

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/380547509>

Sample Selection Bias in Machine Learning for Healthcare

Preprint · May 2024

DOI: 10.48550/arXiv.2405.07841

CITATIONS

0

READS

67

8 authors, including:



Vinod Kumar Chauhan

University of Oxford

52 PUBLICATIONS 591 CITATIONS

SEE PROFILE



Gaurav Nigam

University of Oxford

82 PUBLICATIONS 164 CITATIONS

SEE PROFILE



David A. Clifton

University of Oxford

342 PUBLICATIONS 10,865 CITATIONS

SEE PROFILE

SAMPLE SELECTION BIAS IN MACHINE LEARNING FOR HEALTHCARE

A PREPRINT

Vinod Kumar Chauhan^{1*}, Lei Clifton^{1,2}, Achille Salaün¹, Huiqi Yvonne Lu¹, Kim Branson³,

Patrick Schwab³, Gaurav Nigam^{1,4}, David A. Clifton^{1,5}

¹Institute of Biomedical Engineering, University of Oxford, UK

²Nuffield Department of Population Health, University of Oxford, UK

³Biomedical AI Group, GSK

⁴Translational Gastroenterology Unit, Nuffield Department of Medicine, University of Oxford, Oxford, UK

⁵Oxford-Suzhou Institute of Advanced Research (OSCAR), Suzhou, China

May 14, 2024

ABSTRACT

While machine learning algorithms hold promise for personalised medicine, their clinical adoption remains limited. One critical factor contributing to this restraint is sample selection bias (SSB) which refers to the study population being less representative of the target population, leading to biased and potentially harmful decisions. Despite being well-known in the literature, SSB remains scarcely studied in machine learning for healthcare. Moreover, the existing techniques try to correct the bias by balancing distributions between the study and the target populations, which may result in a loss of predictive performance. To address these problems, our study illustrates the potential risks associated with SSB by examining SSB's impact on the performance of machine learning algorithms. Most importantly, we propose a new research direction for addressing SSB, based on the target population identification rather than the bias correction. Specifically, we propose two independent networks (T-Net) and a multitasking network (MT-Net) for addressing SSB, where one network/task identifies the target subpopulation which is representative of the study population and the second makes predictions for the identified subpopulation. Our empirical results with synthetic and semi-synthetic datasets highlight that SSB can lead to a large drop in the performance of an algorithm for the target population as compared with the study population, as well as a substantial difference in the performance for the target subpopulations that are representative of the selected and the non-selected patients from the study population. Furthermore, our proposed techniques demonstrate robustness across various settings, including different dataset sizes, event rates, and selection rates, outperforming the existing bias correction techniques.

Keywords Sample Selection Bias · Target Population · Machine Learning · Healthcare · Risk Prediction.

1 Introduction

Machine learning algorithms have demonstrated high diagnostic and prognostic accuracy in identifying and categorising diseases, promising personalised interventions and informed decision-making in healthcare [49, 58]. Their ability to analyse vast datasets and uncover hidden patterns surpasses traditional techniques and sometimes even human experts in task-specific applications, such as image-recognition tasks in radiology [36, 33]. However, despite their compelling

*Corresponding author: Vinod Kumar Chauhan (vinod.kumar@eng.ox.ac.uk)

potential and an increasing number of studies every year, their clinical adoption remains constrained by various issues [61]. One such issue is sample selection bias (SSB), recognised as a fundamental pitfall in the research design of clinical studies [65] but remains scarcely studied in machine learning for healthcare, posing a potential hurdle to its real-world applications.

Inherent in clinical studies, which serve as a source of data for machine learning, is the practice of sample selection, characterised by strict criteria for the inclusion and exclusion of patients [45]. However, these studies don't add any bias as long as the study population is representative of the target population. The study population is defined as data used to develop the model and generally split into training, (internal) validation, and test datasets; and the target population is defined as the independent external data, i.e., deployment data or external validation data, respectively. SSB results from a non-uniform sampling process to select patients in the study population, leading to differences in the distributions of the study population and the target population [48, 6, 67, 7]. For example, in a clinical study, the outcomes of some patients are not available due to a loss to follow-up or competing events, and considering only those patients who completed the study introduces SSB in the dataset [2]. When machine learning algorithms are trained on such a biased dataset and employed on an unbiased dataset, it might lead to inaccurate predictions and suboptimal care for patient subpopulations not represented in the study population [50, 22]. Another example is the self-nomination of patients for a clinical study, which may not represent the entire population, such as the UK Biobank and might lead to SSB [56, 8]. Fig. 1 provides an example illustrating SSB and its impact on a prediction algorithm's performance.

