Preventing dataset shift from breaking machine-learning biomarkers

Jérôme Dockès, Gaël Varoquaux, Jean-Baptiste Poline

▶ **To cite this version:**

Jérôme Dockès, Gaël Varoquaux, Jean-Baptiste Poline. Preventing dataset shift from breaking machine-learning biomarkers. GigaScience, inPress, 10.1093/gigascience/giab055 . hal-03293375

HAL Id: hal-03293375
https://hal.science/hal-03293375

Submitted on 20 Jul 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Preventing dataset shift from breaking machine-learning biomarkers

Jérôme Dockès[1*], Gaël Varoquaux[12+], Jean-Baptiste Poline[1+]

[1]McGill University [2]INRIA
[*]Corresponding author [+]JB Poline and Gaël Varoquaux contributed equally to this work.

July 2021

Abstract

Machine learning brings the hope of finding new biomarkers extracted from cohorts with rich biomedical measurements. A good biomarker is one that gives reliable detection of the corresponding condition. However, biomarkers are often extracted from a cohort that differs from the target population. Such a mismatch, known as a dataset shift, can undermine the application of the biomarker to new individuals. Dataset shifts are frequent in biomedical research, e.g. because of recruitment biases. When a dataset shift occurs, standard machine-learning techniques do not suffice to extract and validate biomarkers. This article provides an overview of when and how dataset shifts breaks machine-learning extracted biomarkers, as well as detection and correction strategies.

1 Introduction: dataset shift breaks learned biomarkers

Biomarkers are measurements that provide information about a medical condition or physiological state [Strimbu and Tavel, 2010]. For example, the presence of an antibody may indicate an infection; a complex combination of features extracted from a medical image can help assess the evolution of a tumor. Biomarkers are important for diagnosis, prognosis, and treatment or risk assessments.

Complex biomedical measures may carry precious medical information, as with histopathological images or genome sequencing of biopsy samples in oncology. Identifying quantitative biomarkers from these requires sophisticated statistical analysis. With large datasets becoming accessible, supervised machine learning provides new promises by optimizing the information extracted to relate to a specific output variable of interest, such as a cancer diagnosis [Andreu-Perez et al., 2015, Faust et al., 2018, Deo, 2015]. These methods, cornerstones of artificial intelligence, are starting to appear in clinical practice: a machine-learning based radiological tool for breast-cancer diagnosis has recently been approved by the FDA[1].

Can such predictive biomarkers, extracted through complex data processing, be safely used in clinical practice, beyond the initial research settings? One risk is the potential mismatch, or *dataset shift*, between the distribution of the individuals used to estimate this statistical link and that of the target population that should benefit from the biomarker. In this case, the extracted associations may not apply to the target population [Kakarmath et al., 2020]. Computer aided diagnostic of thoracic diseases from X-ray images has indeed been shown to be unreliable for individuals of a given sex if built from a cohort over-representing the other sex [Larrazabal et al., 2020]. More generally, machine-learning systems may fail on data from different imaging devices, hospitals, populations with a different age distribution, *etc.*. Dataset biases are in fact frequent in medicine. For instance selection biases –*eg* due to volunteering self-selection, non-response, dropout...– [Rothman, 2012, Tripepi et al., 2010] may cause cohorts to capture only a small range of possible patients and disease manifestations in the presence of spectrum effects [Ransohoff and Feinstein, 1978, Mulherin and Miller, 2002]. Dataset shift or dataset bias can cause systematic errors that cannot

[1]`https://fda.report/PMN/K192854`

be fixed by acquiring larger datasets and require specific methodological care.

In this article, we consider predictive biomarkers identified with supervised machine learning. We characterize the problem of dataset shift, show how it can hinder the use of machine learning for health applications [Woo et al., 2017, Wynants et al., 2020], and provide mitigation strategies.

2 A primer on machine learning for biomarkers

2.1 Empirical Risk Minimization

Let us first introduce the principles of machine learning used to identify biomarkers. Supervised learning captures, from observed data, the link between a set of input measures (features) X and an output (e.g. a condition) Y: for example the relation between the absorption spectrum of oral mucosa and blood glucose concentration [Kasahara et al., 2018]. A supervised learning algorithm finds a function f such that $f(X)$ is as close as possible to the output Y. Following machine-learning terminology, we call the system's best guess $f(X)$ for a value X a *prediction*, even when it does not concern a measurement in the future.

Empirical Risk Minimization, central to machine learning, uses a loss function L to measure how far a prediction $f(X)$ is from the true value Y, for example the squared difference:

$$L(Y, f(X)) = (Y - f(X))^2 \, . \tag{1}$$

The goal is to find a function f that has a small *risk*, which is the *expected* loss on the true distribution of X and Y, i.e. on *unseen individuals*. The true risk cannot be computed in practice: it would require having seen all possible patients, the true distribution of patients. The *empirical* risk is used instead: the average error over available examples,

$$\hat{R}(f) = \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)) \, , \tag{2}$$

where $\{(x_i, y_i)\, , \, i = 1, \ldots, n\}$ are available (X, Y) data, called *training* examples. The statistical link of interest is then approximated by choosing f within a family of candidate functions as the one that minimizes the empirical risk $\hat{R}(f)$.

The crucial assumption underlying this very popular approach is that the prediction function f will then be applied to individuals drawn from the same population as the training examples $\{x_i, y_i\}$. It can be important to distinguish the *source* data, used to fit and evaluate a machine-learning model (e.g. a dataset collected for research), from the *target* data, on which predictions are meant to be used for clinical applications (e.g. new visitors of a hospital). Indeed, if the training examples are not representative of the target population – if there is a dataset shift – the empirical risk is a poor estimate of the expected error, and f will not perform well on individuals from the target population.

2.2 Evaluation: Independent test set and cross-validation

Once a model has been estimated from training examples, measuring its error on these same individuals results in a (sometimes wildly) optimistic estimate of the expected error on *unseen* individuals (Friedman et al. [2001, Sec. 7.4], Poldrack et al. [2020, Sec. 1, "Association vs Prediction"]). Indeed, predictors chosen from a rich family of functions are very flexible and can learn rules that fit tightly the training examples but fail to generalize to new individuals. This is called *overfitting*.

To obtain valid estimates of the expected performance on new data, the error is measured on an independent sample held out during training, called the test set. The most common approach to obtain such a test set is to randomly split the available data. This process is usually repeated with several splits, a procedure called cross-validation [Arlot et al., 2010, Friedman et al., 2001, Sec. 7].

When training and test examples are chosen uniformly from the same sample, they are drawn from the same distribution (i.e. the same population): there is no dataset shift. Some studies also measure the error on an *independent* dataset [e.g. Beck et al., 2011, Jin et al., 2020]. This helps establishing external validity, assessing whether the predictor will perform well outside of the dataset used to define it [Bleeker et al., 2003]. Unfortunately, the biases in participant recruitment may be similar in independently collected datasets. For example if patients with severe symptoms are difficult to recruit, this is likely to distort all datasets similarly. Testing on a dataset collected independently is therefore a useful check, but no silver bullet to rule out dataset shift issues.

