0.26 to 0.51. As a point of comparison, prior work on gender shortcuts in pronoun resolution found correlations ranging between 0.31 and 0.55 for different model architectures (Rudinger et al., 2018).

2. There is a weak relationship between test accuracy performance and gendered correlation (STS-B: Spearman $\rho = 0.21$; 95% CI $= (0.00, 0.39)$, Pronoun resolution: Spearman $\rho = 0.08$; 95% CI $= (-0.13, 0.29)$). This indicates that learning accurate predictors does *not* require learning strong gendered correlations.

3. Sensitivity to gender correlations is affected by the random seed at pretraining, and not just fine-tuning. Especially in the pronoun resolution task (Figure 8 (Right)), predictors with the same pretraining seed have similar behavior (i.e., similar sensitivity to gender information) relative to predictors trained with different pretraining seeds, even after fine-tuning on the same data.

In Table 3, we numerically summarize the variance with respect to pretraining and fine-tuning using an F-statistic — the ratio of between-pretraining to within-pretraining variance. The pretraining seed has an effect on both the main fine-tuning task and the stress test, but the small correlation between the fine-tuning tasks and stress test metrics suggests that this random seed affects these metrics independently.

7.2 Stereotypical Associations in Pretrained Language Models

Underspecification in supervised NLP systems can occur at both the fine-tuning and pretraining stages. In the previous section, we gave suggestive evidence that underspecification allows identically

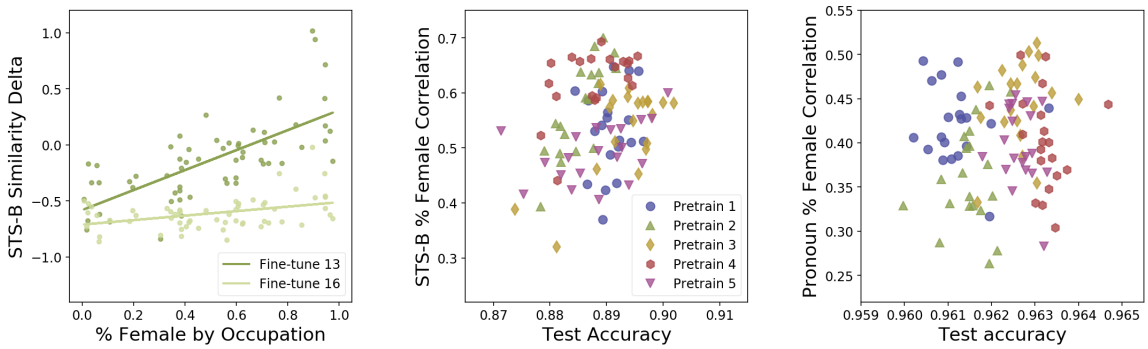

Figure 8: **Reliance on gendered correlations is affected by random initialization. (Left)** The gap in similarity for female and male template sentences is correlated with the gender statistics of the occupation, shown in two randomly-initialized fine-tunes. **(Right)** Pretraining initialization significantly affects the distribution of gender biases encoded at the fine-tuning stage.

pretrained BERT checkpoints to encode substantively different language representations. Here, we examine pretraining underspecification more directly, considering again its impact on reliance on stereotypical shortcuts. Specifically, we examine the performance of our ensemble of five BERT checkpoints on the StereoSet benchmark (Nadeem et al., 2020).

StereoSet is a set of stress tests designed to directly assess how the predictions of pretrained language models correlate with well-known social stereotypes. The test inputs are spans of text with sentences or words masked out, and the task is to score a set of choices for the missing piece of text. The choice set contains one nonsense option, and two plausible options, of which one conforms to a stereotype and the other does not. The benchmark probes stereotypes along the axes of gender, profession, race, and religion. For example, a span targeted at the gender domain reads "Girls tend to be more `[mask]` than boys", with options "soft" (stereotypical), "determined" (anti-stereotypical), and "fish" (non-sensical). Predictors are scored based on both whether they are able to exclude the non-sensical option (LM Score) and whether they consistently choose the option that conforms with the stereotype (Stereotype Score). These scores are averaged together to produce an Idealized Context Association Test (ICAT) score, which can be applied to any language model.

In Figure 9, we show the results of evaluating our five pretrained BERT checkpoints, which differ only in random seed, across all StereoSet metrics. The variation across checkpoints is large. The range of overall ICAT score between our identically trained checkpoints is 3.35. For context, this range is larger than the gap between the top six models on the public leaderboard,[2] which differ in size, architecture, and training data (GPT-2 (small), XLNet (large), GPT-2 (medium), BERT (base), GPT-2 (large), BERT (large)). On the disaggregated metrics, the score range between checkpoints is narrower on the LM score (sensible vs. non-sensible sentence completions) than on the Stereotype score (consistent vs. inconsistent with social stereotypes). This is consistent with underspecification, as the LM score is more closely aligned to the task on which the checkpoints are validated. Interestingly, score ranges are also lower on overall metrics compared to by-demographic metrics, suggesting that even when model performance looks stable in aggregate, checkpoints can encode different social stereotypes.

7.3 Spurious Correlations in Natural Language Inference

Underspecification also affects more general aspects of language representations that align with notions of "semantic understanding" in NLP systems. One task that probes such notions is natural language

2. `https://stereoset.mit.edu` retrieved October 28, 2020.

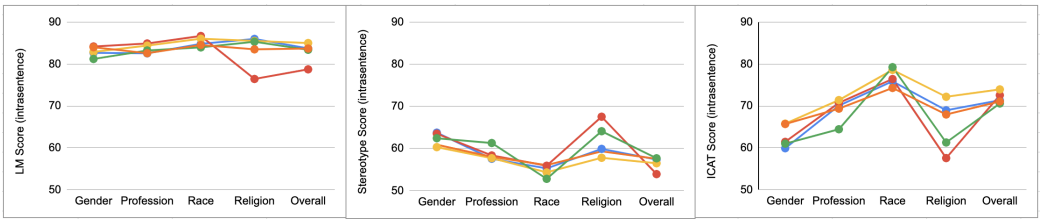

Figure 9: **Different pretraining seeds produce different stereotypical associations.** Results across five identically trained BERT large cased pretraining checkpoints on StereoSet (Nadeem et al., 2020). The ICAT score combines a language model (LM) score measuring "sensibility" and a stereotype score measuring correlations of language model predictions with known stereotypes. A leaderboard featuring canonical pretrainings is available at `https://stereoset.mit.edu/`. Variability in stereotype scores is large between BERT checkpoints that were pretrained identically up to random seed, even among those that achieve similar LM scores.

inference (NLI). The NLI task is to classify sentence pairs (called the **premise** and **hypothesis**) into one of the following semantic relations: entailment (the hypothesis is true whenever the premise is), contradiction (the hypothesis is false when the premise is true), and neutral (Bowman et al., 2015). Typically, language models are fine-tuned for this task on labeled datasets such as the MultiNLI training set (Williams et al., 2018). While test set performance on benchmark NLI datasets approaches human agreement (Wang et al., 2018), it has been shown that there are shortcuts to achieving high performance on many NLI datasets (McCoy et al., 2019b; Zellers et al., 2018, 2019). In particular, on stress tests that are designed to probe semantic representations more directly, these models are still far below human performance.

Notably, previous work has shown that performance on these stronger stress tests is also unstable with respect to the fine-tuning seed (Zhou et al., 2020; McCoy et al., 2019a; Dodge et al., 2020). We interpret this to be a symptom of underspecification. Here, we replicate and extend this prior work by assessing sensitivity to both fine-tuning *and* pretraining random seeds. Here we use the same five pretrained BERT Large cased checkpoints, and fine-tune each on the MultiNLI training set (Williams et al., 2018) 20 times. Across all pretrainings and fine-tunings, accuracy on the standard MNLI matched and mismatched test sets are in tightly constrained ranges of $(83.4\% - 84.4\%)$ and $(83.8\% - 84.7\%)$, respectively.[3]

We evaluate our ensemble of predictors on the HANS stress test (McCoy et al., 2019b) and the StressTest suite from Naik et al. (2018). The HANS stress tests are constructed by identifying spurious correlations in the training data—for example, that hypotheses that are entailed by the premise tend to have high lexical overlap with the premise—and then generating a test set such that the spurious correlations no longer hold. The Naik et al. (2018) stress tests are constructed by perturbing examples, for example by introducing spelling errors or meaningless expressions (e.g., "and true is true").

We again find strong evidence that the extent to which a trained model relies on shortcuts is underspecified, as demonstrated by sensitivity to the choice of random seed at both fine-tuning and pretraining time. Here, we report several broad trends of variation on these stress tests: first, the magnitude of the variation is large; second, the variation is sensitive to the fine-tuning seed, replicating Zhou et al. (2020); third, the variation is also sensitive to the pretraining seed; fourth, the variation is difficult to predict based on performance on the standard MNLI validation sets; and finally, the variation on different stress tests tends to be weakly correlated.

3. The "matched" and "mismatched" conditions refer to whether the test data is drawn from the same genre of text as the training set.

Dataset	F (p-value)	Spearman ρ (95% CI)
MNLI, matched	1.71 (2e-01)	—
MNLI, mismatched	20.18 (5e-12)	0.11 (-0.10, 0.31)
Naik et al. (2018) stress tests		
Antonym, matched	15.46 (9e-10)	0.05 (-0.16, 0.26)
Antonym, mismatched	7.32 (4e-05)	0.01 (-0.20, 0.21)
Length Mismatch, matched	4.83 (1e-03)	0.33 (0.13, 0.50)
Length Mismatch, mismatched	5.61 (4e-04)	-0.03 (-0.24, 0.18)
Negation, matched	19.62 (8e-12)	0.17 (-0.04, 0.36)
Negation, mismatched	18.21 (4e-11)	0.09 (-0.12, 0.29)
Spelling Error, matched	25.11 (3e-14)	0.40 (0.21, 0.56)
Spelling Error, mismatched	14.65 (2e-09)	0.43 (0.24, 0.58)
Word Overlap, matched	9.99 (9e-07)	0.08 (-0.13, 0.28)
Word Overlap, mismatched	9.13 (3e-06)	-0.07 (-0.27, 0.14)
Numerical Reasoning	12.02 (6e-08)	0.18 (-0.03, 0.38)
HANS (McCoy et al., 2019b)	4.95 (1E-03)	0.07 (-0.14, 0.27)

Table 4: **Summary statistics for structure of variation in predictor accuracy across NLI stress tests.** For each dataset, we measure the accuracy of 100 predictors, corresponding to 20 randomly initialized fine-tunings from 5 randomly initialized pretrained BERT checkpoints. All models are fine-tuned on the MNLI training set, and validated on the MNLI matched test set (Williams et al., 2018). See the caption of Table 3 for a description of F-statistics. The Spearman ρ statistic quantifies how ranked performance on the MNLI matched test set correlates with ranked performance on each stress test. For most stress tests, there is only a weak relationship, such that choosing models based on test performance alone would not yield the best models on stress test performance.

Figure 10 shows our full set of results, broken down by pretraining seed. These plots show evidence of the influence of the pretraining seed; for many tests, there appear to be systematic differences in performance from fine-tunings based on checkpoints that were pretrained with different seeds. We report one numerical measurement of these differences with F-statistics in Table 4, where the ratio of between-group variance to within-group variance is generally quite large. Table 4 also reports Spearman rank correlations between stress test accuracies and accuracy on the MNLI matched validation set. The rank correlation is typically small, suggesting that the variation in stress test accuracy is largely orthogonal to validation set accuracy enforced by the training pipeline. Finally, in Figure 11, we show that the correlation between stress tests performance is also typically small (with the exception of some pairs of stress tests meant to test the same aspects of a representation), suggesting that the space of underspecified representations spans many dimensions.

7.4 Conclusions

There is increasing concern about whether natural language processing systems are learning general linguistic principles, or whether they are simply learning to use surface-level shortcuts (e.g., Bender and Koller, 2020; Linzen, 2020). Particularly worrying are shortcuts that reinforce societal biases around protected attributes such as gender (e.g., Webster et al., 2020). The results in this section replicate prior findings that highly-parametrized NLP models do learn spurious correlations and shortcuts. However, this reliance is underspecified by the standard ML pipeline: merely changing the random seed can induce large variation in the extent to which spurious correlations are learned. Furthermore, this variation is demonstrated in both pretraining and fine-tuning, indicating that

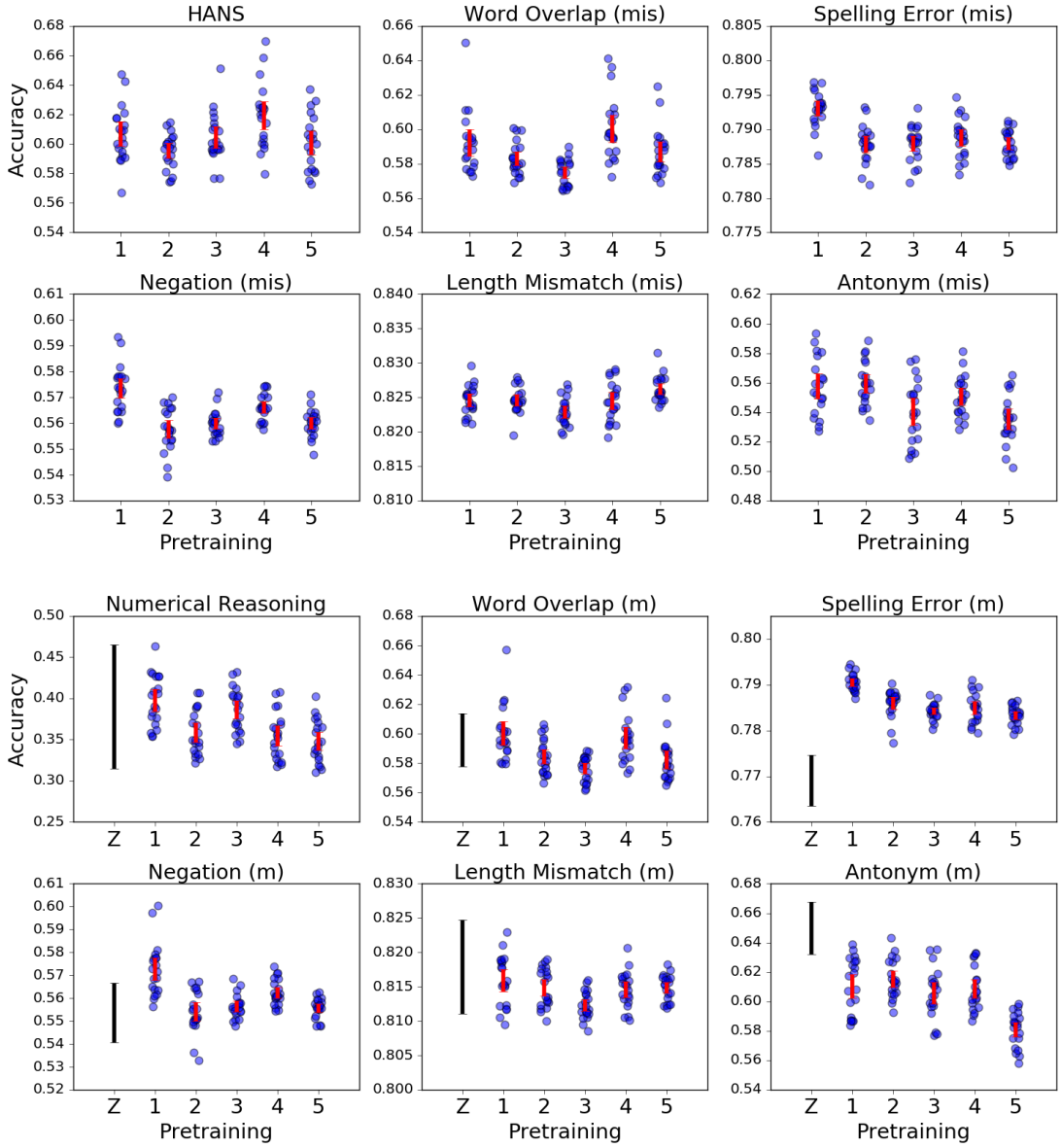

Figure 10: **Predictor performance on NLI stress tests varies both within and between pretraining checkpoints.** Each point corresponds to a fine-tuning of a pretrained BERT checkpoint on the MNLI training set, with pretraining distinguished on the x-axis. All pretrainings and fine-tunings differ only in random seed at their respective training stages. Performance on HANS (McCoy et al., 2019b) is shown in the top left; remaining results are from the StressTest suite (Naik et al., 2018), with the suffixes (m) and (mis) indicating the genre-matched and mismatched conditions respectively. Red bars show a 95% CI around for the mean accuracy within each pretraining. The tests in the bottom group of panels were also explored in Zhou et al. (2020) across fine-tunings from the public BERT large cased checkpoint (Devlin et al., 2019); for these, we also plot the mean +/- 1.96 standard deviations interval, using values reported in Zhou et al. (2020). The magnitude of variation is substantially larger on most stress tests than the MNLI test sets ($< 1\%$ on both MNLI matched and unmatched). There is also substantial variation between some pretrained checkpoints, even after fine-tuning.

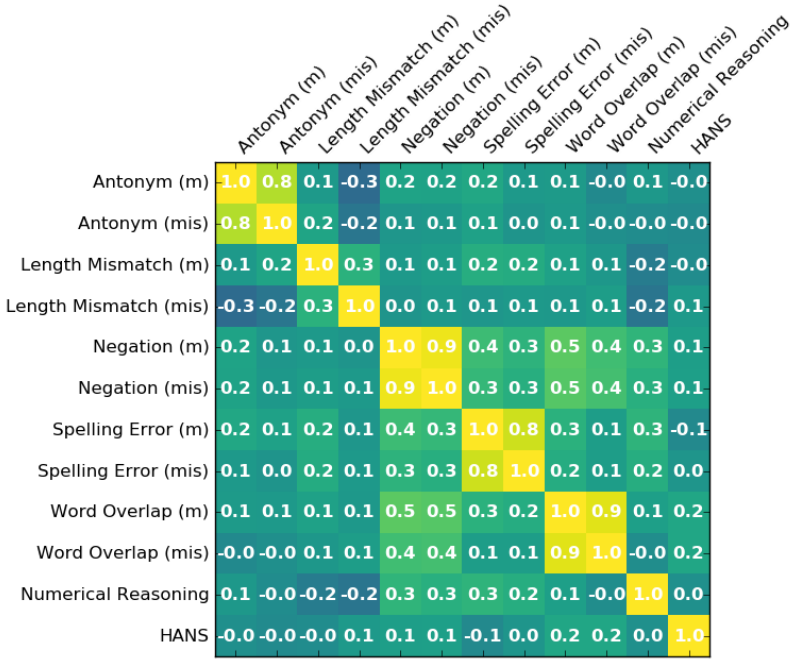

Figure 11: **Predictor performance across stress tests are typically weakly correlated.** Spearman correlation coefficients of 100 predictor accuracies from 20 fine-tunings of five pretrained BERT checkpoints. The suffixes (m) and (mis) indicate the genre-matched and mismatched conditions respectively (Naik et al., 2018).

pretraining alone can produce variation in representations that translates to more or less robustness. This implies that individual stress test results should be viewed as statements about individual model checkpoints, and not about architectures or learning algorithms. More general comparisons require evaluation across many pipeline degrees of freedom (e.g., random seeds).