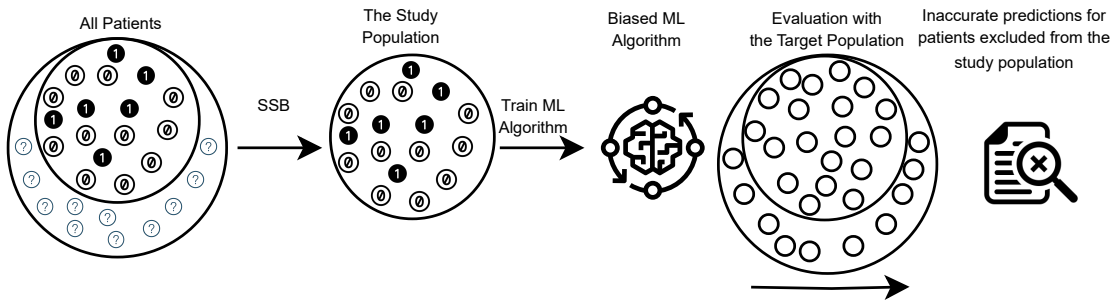


Figure 1: Illustration of SSB and its impact: We start with all patients who participated in the study, where we have known outcomes for patients inside the inner circle (denoted using 0 and 1) while the rest have unknown outcomes (denoted using ‘?’). The study population includes patients only from the inner circle, i.e., those patients who have known outcomes, resulting in SSB. A machine learning (ML) algorithm trained on such a dataset does not have information about the non-selected patients. When such an algorithm is deployed and evaluated with the target population which contains patients that are representative of the selected and the non-selected from the study population, then it may lead to inaccurate predictions and harmful decisions for the subpopulations representative of the non-selected patients.

SSB is well known for decades [25, 26, 3], and it has been discussed across several disciplines, including social science and econometrics [27, 28], environmental studies [62], finance [35], causality [31, 4], fairness [20], healthcare [44], and machine learning [66]. It is also an important consideration for clinical study design as the results of a biased study may not apply to a real patient population [6, 48]. Despite being well-known and a threat to the validity of algorithms [14], SSB has been scarcely studied in machine learning for healthcare [65]. This could be because (i) often less proportion of patients are excluded from the study population, making it hard to find enough data to correct the bias, (ii) most of the algorithms are not developed for the deployment, (iii) the bias is sometimes hidden under the assumed randomness in selection [51], (iv) risk estimation primarily belongs to causality, a principled approach to decision-making that examines cause-and-effect relationships. Causality has developed formal techniques, including causal diagrams, to address SSB [63, 53], and (v) SSB might not always lead to inaccurate predictions for the non-selected patients, as they might not influence the algorithm’s decision-making process.

It is not always possible to avoid SSB in the study but to utilise valuable information hidden in such a biased dataset, it is necessary to address the bias to ensure that decisions are made prudently and with reduced bias. Machine learning for healthcare lacked research to address SSB, however, the machine learning community has developed several techniques to address SSB [66] and to address distribution mismatch between the source and the target domains, such as covariate shift [52] or domain-adaptation [39], which have a setting similar to SSB. The existing techniques correct for the bias by aligning distributions of the study population and the target population, which could lead to a loss of predictive performance of the algorithm [66, 34, 39]. Moreover, in healthcare, the non-selected patients form a subpopulation that

could greatly differ from the study population, making it hard to correct for the bias. Thus, the existing techniques may perform poorly for subpopulations that are representative of the non-selected subpopulation compared to those representative of the selected subpopulation.

To address the above problems, our study presents two-fold contributions. Firstly, we present an empirical study with synthetic and semi-synthetic datasets to highlight the importance of addressing SSB in machine learning for healthcare. Our study demonstrates that SSB leads to a large performance gap for machine learning algorithms between the study and the target populations as well as a substantial gap between the target subpopulations representing the non-selected and the selected patients. Secondly, we propose a new research direction for addressing SSB, focusing on identifying the target population rather than correcting the bias. This approach to address SSB involves first identifying patients from the target population who are representative of the study population, followed by predictions for the identified patients. The proposed approach for addressing SSB makes sense in healthcare settings where the non-selected patients (e.g., those lost to follow-up) form a small subpopulation of the study population. This is in contrast to domain adaptation techniques, where the target domain (i.e., the target population) is assumed to differ from the source domain (i.e., the study population) and the domain adaptation algorithms are supposed to work for the target domain. Further, as concrete implementations of the proposed research direction, we propose two independent networks (T-Net) and a multitasking network (MT-Net) techniques for addressing SSB. T-Net develops two networks, one for identification, i.e., to predict the selection of a patient to the study population, and the other for the underlying prediction problem. Similarly, MT-Net develops a multitasking network for two tasks, one for the identification and the other for the underlying prediction problem. Finally, we analyse the performance of the proposed techniques against the state-of-the-art baselines using synthetic and semi-synthetic datasets to prove their effectiveness.

The contributions of the paper are summarised as given below.

- We present an empirical study with synthetic and semi-synthetic datasets to highlight SSB in machine learning for healthcare, which is otherwise ignored but can lead to biased and harmful decisions.
- We propose a new research direction to address SSB in machine learning for healthcare, focusing on identification of the target subpopulation representative of the study population rather than the bias correction, and hence avoiding loss of predictive performance due to balancing of distributions.
- We propose T-Net and MT-Net, as concrete techniques under the proposed research direction, for addressing SSB, which have two networks/tasks for the population identification and the underlying prediction task. We present an empirical analysis of the proposed techniques against the state-of-the-art baselines using synthetic and semi-synthetic datasets, which show the robustness of the proposed T-Net and MT-Net under different settings.