3 False solutions to tackling dataset shift

We now discuss some misconceptions and confusions with problems not directly related to dataset shift.

"Deconfounding" does not correct dataset shift for predictive models Dataset shift is sometimes confused with the notion of *confounding*, as both settings arise from an undesired effect in the data. Confounding comes from *causal analysis*, estimating the effect of a *treatment* –an intervention, sometimes fictional– on an outcome. A confounder is a third variable –for example age, or a comorbidity– that influences both the treatment and the outcome. It can produce a non-causal association between the two [See Hernán and Robins, 2020, Chap. 7, for a precise definition]. However, the machine-learning methods we consider here capture statistical associations, but *do not target causal effects*. Indeed, for biomarkers, the association itself is interesting, whether causal or not. Elevated body temperature may be the consequence of a condition, but also cause a disorder. It is a clinically useful measure in both settings.

Tools for causal analysis are not all useful for prediction, as pointed out by seminal textbooks: "if the goal of the data analysis is purely predictive, no adjustment for confounding is necessary [...] the concept of confounding does not even apply."[Hernán and Robins, 2020, Sec. 18.1], or Pearl [2019]. In prediction settings, applying procedures meant to adjust for confounding generally degrades prediction performance without solving the dataset shift issue. Figure 1 demonstrates the detrimental effect of "deconfounding" on simulated data: while the target population is shifted due to a different age distribution, removing the effect of age also removes the separation between the two outcomes of interest. The same behavior is visible on real epidemiologic data with age shifts, such as predicting the smoking status of participants in the UKBiobank study [Sudlow et al., 2015], as shown in Figure 2. Drawing training and testing samples with different age distributions highlights the effect of these age shifts on prediction performance (see Appendix B for details on the procedure). For a given learner and test population, training on a different population degrades prediction. For example, predictions on the old population are degraded when the model is trained on the young population. A flexible model (Gradient Boosting) outperforms the linear model with or without dataset shift. "Regressing out" the age (as in the second column of Figure 1, "+ regress-out" strategy in Figure 2) degrades the predictions in *all* configurations.

For both illustrations on simulated and real data (Figure 1 and 2), we also demonstrate an approach suitable for predictive models: reweighting training examples giving more importance to those more likely in the test population. This approach improves the predictions of the overconstrained (misspecified) linear model in the presence of dataset shift, but degrades the predictions of the powerful learner. The non-linear model already captures the correct separation for both young and old individuals, thus reweighting examples does not bring any benefit but only increases the variance of the empirical risk. A more detailed discussion of this approach, called *importance weighting*, is provided in Section 5.

Training examples should not be selected to be homogeneous To obtain valid predictive models that perform well beyond the training sample, it is crucial to collect datasets that represent the whole population and reflect its diversity as much as possible [Kakarmath et al., 2020, England and Cheng, 2019, O'neil, 2016]. Yet clinical research often emphasizes the opposite: very homogeneous datasets and carefully selected participants. While this may help reduce variance and improve statistical testing, it degrades prediction performance and fairness. In other words, the machine-learning system may perform worse for segments of the population that are under-represented in the dataset, resulting in uneven quality of care if it is deployed in clinical settings. Therefore in *predictive* settings, where the goal is machine-learning models that generalize well, large and diverse datasets are desirable.

Simpler models are not less sensitive to dataset shift Often, flexible models can be more robust to dataset shifts, and thus generalize better, than linear models [Storkey, 2009], as seen in Figures 1 and 2. Indeed, an over-constrained (ill-specified) model may only fit well a restricted region of the feature space, and its performance can degrade if the distribution of inputs changes, even if the relation to the output stays the same (i.e. when covariate shift occurs, Section 7.1).

Dataset shift does not call for simpler models as it is not a small-sample issue. Collecting more data from the same

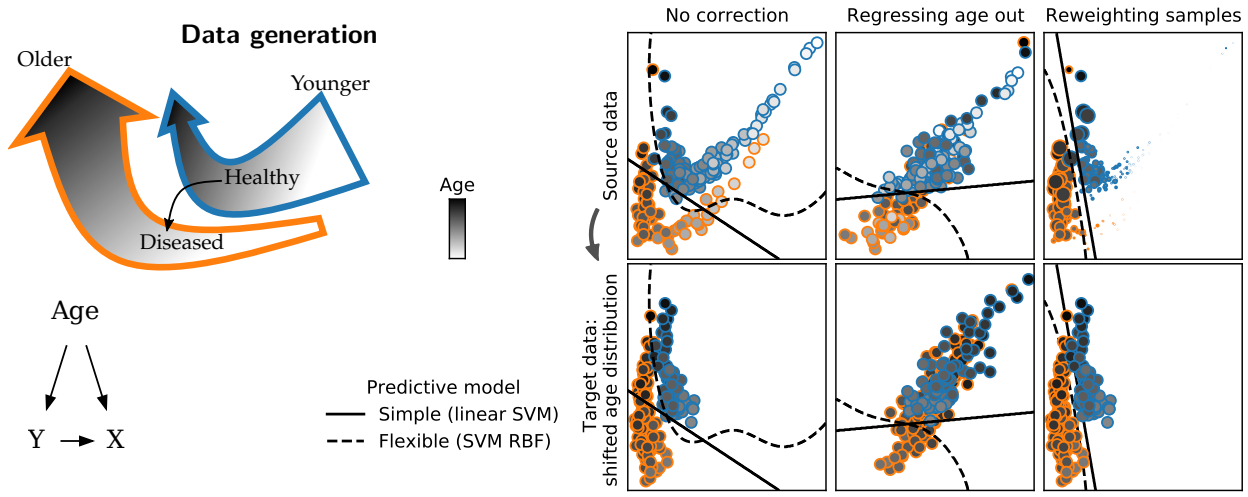

Figure 1. Classification with dataset shift – regressing out a correlate of the shift does not help generalization. The task is to classify patients (orange) from healthy controls (blue), using 2-dimensional features. Age, indicated by the shade of gray, influences both the features and the probability of disease. **Left: generative process for the simulated data.** Age influences both the target Y and the features X, and Y also has an effect on X. Between the source and target datasets, the distribution of age changes. The two arrows point towards increasing age and represent the Healthy and Diseased populations, corresponding to the orange and blue clouds of points in the right panel. The grayscale gradient in the arrows represents the increasing age of the individuals (older individuals correspond to a darker shade). Throughout their life, individuals can jump from the Healthy trajectory to the Diseased trajectory, which is slightly offset in this 2-dimensional feature space. As age increases, the prevalence of the disease increases, hence the Healthy trajectory contains more individuals of young ages (its wide end), and less at older ages (its narrow end) – and vice-versa for the Diseased trajectory. **Right: predictive models** In the target data (bottom row), the age distribution is shifted: individuals tend to be older. Elderly are indeed often less likely to participate in clinical studies [Heiat et al., 2002]. **First column:** no correction is applied. As the situation is close to a covariate shift (Section 7.1), a powerful learner (RBF-SVM) generalizes well to the target data. An over-constrained model – Linear-SVM – generalizes poorly. **Second column:** wrong approach. To remove associations with age, features are replaced by the residuals after regressing them on age. This destroys the signal and results in poor performance for both models and datasets. **Third column:** Samples are weighted to give more importance to those more likely in the target distribution. Small circles indicate younger individuals, with less influence on the classifier estimation. This reweighting improves prediction for the linear model on the older population.

sources will not correct systematic dataset bias.