8. Case Study in Clinical Predictions from Electronic Health Records

The rise of Electronic Health Record (EHR) systems has created an opportunity for building predictive ML models for diagnosis and prognosis (e.g. Ambrosino et al. (1995); Brisimi et al. (2019); Feng et al. (2019)). In this section, we focus on one such model that uses a Recurrent Neural Network (RNN) architecture with EHR data to predict acute kidney injury (AKI) during hospital admissions (Tomašev et al., 2019a). AKI is a common complication in hospitalized patients and is associated with increased morbidity, mortality, and healthcare costs (Khwaja, 2012). Early intervention can improve outcomes in AKI (National Institute for Health and Care Excellence (NICE), 2019), which has driven efforts to predict it in advance using machine learning. Tomašev et al. (2019a) achieve state-of-the-art performance, detecting the onset of AKI up to 48 hours in advance with an accuracy of 55.8% across all episodes and 90.2% for episodes associated with dialysis administration.

Despite this strong discriminative performance, there have been questions raised about the associations being learned by this model and whether they conform with our understanding of physiology (Kellum and Bihorac, 2019). Specifically, for some applications, it is desirable to disentangle physiological signals from operational factors related to the delivery of healthcare, both of which appear in EHR data. As an example, the value of a lab test may be considered a physiological signal; however the timing of that same test may be considered an operational one (e.g. due to staffing constraints during the night or timing of ward rounds). Given the fact that operational signals may be institution-specific and are likely to change over time, understanding to what extent a predictor relies on different signals can help practitioners determine whether the predictor meets their specific generalization requirements (Futoma et al., 2020).

Here, we show that underspecification makes the answer to this question ambiguous. Specifically, we apply our experimental protocol to the Tomašev et al. (2019a) AKI modeling pipeline, which produces a predictor that predicts the continuous risk (every 6 hours) of AKI in a 48h lookahead time window (see Supplement for details).

8.1 Data, Predictor Ensemble, and Metrics

The pipeline and data used in this study are described in detail in Tomašev et al. (2019a). Briefly, the data consists of de-identified EHRs from 703,782 patients across multiple sites in the United States collected at the US Department of Veterans Affairs[4] between 2011 and 2015. Records include structured data elements such as medications, labs, vital signs, diagnosis codes, etc, aggregated in six hour time buckets (time of day 1: 12am-6am, 2: 6am-12pm, 3: 12pm-6pm, 4: 6pm-12am). In addition, precautions beyond standard de-identification have been taken to safeguard patient privacy: free text notes and rare diagnoses have been excluded; many feature names have been obfuscated; feature values have been jittered; and all patient records are time-shifted, respecting relative temporal relationships for individual patients. Therefore, this dataset is only intended for methodological exploration.

The model consists of embedding layers followed by a 3 layer-stacked RNN before a final dense layer for prediction of AKI across multiple time horizons. Our analyses focus on predictions with a 48h lookahead horizon, which have been showcased in the original work for their clinical actionability. To examine underspecification, we construct a predictor ensemble by training the model from 5 random seeds for each of three RNN cell types: Simple Recursive Units (SRU, Lei et al. (2018)), Long

4. Disclaimer: Please note that the views presented in this manuscript are that of the authors and not that of the Department of the Veterans Affairs.

Short-Term Memory (LSTM, Hochreiter and Schmidhuber (1997)) or Update Gate RNN (UGRNN, Collins et al. (2017)). This yields an ensemble of 15 predictors in total.

The primary validation metric that we use to evaluate predictive performance is normalized area under the precision-recall curve (PRAUC) (Boyd et al., 2012), evaluated across all patient-timepoints where the model makes a prediction. This is a PRAUC metric that is normalized for prevalence of the positive label (in this case, AKI events). Our ensemble of predictors achieves tightly constrained normalized PRAUC values between 34.59 and 36.61.

8.2 Reliance on Operational Signals

We evaluate these predictors on stress tests designed to probe the sensitivity to specific operational signals in the data: the timing and number of labs recorded in the EHR. [5] In this dataset, the prevalence of AKI is largely the same across different times of day (see Table 1 of Supplement). However, AKI is diagnosed based on lab tests, [6] and there are clear temporal patterns in how tests are ordered. For most patients, creatinine is measured in the morning as part of a 'routine', comprehensive panel of lab tests. Meanwhile, patients requiring closer monitoring may have creatinine samples taken at additional times, often ordered as part of an 'acute', limited panel (usually, the basic metabolic panel.) [7] Thus, both the time of day that a test is ordered, and the panel of tests that accompany a given measurement may be correlated with AKI risk, but are primarily operational factors.

We test for reliance on this signal by applying two interventions to the test data that modify (1) the time of day of all features (aggregated in 6h buckets) and (2) the selection of lab tests. The first intervention shifts the patient timeseries by a fixed offset, while the second intervention additionally removes all blood tests that are not directly relevant to the diagnosis of AKI. We hypothesize that if the predictor encodes physiological signals rather than these operational cues, the predictions would be invariant to these interventions. More importantly, if the model's reliance on these operational signals is underspecified, we would expect the predictors in our ensemble to respond differently to these modified inputs.

We begin by examining overall performance on this shifted test set across our ensemble. In Figure 12, we show that performance on the intervened data is both worse and more widely dispersed than in the standard test set, especially when both interventions are applied. This shows that the model incorporates time of day and lab content signals, and that the extent to which it relies on these signals is sensitive to both the recurrent unit and random initialization.

The variation in performance reflects systematically different signals encoded by the predictors in the ensemble. We examine this directly by measuring how individual model predictions change under the timeshift and lab intervensions. Here, we focus on two trained LSTM models that differ only in their random seeds, and examine patient-timepoints at which creatinine measurements were taken. In Figure 13 (right), we show distributions of predicted risk on the original patient-timepoints observed in the "early morning" (12am-6am) time range, and proportional changes to these risks when the timeshift and lab interventions were applied. Both predictors exhibit substantial changes in predicted risk under both interventions, but the second predictor is far more sensitive to these changes than the first, with the predicted risks taking on substantially different distributions depending on the time range to which the observation is shifted.

These shifts in risk are consequential for decision-making and can result in AKI episodes being predicted tardily or missed. In Figure 14, we illustrate the number of patient-timepoints where the changed risk score crosses each model's calibrated decision threshold. In addition to substantial

5. Neither of these factors are purely operational—there is known variation in kidney function across the day and the values of accompanying lab tests carry valuable information about patient physiology. However, we use these here as approximations for an operational perturbation.
6. specifically a comparison of past and current values of creatinine (Khwaja, 2012)
7. This panel samples Creatinine, Sodium, Potassium, Urea Nitrogen, CO2, Chloride and Glucose.

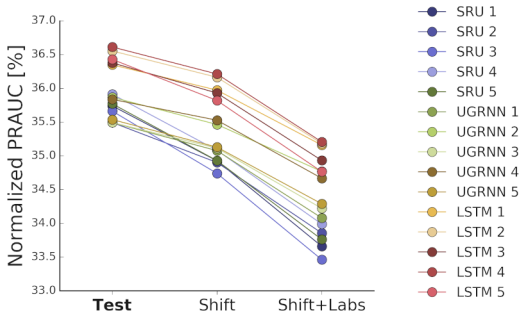

Figure 12: **Variability in performance from ensemble of RNN models processing electronic health records (EHR)**. Model sensitivity to time of day and lab perturbations. The x-axis denotes the evaluation set: "Test" is the original test set; "Shift" is the test set with time shifts applied; "Shift + Labs" applies the time shift and subsets lab orders to only include the basic metabolic panel CHEM-7. The y-axis represents the normalized PRAUC, and each set of dots joined by a line represents a model instance.

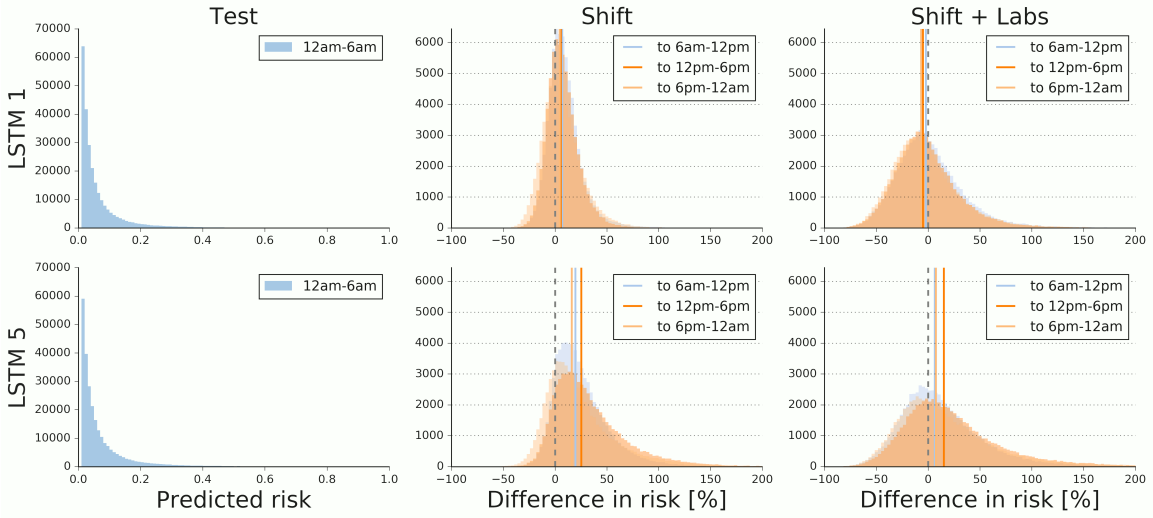

Figure 13: **Variability in AKI risk predictions between two LSTM models processing electronic health records (EHR)**. Histograms showing showing risk predictions from two models, and changes induced by time of day and lab perturbations. Histograms show counts of patient-timepoints where creatinine measurements were taken in the early morning (12am-6am). LSTM 1 and 5 differ only in random seed. "Test" shows histogram of risk predicted in original test data. "Shift" and "Shift + Labs" show histograms of proportional changes (in %) $\frac{\text{Perturbed}-\text{Baseline}}{\text{Baseline}}$ induced by the time-shift perturbation and the combined time-shift and lab perturbation, respectively.

differences in the number of flipped decisions, we also show that most of these flipped decisions occur at different patient-timepoints across predictors.

8.3 Conclusions

Our results here suggest that predictors produced by this model tend to rely on the pattern of lab orders in a substantial way, but the extent of this reliance is underspecified. Depending on how stable this signal is in the deployment context, this may or may not present challenges. However, this result

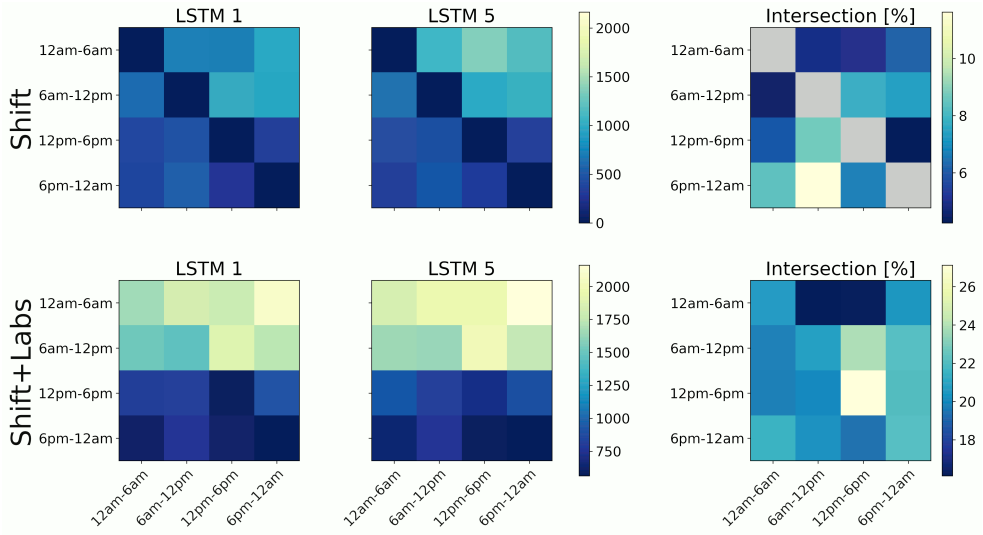

Figure 14: **Variability in AKI predictions between two LSTM models processing electronic health records (EHR)**. Counts (color-coded) of decisions being flipped due to the stress tests, from the LSTM 1 and LSTM 5 models, as well as the proportions of those flipped decision intersecting between the two models (in %). Rows represent the time of day in the original test set, while columns represent the time of day these samples were shifted to. LSTM 1 and 5 differ only in random seed. "Shift" represents the flipped decisions (both positive to negative and negative to positive) between the predictions on the test set and the predictions after time-shift perturbation. "Shift + Labs" represents the same information for the combined time-shift and labs perturbation.

also shows that the reliance on this signal is not *enforced* by the model specification or training data, suggesting that the reliance on lab ordering patterns could be modulated by adding constraints to the training procedure, and without sacrificing iid performance. In the Supplement, we show one such preliminary result, where a model trained with the timestamp feature completely ablated was able to achieve identical iid predictive performance. This is compatible with previous findings that inputting medical/domain relational knowledge has led to better out of domain behavior (Nestor et al., 2019), performance (Popescu and Khalilia, 2011; Choi et al., 2017; Tomašev et al., 2019a,b) and interpretability (Panigutti et al., 2020).

9. Discussion: Implications for ML Practice

Our results show that underspecification is a key failure mode for applied machine learning, where we often have requirements for predictors that extend beyond iid generalization. We have used between-predictor variation in stress test performance as an observable signature of underspecification. This failure mode is distinct from generalization failures due to structural mismatch between training and deployment domains. We have seen that underspecification is ubiquitious in practical machine learning pipelines across many domains. Indeed, thanks to underspecification, substantively important aspects of the decisions are determined by arbitrary choices made in the training pipeline, such as the random seed used for parameter initialization. We close with a discussion of some of the implications of the study, which broadly suggest a need to find better interfaces for domain knowledge in ML pipelines.

First, we note that the methodology in this study underestimates the impact of underspecification: our goal was to detect rather than fully characterize underspecification, and in most examples, we

only explored underspecification through the subtle variation that can result from modifying random seeds in training. However, modern deep learning pipelines incorporate a wide variety of *ad hoc* practices, each of which may carry their own implicit regularizations, which in turn can translate into substantive differences in the real-world behavior of predictors. These include the particular scheme used for initialization; conventions for parameterization; choice of optimization algorithm; conventions for representing data; compression schemes; and choices of batch size, learning rate, and other hyperparameters, all of which may interact with the infrastructure available for training and serving models (Hooker, 2020). We conjecture that many combinations of these choices would reveal a far larger validation-equivalent set of high-performing predictors $\mathcal{F}^*$, a conjecture that has been partially borne out by concurrent work (Wenzel et al., 2020). However, we believe that there would be value in more systematically mapping out the set of validation-equivalent predictors that a pipeline could return as a true measurement of the uncertainty entailed by underspecification. Current efforts to design more effective methods for exploring loss landscapes (Fort et al., 2019; Garipov et al., 2018) could play an important role here, and there are opportunities to import ideas from the sensitivity analysis and partial identification subfields in causal inference and inverse problems.

Second, our findings underscore the need to thoroughly test models on application-specific tasks, and in particular to check that the performance on these tasks is stable. The extreme complexity of modern ML models pipelines ensures that some aspect of predictor behavior will almost certainly be underspecified; thus, the challenge is to ensure that this underspecification does not jeopardize the specific behaviors that are required by an application. In this vein, designing stress tests that are well matched to applied requirements, and that provide good "coverage" of potential failure modes is a major challenge that requires incorporating domain knowledge. This can be particularly challenging, given that our results show that there is often low correlation between performance on distinct stress tests when iid performance is held constant, and the fact that many applications will have fine-grained requirements that require more customized stress testing. For example, within the medical risk prediction domain, the dimensions that a model is required to generalize across (e.g., temporal, demographic, operational, etc.) will depend on the details of the deployment and the goals of the practitioners (Futoma et al., 2020). For this reason, developing best practices for *building* stress tests that crisply represent requirements, rather than standardizing on specific benchmarks, may be an effective approach. This approach has gained traction in the NLP subfiled, where several papers now discuss the process by which stress tests datasets should iterate continuously (Zellers et al., 2019), and new systems for developing customized stress tests have been proposed (Ribeiro et al., 2020; Kaushik et al., 2020).

Third, our results suggest some new strategies for training models that exhibit appropriate real-world behavior when underspecification plays a role. By definition, underspecification can be resolved by specifying additional criteria for selecting predictors from the equivalence class of validation-equivalent predictors $\mathcal{F}^*$. Importantly, this suggests a departure from a popular strategy of improving iid performance of predictors by marginalizing across $\mathcal{F}^*$ (Wilson and Izmailov, 2020). Here, because it is known that some predictors in $\mathcal{F}^*$ behave poorly on stress tests, simply averaging them together is not guaranteed to produce better results on stress tests than carefully choosing a specific predictor from the equivalence class (see Appendix A for some examples). Of course, these approaches can be reconciled if the marginalization is restricted to predictors that satisfy required constraints.

Some central questions remain, however, for designing constraints to mitigate underspecification. First, because they are meant to enforce application-specific requirements on predictor behavior, such criteria or constraints must also be application-specific, presenting a challenge for the development of general methods (similarly to our discussion of testing above). Although some general-purpose heuristics have been proposed (e.g., Bengio, 2017), proposals for expressing application-specific requirements with flexible but unified frameworks may be a promising middle ground solution. Causal DAGs (Schölkopf, 2019) and explanations (Ross et al., 2017) are both promising candidates for such frameworks.