The rest of the paper is organized as: Section 2 provides a problem definition of SSB, while Section 3 presents a brief overview of literature related to SSB. The proposed research direction and T-Net and MT-Net are discussed in Section 4, followed by performance analysis with synthetic and semi-synthetic datasets in Section 5. Discussion on the significance of results is presented in Section 6, and finally, the concluding remarks are presented in Section 7.

2 Sample Selection Bias

In SSB, a study population is selected non-randomly from the general population. In contrast, the target population is selected uniformly from the general population as at the time of deployment patients can belong to any subpopulation. Generally, the selected patients have defined outcomes, i.e., labels available along with features but for the non-selected patients labels are missing and that's why they are not included in the study population. For example, a loss to follow-up or survival bias may lead to SSB as patients lacking labels are not included in the study [2]. A naive approach to developing a machine learning algorithm using a biased dataset could result in a biased algorithm. Such an algorithm may fail to accurately make predictions for the patients representative of the non-selected subpopulation at the time of deployment, as the algorithm lacks information about this subpopulation. So, the challenge with SSB involves developing predictive algorithms using biased labelled samples of the selected patients and unlabelled samples of the non-selected patients, to develop accurate and useful algorithms for making predictions. While the ideal scenario involves mitigating the bias to ensure fair and representative models, the primary objective remains the development of algorithms that effectively leverage available data to provide accurate predictions, despite the presence of bias.

Suppose $\mathbf{x} \in \mathbf{X}$ is a d -dimensional feature vector representing a patient, $y \in \mathbf{y}$ is the corresponding outcome and $s \in \mathbf{s}$ is a random selection variable which takes value 1 if patient \mathbf{x} is selected in the study population or 0 otherwise. For the sake of clarity and simplicity, we focus on binary classification, i.e., risk prediction task. However, our work can be extended to other problem settings, such as regression. Let $\mathbf{X} = \mathbf{X}^0 \cup \mathbf{X}^1 \in \mathbb{R}^d$ be a random sample from the general population, where $\mathbf{x} \in \mathbf{X}^1$ has corresponding label $y \in \mathbf{y}^1$ and are selected in the study population, i.e., $s = 1$, and

$\mathbf{x} \in \mathbf{X}^0$ does not have access to corresponding label and are not selected into the study population, i.e., $s = 0$. We assume access to labelled and unlabelled samples as $\mathbf{D}^1 = \{(\mathbf{x}_1^1, s_1 = 1, y_1^1), (\mathbf{x}_2^1, s_2 = 1, y_2^1), \dots, (\mathbf{x}_n^1, s_n = 1, y_n^1)\}$ of n data points, and $\mathbf{D}^0 = \{(\mathbf{x}_1^0, s_1 = 0), (\mathbf{x}_2^0, s_2 = 0), \dots, (\mathbf{x}_m^0, s_m = 0)\}$ of m data points, respectively. If patients are selected non-randomly in the study population then \mathbf{D}^1 forms a biased distribution of true distribution $\mathbf{X} \times \mathbf{s} \times \mathbf{y}$, and leads to SSB. Thus, SSB has a distribution mismatch between the study population and the target population. Zadrozny [66] first formalised SSB in machine learning and discussed four cases based on the dependence of s on (\mathbf{x}, y) , and also aligned these cases to missingness, treating SSB as a missingness handling issue, since the outcome is missing for the non-selected patients.

- i. If s is independent of (\mathbf{x}, y) then the study population is representative of the target population, and there is no SSB [51]. This is a case of outcomes of the non-selected patients missing-completely-at-random.
- ii. If s depends only on \mathbf{x} , i.e., $P(s|\mathbf{x}, y) = P(s|\mathbf{x})$ then SSB exists [41], and this is a case of outcomes of the non-selected patients missing-at-random.
- iii. If s depends only on y , i.e., $P(s|\mathbf{x}, y) = P(s|y)$ then SSB exists, and this is a case where prior probabilities of the outcomes change [5].
- iv. If there is no independence between \mathbf{x} , y , and s , and the selection depends on an observed set of features \mathbf{x}_s , then SSB exists [10]. This hinders our ability to construct a reliable mapping from features to outcomes unless we can access an additional set of features \mathbf{x}_s , ensuring that $P(s|\mathbf{x}_s, \mathbf{x}, y) = P(s|\mathbf{x}_s)$ for all examples, including those with $s = 0$. This is a case of outcomes of the non-selected patients missing-not-at-random.