4 Preferential sample selection: a common source of shift

In 2017, competitors in the million-dollar-prize data science bowl used machine learning to predict if individuals would be diagnosed with lung cancer within one year, based on a CT scan. Assuming that the winning model achieves satisfying accuracy on left-out examples from this dataset, is it ready to be deployed in hospitals? Most likely not. Selection criteria may make this dataset not representative of the potential lung cancer patients general population. Selected participants verified many criteria, including being a smoker and not having recent medical problems such as pneumonia. How would the winning predictor perform on a more diverse population? For example, another disease could present features that the classifier could mistakenly take for signs of lung cancer. Beyond explicit selection criteria, many factors such as age, ethnicity, or socioeconomic status influence participation in biomedical studies [Henrich et al., 2010, Murthy et al., 2004, Heiat et al., 2002, Chastain et al., 2020]. Not only can these shifts reduce overall predictive performance, they can also lead to discriminative clinical decisions for poorly represented populations

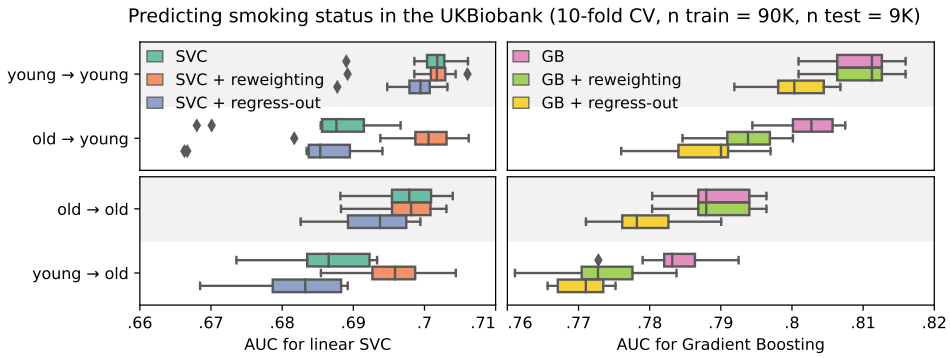

Figure 2. Predicting the smoking status of UKBiobank participants. Different predictive models are trained on 90K UKBiobank participants and tested on 9K participants with a possibly shifted age distribution. "young → old" means the training set was drawn from a younger sample than the testing set. Models perform better when trained on a sample drawn from the same population as the testing set. Reweighting examples that are more likely in the test distribution ("+ reweighting" strategy, known as Importance Weighting, Section 5) alleviates the issue for the simple linear model, but is detrimental for the Gradient Boosting. Regressing out the age ("+ regress-out" strategy) is a bad idea and degrades prediction performance in all configurations.

[Oakden-Rayner et al., 2020, Gianfrancesco et al., 2018, Barocas et al., 2019, Abbasi-Sureshjani et al., 2020, Cirillo et al., 2020].

The examples above are instances of preferential selection, which happens when members of the population of interest do not have equal probabilities of being included in the source dataset: the selection S is not independent of (X, Y). Preferential sample selection is ubiquitous and cannot always be prevented by careful study design [Bareinboim and Pearl, 2012]. It is therefore a major challenge to the identification of reliable and fair biomarkers. Beyond preferential sample selection, there are many other sources of dataset shifts, e.g. population changes over time, interventions such as the introduction of new diagnostic codes in Electronic Health Records [Sáez et al., 2020], and the use of different acquisition devices.

4.1 The selection mechanism influences the type of dataset shift

The correction for a dataset shift depends on the nature of this shift, characterized by which and how distributions are modified [Storkey, 2009]. Knowledge of the mechanism producing the dataset shift helps formulate hypotheses about distributions that remain unchanged in the target data [Schölkopf et al., 2012, Peters et al., 2017, Chap. 5].

Figure 3 illustrates this process with a simulated example of preferential sample selection. We consider the problem of predicting the volume Y of a tumor from features X extracted from contrast CT images. These features can be influenced not only by the tumor size, but also by the dosage of a contrast agent M. The first panel of Figure 3 shows a selection of data independent of the image and tumor volume: there is no dataset shift. In the second panel, selection depends on the CT image itself (for example images with a low signal-to-noise ratio are discarded). As selection is independent of the tumor volume Y given the image X, the distribution of images changes but the conditional distribution $P(Y \mid X)$ stays the same: we face a *covariate shift* (Section 7.1). The learned association remains valid. Moreover, reweighting examples to give more importance to those less likely to be selected can improve predictions for target data (Section 5), and it can be done with only *unlabelled* examples from the target data. In the third panel, individuals who received a low contrast agent dose are less likely to enter the training dataset. Selection is therefore not independent of tumor volume (the output) given the image values (the input features). Therefore we have sample selection bias: the relation $P(Y \mid X)$ is different in source and target data, which will affect the performance of the prediction.

As these examples illustrate, the causal structure of the data helps identify the type of dataset shift and what information is needed to correct it. When such information

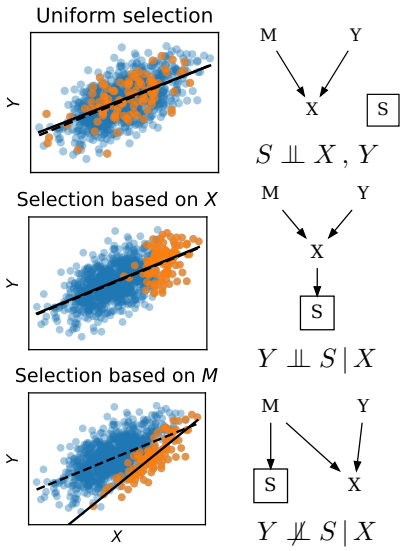

Figure 3. Sample selection bias: three examples. On the right are graphs giving conditional independence relations [Pearl et al., 2016]. Y is the lesion volume to be predicted (i.e. the output). M are the imaging parameters, e.g. contrast agent dosage. X is the image, and depends both on Y and M (in this toy example X is computed as $X \coloneqq Y+M+\epsilon$, where ϵ is additive noise. S indicates that data is selected to enter the source dataset (orange points) or not (blue points). The symbol $\perp\!\!\!\perp$ means independence between variables. Preferentially selecting samples results in a dataset shift (middle and bottom row). Depending on whether $Y \perp\!\!\!\perp S \mid X$, the conditional distribution of $Y \mid X$ – here lesion volume given the image – estimated on the selected data may be biased or not.

is available, it may be possible to leverage it in order to improve robustness to dataset shift [e.g. Subbaswamy et al., 2019].

5 Importance weighting: a generic tool against dataset shift

Importance weighting is a simple approach to dataset shift that applies to many situations and can be easy to implement.

Dataset shift occurs when the joint distribution of the features and outputs is different in the source (data used to fit the machine-learning model) and in the target data. Informally, importance weighting consists in *reweighting* the available data to create a pseudo-sample that follows the same distribution as the target population.