Finally, there is a question of whether imposing constraints on real-world behavior should necessarily trade off with iid generalization. This question comes down to whether the behavioral failures of the predictor are driven by structural conflict or underspecification; indeed, in many problems, both may play a role. However, a major point of this study is that, in many problems, there is significant slack introduced by underspecification where tradeoff-free improvement is possible. Some recent work supports this observation: in the robustness literature Raghunathan et al. (2020) show that robustness / accuracy tradeoffs need not be fundamental, but are often driven by finite-sample phenomena; in the NLP literature, Webster et al. (2020) show that reliance on gendered correlations can be reduced in BERT-derived predictors with little-to-no tradeoff (Webster et al., 2020); and in Appendix C, we show a similar preliminary result from our EHR example. Constraints that are specifically tailored to the application can be particularly effective in this regard. For example, Makar et al. (2021) and Veitch et al. (2021) show that regularization schemes that are designed to respect the causal structure of scenarios that are expected to be encountered in practice can produce models that generalize well in deployment; in particular, Makar et al. (2021) highlight a case where such constraints can actually improve finite-sample iid generalization and real-world behavior, while Veitch et al. (2021) provides a precise characterization of some cases where tradeoffs will and will not arise. Overall, the design of application-specific constraints is a promising direction for incorporating domain expertise without fully compromising the powerful prediction abilities of modern ML models.

Acknowledgements

This research has been conducted using the UK Biobank Resource under Application Number 17643. We would also like to thank our partners—EyePACS in the United States and Aravind Eye Hospital and Sankara Nethralaya in India for providing the datasets used to train the models for predicting diabetic retinopathy from fundus images. We also appreciate the advice of our DeepMind collaborator Dr. Nenad Tomasev, Prof. Finale Doshi-Velez and the wider Google Health Research UK team led by Dr. Alan Karthikesalingam.

References

Adewole S Adamson and Avery Smith. Machine learning and health care disparities in dermatology. *JAMA dermatology*, 154(11):1247–1248, 2018.

Ademide Adelekun, Ginikanwa Onyekaba, and Jules B Lipoff. Skin color in dermatology textbooks: An updated evaluation and analysis. *Journal of the American Academy of Dermatology*, 2020.

R Ambrosino, B G Buchanan, G F Cooper, and M J Fine. The use of misclassification costs to learn rule-based decision support models for cost-effective hospital admission strategies. *Proceedings. Symposium on Computer Applications in Medical Care*, pages 304–8, 1995. ISSN 0195-4210. URL `http://www.ncbi.nlm.nih.gov/pubmed/8563290http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2579104`.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Susan Athey. Beyond prediction: Using big data for policy problems. *Science*, 355(6324):483–485, 2017.

Marzieh Babaeianjelodar, Stephen Lorenz, Josh Gordon, Jeanna Matthews, and Evan Freitag. Quantifying gender bias in different corpora. In *Companion Proceedings of the Web Conference 2020*, pages 752–759, 2020.

Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Advances in Neural Information Processing Systems*, pages 9448–9458, 2019.

Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M Vardoulakis. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2020.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine learning and the bias-variance trade-off. *arXiv preprint arXiv:1812.11118*, 2018.

Emily M. Bender and Alexander Koller. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.463. URL `https://www.aclweb.org/anthology/2020.acl-main.463`.

Yoshua Bengio. The consciousness prior. *arXiv preprint arXiv:1709.08568*, 2017.

Jeremy J Berg, Arbel Harpak, Nasa Sinnott-Armstrong, Anja Moltke Joergensen, Hakhamanesh Mostafavi, Yair Field, Evan August Boyle, Xinjun Zhang, Fernando Racimo, Jonathan K Pritchard, and Graham Coop. Reduced signal for polygenic adaptation of height in UK biobank. *Elife*, 8, March 2019.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357, 2016.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1075. URL `https://www.aclweb.org/anthology/D15-1075`.

Kendrick Boyd, Vítor Santos Costa, Jesse Davis, and C. David Page. Unachievable region in precision-recall space and its effect on empirical evaluation. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012*, volume 1, pages 639–646, 2012. ISBN 9781450312851.

Leo Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231, 2001.

Theodora S. Brisimi, Tingting Xu, Taiyao Wang, Wuyang Dai, and Ioannis Ch Paschalidis. Predicting diabetes-related hospitalizations based on electronic health records. *Statistical Methods in Medical Research*, 28(12):3667–3682, dec 2019. ISSN 14770334. doi: 10.1177/0962280218810911.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91, 2018.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

CARDIoGRAMplusC4D Consortium, Panos Deloukas, Stavroula Kanoni, Christina Willenborg, Martin Farrall, Themistocles L Assimes, John R Thompson, Erik Ingelsson, Danish Saleheen, Jeanette Erdmann, Benjamin A Goldstein, Kathleen Stirrups, Inke R König, Jean-Baptiste Cazier,

Asa Johansson, Alistair S Hall, Jong-Young Lee, Cristen J Willer, John C Chambers, Tõnu Esko, Lasse Folkersen, Anuj Goel, Elin Grundberg, Aki S Havulinna, Weang K Ho, Jemma C Hopewell, Niclas Eriksson, Marcus E Kleber, Kati Kristiansson, Per Lundmark, Leo-Pekka Lyytikäinen, Suzanne Rafelt, Dmitry Shungin, Rona J Strawbridge, Gudmar Thorleifsson, Emmi Tikkanen, Natalie Van Zuydam, Benjamin F Voight, Lindsay L Waite, Weihua Zhang, Andreas Ziegler, Devin Absher, David Altshuler, Anthony J Balmforth, Inês Barroso, Peter S Braund, Christof Burgdorf, Simone Claudi-Boehm, David Cox, Maria Dimitriou, Ron Do, DIAGRAM Consortium, CARDIOGENICS Consortium, Alex S F Doney, Noureddine El Mokhtari, Per Eriksson, Krista Fischer, Pierre Fontanillas, Anders Franco-Cereceda, Bruna Gigante, Leif Groop, Stefan Gustafsson, Jörg Hager, Göran Hallmans, Bok-Ghee Han, Sarah E Hunt, Hyun M Kang, Thomas Illig, Thorsten Kessler, Joshua W Knowles, Genovefa Kolovou, Johanna Kuusisto, Claudia Langenberg, Cordelia Langford, Karin Leander, Marja-Liisa Lokki, Anders Lundmark, Mark I McCarthy, Christa Meisinger, Olle Melander, Evelin Mihailov, Seraya Maouche, Andrew D Morris, Martina Müller-Nurasyid, MuTHER Consortium, Kjell Nikus, John F Peden, N William Rayner, Asif Rasheed, Silke Rosinger, Diana Rubin, Moritz P Rumpf, Arne Schäfer, Mohan Sivananthan, Ci Song, Alexandre F R Stewart, Sian-Tsung Tan, Gudmundur Thorgeirsson, C Ellen van der Schoot, Peter J Wagner, Wellcome Trust Case Control Consortium, George A Wells, Philipp S Wild, Tsun-Po Yang, Philippe Amouyel, Dominique Arveiler, Hanneke Basart, Michael Boehnke, Eric Boerwinkle, Paolo Brambilla, Francois Cambien, Adrienne L Cupples, Ulf de Faire, Abbas Dehghan, Patrick Diemert, Stephen E Epstein, Alun Evans, Marco M Ferrario, Jean Ferrières, Dominique Gauguier, Alan S Go, Alison H Goodall, Villi Gudnason, Stanley L Hazen, Hilma Holm, Carlos Iribarren, Yangsoo Jang, Mika Kähönen, Frank Kee, Hyo-Soo Kim, Norman Klopp, Wolfgang Koenig, Wolfgang Kratzer, Kari Kuulasmaa, Markku Laakso, Reijo Laaksonen, Ji-Young Lee, Lars Lind, Willem H Ouwehand, Sarah Parish, Jeong E Park, Nancy L Pedersen, Annette Peters, Thomas Quertermous, Daniel J Rader, Veikko Salomaa, Eric Schadt, Svati H Shah, Juha Sinisalo, Klaus Stark, Kari Stefansson, David-Alexandre Trégouët, Jarmo Virtamo, Lars Wallentin, Nicholas Wareham, Martina E Zimmermann, Markku S Nieminen, Christian Hengstenberg, Manjinder S Sandhu, Tomi Pastinen, Ann-Christine Syvänen, G Kees Hovingh, George Dedoussis, Paul W Franks, Terho Lehtimäki, Andres Metspalu, Pierre A Zalloua, Agneta Siegbahn, Stefan Schreiber, Samuli Ripatti, Stefan S Blankenberg, Markus Perola, Robert Clarke, Bernhard O Boehm, Christopher O'Donnell, Muredach P Reilly, Winfried März, Rory Collins, Sekar Kathiresan, Anders Hamsten, Jaspal S Kooner, Unnur Thorsteinsdottir, John Danesh, Colin N A Palmer, Robert Roberts, Hugh Watkins, Heribert Schunkert, and Nilesh J Samani. Large-scale association analysis identifies new risk loci for coronary artery disease. *Nat. Genet.*, 45(1):25–33, January 2013.

Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible Models for HealthCare. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15*, pages 1721–1730, 2015. ISBN 9781450336642. doi: 10.1145/2783258.2788613. URL `http://dx.doi.org/10.1145/2783258.2788613http://dl.acm.org/citation.cfm?doid=2783258.2788613`.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/S17-2001. URL `https://www.aclweb.org/anthology/S17-2001`.

Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-sgd: Biasing gradient descent into wide valleys. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124018, 2019.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. 2019.

Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F Stewart, and Jimeng Sun. GRAM: Graph-based Attention Model for Healthcare Representation Learning. 2017. doi: 10.1145/3097983.3098126. URL http://dx.doi.org/10.1145/3097983.3098126.

Gary S Collins, Johannes B Reitsma, Douglas G Altman, and Karel GM Moons. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): the tripod statement. *British Journal of Surgery*, 102(3):148–158, 2015.

Jasmine Collins, Jascha Sohl-Dickstein, and David Sussillo. Capacity and trainability in recurrent neural networks. In *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, 2017.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 120–128, 2019.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

Josip Djolonga, Jessica Yung, Michael Tschannen, Rob Romijnders, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Matthias Minderer, Alexander D'Amour, Dan Moldovan, et al. On robustness and transferability of convolutional neural networks. *arXiv preprint arXiv:2007.08558*, 2020.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*, 2020.

L Duncan, H Shen, B Gelaye, J Meijsen, K Ressler, M Feldman, R Peterson, and B Domingue. Analysis of polygenic risk score usage and performance in diverse human populations. *Nat. Commun.*, 10(1):3328, July 2019.

Michael W Dusenberry, Dustin Tran, Edward Choi, Jonas Kemp, Jeremy Nixon, Ghassen Jerfel, Katherine Heller, and Andrew M Dai. Analyzing the role of model uncertainty for electronic health records. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pages 204–213, 2020.

Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. Neural architecture search: A survey. *Journal of Machine Learning Research*, 20(55):1–21, 2019. URL http://jmlr.org/papers/v20/18-598.html.

Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.

Chenchen Feng, David Le, and Allison B. McCoy. Using Electronic Health Records to Identify Adverse Drug Events in Ambulatory Care: A Systematic Review. *Applied Clinical Informatics*, 10 (1):123–128, 2019. ISSN 18690327. doi: 10.1055/s-0039-1677738.

Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81, 2019.

TB Fitzpatrick. Sun and skin. *Journal de Medecine Esthetique*, 2:33–34, 1975.

Seth Flaxman, Swapnil Mishra, Axel Gandy, H Juliette T Unwin, Thomas A Mellan, Helen Coupland, Charles Whittaker, Harrison Zhu, Tresnia Berah, Jeffrey W Eaton, et al. Estimating the effects of non-pharmaceutical interventions on covid-19 in europe. *Nature*, 584(7820):257–261, 2020.

Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*, 2019.

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel M Roy, and Michael Carbin. Linear mode connectivity and the lottery ticket hypothesis. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Joseph Futoma, Morgan Simons, Trishan Panch, Finale Doshi-Velez, and Leo Anthony Celi. The myth of generalisability in clinical research and machine learning in health care. *The Lancet Digital Health*, 2(9):e489 – e492, 2020. ISSN 2589-7500. doi: https://doi.org/10.1016/S2589-7500(20)30186-2. URL `http://www.sciencedirect.com/science/article/pii/S2589750020301862`.

Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H Chi, and Alex Beutel. Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 219–226, 2019.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P Vetrov, and Andrew G Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. In *Advances in Neural Information Processing Systems*, pages 8789–8798, 2018.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=Bygh9j09KX`.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *arXiv preprint arXiv:2004.07780*, 2020.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*, 316(22):2402–2410, 2016.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Christina Heinze-Deml, Jonas Peters, and Nicolai Meinshausen. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6(2), 2018.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=HJz6tiCqYm`.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *arXiv preprint arXiv:1907.07174*, 2019.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. *arXiv preprint arXiv:2006.16241*, 2020.

Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 1997. URL `http://www7.informatik.tu-muenchen.de/{~}hochreithttp://www.idsia.ch/{~}juergen`.

Wolfgang Hoffmann, Ute Latza, Sebastian E Baumeister, Martin Brünger, Nina Buttmann-Schweiger, Juliane Hardt, Verena Hoffmann, André Karch, Adrian Richter, Carsten Oliver Schmidt, et al. Guidelines and recommendations for ensuring good epidemiological practice (gep): a guideline developed by the german society for epidemiology. *European journal of epidemiology*, 34(3):301–317, 2019.

Sara Hooker. The hardware lottery. *arXiv preprint arXiv:2009.06489*, 2020.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. OntoNotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60, New York City, USA, June 2006. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/N06-2015`.

Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, 2018.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems*, pages 125–136, 2019.

International Schizophrenia Consortium, Shaun M Purcell, Naomi R Wray, Jennifer L Stone, Peter M Visscher, Michael C O'Donovan, Patrick F Sullivan, and Pamela Sklar. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, 460(7256):748–752, August 2009.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018.

Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai. *arXiv preprint arXiv:2010.07487*, 2020.

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=Sklgs0NFvr`.

John A Kellum and Azra Bihorac. Artificial intelligence to predict aki: is it a breakthrough? *Nature Reviews Nephrology*, pages 1–2, 2019.

Christopher J Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. Key challenges for delivering clinical impact with artificial intelligence. *BMC medicine*, 17(1):195, 2019.

Amit V Khera, Mark Chaffin, Krishna G Aragam, Mary E Haas, Carolina Roselli, Seung Hoan Choi, Pradeep Natarajan, Eric S Lander, Steven A Lubitz, Patrick T Ellinor, and Sekar Kathiresan. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.*, 50(9):1219–1224, September 2018.

Arif Khwaja. KDIGO clinical practice guidelines for acute kidney injury. *Nephron - Clinical Practice*, 120(4), oct 2012. ISSN 16602110. doi: 10.1159/000339789.

Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. Prediction policy problems. *American Economic Review*, 105(5):491–95, 2015.

Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Large scale learning of general visual representations for transfer. *arXiv preprint arXiv:1912.11370*, 2019.

Jonathan Krause, Varun Gulshan, Ehsan Rahimy, Peter Karth, Kasumi Widner, Greg S Corrado, Lily Peng, and Dale R Webster. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. *Ophthalmology*, 125(8):1264–1272, 2018.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in neural information processing systems*, pages 4066–4076, 2017.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6402–6413. Curran Associates, Inc., 2017. URL `http://papers.nips.cc/paper/7219-simple-and-scalable-predictive-uncertainty-estimation-using-deep-ensembles.pdf`.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

Olivier Ledoit and Sandrine Péché. Eigenvectors of some large sample covariance matrix ensembles. *Probability Theory and Related Fields*, 151(1-2):233–264, 2011.

Tao Lei, Yu Zhang, Sida I. Wang, Hui Dai, and Yoav Artzi. Simple recurrent units for highly parallelizable recurrence. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMzhou 2018*, pages 4470–4481. Association for Computational Linguistics, sep 2018. ISBN 9781948087841. doi: 10.18653/v1/d18-1477. URL `http://arxiv.org/abs/1709.02755`.

Tal Linzen. How can we accelerate progress towards human-like linguistic generalization? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5210–5217, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.465. URL `https://www.aclweb.org/anthology/2020.acl-main.465`.

Xiaoxuan Liu, Samantha Cruz Rivera, David Moher, Melanie J Calvert, and Alastair K Denniston. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the consort-ai extension. *bmj*, 370, 2020a.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Yuan Liu, Ayush Jain, Clara Eng, David H Way, Kang Lee, Peggy Bui, Kimberly Kanada, Guilherme de Oliveira Marinho, Jessica Gallegos, Sara Gabriele, et al. A deep learning system for differential diagnosis of skin diseases. *Nature Medicine*, pages 1–9, 2020b.

Sara Magliacane, Thijs van Ommen, Tom Claassen, Stephan Bongers, Philip Versteeg, and Joris M Mooij. Domain adaptation by using causal inference to predict invariant conditional distributions. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31 (NeurIPS2018)*, pages 10869–10879. Curran Associates, Inc., 2018. URL `http://papers.nips.cc/paper/8282-domain-adaptation-by-using-causal-inference-to-predict-invariant-conditional-distributions.pdf`.

Maggie Makar, Ben Packer, Dan Moldovan, Davis Blalock, Yoni Halpern, and Alexander D'Amour. Causally-motivated shortcut removal using auxiliary labels. *arXiv preprint arXiv:2105.06422*, 2021.

Alicia R Martin, Christopher R Gignoux, Raymond K Walters, Genevieve L Wojcik, Benjamin M Neale, Simon Gravel, Mark J Daly, Carlos D Bustamante, and Eimear E Kenny. Human demographic history impacts genetic risk prediction across diverse populations. *Am. J. Hum. Genet.*, 100(4):635–649, April 2017.

Alicia R Martin, Masahiro Kanai, Yoichiro Kamatani, Yukinori Okada, Benjamin M Neale, and Mark J Daly. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.*, 51(4):584–591, April 2019.

Charles T Marx, Flavio du Pin Calmon, and Berk Ustun. Predictive multiplicity in classification. *arXiv preprint arXiv:1909.06677*, 2019.

R Thomas McCoy, Junghyun Min, and Tal Linzen. Berts of a feather do not generalize together: Large variability in generalization across models with similar test set performance. *arXiv preprint arXiv:1911.02969*, 2019a.

R Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *arXiv preprint arXiv:1902.01007*, 2019b.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv:1908.05355*, 2019.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Joannella Morales, Danielle Welter, Emily H Bowler, Maria Cerezo, Laura W Harris, Aoife C McMahon, Peggy Hall, Heather A Junkins, Annalisa Milano, Emma Hastings, Cinzia Malangone, Annalisa Buniello, Tony Burdett, Paul Flicek, Helen Parkinson, Fiona Cunningham, Lucia A Hindorff, and Jacqueline A L MacArthur. A standardized framework for representation of ancestry data in genomics studies, with application to the NHGRI-EBI GWAS catalog. *Genome Biol.*, 19 (1):21, February 2018.

Sendhil Mullainathan and Jann Spiess. Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106, 2017.

Moin Nadeem, Anna Bethke, and Siva Reddy. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*, 2020.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. Stress test evaluation for natural language inference. *arXiv preprint arXiv:1806.00692*, 2018.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=B1g5sA4twr`.

National Institute for Health and Care Excellence (NICE). Acute kidney injury: prevention, detection and management. *NICE Guideline NG148*, 2019.

Radford M Neal. Priors for infinite networks. In *Bayesian Learning for Neural Networks*, pages 29–53. Springer, 1996.

Anna C Need and David B Goldstein. Next generation disparities in human genomics: concerns and remedies. *Trends Genet.*, 25(11):489–494, November 2009.

Bret Nestor, Matthew B. A. McDermott, Willie Boag, Gabriela Berner, Tristan Naumann, Michael C Hughes, Anna Goldenberg, and Marzyeh Ghassemi. Feature Robustness in Non-stationary Health Records: Caveats to Deployable Model Performance in Common Clinical Machine Learning Tasks. *Proceedings of Machine Learning Research*, 106:1–23, 2019. URL `https://mimic.physionet.org/mimicdata/carevue/http://arxiv.org/abs/1908.00690`.

Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pages 151–159, 2020.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, oct 2019. ISSN 10959203. doi: 10.1126/science.aax2342.

Cecilia Panigutti, Alan Perotti, and Dino Pedreschi. Doctor XAI An ontology-based approach to black-box sequential data classification explanations. In *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 629–639, 2020. ISBN 9781450369367. doi: 10.1145/3351095.3372855. URL `https://doi.org/10.1145/3351095.3372855`.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016. doi: 10.1111/rssb.12167. URL `https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12167`.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, 2018.

Alice B Popejoy and Stephanie M Fullerton. Genomics is failing on diversity. *Nature*, 538(7624): 161–164, October 2016.

Mihail Popescu and Mohammad Khalilia. Improving disease prediction using ICD-9 ontological features. In *IEEE International Conference on Fuzzy Systems*, pages 1805–1809, 2011. ISBN 9781424473175. doi: 10.1109/FUZZY.2011.6007410.

Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, 38(8):904–909, August 2006.

Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A R Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I W de Bakker, Mark J Daly, and Pak C Sham. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, 81(3):559–575, September 2007.

Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John Duchi, and Percy Liang. Understanding and mitigating the tradeoff between robustness and accuracy. *arXiv preprint arXiv:2002.10716*, 2020.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pages 1177–1184, 2008.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.442. URL `https://www.aclweb.org/anthology/2020.acl-main.442`.

Samantha Cruz Rivera, Xiaoxuan Liu, An-Wen Chan, Alastair K Denniston, and Melanie J Calvert. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the spirit-ai extension. *bmj*, 370, 2020.

Andrew Slavin Ross, Michael C Hughes, and Finale Doshi-Velez. Right for the right reasons: training differentiable models by constraining their explanations. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 2662–2670. AAAI Press, 2017.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, 2018.

Bernhard Schölkopf. Causality for machine learning. *arXiv preprint arXiv:1911.10500*, 2019.

Montgomery Slatkin. Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*, 9:477–485, 2008.

Jasper Snoek, Yaniv Ovadia, Emily Fertig, Balaji Lakshminarayanan, Sebastian Nowozin, D Sculley, Joshua Dillon, Jie Ren, and Zachary Nado. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, pages 13969–13980, 2019.

Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, Bette Liu, Paul Matthews, Giok Ong, Jill Pell, Alan Silman, Alan Young, Tim Sprosen, Tim Peakman, and Rory Collins. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.*, 12(3):e1001779, March 2015.

Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017.

Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*, 2017.

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *arXiv preprint arXiv:2007.00644*, 2020.

Daniel Shu Wei Ting, Carol Yim-Lui Cheung, Gilbert Lim, Gavin Siew Wei Tan, Nguyen D Quang, Alfred Gan, Haslina Hamzah, Renata Garcia-Franco, Ian Yew San Yeo, Shu Yen Lee, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *Jama*, 318(22):2211–2223, 2017.

Nenad Tomašev, Xavier Glorot, Jack W Rae, Michal Zielinski, Harry Askham, Andre Saraiva, Anne Mottram, Clemens Meyer, Suman Ravuri, Ivan Protsyuk, Alistair Connell, Cían O Hughes, Alan Karthikesalingam, Julien Cornebise, Hugh Montgomery, Geraint Rees, Chris Laing, Clifton R Baker, Kelly Peterson, Ruth Reeves, Demis Hassabis, Dominic King, Mustafa Suleyman, Trevor Back, Christopher Nielson, Joseph R Ledsam, and Shakir Mohamed. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature*, 572(7767):116–119, aug 2019a. ISSN 0028-0836. doi: 10.1038/s41586-019-1390-1.

Nenad Tomašev, Xavier Glorot, Jack W. Rae, Michal Zielinski, Harry Askham, Andre Saraiva, Anne Mottram, Clemens Meyer, Suman Ravuri, Ivan Protsyuk, Alistair Connell, Cian O. Hugues, Alan Kathikesalingam, Julien Cornebise, Hugh Montgomery, Geraint Rees, Chris Laing, Clifton R. Baker, Kelly Peterson, Ruth Reeves, Demis Hassabis, Dominic King, Mustafa Suleyman, Trevor Back, Christopher Nielson, Joseph R. Ledsam, and Shakir Mohamed. Developing Deep Learning Continuous Risk Models for Early Adverse Event Prediction in Electronic Health Records: an AKI Case Study. *PROTOCOL available at Protocol Exchange*, version 1, jul 2019b. doi: 10.21203/RS.2.10083/V1.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. Counterfactual invariance to spurious correlations: Why and how to pass stress tests. *arXiv preprint arXiv:2106.00545*, 2021.

Bjarni J Vilhjálmsson, Jian Yang, Hilary K Finucane, Alexander Gusev, Sara Lindström, Stephan Ripke, Giulio Genovese, Po-Ru Loh, Gaurav Bhatia, Ron Do, Tristan Hayeck, Hong-Hee Won, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Discovery, Biology, and Risk of Inherited Variants in Breast Cancer (DRIVE) study, Sekar Kathiresan, Michele Pato, Carlos Pato, Rulla Tamimi, Eli Stahl, Noah Zaitlen, Bogdan Pasaniuc, Gillian Belbin, Eimear E Kenny, Mikkel H Schierup, Philip De Jager, Nikolaos A Patsopoulos, Steve McCarroll, Mark Daly, Shaun Purcell, Daniel Chasman, Benjamin Neale, Michael Goddard, Peter M Visscher, Peter Kraft, Nick Patterson, and Alkes L Price. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am. J. Hum. Genet.*, 97(4):576–592, October 2015.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, 2018.

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pages 10506–10518, 2019.

Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P Xing. High-frequency component helps explain the generalization of convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8684–8694, 2020.

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, and Slav Petrov. Measuring and reducing gendered correlations in pre-trained models. *arXiv preprint arXiv:2010.06032*, 2020.

Florian Wenzel, Jasper Snoek, Dustin Tran, and Rodolphe Jenatton. Hyperparameter ensembles for robustness and uncertainty quantification. *arXiv preprint arXiv:2006.13570*, 2020.

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, 2018.

Andrew Gordon Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. *arXiv preprint arXiv:2002.08791*, 2020.

Julia K Winkler, Christine Fink, Ferdinand Toberer, Alexander Enk, Teresa Deinlein, Rainer Hofmann-Wellenhof, Luc Thomas, Aimilios Lallas, Andreas Blum, Wilhelm Stolz, et al. Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. *JAMA dermatology*, 155(10):1135–1141, 2019.

Naomi R Wray, Michael E Goddard, and Peter M Visscher. Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Res.*, 17(10):1520–1528, October 2007.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5753–5763, 2019.

Dong Yin, Raphael Gontijo Lopes, Jon Shlens, Ekin Dogus Cubuk, and Justin Gilmer. A fourier perspective on model robustness in computer vision. In *Advances in Neural Information Processing Systems*, pages 13255–13265, 2019.

Bin Yu et al. Stability. *Bernoulli*, 19(4):1484–1500, 2013.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. Swag: A large-scale adversarial dataset for grounded commonsense inference. *arXiv preprint arXiv:1808.05326*, 2018.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, 2019.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, 2018.

Xiang Zhou, Yixin Nie, Hao Tan, and Mohit Bansal. The curse of performance instability in analysis datasets: Consequences, source, and suggestions. *arXiv preprint arXiv:2004.13606*, 2020.

Appendix A. Computer Vision: Marginalization versus Model Selection

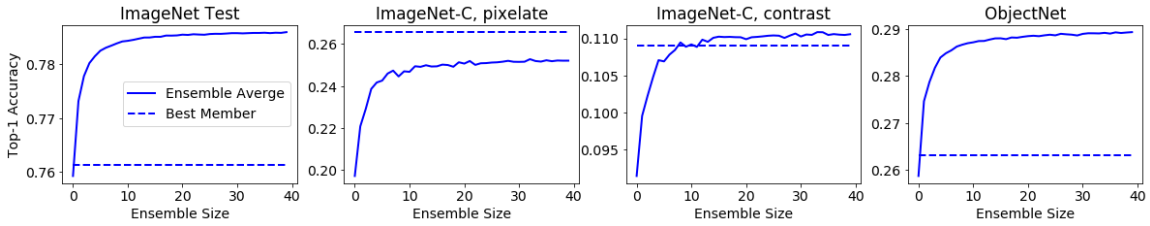

Figure 15: **Comparison of performance of the "best" ensemble member from an ensemble of 50 ResNet-50 predictor (dashed) against the average performance from averaging the predictions of differently-sized subsets of the ensemble (solid).** ImageNet Test is an iid evaluation; the other three panels show stress tests. See main text for full description. Iid performance on ImageNet improves as ensemble size increases, and this is associated with correlated improvements in stress tests. The larger the variability in stress test performance within the ensemble, the larger the ensemble needs to be to out-perform the best single ensemble member. In some cases, the ensemble average never out-performs the best single model.

In the discussion in the main text, we argue that marginalization may not be the best response to underspecification when the goal is to obtain predictors that encode the "right" structure for a given application. We suggest instead that model selection may be a reasonable approach here. This is because, by the nature of underspecification, some predictors returned by the pipeline will exhibit worse behavior in deployment domains than others, so averaging them together does not guarantee that the ensemble average will out-perform the best member. Notably, this represents a departure from the argument made in favor of marginalization for improving iid performance: in the training domain, all of the models in the validation-equivalent class $\mathcal{F}^*$ (recall this definition from Section 2 of the main text) contain a "right" answer for iid generalization, so one would expect that averaging them could only lead to improvements.

In this section, we provide some empirical support for this argument. Broadly, there is an interplay between iid performance and performance on stress tests that can make marginalization beneficial, but when there is large variability in stress test performance across an ensemble, selecting the best single model can out-perform large ensemble averages.

In Figure 15, we show a comparison between performance of individual ensemble members and ensemble averages on several test sets. We calculate these metrics with respect to the ensemble of 50 ResNet-50 models used to produce the result in Section 5. The dashed line shows the performance of the best model from this ensemble, while the solid line shows the average performance from marginalizing across differently-sized subsets of models in this ensemble. The ImageNet test set is the iid evaluation, while the other test sets are from the ImageNet-C and ObjectNet benchmarks. As expected, performance on the ImageNet test set improves substantially as more ensemble members are averaged together. This translates to correlated performance improvements on stress test benchmarks, which is a well-known phenomenon in the image robustness literature (see, e.g. Taori et al., 2020; Djolonga et al., 2020). Interestingly, however, it takes marginalizing across a larger subset of models to surpass the performance of the best predictor on stress tests compared to the iid evaluation. In the case of the iid evaluation, even a small ensemble is sufficient to surpass the best member. However, for the stress tests, the higher the variance of performance across the ensemble, the more predictors need to be averaged to beat the best single model. In the case of the pixealate task, the full ensemble of 50 models is never able to surpass the best single model.

Appendix B. Natural Language Processing

B.1 Analysis of Static Embeddings

In the main text, we showed that underspecification plays a key role in shortcut learning in BERT-based NLP models. However, highly parameterized models pre-date this approach, and here we provide a supplementary analysis suggesting that underspecification is also present in static word embeddings such as word2vec (Mikolov et al., 2013). Here, we examine stereotypical associations with respect to demographic attributes like race, gender, and age, which have been studied in the past (Bolukbasi et al., 2016).

We train twenty different 500-dimensional word2vec models on large news and wikipedia datasets using the `demo-train-big-model-v1.sh` script from the canonical word2vec repository,[8] varying only the random seeds. These models obtain very consistent performance on a word analogy task, scoring between 76.2% and 76.7%.

As a stress test, we apply the Word Embedding Association Test, which quantifies the extent to which these associations are encoded by a given set of embeddings (Caliskan et al., 2017). Specifically, the WEAT score measures the relative similarity of two sets of *target* words (e.g., types of flowers, types of insects) to two sets of *attribute* words (e.g., pleasant words, unpleansant words). Let X and Y be the sets of target words, and A and B the sets of attribute words. The test statistic is then:

$$s(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B)$$

where

$$s(w, A, B) = \text{mean}_{a \in A} \cos(w, a) - \text{mean}_{b \in B} \cos(w, b)$$

and $\cos(a, b)$ is the cosine distance between two word vectors. This score is then normalized by the standard deviation of $s(w, A, B)$ for all w in $X \cup Y$. If the score is closer to zero, the relative similarity difference (i.e., bias) is considered to be smaller. Caliskan et al. (2017) provide a number of wordsets to probe biases along socially salient axes of gender, race, disability, and age.

For each model and test, we compute statistical significance using a permutation test that compares the observed score against the score obtained under a random shuffling of target words. As shown in Figure 17, we find strong and consistent gender associations, but observe substantial variation on the three tests related to associations with race: in many cases, whether an association is statistically significant depends on the random seed. Finally, we note that the particular axes along which we observe sensitivity depends on the dataset used to train the embeddings, and could vary on embeddings trained on other corpora (Babaeianjelodar et al., 2020).

B.2 Exploratory Analysis of Gendered Correlations in STS Task

Here, we present an additional exploratory analysis of how different predictors in the semantic text similarity prediction ensemble in Section 7.1.1 process gender information. Specifically, we use principal component analysis to analyze the structure in how similarity scores produced by the predictors in our ensemble deviate from the ensemble mean. Here, we find that the main axis of variation aligns, at least at its extremes, with differences in how predictors represent sterotypical associations between profession and gender. Specifically, we perform principal components analysis (PCA) over similarity score produced by 20 fine-tunings of a single BERT checkpoint. We plot the first principal components, which contains 22% of the variation in score deviations, against BLS female participation percentages in Figure 16. Notably, examples in the region where the first principal component values are strongly negative include some of the strongest gender imbalances.

8. Canonical codebase is `https://code.google.com/archive/p/word2vec/`; a GitHub export of this repository is available at `https://github.com/tmikolov/word2vec`.

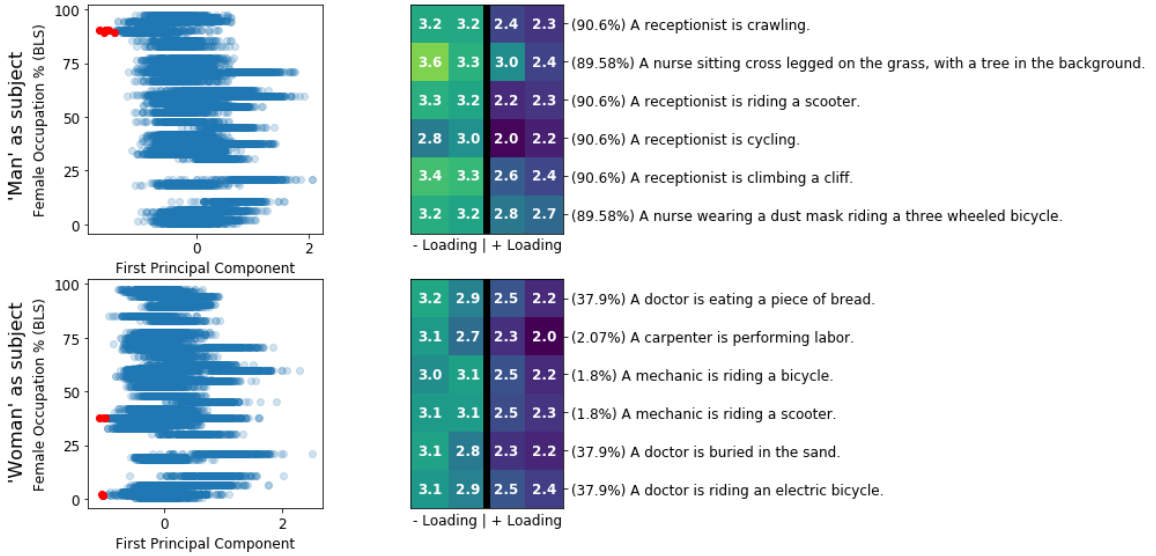

Figure 16: **The first principal axis of model disagreement predicts differences in handling stereotypes.** (Left) A plot of (profession, gender, template) examples from our STS stress test set. Each point is an example; the x-axis is the example's coordinate along first principal axis of disagreement between similarity predictions from our ensemble fine-tuned for STS-B; the y-axis is the % female participation of a profession in the BLS data. The top panel shows examples with a male subject (e.g., "a man") and the bottom panel shows examples with a female subject. The region to the far left shows that the first principal component aligns with apparent gender contradictions: 'man' partnered with a female-dominated profession (top) or 'woman' partnered with a male-dominated profession (bottom). (Right) Predicted similarities from four predictors in our ensemble on the examples highlighted in red on the left plot. These red examples are shown along with their BLS percentages in parentheses. Each column in the table next to the examples corresponds to similarity predictions from a particular predictor. The '- Loading' columns show predicted similarities produced by the two predictors with the most negative loadings on the first PC; the '+ Loading' columnes show predicted similarities from the two predictors with the most positive loadings. These predictors produce systematically different predictions on examples with apparent gender contradictions.