In our study, we focus on the case (ii) since it is the most important SSB case in practice and it is more realistic to assume the case (ii) than to assume the case (i) [66]. Secondly, this is the most widely assumed case, while using multiple imputations to address SSB, in statistical and clinical literature [41]. SSB has connections to other machine learning problems, including missingness and covariate shift or domain adaption. However, SSB is a bit different to standard missingness handling problems where some attributes of a patient are missing while other patients from the same population have those attributes [12]. SSB is a special case of missingness problem where the outcome is missing rather than attributes [19], and it may be difficult to impute the missing labels of the non-selected patients as labels of the entire subpopulation are missing. SSB also intertwines with the broader concept of domain adaptation. Domain adaptation ensures machine learning algorithms trained on a source domain can perform well in a different target domain with varying feature distributions [39]. Specifically, covariate shift addresses differences in input feature distributions between source and target domains, i.e., a setting where $P^{\text{source}}(x) \neq P^{\text{target}}(x)$ while everything else remains unchanged. Since for SSB $P(s|\mathbf{x}, y) = P(s|\mathbf{x})$ and $P(\mathbf{x}|s = 1) \neq P(\mathbf{x}|s)$, this implies $P^{\text{study}}(\mathbf{x}) \neq P^{\text{target}}(\mathbf{x})$. Therefore, SSB and covariate shift share fundamental similarities in addressing mismatches in feature distributions between the study population (source domain) and the target population (target domain). However, they originate from different settings because for SSB the target population is supposed to be a super-set of the study population while for covariate shift the target domain is supposed to be different than the source domain.

3 Brief Literature Review

Here, we discuss work on SSB related to machine learning for healthcare as well as to other fields of machine learning.

3.1 SSB in Machine Learning for Healthcare

Wolff et al. [64] developed a critical appraisal tool called the Prediction Algorithm Risk of Bias Assessment Tool (PROBAST), to assess the risk of bias and applicability of diagnostic and prognostic prediction algorithms. This tool poses a set of questions across four domains to identify biases, with the *participant domain* being particularly relevant to SSB. It considers to what extent the study and the target populations match. A question specifically related to SSB is “1.2 Predictors domain: Were all inclusions and exclusions of participants appropriate?”. Several reviews and systematic reviews have recently employed PROBAST to study the risk of bias in risk prediction studies, treating causal and machine learning risk prediction studies similarly. For example, Navarro et al. [44] conducted a systematic review of different biases in machine learning prediction techniques. Out of 152 studies, 63 had a study population not representative of the target population or lacked information for verification, and 89 did not handle censoring and related events or had no available information. de Jong et al. [17] performed a meta-review of 50 reviews, with 2104 models and found that information on the risk of bias per domain was available for 1039 (47%) studies, of which 25% had a high risk of bias in participant selection.

To address SSB in machine learning for healthcare, some studies have developed study-specific techniques. For example, Thakur et al. [57] identified selection bias in early-stage breast cancer prediction due to different observers scoring

different regions of the same tumour. They proposed an automated method using whole-slide analysis to mitigate the bias. Mei and Xia [43] considered selection bias due to the secondary use of regional electronic health records data and proposed a knowledge learning symbiosis framework to incorporate domain knowledge to address the bias. Additionally, in medical statistics and clinical studies, SSB is commonly treated as a missingness problem under missingness-at-random. Missing outcomes, along with other missing attributes of a patient, are imputed using multiple imputations. For example, Liu et al. [41] used imputation for post-menopausal breast cancer in the UK Biobank.

A search on PubMed² for “(risk prediction) AND (machine learning)” lists 14,023 articles. In contrast, a search for “(risk prediction) AND (machine learning) AND ((selection bias) OR (participant bias))” lists only 138 articles, of which 81 were reviews/systematic-reviews/meta-reviews, 45 did not discuss selection bias, seven were irrelevant, two were related to causality and only three discussed SSB with solutions specific to their problem settings. Recently, Yu and Eng [65] reviewed SSB in radiology and discussed two approaches to address SSB. First to use an independent external test dataset to determine if SSB exists, and second to provide as much detail as possible about the target population and the study population. Interestingly, both approaches are based on preventing SSB. These instances highlight that despite the significance of SSB, machine learning for healthcare has scarcely studied SSB.

3.2 SSB in other Fields

SSB was initially discussed in econometrics and received great attention [3, 26, 59] because in econometrics data collection often relies on surveys where participants are self-selected and are not representative of the population [25]. Heckman [28], in a Nobel Prize-winning work, developed a method to correct SSB, using the probability of selection into the sample. However, this method is only applicable to linear regression algorithms. Zdrozny [66] formalised SSB in machine learning and proposed a bias correction technique for classifiers. They used selection ratio $P(s = 1)/P(s = 1|x)$ as a weight for each example in a cost-sensitive learning approach [21, 60]. This is similar to inverse propensity score or inverse probability of weighting (IPW) as used in treatment effects estimation for reweighting examples [30]. IPW is known to destabilise the training for extreme probabilities. Recently, Alaimo Di Loro et al. [1] employed IPW for SSB correction in tax audits and Bradley and Nichols [8] employed IPW for SSB correction in the UK Biobank dataset for neurological imaging cohort. Huang et al. [34] proposed kernel mean matching (KMM) which does not need selection probabilities, and performs reweighting of the study population such that the means of the study population and the target population are close in a reproducing kernel Hilbert space. Later, Cortes et al. [16] presented an SSB correction theory based on distributional stability and analysed bias correction techniques, including KMM.