To do so, examples are reweighted by their *importance weights* – the ratio of their likelihood in target data over source data. Examples that are rare in the source data but are likely in the target data are more relevant and therefore receive higher weights. A related approach is importance *sampling* – resampling the training data according to the importance weights. Many statistical learning algorithms – including Support Vector Machines, decision trees, random forests, neural networks – naturally support weighting the training examples. Therefore, the challenge relies mostly in the estimation of the appropriate sample weights and the learning algorithm itself does not need to be modified.

To successfully use importance weighting, no part of the target distribution should be completely unseen. For example, if sex (among other features) is used to predict heart failure and the dataset only includes men, importance weighting cannot transform this dataset and make its sex distribution similar to that of the general population (Figure 4). Conversely, the source distribution may be broader than the target distribution (as seen for example in Figure 1).

Importance weights can also be applied to validation examples, which may produce a more accurate estimation of generalization error on target data.

Importance weighting is a well-known approach and an important body of literature focuses on its application and the estimation of importance weights. It is illustrated on small datasets for the prediction of breast cancer in Dudík et al. [2006] and heart disease in Kouw and Loog [2019].

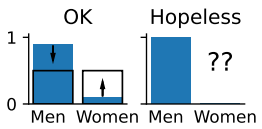

Figure 4. Dataset shifts that may or may not be compensated by reweighting – Left: distribution of sex can be balanced by downweighting men and upweighting women. **Right:** women are completely missing; the dataset shift cannot be fixed by importance weighting.

However, it cannot always be applied: some knowledge of the target distribution is required, and the source distribution must cover its support. Moreover, importance weighting can increase the variance of the empirical risk estimate, and thus sometimes *degrades* performance – as seen in Figure 2. It is therefore a straightforward and popular approach to consider, but not a complete solution. It is particularly beneficial when using a simple learning model which cannot capture the full complexity of the data, such as the linear models in Figure 1. Indeed, simple models are often prefered in biomedical applications because they are easy to interpret and audit.

In Appendix A, we provide a more precise definition of the importance weights, as well as an overview of how they can be estimated and used.

6 Other approaches to dataset shift

Beyond importance weighting, many other solutions to dataset shift have been proposed. They are typically more difficult to implement, as they require adapting or desiging new learning algorithms. However, they may be more effective, or applicable when information about the target distribution is lacking. We summarize a few of these approaches here. A more systematic review can be found in Kouw and Loog [2019]. Weiss et al. [2016] and Pan and Yang [2009] give systematic reviews of transfer learning (a wider family of learning problems which includes dataset shift).

The most obvious solution is to do nothing – ignoring the dataset shift. This approach should be included as a baseline when testing on a sample of target data – which is a prerequisite to clinical use of a biomarker [Storkey, 2009, Woo et al., 2017]. With flexible models, this is a strong baseline that can outperform importance weighting, as in the right panel of Figure 2.

Another approach is to learn representations– transformations of the signal— that are invariant to the shift [Achille and Soatto, 2018]. Some deep-learning methods strive to extract features that are predictive of the target while having similar distributions in the source and target domains [e.g. Long et al., 2015], or while preventing an adversary to distinguish source and target data ["domain-adversarial" learning, e.g. Tzeng et al., 2017]. When considering such methods, one must be aware of the fallacy shown in Figure 1: making the features invariant to the effect driving the dataset shift can remove valuable signal if this effect is not independent of the outcome of interest.

It may also be possible to explicitly model the mapping from source to target domains, e.g. by training a neural network to translate images from one modality or imaging device to another, or by relying on optimal transport [Courty et al., 2016].

Finally, synthetic data augmentation sometimes helps – relying on known invariances e.g. for images by applying affine transformations, resampling, *etc.*. or with learned generative models [e.g. Antoniou et al., 2017].

6.1 Performance heterogeneity and fairness

It can be useful not to target a specific population, but rather find a predictor robust to certain dataset shifts. Distributionally robust optimization tackles this goal by defining an ambiguity, or uncertainty set – a set of distributions to which the target distribution might belong – then minimizing the worse risk across all distributions in this set [see Rahimian and Mehrotra, 2019, for a review]. The uncertainty set is often chosen centered on the empirical (source) distribution for some divergence between distributions. Popular choices for this divergence are the Wasserstein distance, f-divergences (e.g. the KL divergence) [Duchi and Namkoong, 2018], and the Maximum Mean Discrepancy [Zhu et al., 2020]. If information about the target distribution is available, it can be incorporated in the definition of the uncertainty set. An approach related to robust optimization is to strive not only to minimize the empirical loss $L(Y, f(X))$ but also its variance Maurer and Pontil [2009], Namkoong and Duchi [2017].

It is also useful to assess model performance across values

of demographic variables such as age, socioeconomic status or ethnicity. Indeed, a good overall prediction performance can be achieved despite a poor performance on a minority group. Ensuring that a predictor performs well for all subpopulations reduces sensitivity to potential shifts in demographics and is essential to ensure fairness [Abbasi-Sureshjani et al., 2020]. For instance, there is a risk that machine-learning analysis of dermoscopic images under-diagnoses malignant moles on skin tones that are typically under-represented in the training set Adamson and Smith [2018]. Fairness is especially relevant when the model output could be used to grant access to some treatment. As similar issues arise in many applications of machine learning, there is a growing literature on fairness [see e.g. Barocas et al., 2019, for an overview]. For instance, Duchi and Namkoong [2018] show that distributionally robust optimization can help performance on under-represented subpopulations.

6.2 Multi-site datasets

Often datasets are collected across several sites or hospitals, or with different measurement devices. This heterogeneity provides an opportunity to train models that generalize to unseen sites or devices. Some studies attempt to remove site effects by regressing all features on the site indicator variable. For the same reasons that regressing out age is detrimental in Figure 1, this strategy often gives worse generalization across sites.

Data harmonization, such as compensating differences across measurement devices, is crucial, but remains very difficult and cannot correct these differences perfectly [Glocker et al., 2019]. Removing too much inter-site variance can lead to loss of informative signal. Rather, it is important to model it well, accounting for the two sources of variance, across participants and across sites. A good model strives to yield good results on all sites. One solution is to adapt ideas from robust optimization: on data drawn from different distributions (e.g. from several sites), Krueger et al. [2020] show the benefits of minimizing the empirical risk on the worse site or adding penalties on the variance of the loss across sites.

Measures of prediction performance should aggregate scores at the site level (not pooling all individuals), and check the variance across sites and the performance on the worse site. Cross-validation schemes should hold out entire sites [Woo et al., 2017, Little et al., 2017].

7 Special cases of dataset shift

Categorizing dataset shift helps finding the best approach to tackle it Storkey [2009], Moreno-Torres et al. [2012]. We summarize two frequently-met scenarios that are easier to handle than the general case and can call for different adjustments: covariate shift (Section 7.1) and prior probability shift (Section 7.2).