The right side of Figure 16 shows some of these examples (marked in red on the scatterplots), along with the predicted similarities from models that have strongly negative or strongly positive loadings on this principal axis. The similarity scores between these models are clearly divergent, with the positive-loading models encoding a sterotypical contradiction between gender and profession—that is, a contradiction between 'man' and 'receptionist' or 'nurse'; or a contradiction between 'woman' and 'mechanic', 'carpenter', and 'doctor'—that the negative-loading models do not.

Appendix C. Clinical Prediction with EHR: Additional Details and Supplementary Ablation Experiment

This section provides additional details and results for the analysis of the model in Tomašev et al. (2019a) performed in the main text. In particular, we provide some descriptive statistics regarding AKI prevalence, and additional summaries of model performance across different time slices and dataset shifts.

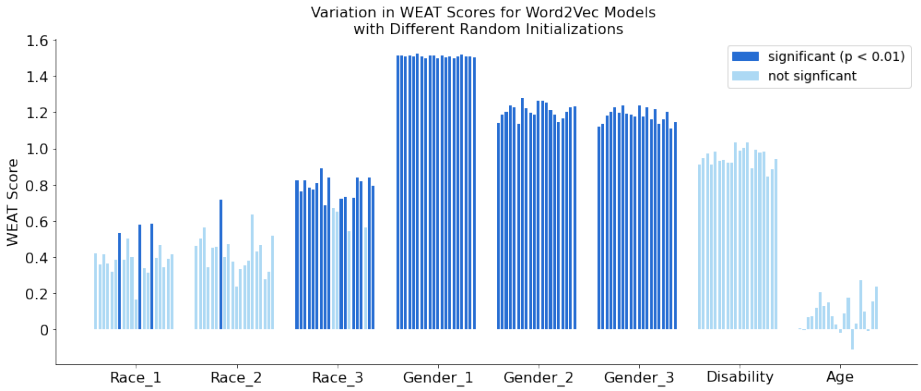

Figure 17: **Static word embeddings also show evidence of stereotype-aligned underspecification.** Word Embedding Association Test (WEAT) scores across twenty word2vec models. Each group corresponds to a specific association test, and each bar corresponds to the score on the test for a specific word2vec model.

C.1 Lab Order Patterns and Time of Day

In the main text, we investigate how reliant predictors can be on signals related to the timing and composition of lab tests. Here, we show some descriptive statistics for how these tests tend to be distributed in time, and some patterns that emerge as a result.

Table 5 shows patterns of AKI prevalence and creatinine sampling. Even though AKI prevalence is largely constant across time windows, creatinine is sampled far more frequently in between 12am and 6am. Thus, when creatinine samples are taken in other time windows, the prevalence of AKI conditional on that sample being taken is higher.

Figure 18 shows the distributions of number of labs taken at different times of day. The distribution of labs in the first two time buckets is clearly distinct from the distributions in the second two time buckets.

Table 5: **Patterns of creatinine sampling can induce a spurious relationship between time of day and AKI.** Prevalence of AKI is stable across times of day in the test set (test set prevalence is 2.269%), but creatinine samples are taken more frequently in the first two time buckets. As a result, *conditional on a sample being taken*, AKI prevalence is higher in the latter two time buckets.

Metric	**12am-6am**	**6am-12pm**	**12pm-6pm**	**6pm-12am**
Prevalence of AKI (%)	2.242	2.153	2.287	2.396
Creatinine samples	332743	320068	89379	67765
Creatinine samples (%)	3.570	3.433	0.959	0.727
Prevalence of AKI (%) in creatinine samples	6.236	5.425	8.824	9.154

C.2 Details of Predictor Performance on Intervened Data

Here, we perform a stratified evaluation of model performance across different time buckets in the standard and intervened test data. This analysis is repeated for each of the 15 models trained, and acoss the Shift and Shift+Labs intevened datasets described in the main text. Table 6 displays model performance on each model instance for the test set as well as for time windows where creatinine samples were taken.

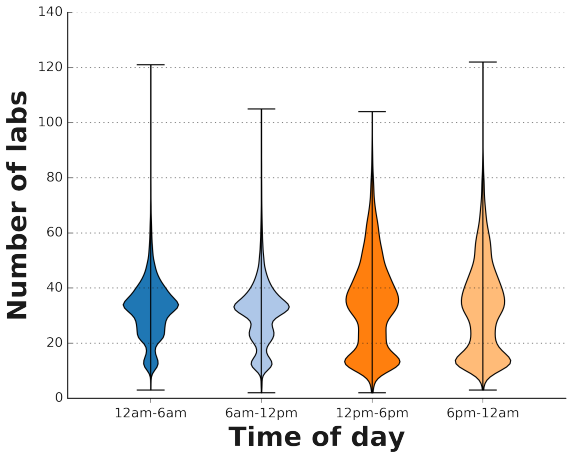

Figure 18: Distribution of number of lab values observed on average across 100000 time steps (random sample in the test set), per time of day. Each violin plot represent a time of day while the y-axis represents the number of lab values observed.

Table 6: **Model sensitivity per time of day:** Normalized PRAUC on the test set, per time of day, and per time of day for creatinine samples for each model instance when the data is not perturbed ('Test'), when time of day is shifted ('Shift') and when time of day is shifted and only CHEM-7 labs are considered for creatinine samples ('Shift+Labs'). 'Diff.' refers to the maximum difference in value between instances.

cell	**SRU**					**UGRNN**					**LSTM**					
seed	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**	**Diff.**
Test																
12am - 6am	34.69	34.42	34.64	34.99	34.74	34.61	34.88	34.46	34.86	34.62	35.43	35.55	35.42	35.63	35.51	1.21
6am - 12pm	34.26	34	34.14	34.32	34.26	33.98	34.39	34.05	34.3	34.01	34.86	35.06	34.91	35.12	34.94	1.14
12pm - 6pm	36.51	36.26	36.43	36.67	36.55	36.13	36.57	36.24	36.57	36.25	37.05	37.28	37.07	37.35	37.11	1.22
6pm - 12am	37.22	37	37.17	37.41	37.24	36.97	37.34	36.93	37.32	36.99	37.78	38.03	37.82	38.07	37.89	1.14
Creatinine samples																
12am - 6am	36.51	36.32	36.66	37.06	36.57	36.52	36.92	36.31	36.82	36.55	37.43	37.64	37.5	37.57	37.48	1.33
6am - 12pm	34.4	34.53	34.24	34.75	34.55	34.61	34.85	34.37	34.59	34.58	35.42	35.52	35.28	35.48	35.41	1.27
12pm - 6pm	42.68	42.5	42.68	42.34	42.23	42.17	42.49	41.64	42.44	42.25	43	43.2	42.93	43.77	43.38	2.13
6pm - 12am	41.81	41.86	41.72	42.17	41.91	41.87	42.52	41.05	41.7	41.75	42.58	42.79	42.86	42.84	43.19	2.14
Shift																
Population	34.93	34.9	34.74	35.08	34.93	35.08	35.46	35.11	35.53	35.13	35.97	36.16	35.93	36.21	35.82	1.47
Test																
12am - 6am	33.75	33.63	33.67	33.79	33.52	34.06	34.51	34.05	34.55	34.09	35.03	35.11	34.82	35.16	34.85	1.64
6am - 12pm	33.55	33.54	33.32	33.74	33.62	33.7	33.89	33.7	34	33.62	34.44	34.62	34.51	34.73	34.3	1.4
12pm - 6pm	35.72	35.78	35.57	35.94	35.88	35.79	36.22	35.92	36.28	35.97	36.74	36.97	36.71	37.03	36.56	1.45
6pm - 12am	36.49	36.45	36.28	36.71	36.45	36.52	36.99	36.54	37.04	36.61	37.45	37.68	37.41	37.69	37.32	1.41
Creatinine samples																
12am - 6am	36.15	36.1	36.22	36.66	36.12	36.18	36.72	36.12	36.63	36.24	37.2	37.43	37.12	37.41	37.2	1.33
6am - 12pm	33.72	33.97	33.64	34.06	33.97	34.44	34.5	34.05	34.39	34.23	35.03	35.14	34.97	35.15	34.97	1.51
12pm - 6pm	41.81	41.98	41.99	41.68	41.55	41.73	42.2	41.03	42.27	41.95	42.77	42.87	42.67	43.51	42.95	2.48
6pm - 12am	41	41.27	41.1	41.49	41.13	41.27	42.11	40.6	41.31	41.36	42.11	42.33	42.43	42.51	42.7	2.1
Shift+Labs																
Population	33.66	33.86	33.46	33.99	33.76	34.08	34.77	34.22	34.67	34.29	35.16	35.18	34.93	35.21	34.76	1.74
Test																
12am - 6am	32.5	32.62	32.44	32.74	32.29	32.94	33.79	33.19	33.72	33.18	34.25	34.05	33.81	34.2	33.72	1.96
6am - 12pm	32.22	32.44	32.09	32.55	32.43	32.76	33.23	32.84	33.07	32.77	33.68	33.72	33.52	33.73	33.34	1.64
12pm - 6pm	34.45	34.71	34.24	34.85	34.78	34.82	35.5	35	35.39	35.13	35.89	35.98	35.69	36.01	35.49	1.77
6pm - 12am	35.27	35.43	34.96	35.68	35.34	35.57	36.33	35.64	36.26	35.86	36.62	36.69	36.42	36.66	36.27	1.73
Creatinine samples																
12am - 6am	34.32	34.55	34.58	35.03	34.43	34.56	35.7	35.03	35.45	34.64	36.05	35.99	35.69	36.08	35.73	1.76
6am - 12pm	32.58	33.03	32.65	33.1	32.91	33.5	33.85	33.37	33.65	33.45	34.29	34.26	34.03	34.16	34.28	1.71
12pm - 6pm	40.05	40.37	40.1	40.13	39.95	39.97	40.91	39.59	40.85	40.35	41.1	40.85	40.83	41.81	41.03	2.23
6pm - 12am	38.91	39.43	38.87	39.53	39.18	39.33	40.63	39.13	39.68	39.7	40.42	40.23	40.35	40.79	40.68	1.92

When perturbing the correlation between AKI label, time of day and number of labs on a per patient basis, we observe a decrease in model performance, as well as a widening of the performance bounds (reported in the main text). This widening of performance bounds displays a differential effect of the shifts on the different models of the ensemble, both on the individual patient timepoints risk (Figures 19, 20, 21) and on the model's decisions (Table 7).

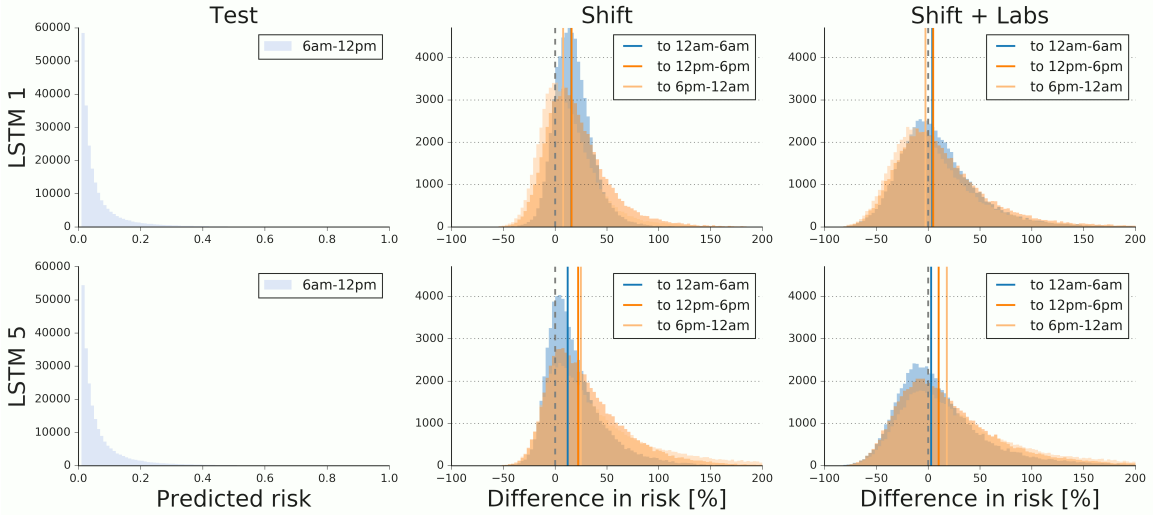

Figure 19: **Variability in AKI risk predictions from ensemble of RNN models processing electronic health records (EHR)**. Histograms showing showing risk predictions from two models, and changes induced by time of day and lab perturbations. Histograms show counts of patient-timepoints where creatinine measurements were taken in the morning (6am-12pm). LSTM 1 and 5 differ only in random seed. "Test" shows histogram of risk predicted in original test data. "Shift" and "Shift + Labs" show histograms of proportional changes (in %) $\frac{\text{Perturbed}-\text{Baseline}}{\text{Baseline}}$ induced by the time-shift perturbation and the combined time-shift and lab perturbation, respectively.

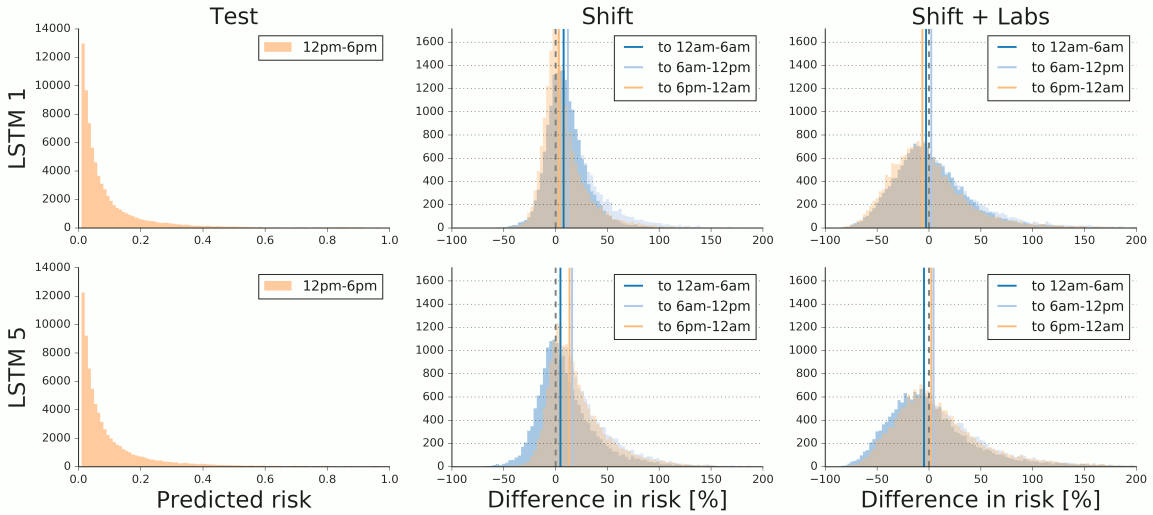

Figure 20: Same as Figure 19 for the afternoon (12pm-6pm).

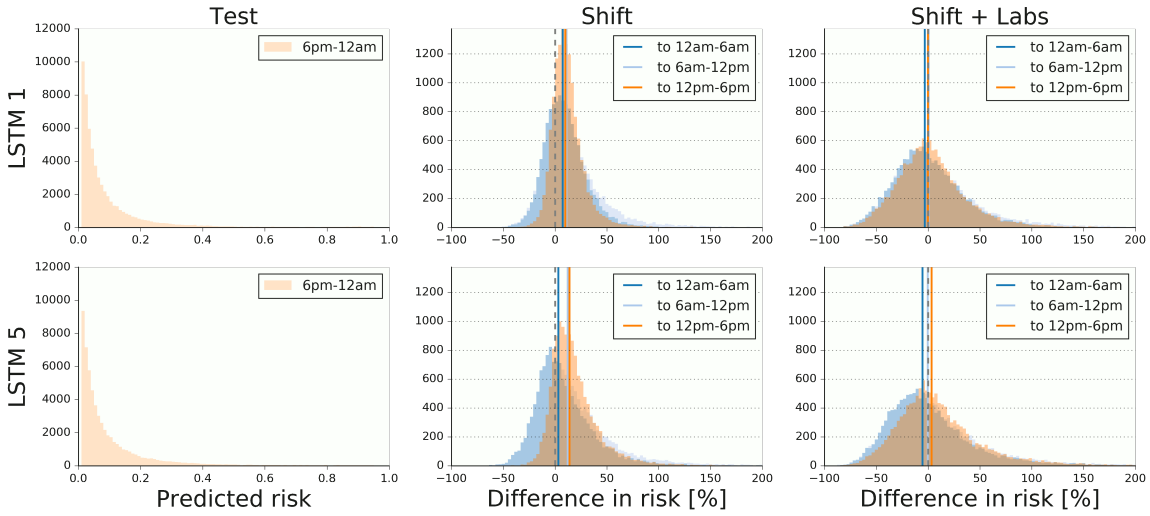

Figure 21: Same as Figure 19 for the evening (6pm-12am).

Table 7: **Flipped decisions under time-shift and lab order composition interventions depend on random seed.** Each cell is number of patient-timepoints at which decisions changed when the time range feature and lab order composition were changed, for patient timepoints with creatinine measured. "+ to -" indicates a change from the "at risk of AKI in next 48 hrs" to "not at risk"; "- to +" indicates the opposite change. Model 1 and model 2 are LSTM models that differ only in random seed. Overlap indicates the number of patient-timepoint flips shared between the two models. The number of flips in each direction changes as a function of random seed, and the patient-timepoints that flip are largely disjoint between random seeds.

	Shifted	E. Morning (12am-6am)		Morning (6am-12pm)		Afternoon (12pm-6pm)		Night (6pm-12am)	
Original		**+ to -**	**- to +**	**+ to -**	**- to +**	**+ to -**	**- to +**	**+ to -**	**- to +**
E. Morning (12am-6am)	model 1	1417	251	1378	456	1400	394	1700	396
	model 2	1560	285	1425	553	1206	764	1663	499
	overlap	570	31	459	70	423	103	646	76
Morning (6am-12pm)	model 1	1202	321	1279	188	1616	278	1523	215
	model 2	1368	281	1437	193	1694	297	1395	366
	overlap	483	42	508	26	700	49	585	50
Afternoon (12pm-6pm)	model 1	588	221	509	318	601	161	767	150
	model 2	738	191	555	271	607	109	708	189
	overlap	255	32	209	67	258	23	294	33
Night (6pm-12am)	model 1	422	199	423	349	410	223	441	124
	model 2	520	124	542	233	417	181	468	95
	overlap	189	35	188	73	159	41	186	19

C.3 Preliminary Ablation Experiment

Finally, to test the hypothesis that our results point to the possibility of modifying the signals a predictor uses to make its predictions *without* affecting iid performance, we perform an experiment

where we ablate the timestamp feature entirely while training a predictor. In particular, we rerun the pipeline with an LSTM architecture. This simple ablation leads to a test set population performance similar to the rest of our ensemble of predictors where that feature was included (normalized PRAUC of 0.368, compared to a range of 0.346 to 0.366).