As discussed in Section 2, covariate shift or domain adaptation is closely related to SSB and techniques developed in domain adaptation can be applied to address SSB. Domain adaptation techniques can be broadly categorised as feature-based learning and instance-based learning [18]. Feature-based learning approach for domain adaptation aligns representations of the study and target populations in the latent space, e.g., Ganin et al. [23] proposed domain-adversarial training of neural networks (DANN) where patient representations are learned agnostic to selection of patient into the study. On the other hand, the instance-based learning approach for domain adaptation reweights loss function to achieve a balance between the study and the target study distributions, e.g., Sugiyama et al. [55] proposed Kullback–Leibler importance estimation procedure (KLIEP) that finds weights which minimise the Kullback-Leibler divergence between the study and the target distributions.

In healthcare, SSB is discussed mostly in the context of causality [9, 53, 42, 40], which is a fundamentally different subject to machine learning as machine learning generally applies a correlational approach [47]. Causality has some overlaps with machine learning as it employs statistical algorithms for causal effects estimation of interventions but these problems are different than prediction problems [11]. Causality has developed several techniques to address SSB, including causal diagrams [24] and inverse propensity score weighting [63]. Banack et al. [2] discusses SSB in ageing research due to survival bias and loss to follow-up for geriatrics research in a causal setting and presents an SSB toolkit consisting of different approaches to address SSB in causality, including causal diagrams, instrument variables, multiple imputation and sensitivity analysis.

Fairness in machine learning predictions is another area linked to SSB, as biased data can perpetuate discriminatory outcomes. However, fairness considers biases based on certain sensitive attributes only. For example, Du et al. [20] discusses SSB for fairness when dependent variable values of a set of samples from training data are missing as a result of another hidden process. Their framework adopts the classic Heckman model [28] for bias correction with Lagrange duality, achieving fairness in regression setting based on a variety of fairness notions. Similarly, to address systematic biases in risk prediction across demographic groups, Hong et al. [32] employed imbalance learning, transfer learning, and federated learning techniques.

²search performed on February 01, 2024.

From the above brief discussion, we find that SSB is well recognised as a threat to the generalisation of machine learning algorithms. Despite one of the major hurdles in the clinical adoption of machine learning algorithms, SSB is scarcely discussed. Moreover, the existing techniques to address SSB are based on bias correction which aligns the distributions between the study and the target populations and may lead to a loss of predictive performance of the prediction algorithms. Thus, there is a need to highlight the potential dangers of SSB in machine learning for healthcare and develop techniques that effectively leverage available data to provide accurate predictions, despite the presence of bias.

4 Target Population Identification

In this section, we present the proposed target population identification (TPI) approach – a new research direction – to address SSB in machine learning for healthcare for developing algorithms that utilise the available biased data to make accurate predictions. The key idea of our proposed approach is that we constrain the underlying machine learning algorithms, at deployment time, to a subpopulation seen by the algorithm during the training process, i.e., to a target subpopulation which is representative of the study population, without requiring to balance distributions of the two populations. This differs from the existing SSB correction approaches which align the distributions of the study and the target populations, resulting in a loss of predictive performance of algorithms. Our novel approach is motivated by the observation that in healthcare the study and the target populations differ in terms of the non-selected patients. This is unlike domain adaptation where the target dataset is different to the source dataset and the aim is to perform better on the target dataset using information from the source dataset [39]. So for healthcare, an approach based on the identification of the target subpopulation representative of the study population can help in addressing SSB.

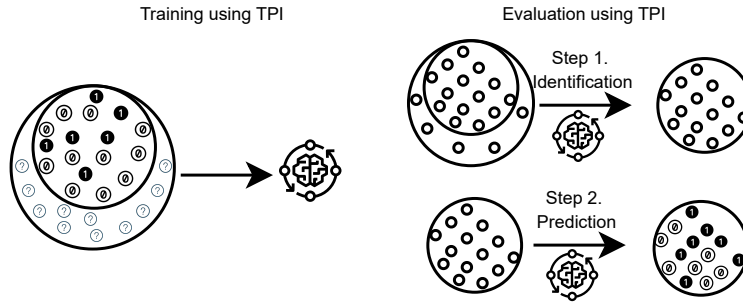


Figure 2: An illustration of the proposed target population identification (TPI) approach which utilises all patients during training and employs a two-step evaluation process where first it identifies the target subpopulation representative of the study population, followed by predictions for the selected subpopulation.