7.1 Covariate shift

Covariate shift occurs when the marginal distribution of X changes between the source and target datasets (i.e. $p_t(x) \neq p_s(x)$), but $P(Y \mid X)$ stays the same. This happens for example in the second scenario in Figure 3, where sample selection based on X (but not Y) changes the distribution of the inputs. If the model is correctly specified, an estimator trained with uniform weights will lead to optimal predictions given sufficient training data [prediction consistency Shimodaira, 2000, Lemma 4]. However the usual (unweighted) estimator is not consistent for an over-constrained (misspecified) model. Indeed, a over-constrained model may be able to fit the data well only in some regions of the input feature space (Figure 1). In this case reweighting training examples (Section 5) to give more importance to those that are more representative of the target data is beneficial [Storkey, 2009, Schölkopf et al., 2012]. Figure 5 illustrates covariate shift.

7.2 Prior probability shift

Another relatively simple case of dataset shift is *prior probability shift*. With prior probability shift (a.k.a. label shift or target shift), the distribution of Y changes but not $P(X \mid Y)$. This happens for example when disease prevalence changes in the target population but manifests itself in the same way. Even more frequently, prior probability shift arises when one rare class is over-represented in the training data so that the dataset is more balanced, as when extracting a biomarker from a case-control cohort, or when the dataset is resampled as a strategy to handle the *class imbalance* problem [He and Garcia, 2009]. Prior probability shift can be corrected without extracting a new biomarker, simply by adjusting a model's predicted probabilities using Bayes' rule [as noted for example in Storkey, 2009, Schölkopf et al., 2012]. When the classes are well separated, the effect of this correction may be small, i.e. the uncorrected classifier

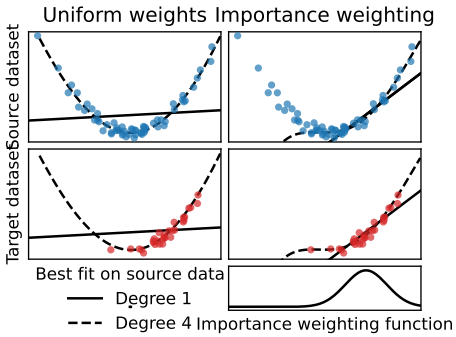

Figure 5. Covariate shift: $P(Y \mid X)$ stays the same but the feature space is sampled differently in the source and target datasets. A powerful learner may generalize well as $P(Y \mid X)$ is correctly captured [Storkey, 2009]. Thus the polynomial fit of degree 4 performs well on the new dataset. However, an overconstrained learner such as the linear fit can benefit from reweighting training examples to give more importance to the most relevant region of the feature space.

may generalize well without correction. Figure 6 illustrates prior probability shift.

8 Conclusion

Ideally, machine learning biomarkers would be designed and trained using datasets carefully collected to be representative of the targeted population – as in Liu et al. [2020]. To be trusted, biomarkers ultimately need to be evaluated rigorously on one or several independent and representative samples. However, such data collection is expensive. It is therefore useful to exploit existing datasets in an opportunistic way as much as possible in the early stages of biomarker development. When doing so, correctly accounting for dataset shift can prevent wasting important resources on machine-learning predictors that have little chance of performing well outside of one particular dataset.

We gave an overview of importance weighting, a simple tool against dataset shift. Importance weighting needs a clear definition of the targeted population and access to a diverse training dataset. When this is not possible, distributionally robust optimization may be promising alternative, though it is a more recent approach and more difficult to implement. Despite much work and progress, dataset shift remains a difficult problem. Characterizing its impact

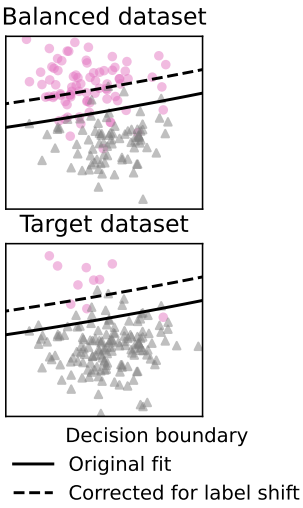

Figure 6. Prior probability shift: when $P(Y)$ changes but $P(X \mid Y)$ stays the same. This can happen for example when participants are selected based on Y – possibly to have a dataset with a balanced number of patients and healthy participants: $X \leftarrow Y \rightarrow \boxed{S}$. When the prior probability (marginal distribution of Y) in the target population is known, this is easily corrected by applying Bayes' rule. The output Y is typically low-dimensional and discrete (often it is a single binary value), so $P(Y)$ can often be estimated precisely from few examples.

and the effectiveness of existing solutions for biomarker discovery will be important for machine learning models to become more reliable in healthcare applications.

We conclude with the following recommendations:

- be aware of the dataset shift problem and the difficulty of out-of-dataset generalization. Do not treat cross-validation scores on one dataset as a guarantee that a model will perform well on clinical data.
- collect diverse, representative data.
- use powerful machine-learning models and large datasets.
- consider using importance weighting to correct biases in the data collection, especially if the learning model may be over-constrained (e.g. when using a linear model).
- look for associations between prediction performance and demographic variables in the validation set to detect potential generalization or fairness issues.
- *do not* remove confounding signal in a predictive setting.

These recommendations should help designing fair biomarkers and their efficient application on new cohorts.

Author contributions Jérôme Dockès, Gaël Varoquaux and Jean-Baptiste Poline participated in conception, literature search, data interpretation, and editing the manuscript. Jérôme Dockès wrote the software and drafted the manuscript. Both Gaël Varoquaux and Jean-Baptiste Poline contributed equally to this work (as last authors).

Competing interests statement The authors declare that there are no competing interests.

Software and data availability The source files used to create this publication can be found in this repository: `https://github.com/neurodatascience/dataset_shift_biomarkers`. UKBiobank data can be obtained from `https://www.ukbiobank.ac.uk`.

References

S. Abbasi-Sureshjani, R. Raumanns, B. E. J. Michels, G. Schouten, and V. Cheplygina. Risk of training diagnostic algorithms on data with demographic bias. In *Interpretable and Annotation-Efficient Learning for Medical Image Computing*, pages 183–192, Cham, 2020. Springer International Publishing. ISBN 978-3-030-61166-8.

A. Achille and S. Soatto. Emergence of invariance and disentanglement in deep representations. *The Journal of Machine Learning Research*, 19(1):1947–1980, 2018.

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

J. Andreu-Perez, C. C. Poon, R. D. Merrifield, S. T. Wong, and G.-Z. Yang. Big data for health. *IEEE journal of biomedical and health informatics*, 19(4):1193–1208, 2015.

A. Antoniou, A. Storkey, and H. Edwards. Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340*, 2017.

S. Arlot, A. Celisse, et al. A survey of cross-validation procedures for model selection. *Statistics surveys*, 4: 40–79, 2010.

P. C. Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 46(3):399–424, 2011.

E. Bareinboim and J. Pearl. Controlling selection bias in causal inference. In *Artificial Intelligence and Statistics*, pages 100–108, 2012.

S. Barocas, M. Hardt, and A. Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. `http://www.fairmlbook.org`.

A. H. Beck, A. R. Sangoi, S. Leung, R. J. Marinelli, T. O. Nielsen, M. J. Van De Vijver, R. B. West, M. Van De Rijn, and D. Koller. Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Science translational medicine*, 3(108):108ra113–108ra113, 2011.