In addition, there is evidence that underspecification here results from a collinearity between features, similar to that discussed in the Genomics example in the main text. In particular, this ablated model can predict time of day with an accuracy of 85% using an auxiliary head (without backpropagation). These results suggest that the signal related to time of day is present through different correlations that the training pipeline is unable to pull apart.

Appendix D. Genomics: Full Experimental Details

In this section we provide full details of the random featurization experiment using linear models in genomic medicine, along with a brief overview of the relevant research areas.

D.1 Background

In genetics research, a *genome-wide association study* (*GWAS*) is an observational study of a large group of individuals to identify genetic variants (or *genotypes*) associated with a particular trait (*phenotype*) of interest. One application of GWAS results is for construction of a *polygenic risk score* (*PRS*) Wray et al. (2007); International Schizophrenia Consortium et al. (2009) for the phenotype for each individual, generally defined as a weighted sum of the associated genotypes, where the weights are derived from the GWAS. One crucial factor to consider in this construction is that genetic variants are not independent and may contain highly correlated pairs due to a phenomenon called *linkage disequilibrium* (*LD*) Slatkin (2008). The most common way to correct for LD is to partition the associated variants into clusters of highly-correlated variants and to only include one representative of each cluster for the PRS (e.g. International Schizophrenia Consortium et al. (2009); CARDIoGRAMplusC4D Consortium et al. (2013)). There are other more advanced methods (e.g. Bayesian modeling of LD Vilhjálmsson et al. (2015)) which will not be discussed here.

While PRS show potential for identifying high-risk individuals for certain common diseases (e.g. Khera et al. (2018)) when derived and tested within one ancestry group (mostly European), recent work has shown that the prediction accuracy of PRS from one ancestry group does not necessarily generalize to other ancestry groups Martin et al. (2017); Duncan et al. (2019); Berg et al. (2019). When combined with the fact that more than three quarters of individuals in widely-used GWAS are of European ancestry Morales et al. (2018) (while representing less than a quarter of global population), this has raised scientific and ethical concerns about the clinical use of PRS and GWAS in the community Martin et al. (2019); Need and Goldstein (2009); Popejoy and Fullerton (2016).

D.2 Methods

In this work we investigate the issue of generalizability of PRS from a slightly different angle. Instead of focusing on the loss of predictive accuracy of a PRS when transferred to a different ancestry group, we investigate the *sensitivity* of the PRS to the choice of genotypes used in the derivation of the score, when evaluated within the same ancestry group versus outside of the group. Our phenotype of interest is the *intraocular pressure* (*IOP*), a continuous phenotype representing the fluid pressure inside the eye. This metric is an important aspect in the evaluation of risk of eye diseases such as glaucoma. We aim to predict IOP of individuals with their demographic information (age, sex, and BMI) and their genomic variants only, using the *UK Biobank* dataset Sudlow et al. (2015), a large, de-identified biobank study in the United Kingdom.

We first performed a GWAS on IOP and identified 4,054 genetic variants significantly associated with IOP distributed over 16 human chromosomes. We partitioned the variants into 129 clusters,

Table 8: **Distribution of IOP associated variants**. 129 variant clusters are distributed over 16 chromosomes.

chrom	clusters	chrom	clusters
chr1	19	chr9	10
chr2	10	chr11	16
chr3	10	chr13	2
chr4	13	chr14	2
chr5	1	chr16	2
chr6	9	chr17	4
chr7	15	chr20	2
chr8	11	chr22	3

where variants in the same cluster are highly-correlated and the ones in different clusters are relatively less correlated, and constructed the set of "index variants", consisting of the best representative of each cluster. We identified and clustered the IOP-associated variants with PLINK v1.9 Purcell et al. (2007), a standard tool in population genetics, using the `--clump` command. We used 5×10^{-8} for the index variant p-value threshold, 5×10^{-6} for the p-value threshold for the rest of the associated variants, 0.5 for the r^2 threshold, and 250 kb for the clumping radius. Table 8 summarizes the distribution of IOP-associated variant clusters over chromosomes.

After identifying the 129 clusters of variants, we created 1,000 *sets* of variants, each set consisting of 129 variants, exactly one variant in each cluster. The first of those sets is the set of index variants identified by PLINK. We then sampled 999 sets of cluster representatives by sampling one variant in each cluster uniformly at random. Each of these 1,000 sets defines a set of 129 genomic features to be used in our regression models.

For training and evaluation we partitioned the UK Biobank population into British and "non-British" individuals, and then we randomly partitioned the British individuals into British training and evaluation set. We leave the "non-British" individuals out of training and use them solely for evaluation. We measured the "British-ness" of an individual by the distance from the coordinate-wise median of the *self-reported* British individuals, in the 10-dimensional vector space of the top 10 principal components (PCs) of genetic variation Price et al. (2006). Individuals whose z-scored distance from the coordinate-wise median are no greater than 4 in this PC space, are considered British. We then randomly partitioned 91,971 British individuals defined as above into a British training set (82,309 individuals) and a British evaluation set (9,662 individuals). All remaining "non-British" set (14,898 individuals) was used for evaluation.

We trained linear regression models for predicting IOP with (a) demographics and a set of 129 genomic features (one of the 1,000 sets created above) and (b) demographic features only, using the British training set. We used L_2 regularization whose strength was determined by 10-fold cross validation in the training set.

We observed drastically increased sensitivity (Figure 3, left, in the main text) for the genomic models (blue dots) in the "non-British" evaluation set compared to the British evaluation set or the training set. Genomic models' margin of improvement from the baseline demographic model (gray line) is also decreased in the "non-British" evaluation set. In the "non-British" evaluation set we still see in general some improvement over the baseline, but the margins are highly variable. We also observed that the performance in British and "non-British" evaluation sets are mostly uncorrelated ($r = 0.131$) given the same set of genomic features (Figure 3, middle, in the main text).

Another interesting point is that the model using the original index variants (red dot) outperforms models with other choices of cluster representative in the British evaluation set, but not in the

"non-British" evaluation set. This implies the cluster representatives chosen in the training domain are not always the best representatives outside of the British ancestry group.

In summary, we investigated the sensitivity to the choice of features (variants) representing clusters of highly correlated features in the context of genomics in this section. We observed that the prediction errors become highly sensitive when the evaluation domain is distinct from the training domain, in addition to being higher in magnitude. As previously discussed the robust generalization of PRS in underrepresented ancestry groups is one of the major open questions for its real-world application in clinical settings.

Appendix E. Random Feature Model: Complete Theoretical Analysis

In this section we presents definitions and asymptotic formulas for the random features model that is presented in Section 3 of the main text. Before focusing on the random features model, we will describe the general setting and define some quantities of interest.

Our general focus is to investigate the validity of two main hypotheses that arose from our empirical case studies: (*i*) The model learnt depends strongly on the arbitrary or random choices in the training procedure; (*ii*) As a consequence, for most choices of the training procedure, there exist test distributions that are close to the train distribution and have much higher test error, while the test error on the same test distribution is unchanged for other choices of the training procedure.

E.1 General Definitions

We consider for simplicity a regression problem: we are given data $\{(\boldsymbol{x}_i, y_i)\}_{i\le n}$, with $\boldsymbol{x}_i \in \mathbb{R}^d$ vector of covariates and $y_i \in \mathbb{R}$ a response. We learn a model $f_\tau : \mathbb{R}^d \to \mathbb{R}$, where τ captures the arbitrary choices in the training procedure, such as initialization, stepsize schedule, and so on. Also, to be concrete, we consider square loss and hence the test error reads

$$R(\tau, \mathsf{P}_{\text{test}}) := \mathsf{E}_{\text{test}}\{(y - f_\tau(\boldsymbol{x}))^2\}\,. \tag{1}$$

Our definitions are easily generalized to other loss functions. The notation emphasizes that the test error is computed with respect to a distribution P_{test} that is not necessarily the same as the training one (which will be denoted simply by P). The classical in-distribution test error reads $R(\tau, \mathsf{P})$.

As a first question, we want to investigate to what extent the model $f_\tau(\boldsymbol{x})$ is dependent on the arbitrary choice of τ, in particular when this is random. In order to explore this point, we define the model sensitivity as

$$S(\tau_1, \tau_2; \mathsf{P}_{\text{test}}) := \mathsf{E}_{\text{test}}\big\{[f_{\tau_1}(\boldsymbol{x}) - f_{\tau_2}(\boldsymbol{x})]^2\big\}\,. \tag{2}$$

We next want to explore the effect to this sensitivity on the out-of-distribution test error. In particular, we want to understand whether the out-of-distribution error can increase significantly, even when the in-distribution error does not change much. Normally, the out-of-distribution risk is defined by constructing a suitable neighborhood of the train distribution P, call it $\mathcal{N}(\mathsf{P})$, and letting $R_{\text{shift}}(\tau_0) := \sup_{\mathsf{P}_{\text{test}} \in \mathcal{N}(\mathsf{P})} R(\tau_0; \mathsf{P}_{\text{test}})$.

Here we extend this classical definition, as to incorporate the constraint that the distribution shift should not damage the model constructed with an average choice of τ:

$$R_{\text{shift}}(\tau_0; \delta) := \sup_{\mathsf{P}_{\text{test}} \in \mathcal{N}(\mathsf{P})} \big\{R(\tau_0; \mathsf{P}_{\text{test}}) :\ \mathsf{E}_\tau R(\tau; \mathsf{P}_{\text{test}}) \le \delta\big\}\,. \tag{3}$$

E.2 Random Featurization Maps

A broad class of overparametrized models is obtained by constructing a featurization map $\boldsymbol{\phi}_\tau : \mathbb{R}^d \to \mathbb{R}^N$. We then fit a model that is linear in $\boldsymbol{\phi}_\tau(\boldsymbol{x})$, e.g. via min-norm interpolation

$$\text{minimize} \quad \|\boldsymbol{\theta}\|_2 \,, \tag{4}$$

$$\text{subject to} \quad \boldsymbol{y} = \boldsymbol{\Phi}_\tau(\boldsymbol{X})\boldsymbol{\theta} \,. \tag{5}$$

(Other procedures make sense as well.) Here $\boldsymbol{\Phi}_\tau(\boldsymbol{X}) \in \mathbb{R}^{n\times N}$ is the matrix whose i-th row is $\boldsymbol{\phi}_\tau(\boldsymbol{x}_i)$. The corresponding estimator is denoted by $\hat{\boldsymbol{\theta}}_\tau$, and the predictive model is $f_\tau(\boldsymbol{x}) = \langle \hat{\boldsymbol{\theta}}_\tau, \boldsymbol{\phi}_\tau(\boldsymbol{x})\rangle$. It is useful to consider a couple of examples.

Example 1. Imagine training a highly overparametrized neural network using SGD. Let $F(\,\cdot\,;\boldsymbol{w}) : \mathbb{R}^d \to \mathbb{R}$ be the input-output relation of the network. In the lazy training regime, this is well approximated by its first-order Taylor expansion around the initialization $\boldsymbol{w}_0$ (Chizat et al., 2019). Namely $F(\boldsymbol{x};\boldsymbol{w}) \approx F(\boldsymbol{x};\boldsymbol{w}_0)+\langle \nabla_{\boldsymbol{w}} F(\boldsymbol{x};\boldsymbol{w}_0), \boldsymbol{w}-\boldsymbol{w}_0\rangle$. If the initialization is symmetric, we can further neglect the zero-th order term, and, by letting $\boldsymbol{\theta} = \boldsymbol{w} - \boldsymbol{w}_0$, we obtain $F(\boldsymbol{x};\boldsymbol{w}) \approx \langle \nabla_{\boldsymbol{w}} F(\boldsymbol{x};\boldsymbol{w}_0), \boldsymbol{\theta}\rangle$. We can therefore identify $\tau = (\boldsymbol{w}_0)$ and $\boldsymbol{\phi}_\tau(\boldsymbol{x}) = \nabla_{\boldsymbol{w}} F(\boldsymbol{x};\boldsymbol{w}_0)$.

Example 2. Imagine $\boldsymbol{x}_i \in \mathbb{R}^d$ represents the degree of activity of d biological mechanism in patient i. We do not have access to $\boldsymbol{x}_i$, and instead we observe the expression levels of $N_0 \gg d$ genes, which are given by $\boldsymbol{u}_{0,i} = \boldsymbol{W}_0\boldsymbol{x}_i + \boldsymbol{z}_{0,i}$, where $\boldsymbol{W}_0 \in \mathbb{R}^{N_0\times d}$ and $\boldsymbol{z}_{0,i}$ are unexplained effects. We do not fit a model that uses the whole vector $\boldsymbol{u}_{0,i}$, and instead select by a clustering procedure τ a subset of N genes, hence obtaining a vector of features $\boldsymbol{u}_i = \boldsymbol{W}_\tau\boldsymbol{x}_i + \boldsymbol{z}_i \in \mathbb{R}^N$. In this case we identify the featurization map with the random map $\boldsymbol{\phi}_\tau(\boldsymbol{x}_i) := \boldsymbol{W}_\tau\boldsymbol{x}_i + \boldsymbol{z}_i$.

As a mathematically rich and yet tractable model, we consider the random features model of Rahimi and Recht (2008), whereby

$$\boldsymbol{\phi}_\tau(\boldsymbol{x}) = \sigma(\boldsymbol{W}\boldsymbol{x}) \,, \qquad \boldsymbol{W} \in \mathbb{R}^{N\times d} \,. \tag{6}$$

Here $\sigma : \mathbb{R} \to \mathbb{R}$ is an activation function: it is understood that this applied to vectors entrywise. Further, $\boldsymbol{W}$ is a matrix of first-layer weights which are drawn randomly and are not optimized over. We will draw the rows $(\boldsymbol{w}_i)_{i\le N}$ independently with $\boldsymbol{w}_i \sim \text{Unif}(\mathbb{S}^{d-1}(1))$. (Here and below $\mathbb{S}^{d-1}(r)$ is the sphere of radius r in d dimensions.) We identify the arbitrary training choice with the choice of this first-layer weights $\tau = \boldsymbol{W}$.

We assume an extremely simple data distribution, namely $\boldsymbol{x}_i \sim \text{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$ and $y_i = f_*(\boldsymbol{x}_i) = \langle \boldsymbol{\beta}_0, \boldsymbol{x}_i\rangle$. Note that $\|f_*\|_{L^2} = \|\boldsymbol{\beta}_0\|_2$.

We will derive exact characterizations of the sensitivity and risk under shifted test distribution in the proportional asymptotics $N, n, d \to \infty$, with

$$\frac{N}{d} \to \psi_1 \,, \quad \frac{n}{d} \to \psi_2 \,, \tag{7}$$

for some $\psi_1, \psi_2 \in (0,\infty)$. In what follows we will assume to to be given sequences of triples (N, n, d) which without loss of generality we can think to be indexed by d. When we write $d \to \infty$, it is understood that $N, n \to \infty$ as well , with Eq. (7) holding. Finally, we assume $\lim_{d\to\infty} \|\boldsymbol{\beta}_0\|_2^2 = r^2 \in (0,\infty)$.

E.3 Random features model: Risk

We begin by recalling some results and notations from Mei and Montanari (2019). We refer to the original paper for formal statements.