At training time (or development time), the proposed TPI approach, in addition to learning the underlying risk prediction task from the study population, learns the selection process of including patients in the study population using patient attributes, i.e., $P(s = 1|\mathbf{x})$ using included as well as excluded patients. At test time (or deployment time), the proposed approach employs a two-step evaluation process. First, it identifies the target patient subpopulation that is representative of the study population, and then it applies the risk prediction algorithm to the patients identified in the first step, as depicted in Fig. 2. Thus, the proposed approach is based on the assumption that the selection process can be identified from patient characteristics \mathbf{x} . In other words, the proposed approach assumes that the outcomes of the non-selected patients are missing-at-random and it refers to case (ii) as discussed in Section 2. Moreover, the learning of the selection process has access to more data than the underlying risk prediction task because it considers included as well as excluded patients.

To implement the TPI approach, we propose two techniques called T-Net and MT-Net, as described in the following subsections.

4.1 T-Net

To address SSB using the TPI approach, T-Net develops two independent neural networks, one for identification, i.e., to predict the selection/inclusion of a patient in the study population and the other network for the risk prediction task. The selection network trains a classifier $s = f : \mathbf{x} \rightarrow \{0, 1\}$ over patients included as well as excluded from the study, where $s = 1$ for patients included in the study population, i.e., $\mathbf{x} \in \mathbf{X}^1$ and $s = 0$ for patients excluded from the study due to missing outcomes, i.e., $\mathbf{x} \in \mathbf{X}^0$. On the other hand, the risk prediction network trains a separate

neural network $y = g : \mathbf{x} \rightarrow y$ over the study population (only patients included in the study), i.e., $\mathbf{x} \in \mathbf{X}^1$. T-Net employs a two-step evaluation where first the selection network identifies the target subpopulation belonging to the study population, followed by predictions by the risk predictor for the patients selected in the first step. Fig. 3(a) depicts the working of T-Net. Conversely, employing a naive machine learning algorithm would involve training the risk predictor $g(\mathbf{x})$ solely on data from selected patients, intending to make predictions for the entire target population. However, this approach may result in misclassifications for patients within the non-selected subpopulation, potentially leading to unnecessary or inadequate interventions. The risk predictor, developed on a biased dataset, lacks information about the excluded subpopulation, contributing to these shortcomings.

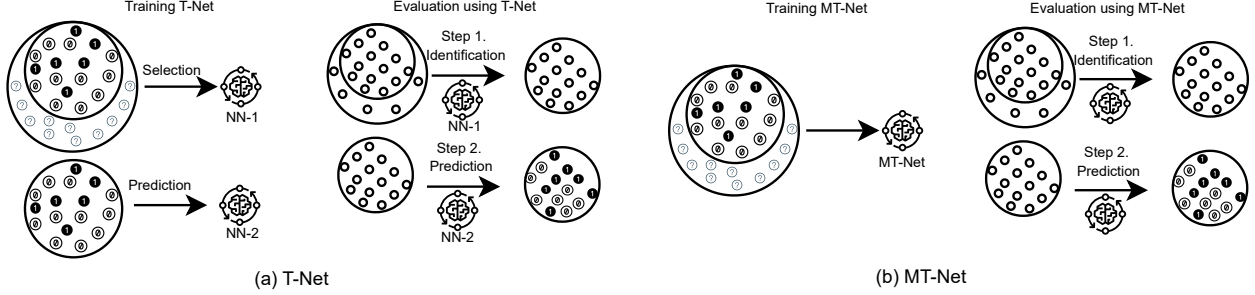


Figure 3: An illustration of the proposed T-Net and MT-Net, as specific techniques under target population identification (TPI), to address SSB. T-Net trains two independent neural networks (NN) while MT-Net uses a single multitasking network.

Hence, T-Net makes predictions specifically for the subpopulation on which it was trained, leveraging available information. In contrast, a naive algorithm is compelled to make predictions for a subpopulation it may not have encountered during training, leading to potential inaccuracies due to lack of relevant information.

4.2 MT-Net

To address SSB using the TPI approach, MT-Net develops a multitasking neural network which has representation layers $\phi : \mathbf{x} \rightarrow \mathbb{R}^{d'}$, learning shared patient representation of d' dimensions, followed by two task-specific heads, one to predict the selection $s = f : \phi(\mathbf{x}) \rightarrow \{0, 1\}$ and the other for risk prediction $y = g : \phi(\mathbf{x}) \rightarrow y$. Thus, MT-Net facilitates shared learning between the selection and the risk prediction and benefits from inductive learning (knowledge transfer between tasks with a shared domain) between the tasks. Similar to T-Net, MT-Net employs a two-step evaluation process where the selection task identifies the subpopulation of the target population which is representative of the study population, followed by predictions by the risk prediction task for the patients selected in the first step. Fig. 3(b) depicts the working of MT-Net. On the other hand, a naive multitasking approach, which employs one task each for selection and risk prediction, benefits from shared learning but predicts for the entire target population. Similar to the naive single-task risk predictor, a naive multitasking predictor might lead to unnecessary or insufficient interventions because it's trained on the biased study population. Similar to T-Net, MT-Net makes predictions specifically for the subpopulation on which it was trained. In contrast, a naive multitasking algorithm is compelled to make predictions for a subpopulation it may not have encountered during training, leading to potential inaccuracies due to a lack of relevant information.