S. Bleeker, H. Moll, E. Steyerberg, A. Donders, G. Derksen-Lubsen, D. Grobbee, and K. Moons. External validation is necessary in prediction research:: A clinical example. *Journal of clinical epidemiology*, 56(9):826–832, 2003.

D. B. Chastain, S. P. Osae, A. F. Henao-Martínez, C. Franco-Paredes, J. S. Chastain, and H. N. Young. Racial disproportionality in covid clinical trials. *New England Journal of Medicine*, 383(9):e59, 2020.

D. Cirillo, S. Catuara-Solarz, C. Morey, E. Guney, L. Subirats, S. Mellino, A. Gigante, A. Valencia, M. J. Rementeria, A. S. Chadha, et al. Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. *NPJ Digital Medicine*, 3(1):1–11, 2020.

C. Cortes, M. Mohri, M. Riley, and A. Rostamizadeh. Sample selection bias correction theory. In *International conference on algorithmic learning theory*, pages 38–53. Springer, 2008.

N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39 (9):1853–1865, 2016.

R. C. Deo. Machine learning in medicine. *Circulation*, 132 (20):1920–1930, 2015.

J. Duchi and H. Namkoong. Learning models with uniform performance via distributionally robust optimization. *arXiv preprint arXiv:1810.08750*, 2018.

M. Dudík, S. J. Phillips, and R. E. Schapire. Correcting sample selection bias in maximum entropy density estimation. In *Advances in neural information processing systems*, pages 323–330, 2006.

J. R. England and P. M. Cheng. Artificial intelligence for medical image analysis: a guide for authors and reviewers. *American Journal of Roentgenology*, 212(3): 513–519, 2019.

O. Faust, Y. Hagiwara, T. J. Hong, O. S. Lih, and U. R. Acharya. Deep learning for healthcare applications based on physiological signals: A review. *Computer methods and programs in biomedicine*, 161:1–13, 2018.

J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*. Springer series in statistics New York, 2001.

M. A. Gianfrancesco, S. Tamang, J. Yazdany, and G. Schmajuk. Potential biases in machine learning algorithms using electronic health record data. *JAMA internal medicine*, 178(11):1544–1547, 2018.

B. Glocker, R. Robinson, D. C. Castro, Q. Dou, and E. Konukoglu. Machine learning with multi-site imaging data: An empirical study on the impact of scanner effects. *arXiv preprint arXiv:1910.04597*, 2019.

H. He and E. A. Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

A. Heiat, C. P. Gross, and H. M. Krumholz. Representation of the elderly, women, and minorities in heart failure clinical trials. *Archives of internal medicine*, 162(15), 2002.

J. Henrich, S. J. Heine, and A. Norenzayan. Most people are not weird. *Nature*, 466(7302):29–29, 2010.

M. Hernán and J. Robins. Causal inference: What if. *Boca Raton: Chapman & Hill/CRC*, 2020.

M. A. Hernán, S. Hernández-Díaz, and J. M. Robins. A structural approach to selection bias. *Epidemiology*, pages 615–625, 2004.

J. Huang, A. Gretton, K. Borgwardt, B. Schölkopf, and A. J. Smola. Correcting sample selection bias by unlabeled data. In *Advances in neural information processing systems*, pages 601–608, 2007.

D. Jin, B. Zhou, Y. Han, J. Ren, T. Han, B. Liu, J. Lu, C. Song, P. Wang, D. Wang, et al. Generalizable, reproducible, and neuroscientifically interpretable imaging biomarkers for alzheimer's disease. *Advanced Science*, page 2000675, 2020.

S. Kakarmath, A. Esteva, R. Arnaout, H. Harvey, S. Kumar, E. Muse, F. Dong, L. Wedlund, and J. Kvedar. Best practices for authors of healthcare-related artificial intelligence manuscripts. *NPJ Digital Medicine*, 2020.

T. Kanamori, S. Hido, and M. Sugiyama. A least-squares approach to direct importance estimation. *The Journal of Machine Learning Research*, 10:1391–1445, 2009.

R. Kasahara, S. Kino, S. Soyama, and Y. Matsuura. Noninvasive glucose monitoring using mid-infrared absorption spectroscopy based on a few wavenumbers. *Biomedical optics express*, 9(1):289–302, 2018.

W. M. Kouw and M. Loog. A review of domain adaptation without target labels. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

D. Krueger, E. Caballero, J.-H. Jacobsen, A. Zhang, J. Binas, R. L. Priol, and A. Courville. Out-of-distribution generalization via risk extrapolation (rex). *arXiv preprint arXiv:2003.00688*, 2020.

A. J. Larrazabal, N. Nieto, V. Peterson, D. H. Milone, and E. Ferrante. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, 117:12592, 2020.

M. A. Little, G. Varoquaux, S. Saeb, L. Lonini, A. Jayaraman, D. C. Mohr, and K. P. Kording. Using and understanding cross-validation strategies. perspectives on saeb et al. *GigaScience*, 6(5):gix020, 2017.

M. Liu, G. Oxnard, E. Klein, C. Swanton, M. Seiden, M. C. Liu, G. R. Oxnard, E. A. Klein, D. Smith, D. Richards, et al. Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free dna. *Annals of Oncology*, 2020.

M. Long, Y. Cao, J. Wang, and M. Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR, 2015.

A. Maurer and M. Pontil. Empirical bernstein bounds and sample variance penalization. *stat*, 1050:21, 2009.

J. G. Moreno-Torres, T. Raeder, R. Alaiz-RodríGuez, N. V. Chawla, and F. Herrera. A unifying view on dataset shift in classification. *Pattern recognition*, 45(1):521–530, 2012.

S. A. Mulherin and W. C. Miller. Spectrum bias or spectrum effect? subgroup variation in diagnostic test evaluation. *Annals of internal medicine*, 137(7):598–602, 2002.

V. H. Murthy, H. M. Krumholz, and C. P. Gross. Participation in cancer clinical trials: race-, sex-, and age-based disparities. *Jama*, 291(22):2720–2726, 2004.

H. Namkoong and J. C. Duchi. Variance-based regularization with convex objectives. In *Advances in neural information processing systems*, pages 2971–2980, 2017.

A. Niculescu-Mizil and R. Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 625–632, 2005.

L. Oakden-Rayner, J. Dunnmon, G. Carneiro, and C. 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.

C. O'neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books, 2016.

S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10): 1345–1359, 2009.

J. Pearl. The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, 62 (3):54–60, 2019.

J. Pearl, M. Glymour, and N. P. Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

J. Peters, D. Janzing, and B. Schölkopf. *Elements of causal inference: foundations and learning algorithms*. MIT press, 2017.

R. A. Poldrack, G. Huckins, and G. Varoquaux. Establishment of best practices for evidence for prediction: a review. *JAMA psychiatry*, 77(5):534–540, 2020.

H. Rahimian and S. Mehrotra. Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*, 2019.

D. F. Ransohoff and A. R. Feinstein. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *New England Journal of Medicine*, 299(17):926–930, 1978.