In this section we consider the random features model under a slightly more general setting than the one introduced above. Namely, we allow for Gaussian noise $y_i = f_*(\boldsymbol{x}_i) + \varepsilon_i$, $\varepsilon_i \sim \mathsf{N}(0, s^2)$, and perform ridge regression with a positive regularization parameter $\lambda > 0$:

$$\hat{\boldsymbol{\theta}}(\boldsymbol{W}, \boldsymbol{X}, \boldsymbol{y}; \lambda) := \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^N} \Big\{ \frac{1}{n} \sum_{i=1}^{n} \big(y_i - \langle \boldsymbol{\theta}, \sigma(\boldsymbol{W}\boldsymbol{x}_i) \rangle\big)^2 + \frac{N\lambda}{d} \|\boldsymbol{\theta}\|_2^2 \Big\}, \tag{8}$$

In the following we will omit all arguments except $\boldsymbol{W}$, and write $\hat{\boldsymbol{\theta}}(\boldsymbol{W}) := \hat{\boldsymbol{\theta}}(\boldsymbol{W}, \boldsymbol{X}, \boldsymbol{y}; \lambda)$. By specializing our general definition, the risk per realization of the first layer weights $\boldsymbol{W}$ is given by

$$R(\boldsymbol{W}, \mathsf{P}_{\text{test}}) = R(\tau, \mathsf{P}_{\text{test}}) := \mathsf{E}_{\text{test}} \big\{ [y - \langle \hat{\boldsymbol{\theta}}(\boldsymbol{W}), \sigma(\boldsymbol{W}\boldsymbol{x}) \rangle]^2 \big\}. \tag{9}$$

The activation function is assumed to be $\sigma \in L^2(\mathbb{R}, \gamma)$, with γ the Gaussian measure. Such an activation function is characterized via its projection onto constant and linear functions

$$\mu_0 := \mathsf{E}\{\sigma(G)\}, \quad \mu_1 := \mathsf{E}\{G\sigma(G)\}, \quad \mu_*^2 := \mathsf{E}\{\sigma(G)^2\} - \mu_0^2 - \mu_1^2. \tag{10}$$

In particular, we define the following ratio

$$\zeta \equiv \frac{\mu_1}{\mu_*}. \tag{11}$$

Let $\nu_1, \nu_2 : \mathbb{C}_+ \to \mathbb{C}$ be analytic functions such that, for $\Im(\xi) > C$ a large enough constant $\nu_1(\xi), \nu_2(\xi)$ satisfy

$$\begin{aligned} \nu_1 &= \psi_1\Big(-\xi - \nu_2 - \frac{\zeta^2 \nu_2}{1 - \zeta^2 \nu_1 \nu_2} \Big)^{-1}, \\ \nu_2 &= \psi_2\Big(-\xi - \nu_1 - \frac{\zeta^2 \nu_1}{1 - \zeta^2 \nu_1 \nu_2} \Big)^{-1}. \end{aligned} \tag{12}$$

We then let

$$\chi \equiv \nu_1(\boldsymbol{i}(\psi_1\psi_2\overline{\lambda})^{1/2}) \cdot \nu_2(\boldsymbol{i}(\psi_1\psi_2\overline{\lambda})^{1/2}), \tag{13}$$

$$\overline{\lambda} = \frac{\lambda}{\mu_*^2}. \tag{14}$$

Finally, define

$$\begin{aligned} \mathscr{E}_0(\zeta, \psi_1, \psi_2, \overline{\lambda}) \equiv\ & -\chi^5\zeta^6 + 3\chi^4\zeta^4 + (\psi_1\psi_2 - \psi_2 - \psi_1 + 1)\chi^3\zeta^6 - 2\chi^3\zeta^4 - 3\chi^3\zeta^2 \\ & + (\psi_1 + \psi_2 - 3\psi_1\psi_2 + 1)\chi^2\zeta^4 + 2\chi^2\zeta^2 + \chi^2 + 3\psi_1\psi_2\chi\zeta^2 - \psi_1\psi_2, \\ \mathscr{E}_1(\zeta, \psi_1, \psi_2, \overline{\lambda}) \equiv\ & \psi_2\chi^3\zeta^4 - \psi_2\chi^2\zeta^2 + \psi_1\psi_2\chi\zeta^2 - \psi_1\psi_2, \\ \mathscr{E}_2(\zeta, \psi_1, \psi_2, \overline{\lambda}) \equiv\ & \chi^5\zeta^6 - 3\chi^4\zeta^4 + (\psi_1 - 1)\chi^3\zeta^6 + 2\chi^3\zeta^4 + 3\chi^3\zeta^2 + (-\psi_1 - 1)\chi^2\zeta^4 - 2\chi^2\zeta^2 - \chi^2. \end{aligned} \tag{15}$$

We then have, from (Mei and Montanari, 2019, Theorem 1), the following characterization of the *in-distribution* risk[9]

$$R(\boldsymbol{W}, \mathsf{P}) = r^2 \frac{\mathscr{E}_1(\zeta, \psi_1, \psi_2, \overline{\lambda})}{\mathscr{E}_0(\zeta, \psi_1, \psi_2, \overline{\lambda})} + s^2 \frac{\mathscr{E}_2(\zeta, \psi_1, \psi_2, \overline{\lambda})}{\mathscr{E}_0(\zeta, \psi_1, \psi_2, \overline{\lambda})} + o_{\mathsf{P}}(1). \tag{16}$$

Here the $o_P(1)$ term depends on the realization of $\boldsymbol{W}$, and is such that $\mathsf{E}|o_{\mathsf{P}}(1)| \to 0$ as $N, n, d \to \infty$.

Remark 1 *Notice that the right-hand side of Eq.* (16) *is independent of* $\boldsymbol{W}$*. Hence we see that the in-distribution error is (for large* N, n, d*) essentially the same for most choices of* $\boldsymbol{W}$*.*

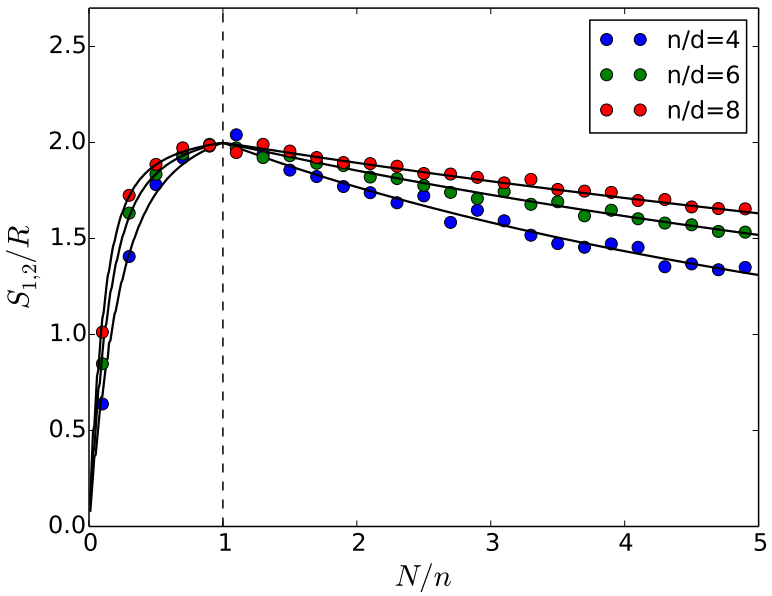

Figure 22: Random features model trained via min-norm least squares: sensitivity to the initial condition, normalized by the risk. Here the input dimension is $d = 40$, N is the number of neurons, and n the number of samples. We use ReLU activations; the ground truth is linear with $\|\boldsymbol{\beta}_0\|_2 = 1$. Circles are empirical results obtained by averaging over 50 realizations. Continuous lines correspond to the analytical prediction of Eq. (34).

E.4 Random features model: Sensitivity to random featurization

Let $\boldsymbol{W}_1, \boldsymbol{W}_2$ be two realizations of the first-layer weights. We can decompose

$$\sigma(\boldsymbol{W}\boldsymbol{x}) = \mu_0 + \mu_1 \boldsymbol{W}\boldsymbol{x} + \mu_* \boldsymbol{z}^\perp \,, \tag{17}$$

where, under P, we have $\mathsf{E}\{\boldsymbol{z}^\perp(\boldsymbol{z}^\perp)^\mathsf{T}\} = \mathbf{I} + c_n \mathbf{1}\mathbf{1}^\mathsf{T} + \boldsymbol{\Delta}$, $\|\boldsymbol{\Delta}\|_{\rm op} = o_\mathsf{P}(1)$, and $\mathsf{E}\{(\boldsymbol{W}\boldsymbol{x})\boldsymbol{z}^\perp\} = 0$. We therefore have (writing $\hat{\boldsymbol{\theta}}_i := \hat{\boldsymbol{\theta}}(\boldsymbol{W}_i)$)

$$\begin{aligned}
S(\boldsymbol{W}_1, \boldsymbol{W}_2) &= \mathsf{E}\big\{(\langle \hat{\boldsymbol{\theta}}(\boldsymbol{W}_1), \sigma(\boldsymbol{W}_1\boldsymbol{x})\rangle - \langle \hat{\boldsymbol{\theta}}(\boldsymbol{W}_2), \sigma(\boldsymbol{W}_2\boldsymbol{x})\rangle)^2\big\} \\
&= \mu_1^2 \mathsf{E}\big\{(\langle \hat{\boldsymbol{\theta}}(\boldsymbol{W}_1), \boldsymbol{W}_1\boldsymbol{x}\rangle - \langle \hat{\boldsymbol{\theta}}(\boldsymbol{W}_2), \boldsymbol{W}_2\boldsymbol{x}\rangle)^2\big\} + \mu_*^2 \mathsf{E}\big\{(\langle \hat{\boldsymbol{\theta}}(\boldsymbol{W}_1) - \hat{\boldsymbol{\theta}}(\boldsymbol{W}_2), \boldsymbol{z}^\perp\rangle)^2\big\} \\
&= \mu_1^2 \mathsf{E}_{\boldsymbol{X},\boldsymbol{y}}\big\{\big\|\boldsymbol{W}_1^\mathsf{T}\hat{\boldsymbol{\theta}}_1 - \boldsymbol{W}_2^\mathsf{T}\hat{\boldsymbol{\theta}}_2\big\|_2^2\big\} + (\mu_*^2 + o_\mathsf{P}(1))\mathsf{E}_{\boldsymbol{X},\boldsymbol{y}}\{\|\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_2\|^2\} \,.
\end{aligned}$$

We consider two random independent choices of $\boldsymbol{W}_1$, $\boldsymbol{W}_2$, and define $S_{\rm av} := \mathsf{E}\{S(\boldsymbol{W}_1, \boldsymbol{W}_2)\}$, thus obtaining:

$$S_{\rm av} = 2\mu_1^2 \Big\{ \mathsf{E}\big[\|\boldsymbol{W}^\mathsf{T}\hat{\boldsymbol{\theta}}(\boldsymbol{W})\|_2^2\big] - \mathsf{E}\big[\big\|\mathsf{E}[\boldsymbol{W}^\mathsf{T}\hat{\boldsymbol{\theta}}(\boldsymbol{W})|\boldsymbol{X}, \boldsymbol{\beta}_0]\big\|_2^2\big]\big] \Big\} + 2\mu_*^2 \mathsf{E}\big\{\|\hat{\boldsymbol{\theta}}(\lambda)\|_2^2\big\} + o(1) \,. \tag{18}$$

9. Theorem 1 in Mei and Montanari (2019) holds for the risk, conditional on the realization of $\boldsymbol{X}, \boldsymbol{y}$. The statement given here is obtained simply my taking expectation over $\boldsymbol{X}, \boldsymbol{y}$.

In order to evaluate the asymptotics of this expression, we recall some formulas that follow from Mei and Montanari (2019):

$$\|\boldsymbol{W}^{\mathsf{T}}\hat{\boldsymbol{\theta}}(\boldsymbol{W})\|_2^2 = \frac{r^2}{\mu_*^2}\cdot\frac{\mathscr{D}_1(\zeta,\psi_1,\psi_2,\overline{\lambda})}{(\chi\zeta^2-1)\mathscr{D}_0(\zeta,\psi_1,\psi_2,\overline{\lambda})} + \frac{s^2}{\mu_*^2}\cdot\frac{\mathscr{D}_2(\zeta,\psi_1,\psi_2,\overline{\lambda})}{\mathscr{D}_0(\zeta,\psi_1,\psi_2,\overline{\lambda})} + o_{\mathsf{P}}(1)\,, \tag{19}$$

$$\mathsf{E}(\boldsymbol{W}^{\mathsf{T}}\hat{\boldsymbol{\theta}}(\boldsymbol{W})|\boldsymbol{\beta}_0) = \left\{\frac{1}{\mu_1}\mathscr{H}(\zeta,\psi_1,\psi_2,\overline{\lambda}) + o(1)\right\}\boldsymbol{\beta}_0\,, \tag{20}$$

$$\|\hat{\boldsymbol{\theta}}(\boldsymbol{W})\|_2^2 = \frac{r^2}{\mu_*^2}\frac{\mathscr{G}_1(\zeta,\psi_1,\psi_2,\overline{\lambda})}{\mathscr{G}_0(\zeta,\psi_1,\psi_2,\overline{\lambda})} + \frac{s^2}{\mu_*^2}\frac{\mathscr{G}_2(\zeta,\psi_1,\psi_2,\overline{\lambda})}{\mathscr{G}_0(\zeta,\psi_1,\psi_2,\overline{\lambda})} + o_{\mathsf{P}}(1)\,, \tag{21}$$

Here the terms $o_{\mathsf{P}}(1)$ converge to 0 in L^1, and we used the following notations:

$$\begin{aligned}\mathscr{D}_0(\zeta,\psi_1,\psi_2,\overline{\lambda}) &= \chi^5\zeta^6 - 3\chi^4\zeta^4 + (\psi_1+\psi_2-\psi_1\psi_2-1)\chi^3\zeta^6 + 2\chi^3\zeta^4 + 3\chi^3\zeta^2 \\ &\quad + (3\psi_1\psi_2-\psi_2-\psi_1-1)\chi^2\zeta^4 - 2\chi^2\zeta^2 - \chi^2 - 3\psi_1\psi_2\chi\zeta^2 + \psi_1\psi_2\,,\end{aligned} \tag{22}$$

$$\begin{aligned}\mathscr{D}_1(\zeta,\psi_1,\psi_2,\overline{\lambda}) &= \chi^6\zeta^6 - 2\chi^5\zeta^4 - (\psi_1\psi_2-\psi_1-\psi_2+1)\chi^4\zeta^6 + \chi^4\zeta^4 \\ &\quad + \chi^4\zeta^2 - 2(1-\psi_1\psi_2)\chi^3\zeta^4 - (\psi_1+\psi_2+\psi_1\psi_2+1)\chi^2\zeta^2 - \chi^2\,,\end{aligned} \tag{23}$$

$$\mathscr{D}_2(\zeta,\psi_1,\psi_2,\overline{\lambda}) = -(\psi_1-1)\chi^3\zeta^4 - \chi^3\zeta^2 + (\psi_1+1)\chi^2\zeta^2 + \chi^2\,, \tag{24}$$

$$\begin{aligned}\mathscr{G}_0(\zeta,\psi_1,\psi_2,\overline{\lambda}) &= -\chi^5\zeta^6 + 3\chi^4\zeta^4 + (\psi_1\psi_2-\psi_2-\psi_1+1)\chi^3\zeta^6 - 2\chi^3\zeta^4 - 3\chi^3\zeta^2 \\ &\quad + (\psi_1+\psi_2-3\psi_1\psi_2+1)\chi^2\zeta^4 + 2\chi^2\zeta^2 + \chi^2 + 3\psi_1\psi_2\chi\zeta^2 - \psi_1\psi_2\,,\end{aligned} \tag{25}$$

$$\mathscr{G}_1(\zeta,\psi_1,\psi_2,\overline{\lambda}) = -\chi^2(\chi\zeta^4 - \chi\zeta^2 + \psi_2\zeta^2 + \zeta^2 - \chi\psi_2\zeta^4 + 1\,, \tag{26}$$

$$\mathscr{G}_2(\zeta,\psi_1,\psi_2,\overline{\lambda}) = \chi^2(\chi\zeta^2-1)(\chi^2\zeta^4 - 2\chi\zeta^2 + \zeta^2 + 1)\,, \tag{27}$$

$$\mathscr{H}(\zeta,\psi_1,\psi_2,\overline{\lambda}) = \frac{\zeta^2\chi}{\zeta^2\chi-1}\,. \tag{28}$$

We also claim that, for $s=0$ we have

$$\mathsf{E}(\boldsymbol{W}^{\mathsf{T}}\hat{\boldsymbol{\theta}}(\boldsymbol{W})|\boldsymbol{X},\boldsymbol{\beta}_0) = \frac{\mu_1}{d}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}\Big(\frac{\mu_1^2}{d}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} + \mu_*^2(1+q)\mathbf{I}_d\Big)^{-1}\boldsymbol{\beta}_0 + \mathbf{err}(d,\lambda)\,, \tag{29}$$

$$q := -\frac{(\psi_2-\psi_1)_+}{\chi_0}\,, \tag{30}$$

$$\chi_0 = \frac{1+\zeta^2-\psi_!\zeta^2 - \sqrt{(1+\zeta^2-\psi_1\zeta^2)^2+4\psi_1\zeta^2}}{2\zeta^2}\,, \tag{31}$$

where the error term $\mathbf{err}(d,\lambda)$ satisfies

$$\lim_{\lambda\to 0}\lim_{d\to\infty}\mathsf{E}\big\{\|\mathbf{err}(d,\lambda)\|_2^2\big\} = 0\,. \tag{32}$$

We refer to Section E.6 for a sketch of the proof of this claim.

Using Eq. (29) and the asymptotics of the Stieltjes transform of the spectrum of Wishart matrices, it follows that

$$\begin{aligned}&\lim_{\lambda\to 0}\lim_{d\to\infty}\mathsf{E}\big[\big\|\mathsf{E}[\boldsymbol{W}^{\mathsf{T}}\hat{\boldsymbol{a}}(\lambda)|\boldsymbol{X},\boldsymbol{\beta}_0]\big\|_2^2\big] = r^2\mathscr{L}(\zeta,\psi_1,\psi_2)\,,\\ &\mathscr{L}(\zeta,\psi_1,\psi_2) := 1 - 2\frac{1+q}{\zeta^2}\,g\Big(-\frac{1+q}{\zeta^2};\psi_2\Big) + \Big(\frac{1+q}{\zeta^2}\Big)^2 g'\Big(-\frac{1+q}{\zeta^2};\psi_2\Big)\,,\\ &g(z;\psi_2) := \frac{\psi_2-1-z-\Delta}{2z}\,,\qquad \Delta := -\sqrt{(\psi_2-1-z)^2-4z}\,,\\ &g'(z;\psi_2) = \frac{-\psi_2+1+z+\Delta}{2z^2} - \frac{-\psi_2-1+z+\Delta}{2z\Delta}\,.\end{aligned} \tag{33}$$

Using Eqs. (19), (21), (33) in Eq. (18), we finally obtain:

$$\lim_{\lambda\to 0}\lim_{d\to\infty} S_{\rm av} = 2r^2\Big\{\frac{\zeta^2\mathscr{D}_1}{(\chi\zeta^2-1)\mathscr{D}_0} - \mathscr{L} + \frac{\mathscr{G}_1}{\mathscr{G}_0}\Big\}\,. \tag{34}$$

In Figure 3 in the main text we compare this asymptotic prediction with numerical simulations for $d = 40$. We report the in-distribution sensitivity $S_{\rm av}$ normalized by the risk $R(\boldsymbol{W},\mathsf{P})$. In the classical underparametrized regime $N/n \to 0$, the sensitivity is small. However, as the number of neurons increases, S/R grows rapidly, with S/R not far from 2 over a large interval. Notice that $S/R = 2$ has a special meaning. Letting $h_i(\boldsymbol{x}) = f_{\tau_i}(\boldsymbol{x}) - f_*(\boldsymbol{x})$, it corresponds to $\|h_1 - h_2\|_{L^2}^2 = 2\|h_1\|_{L^2}^2 = 2\|h_2\|_{L^2}^2$, i.e. $\langle h_1, h_2\rangle_{L^2} = 0$. In other words, two models generated with two random choices of τ as 'as orthogonal as they can be'.

E.5 Random features model: Distribution shift

In order to explore the effect of a distribution shift in the random features model, we consider the case of a mean shift. Namely, $\boldsymbol{x}_{\rm test} = \boldsymbol{x}_0 + \boldsymbol{x}$ where $\boldsymbol{x} \sim {\rm Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$ is again uniform on the sphere, and $\boldsymbol{x}_0$ is deterministic and adversarial for a given choice $\boldsymbol{W}_0$, under the constraint $\|\boldsymbol{x}_0\|_2 \le \Delta$. We denote this distribution by $\mathsf{P}_{\boldsymbol{W}_0,\Delta}$. We will construct a specific perturbation $\boldsymbol{x}_0$ that produces a large increase in $R(\boldsymbol{W}_0, \mathsf{P}_{\boldsymbol{W}_0,\Delta})$ but a small change on $R(\boldsymbol{W}, \mathsf{P}_{\boldsymbol{W}_0,\Delta})$ for a typical random $\boldsymbol{W}$ independent of $\boldsymbol{W}_0$. We leave to future work the problem of determining the worst case perturbation $\boldsymbol{x}_0$.