Therefore, both TPI approaches, T-Net and MT-Net, train on both the selected and the non-selected patients, employing a two-step evaluation. This ensures predictions are made for the target subpopulation encountered during training, yielding similar performances on both study and target populations—essential for the development and deployment of machine learning algorithms. Thus, the proposed approach not only offers a novel method for addressing SSB but also utilises the data which otherwise is not selected and is particularly crucial for data-limited studies. While T-Net provides flexibility with two separate networks, MT-Net benefits from shared learning, proving advantageous in settings with limited data for tackling SSB.

5 Results

This section discusses the experimental setup of the empirical study and presents results with synthetic and semi-synthetic datasets to analyse the performance of different techniques to address SSB.

5.1 Experimental Setting

Baselines: We have utilised commonly used baselines, including IPW and imputation, and baselines from machine learning, including SSB and domain adaptation, as given below.

Oracle: It presents a case where the study and the target population have the same distributions, and its performance reflects the ideal situation without any bias. Here, an algorithm is trained on a biased dataset and tested on a biased dataset, i.e., the algorithm has seen the target population during the training time. So, the closer the performance of the rest of the algorithms to Oracle the better it is.

Naive: As discussed in Section 4, a naive approach ignores SSB and develops an algorithm with a biased dataset and evaluates with an unbiased dataset as the algorithm does not have any mechanism to correct the bias or identify the target population. So, Naive should exhibit the lowest performance in the presence of SSB. Naive, Oracle and T-Net share an algorithm for risk prediction, however, the three differ in their evaluation process. Naive is evaluated on the entire test dataset, Oracle is evaluated only for patients belonging to the study population, and T-Net identifies the target subpopulation representing the study population and is evaluated only for the identified patients.

MT-Naive: It is a naive multitasking technique to address SSB which similar to MT-Net has two tasks for the selection and the risk prediction. Similar to Naive, it trains the risk prediction task on the biased dataset and is evaluated on the unbiased dataset.

IPW: It first learns the selection probability for each data point followed by cost-weighted training for risk prediction where each data point is weighted by the inverse of the selection probability, as discussed above Section 3.

Imputation: This is the most commonly used approach in statistical and clinical literature to address SSB. This treats SSB as a missingness handling problem to impute outcomes for the non-selected patients, followed by a standard supervised machine learning approach. We utilise the state-of-the-art MissForest algorithm [54] for imputation which is based on multiple imputations and utilises random forest.

KMM: It performs reweighting of the study populations such that the means of the study populations and the target populations are close in a reproducing kernel Hilbert space.

KLIEP: It also performs reweighting but finds weights which minimise the Kullback-Leibler divergence of the study and the target study distributions.

DANN: It learns patient representations agnostic to the selection of patients into the study through adversarial training. This technique, similar to MT-Net, has a multitasking network and access to the included as well as excluded patients, however, the selection network is used for adversarial training and not for the identification.

Datasets: For *synthetic data generation*, we generated a dataset with 25 features, where features \mathbf{X} are sampled from a uniform distribution over $(-10, 10)$. For a uniformly sampled vector \mathbf{a} of dimensions equal to features and a scalar c , outcomes y for risk prediction are generated using the following process,

$$y = (\mathbf{X}\mathbf{a} + c) > 0. \quad (1)$$

To add noise to outcomes, we randomly flip one percent of the outcomes. For introducing SSB to the dataset, we take another uniformly sampled vector \mathbf{b} and a scalar d , and generate selection variable s using the following process,

$$s = y \oplus ((\mathbf{X}\mathbf{b} + d) > 0), \quad (2)$$

where \oplus represents an exclusive OR operation. For *semi-synthetic datasets*, we have access to real-world features as well as outcomes. To generate a selection variable, we follow [16]. For a random projection vector $\mathbf{a} \in [-1, 1]^d$ of dimensions equal to the number of features, and mean of features $\bar{\mathbf{x}}$, the selection probability of a patient \mathbf{x} is calculated as given below.

$$p(s = 1|\mathbf{x}) = \frac{1}{1 + e^{-\mathbf{v}}}, \quad \text{where } \mathbf{v} = \frac{4\mathbf{a} \cdot (\mathbf{x} - \bar{\mathbf{x}})}{\sigma_{\mathbf{a} \cdot (\mathbf{x} - \bar{\mathbf{x}})}}, \quad (3)$$

where $\sigma_{\mathbf{a} \cdot (\mathbf{x} - \bar{\mathbf{x}})}$ is standard deviation of expression $\mathbf{a} \cdot (\mathbf{x} - \bar{\mathbf{x}})$.