K. J. Rothman. *Epidemiology: an introduction*. Oxford university press, 2012.

C. Sáez, A. Gutiérrez-Sacristán, I. Kohane, J. M. García-Gómez, and P. Avillach. Ehrtemporalvariability: delineating temporal dataset shifts in electronic health records. *medRxiv*, 2020.

B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. Mooij. On causal and anticausal learning. In *29th International Conference on Machine Learning (ICML 2012)*, pages 1255–1262. International Machine Learning Society, 2012.

H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

A. Storkey. When training and test sets are different: characterizing learning transfer. *Dataset shift in machine learning*, pages 3–28, 2009.

K. Strimbu and J. A. Tavel. What are biomarkers? *Current Opinion in HIV and AIDS*, 5(6):463, 2010.

A. Subbaswamy, P. Schulam, and S. Saria. Preventing failures due to dataset shift: Learning predictive models that transport. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3118–3127, 2019.

C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray, et al. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3), 2015.

M. Sugiyama and M. Kawanabe. *Machine learning in non-stationary environments: Introduction to covariate shift adaptation*. MIT press, 2012.

M. Sugiyama, M. Krauledat, and K.-R. Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(May):985–1005, 2007.

M. Sugiyama, S. Nakajima, H. Kashima, P. V. Buenau, and M. Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in neural information processing systems*, pages 1433–1440, 2008.

B. Sun, J. Feng, and K. Saenko. Return of frustratingly easy domain adaptation. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

G. Tripepi, K. J. Jager, F. W. Dekker, and C. Zoccali. Selection bias and information bias in clinical research. *Nephron Clinical Practice*, 115(2):c94–c99, 2010.

E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017.

K. Weiss, T. M. Khoshgoftaar, and D. Wang. A survey of transfer learning. *Journal of Big data*, 3(1):9, 2016.

C.-W. Woo, L. J. Chang, M. A. Lindquist, and T. D. Wager. Building better biomarkers: brain models in translational neuroimaging. *Nature neuroscience*, 20(3):365, 2017.

L. Wynants, B. Van Calster, M. M. Bonten, G. S. Collins, T. P. Debray, M. De Vos, M. C. Haller, G. Heinze, K. G. Moons, R. D. Riley, et al. Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal. *bmj*, 369, 2020.

B. Zadrozny. Learning and evaluating classifiers under sample selection bias. In *Proceedings of the twenty-first international conference on Machine learning*, page 114, 2004.

B. Zadrozny, J. Langford, and N. Abe. Cost-sensitive learning by cost-proportionate example weighting. In *Third IEEE international conference on data mining*, pages 435–442. IEEE, 2003.

K. Zhang, B. Schölkopf, K. Muandet, and Z. Wang. Domain adaptation under target and conditional shift. In *International Conference on Machine Learning*, pages 819–827, 2013.

J.-J. Zhu, W. Jitkrittum, M. Diehl, and B. Schölkopf. Kernel distributionally robust optimization. *arXiv preprint arXiv:2006.06981*, 2020.

A Definition and estimation of importance weights

We will implicitly assume that all the random variables we consider admit densities and denote p_s and p_t the density of the joint distribution of (X, Y) applied to the source and target populations respectively. If the support of p_t is included in that of p_s (meaning that $p_s > 0$ wherever $p_t > 0$), we have:

$$\mathbb{E}_{\text{source}}[\, L(Y,\, f(X))\,] = \mathbb{E}_{\text{target}}\left[\, \frac{p_t(X,\, Y)}{p_s(X,\, Y)}\, L(Y,\, f(X))\right], \tag{3}$$

where L is the cost function and f is a prediction function, $\mathbb{E}_{\text{source}}$ (resp. $\mathbb{E}_{\text{target}}$) the expectation on the source (resp. target) data. The risk (on target data) can therefore be computed as an expectation on the source distribution where the loss function is reweighted by the *importance weights*:

$$w(x, y) = \frac{p_t(x, y)}{p_s(x, y)}\ . \tag{4}$$

If $\hat{w}$ are empirical estimates of the importance weights w, it is possible to compute the reweighted empirical risk:

$$\hat{R}_{\hat{w}}(f) = \sum_{i=1}^{n} \hat{w}(x_i, y_i)\, L(y_i, f(x_i))\ . \tag{5}$$

Rather than being weighted, examples can also be resampled with importance or rejection sampling [Zadrozny et al., 2003, Zadrozny, 2004]. Importances can also be taken into account for model selection – for example in Sugiyama et al. [2007] examples of the test set are also reweighted when computing cross-validation scores. Cortes et al. [2008] study how errors in the estimation of the weights affect the prediction performance.

A.1 Preferential Sample selection and Inverse Probability weighting

In the case of preferential sample selection (Section 4), the condition that requires for the support of p_t to be included in the support of p_s translates to a requirement that all individuals have a non-zero probability of being selected: $P(S = 1 \,|\, x, y) > 0$ for all (x, y) in the support of p_t. When this is verified, by applying Bayes' rule the definition of importance weights in Equation (4) can be reformulated [see Cortes et al., 2008, Sec. 2.3]:

$$w(x, y) = \frac{P(S = 1)}{P(S = 1 \,|\, X = x, Y = y)} \tag{6}$$

These weights are sometimes called Inverse Probability weights [Hernán et al., 2004] or Inverse Propensity scores [Austin, 2011]. Training examples that had a low probability of being selected receive higher weights, because they have to account for similar individuals who were not selected.

A.2 Computing importance weights

In practice $p_t(x, y)$, which is the joint density of (X, Y) in the target data, is not known. However, it is not needed for the estimation of p_t/p_s. More efficient estimation hinges on two observations: estimation of both densities separately is not necessary to estimate their ratio, and variables that have the same distribution in source and target data can be factored out.

Here we describe methods that estimate the true importance weights p_t/p_s, but we point out that reweighting the training examples reduces the bias of the empirical risk but increases the variance of the estimated model parameters. Even when the importances are perfectly known, it can therefore be beneficial to regularize the weights [Shimodaira, 2000].

A.2.1 Computing importance weights does not require distributions densities estimation

Importance weights can be computed by modelling separately p_s and p_t and then computing their ratio [Sugiyama and Kawanabe, 2012, Sec. 4.1]. However, distribution density estimation is notoriously difficult; non-parametric methods suffer from the curse of dimensionality and parametric methods depend heavily on the correct specification of a parametric form.

But estimating both densities is more information than is needed to compute the sample weights. Instead, one can directly optimize importance weights in order to make the reweighted sample similar to the target distribution, by matching moments [Sun et al., 2016] or mean embeddings [Huang et al., 2007, Zhang et al., 2013], minimizing the KL-divergence [Sugiyama et al., 2008], solving a least-squares estimation problem [Kanamori et al., 2009] or with optimal transport [Courty et al., 2016].