We next consider the risk when the first layer weights are $\boldsymbol{W}$, and the test distribution is $\mathsf{P}_{\boldsymbol{W}_0,\Delta}$. Using the fact that $\|\boldsymbol{x}_0\|_2 = \Delta \ll \|\boldsymbol{x}\|_2$, we get

$$\begin{aligned}
R(\boldsymbol{W}, \mathsf{P}_{\boldsymbol{W}_0,\Delta}) &= \mathsf{E}\big\{(\langle\boldsymbol{\beta}_0, \boldsymbol{x}_{\rm test}\rangle - \langle\hat{\boldsymbol{\theta}}(\boldsymbol{W}), \sigma(\boldsymbol{W}\boldsymbol{x}_{\rm test})\rangle)^2\big\}\\
&= \mathsf{E}\big\{(\langle\boldsymbol{\beta}_0, \boldsymbol{x}\rangle + \langle\boldsymbol{\beta}_0, \boldsymbol{x}_0\rangle - \langle\hat{\boldsymbol{\theta}}(\boldsymbol{W}), \sigma(\boldsymbol{W}\boldsymbol{x})\rangle - \langle\hat{\boldsymbol{\theta}}(\boldsymbol{W}), \sigma'(\boldsymbol{W}\boldsymbol{x})\odot\boldsymbol{W}\boldsymbol{x}_0\rangle)^2\big\} + o_{\mathsf{P}}(1)\\
&\overset{(a)}{=} \mathsf{E}\big\{(\langle\boldsymbol{\beta}_0, \boldsymbol{x}\rangle + \langle\boldsymbol{\beta}_0, \boldsymbol{x}_0\rangle - \langle\hat{\boldsymbol{\theta}}(\boldsymbol{W}), \sigma(\boldsymbol{W}\boldsymbol{x})\rangle - \mu_1\langle\hat{\boldsymbol{\theta}}(\boldsymbol{W}), \boldsymbol{W}\boldsymbol{x}_0\rangle)^2\big\} + o_{\mathsf{P}}(1)\\
&\overset{(b)}{=} R(\boldsymbol{W}, \mathsf{P}_{\boldsymbol{W}_0,\Delta}) + \langle\boldsymbol{\beta}_0 - \mu_1\boldsymbol{W}^{\mathsf{T}}\hat{\boldsymbol{\theta}}(\boldsymbol{W}), \boldsymbol{x}_0\rangle^2 + o_{\mathsf{P}}(1)\,,
\end{aligned}$$

where (a) follows by replacing $\sigma'(\langle\boldsymbol{w}_i, \boldsymbol{x}\rangle)$ by its expectation over $\boldsymbol{x}$ $\mathsf{E}\sigma'(\langle\boldsymbol{x}_i, \boldsymbol{x}\rangle) = \mathsf{E}\sigma'(G) + o(1) = \mu_1 + o(1)$, and (b) since $\mathsf{E}(\boldsymbol{x}) = 0$.

The choice $\boldsymbol{x}_0$ that maximizes the risk $R(\boldsymbol{W}_0, \mathsf{P}_{\boldsymbol{W}_0,\Delta})$ is $\boldsymbol{x}_0 = \Delta(\boldsymbol{\beta}_0 - \mu_1\boldsymbol{W}_0^{\mathsf{T}}\hat{\boldsymbol{\theta}}(\boldsymbol{W}_0))/\|\boldsymbol{\beta}_0 - \mu_1\boldsymbol{W}_0^{\mathsf{T}}\hat{\boldsymbol{\theta}}(\boldsymbol{W}_0)\|_2$. However this mean shift can have a significant component along $\boldsymbol{\beta}_0$, which results in a large increase of $R(\boldsymbol{W}; \mathsf{P}_{\boldsymbol{W}_0,\Delta})$ for other $\boldsymbol{W}$ as well. To avoid this, we project this vector orthogonally to $\boldsymbol{\beta}_0$:

$$\boldsymbol{x}_0 = -\Delta\frac{\mathsf{P}_{\boldsymbol{\beta}_0}^{\perp}\boldsymbol{W}_0^{\mathsf{T}}\hat{\boldsymbol{\theta}}(\boldsymbol{W}_0)}{\|\mathsf{P}_{\boldsymbol{\beta}_0}^{\perp}\boldsymbol{W}_0^{\mathsf{T}}\hat{\boldsymbol{\theta}}(\boldsymbol{W}_0)\|_2}\,. \tag{35}$$

This results in the following expression for the test error on the shifted distribution:

$$R(\boldsymbol{W}, \mathsf{P}_{\boldsymbol{W}_0,\Delta}) = R(\boldsymbol{W}, \mathsf{P}) + \Delta^2\mu_1^2 T(\boldsymbol{W}, \boldsymbol{W}_0) + o_{\mathsf{P}}(1)\,, \tag{36}$$

$$T(\boldsymbol{W}, \boldsymbol{W}_0) := \frac{\langle\mathsf{P}_{\boldsymbol{\beta}_0}^{\perp}\boldsymbol{W}^{\mathsf{T}}\hat{\boldsymbol{\theta}}(\boldsymbol{W}), \mathsf{P}_{\boldsymbol{\beta}_0}^{\perp}\boldsymbol{W}_0^{\mathsf{T}}\hat{\boldsymbol{\theta}}(\boldsymbol{W}_0)\rangle}{\|\mathsf{P}_{\boldsymbol{\beta}_0}^{\perp}\boldsymbol{W}_0^{\mathsf{T}}\hat{\boldsymbol{\theta}}(\boldsymbol{W}_0)\|_2^2}\,. \tag{37}$$

We first consider the case $\boldsymbol{W} = \boldsymbol{W}_0$. We then have

$$\mathsf{E}T(\boldsymbol{W}_0, \boldsymbol{W}_0) := \mathsf{E}\{\|\mathsf{P}_{\boldsymbol{\beta}_0}^{\perp}\boldsymbol{W}_0^{\mathsf{T}}\hat{\boldsymbol{\theta}}(\boldsymbol{W}_0)\|_2^2\} \tag{38}$$

$$= \mathsf{E}\{\|\boldsymbol{W}_0^{\mathsf{T}}\hat{\boldsymbol{\theta}}(\boldsymbol{W}_0)\|_2^2\} - \frac{1}{r^2}\mathsf{E}\{\langle\boldsymbol{\beta}_0, \boldsymbol{W}_0^{\mathsf{T}}\hat{\boldsymbol{\theta}}(\boldsymbol{W}_0)\rangle^2\} \tag{39}$$

$$\overset{(a)}{=} r^2\left\{\frac{\zeta^2\mathscr{D}_1}{(\chi\zeta^2-1)\mathscr{D}_0} - \mathscr{H}^2\right\} + s^2\frac{\zeta^2\mathscr{D}_2}{\mathscr{D}_0} + o(1)\,, \tag{40}$$

where (a) follows by Eqs. (19) and (20).

For $\boldsymbol{W}$ independent of $\boldsymbol{W}_0$, we have (for $s = 0$)

$$\begin{aligned}
\mathsf{E}T(\boldsymbol{W}, \boldsymbol{W}_0) &:= \mathsf{E}\left\{ \frac{(\langle \boldsymbol{W}^{\mathsf{T}}\hat{\boldsymbol{\theta}}(\boldsymbol{W}), \boldsymbol{W}_0^{\mathsf{T}}\hat{\boldsymbol{\theta}}(\boldsymbol{W}_0)\rangle - r^{-2}\langle \boldsymbol{\beta}_0, \boldsymbol{W}_0^{\mathsf{T}}\hat{\boldsymbol{\theta}}(\boldsymbol{W}_0)\rangle\langle \boldsymbol{\beta}_0, \boldsymbol{W}_1^{\mathsf{T}}\hat{\boldsymbol{\theta}}(\boldsymbol{W}_1)\rangle)^2}{\|\boldsymbol{W}_0^{\mathsf{T}}\hat{\boldsymbol{\theta}}(\boldsymbol{W}_0)\|_2^2 - r^{-2}\langle \boldsymbol{\beta}_0, \boldsymbol{W}_0^{\mathsf{T}}\hat{\boldsymbol{\theta}}(\boldsymbol{W}_0)\rangle^2} \right\} \\
&= \frac{\left(\mathsf{E}_{\boldsymbol{X},\boldsymbol{y}}\{\|\mathsf{E}_{\boldsymbol{W}}[\boldsymbol{W}^{\mathsf{T}}\hat{\boldsymbol{\theta}}(\boldsymbol{W})|\boldsymbol{X}, \boldsymbol{y}]\|^2\} - r^{-2}(\mathsf{E}_{\boldsymbol{W},\boldsymbol{X},\boldsymbol{y}}\langle \boldsymbol{\beta}_0, \boldsymbol{W}^{\mathsf{T}}\hat{\boldsymbol{\theta}}(\boldsymbol{W})\rangle)^2\right)^2}{\mathsf{E}_{\boldsymbol{W},\boldsymbol{X},\boldsymbol{y}}\{\|\boldsymbol{W}^{\mathsf{T}}\hat{\boldsymbol{\theta}}(\boldsymbol{W})\|_2^2\} - r^{-2}(\mathsf{E}_{\boldsymbol{W},\boldsymbol{X},\boldsymbol{y}}\langle \boldsymbol{\beta}_0, \boldsymbol{W}^{\mathsf{T}}\hat{\boldsymbol{\theta}}(\boldsymbol{W})\rangle)^2} + o_{\mathsf{P}}(1) \\
&= \frac{\mathscr{T}_1(\zeta, \psi_1, \psi_2, \overline{\lambda})^2}{\mathscr{T}_0(\zeta, \psi_1, \psi_2, \overline{\lambda})} + o_{\mathsf{P}}(1)\,, \qquad (41)
\end{aligned}$$

where

$$\mathscr{T}_0(\zeta, \psi_1, \psi_2, \overline{\lambda}) = \frac{\zeta^2 \mathscr{D}_1(\zeta, \psi_1, \psi_2, \overline{\lambda})}{(\chi\zeta^2 - 1)\mathscr{D}_0(\zeta, \psi_1, \psi_2, \overline{\lambda})} - \mathscr{H}^2(\zeta, \psi_1, \psi_2, \overline{\lambda})\,, \qquad (42)$$

$$\lim_{\overline{\lambda}\to 0} \mathscr{T}_1(\zeta, \psi_1, \psi_2, \overline{\lambda}) = \mathscr{L}(\zeta, \psi_1, \psi_2) - \mathscr{H}^2(\zeta, \psi_1, \psi_2, 0)\,. \qquad (43)$$

In Figure 3 of the main text, we plot the ratios

$$\frac{\mathsf{E}_{\boldsymbol{W}_0} R(\boldsymbol{W}_0; \mathsf{P}_{\boldsymbol{W}_0,\Delta})}{\mathsf{E}_{\boldsymbol{W}} R(\boldsymbol{W}; \mathsf{P})}\,, \qquad \frac{\mathsf{E}_{\boldsymbol{W},\boldsymbol{W}_0} R(\boldsymbol{W}; \mathsf{P}_{\boldsymbol{W}_0,\Delta})}{\mathsf{E}_{\boldsymbol{W}} R(\boldsymbol{W}; \mathsf{P})}\,. \qquad (44)$$

Note that the perturbation introduced here is extremely small in ℓ_∞ norm. Namely $\|\boldsymbol{x}_0\|_\infty \approx \Delta\sqrt{(2\log d)/d}$: as d gets larger, this is much smaller than the typical entry of $\boldsymbol{x}$, which is of order 1.

E.6 Random features model: Derivation Eq. (29)

In this section we outline the derivation of Eq. (29). Let $\boldsymbol{U}_\sigma = \sigma(\boldsymbol{X}\boldsymbol{W}^{\mathsf{T}}) \in \mathbb{R}^{n\times N}$ (where σ is applied entrywise to the matrix $\boldsymbol{X}\boldsymbol{W}^{\mathsf{T}}$). Then

$$\begin{aligned}
\mathsf{E}\{\boldsymbol{W}^{\mathsf{T}}\hat{\boldsymbol{\theta}}(\boldsymbol{W})|\boldsymbol{X}, \boldsymbol{y}\} &= \mathsf{E}\{\boldsymbol{W}^{\mathsf{T}}(\boldsymbol{U}_\sigma^{\mathsf{T}}\boldsymbol{U}_\sigma + du^2\mathbf{I})^{-1}\boldsymbol{U}_\sigma^{\mathsf{T}}\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{y}\} \qquad (45) \\
&= \mathsf{E}_{\boldsymbol{W}}\{\boldsymbol{W}^{\mathsf{T}}(\boldsymbol{U}_\sigma^{\mathsf{T}}\boldsymbol{U}_\sigma + du^2\mathbf{I})^{-1}\boldsymbol{U}_\sigma^{\mathsf{T}}\boldsymbol{X}\boldsymbol{\beta}_0\}\,. \qquad (46)
\end{aligned}$$

where $u^2 = \lambda\psi_1\psi_2$. We will work within the 'noisy linear features model' of Mei and Montanari (2019) which replaces $\boldsymbol{U}_\sigma$ with

$$\boldsymbol{U} = \mu_1\boldsymbol{X}\boldsymbol{W}^{\mathsf{T}} + \mu_*\boldsymbol{Z}\,, \qquad (47)$$

where $(Z_{ij})_{i\le n, j\le N} \sim_{iid} \mathsf{N}(0,1)$, $(X_{ij})_{i\le n, j\le d} \sim_{iid} \mathsf{N}(0,1)$, $(\boldsymbol{W}_{ij})_{i\le N, j\le d} \sim_{iid} \mathsf{N}(0, 1/d)$. The universality results of Mei and Montanari (2019) suggest that this substitution produces an error that is asymptotically negligible, namely:

$$\begin{aligned}
\mathsf{E}\{\boldsymbol{W}^{\mathsf{T}}\hat{\boldsymbol{\theta}}(\boldsymbol{W})|\boldsymbol{X}, \boldsymbol{y}\} &= \tilde{\boldsymbol{Q}}(\boldsymbol{X})\boldsymbol{\beta}_0 + \mathbf{err}_0(d, \lambda)\,, \qquad (48) \\
\tilde{\boldsymbol{Q}}(\boldsymbol{X}) &:= \mathsf{E}_{\boldsymbol{W},\boldsymbol{Z}}\big\{\boldsymbol{W}^{\mathsf{T}}(\boldsymbol{U}^{\mathsf{T}}\boldsymbol{U}_\sigma + du^2\mathbf{I}_N)^{-1}\boldsymbol{U}^{\mathsf{T}}\boldsymbol{X}\big\}\,. \qquad (49)
\end{aligned}$$

Note that, conditional on $\boldsymbol{X}$ $\boldsymbol{U}, \boldsymbol{W}$ are jointly Gaussian, and therefore it is easy to compute the conditional expectation

$$\mathsf{E}\{\boldsymbol{W}^{\mathsf{T}}|\boldsymbol{U}, \boldsymbol{X}\} = \mu_1\Big(\mu_*^2 d\mathbf{I}_d + \mu_1^2\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}\Big)^{-1}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{U}\,. \qquad (50)$$

We therefore get

$$\begin{aligned}
\tilde{\boldsymbol{Q}}(\boldsymbol{X}) &= \mu_1\Big(\mu_*^2 d\mathbf{I}_d + \mu_1^2\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}\Big)^{-1}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{Q}(\boldsymbol{X})\boldsymbol{X}\,, \qquad (51) \\
\boldsymbol{Q}(\boldsymbol{X}) &:= \mathsf{E}_{\boldsymbol{W},\boldsymbol{Z}}\{\boldsymbol{U}(\boldsymbol{U}^{\mathsf{T}}\boldsymbol{U} + du^2\mathbf{I}_N)^{-1}\boldsymbol{U}^{\mathsf{T}}\}\,. \qquad (52)
\end{aligned}$$

At this point we claim that, for $\psi_1 \neq \psi_2$,

$$\boldsymbol{Q}(\boldsymbol{X}) = \Big(\frac{\mu_1^2}{d}\boldsymbol{X}\boldsymbol{X}^{\mathsf{T}} + \mu_*^2(1+q)\mathbf{I}_d\Big)\Big(\frac{\mu_1^2}{d}\boldsymbol{X}\boldsymbol{X}^{\mathsf{T}} + \mu_*^2(1+q)\mathbf{I}_n\Big)^{-1} + \mathbf{Err}(\lambda, d)\,, \tag{53}$$

where the $\mathbf{Err}(\lambda, d)$ is negligible for our purposes (namely $\|\mu_1\Big(\mu_*^2 d\mathbf{I}_d + \mu_1^2\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}\Big)^{-1}\boldsymbol{X}^{\mathsf{T}}\mathbf{Err}\boldsymbol{X}\boldsymbol{\beta}_0\|_2 = o_{\mathsf{P}}(1)$ when $\lambda \to 0$ after $d \to \infty$). The main claim of this section —Eq. (29)— follows by substituting (53) in Eqs. (48), (51) after some algebraic manipulations.

In the overparametrized regime $\psi_1 > \psi_2$ (which is the main focus of our analysis) Eq. (53) is straightforward. Notice that in this case $q = 0$, and therefore this claim amounts to $\boldsymbol{Q}(\boldsymbol{X}) = \mathbf{I}_n + \mathbf{Err}(\lambda, d)$. Indeed this is straightforward since in that case $\boldsymbol{U}$ has full column rank, and minimum singular value bounded away from zero. Therefore

$$\big\|\boldsymbol{Q}(\boldsymbol{X}) - \mathsf{E}_{\boldsymbol{W},\boldsymbol{Z}}\{\boldsymbol{U}(\boldsymbol{U}^{\mathsf{T}}\boldsymbol{U})^{-1}\boldsymbol{U}^{\mathsf{T}}\}\big\|_{\text{op}} \leq Cu^2 = C'\lambda\,, \tag{54}$$

and $\boldsymbol{U}(\boldsymbol{U}^{\mathsf{T}}\boldsymbol{U})^{-1}\boldsymbol{U}^{\mathsf{T}} = \mathbf{I}_n$.

In the underparametrized regime $\psi_1 < \psi_2$ the result can be obtained making use of Ledoit and Péché (2011).