Alternatively, a discriminative model can be trained to distinguish source and target examples. In the specific case of preferential sample selection, this means estimating directly the probability of selection $P(S = 1)$ (cf Equation (6)). In general, the shift is not always due to selection: the source data is not necessarily obtained by subsampling the target population. In this case we denote $T = 1$ if an individual comes from the target data and $T = 0$ if it comes from the source data. Then, a classifier can be trained to predict from which dataset (source or target) a sample is drawn, and the importance weights obtained from the predicted probabilities [Sugiyama and Kawanabe, 2012, Sec. 4.3]:

$$w(x, y) = \frac{P(T = 1 \,|\, X = x, Y = y)\, P(T = 0)}{P(T = 0 \,|\, X = x, Y = y)\, P(T = 1)} \,, \quad (7)$$

The classifier must be calibrated (i.e. produce accurate probability estimates, not only a correct decision), see Niculescu-Mizil and Caruana [2005]. Note that constant factors such as $P(T = 0)/P(T = 1)$ usually do not matter and are easy to estimate if needed. This discriminative approach is effective because the distribution of $(T \,|\, X = x, Y = y)$ is much easier to estimate than the distribution of $(X, Y \,|\, T = t) : T$ is a single binary variable whereas (X, Y) is high-dimensional and often continuous.

The classifier does not need to distinguish source and target examples with high accuracy. In the ideal situation of no dataset shift, the classifier will perform at chance level. On the contrary, a high accuracy means that there is little overlap between the source and target distributions and the model will probably not generalize well.

A.2.2 What distributions differ in source and target data?

When computing importance weights, it is possible to exploit prior knowledge that some distributions are left unchanged in the target data. For example,

$$\frac{p_t(x, y)}{p_s(x, y)} = \frac{p_t(y \,|\, x)\, p_t(x)}{p_s(y \,|\, x)\, p_s(x)} \,. \quad (8)$$

Imagine that the marginal distribution of input X differs in source and target data, but the conditional distribution of the output Y given the input stays the same: $p_t(x) \neq p_s(x)$ but $p_t(y \,|\, x) = p_s(y \,|\, x)$ (a setting known as *covariate shift*). Then, the importance weights simplify to

$$w(x, y) = \frac{p_t(x)}{p_s(x)} \,. \quad (9)$$

In this case, importance weights can be estimated using only unlabelled examples (individuals for whom Y is unkown) from the target distribution.

Often, the variables that influence selection (e.g. demographic variables such as age) are lower-dimensional than the full features (e.g. high-dimensional images), and dataset shift can be corrected with limited information on the target distribution, with importance weights or otherwise. Moreover, even if additional information Z that predicts selection but is independent of (X, Y) is available, it should *not* be used to compute the importance weights. Indeed, this would only increase the weights' variance without reducing the bias due to the dataset shift [Hernán and Robins, 2020, Sec. 15.5].

B Tobacco smoking prediction in the UKBiobank

We consider predicting the smoking status of participants in the UKBiobank study to illustrate the effect of dataset shift on prediction performance.

6,000 participants are used in a preliminary step to identify the 29 most relevant predictive features (listed in appendix B.1), by cross-validating a gradient boosting model and computing permutation feature importances. We then draw two samples of 100K individuals from the rest of the dataset, that have different age distributions. In the young sample, 90% of individuals come from the youngest 20% of the dataset, and the remaining 10% are sampled from the oldest 20% of the dataset. In the old sample, these proportions are reversed. We then perform 10-fold cross validation. For each fold, both the training and testing set can be drawn from either the young or the old population, resulting in four tasks on which several machine-learning estimators are evaluated. We use this experiment to compare 2 machine-learning models: a simple one – regularized linear Support Vector Classifier, and a flexible one – Gradient Boosting. For each classifier, 3 strategies are considered to handle the dataset shift: (i) baseline – the generic algorithm without modifications, (ii) Importance Weighting (Section 5), and (iii) the (unfortunately popular) non-solution: "regressing

out the confounder" – regressing the predictive features on the age and using the residuals as inputs to the classifier.

The results are similar to those seen with simulated data in Figure 1. For a given learner and test population, training on a different population degrades the prediction score. For example, if the learner is to be tested on the young population, it performs best when trained on the young population. Gradient Boosting vastly outperforms the linear model in all configurations. Regressing out the age always degrades the prediction; it is always worse than the unmodified baseline, whether a dataset shift is present or not. Finally, Importance Weighting (Section 5) improves the predictions of the over-constrained (misspecified) linear model in the presence of dataset shift, but degrades the prediction of the powerful learner used in this experiment. This is due to the fact that the Gradient Boosting already captures the correct separation for both young and old individuals, and therefore Importance Weighting does not bring any benefit but only reduces the effective training sample size by increasing the variance of the empirical risk.

B.1 Features used for tobacco smoking status prediction

The 30 most important features were identified in a preliminary experiment with 6,000 participants (that were not used in the subsequent analysis). One of these features, "Date F17 first reported (mental and behavioural disorders due to use of tobacco)", was deemed trivial – too informative, as it directly implies that the participant does smoke tobacco, and removed. The remaining 29 features were used for the experiment described in Section 3.

- Forced expiratory volume in 1-second (FEV1), predicted percentage
- Lifetime number of sexual partners
- Age first had sexual intercourse
- Age when last took cannabis
- Ever taken cannabis
- Forced expiratory volume in 1-second (FEV1), predicted
- Acceptability of each blow result
- Mouth/teeth dental problems
- Coffee intake
- FEV1/ FVC ratio Z-score
- Alcohol intake frequency.
- Date J44 first reported (other chronic obstructive pulmonary disease)
- Former alcohol drinker
- Average weekly spirits intake
- Year of birth
- Acceptability of each blow result
- Date of chronic obstructive pulmonary disease report
- Leisure/social activities
- Morning/evening person (chronotype)
- Mean sphered cell volume
- Lymphocyte count
- Townsend deprivation index at recruitment
- Age hay fever, rhinitis or eczema diagnosed
- Age started oral contraceptive pill
- White blood cell (leukocyte) count
- Age completed full time education
- Age at recruitment
- Workplace had a lot of cigarette smoke from other people smoking
- Wheeze or whistling in the chest in last year

C Glossary

Here we provide a summary of some terms and notations used in the paper.

Target population the population on which the biomarker (machine-learning model) will be applied.

Source population the population from which the sample used to train the machine-learning model is drawn.

Selection in the case that source data are drawn (with non-uniform probabilities) from the target population, we denote by $S = 1$ the fact that an individual is selected to enter the source data (e.g. to participate in a medical study).

Provenance of an individual when samples from both the source and the target populations (e.g. Appendix A.2.1) are available, we also denote $T = 1$ if an individual comes from the target population and $T = 0$ if they come from the source population.

Confounding in *causal inference*, when estimating the effect of a treatment on an outcome, confounding occurs if a third variable (e.g. age, a comorbidity, the seriousness of a condition) influences both the treatment and the outcome, possibly producing a spurious statistical association between the two. This notion is not directly relevant to dataset shift, and we mention it only to insist that it is a different problem. See Hernán and Robins [2020], Chap. 7, for a more precise definition.

Domain adaptation the task of designing machine-learning methods that are resilient to dataset shift – essentially a synonym for dataset shift, i.e. another useful search term for readers looking for further information on this problem.