We used two semi-synthetic datasets, hereon called Diabetes and COVID-19 datasets. Diabetes dataset is a health-related telephone survey dataset³, collected annually by the Behavioral Risk Factor Surveillance System at the Centers for Disease Control and Prevention. We have used a cleaned, consolidated and balanced version from Kaggle⁴, and the dataset is a binary dataset to predict diabetes, and it has 70,692 data points, 21 features with an event rate of 50%. COVID-19 dataset is an anonymised dataset released by the Mexican government⁵ and accessed from Kaggle⁶. Since the dataset has high missingness, we have excluded records with missing values. This resulted in a complete dataset

³<https://www.kaggle.com/datasets/cdc/behavioral-risk-factor-surveillance-system>

⁴<https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset>

⁵<https://www.gob.mx/salud/documentos/datos-abiertos-152127>

⁶<https://www.kaggle.com/datasets/tanmoyx/covid19-patient-precondition-dataset/data>

comprising 20,352 data points and 18 features, used to predict whether an individual has COVID-19 or not, with an event rate of 41%.

Hyperparameters and Performance Metrics: We employed neural networks for analysing SSB due to their flexibility, expressiveness, and popularity [13, 15]. We split the datasets as 20% of the full dataset as a test dataset, 20% of the remaining as a validation dataset and the rest as a training dataset. We utilised Adam optimiser [37] and tuned the learning rate from 0.0001 and 0.0005. A mini-batch size of 64 data points and 1000 epochs were used, and an early stopping with patience of 10 epochs. A dropout rate of 10% served as regularisation to prevent overfitting. The neural network architectures were tuned for hidden layers from [50], [100], [100, 100], where the number of list elements indicates the number of hidden layers, and values represent the number of neurons in the corresponding hidden layer. For task-specific layers, we tuned from [50], [100]. For a performance metric, we use the Area Under the Receiver Operating Characteristic Curve (AUC) to assess an algorithm’s ability to distinguish between positive and negative instances. It quantifies the area under the curve generated by plotting the true positive rate against the false positive rate at various classification thresholds. A higher AUC (closer to 1) indicates better discrimination performance, while a random classifier has an AUC of 0.5.

All the experiments are implemented in Python using PyTorch [46]. The experiments are executed on an Ubuntu machine (64GB RAM, one NVIDIA GeForce GTX 1080) and each experiment is averaged over ten seeds.

5.2 Analysis with the Synthetic Data

Here, we present a comparative study of various techniques for handling SSB with the synthetic dataset. Since the performance could vary across different event rates, selection rates or dataset sizes, first we present a summary of the performance across different settings using boxplots in Fig. 4, and subsequent subsections study the effect of the selection/event rates and dataset sizes. We consider the event rates and the non-selection rates from [5, 10, 20, 30, 40]%, and dataset sizes from [1000, 2000, 3000, 4000, 5000]⁷, where each setting is repeated and averaged over 10 seeds. We observe large variations in the performance of all the techniques due to diverse settings, where extreme cases like low event rates, low non-selection rates or small dataset sizes could lead to a large drop in the performance.

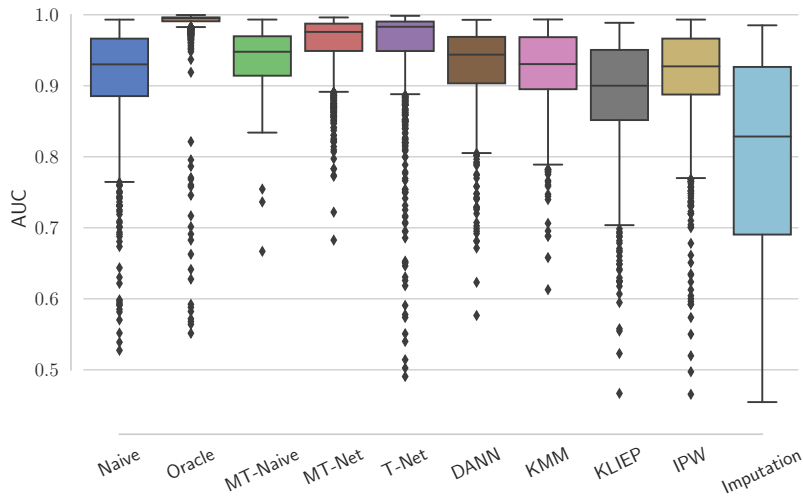


Figure 4: Comparative analysis of the average performance of various techniques for handling SSB using the synthetic dataset across the event rates and the non-selection rates ranging from 5% to 40%, and dataset sizes ranging from 1000 to 5000.

Upon analysing the performance of various techniques, we note an average drop of over seven percent for Naive technique compared to Oracle, which is indicative of reduced performance on the target population compared to the study population. This is because Oracle is evaluated on the target population which is representative of the study population, however, Naive has SSB. Oracle performs the best in terms of median performance, as it represents a bias-free setting, and Imputation technique performs the worst in terms of median performance as well as variation in the performance. Oracle is followed by the proposed T-Net and MT-Net, however, MT-Net shows the least variation

⁷For a dataset size of 1000 with a non-selection rate and an event rate of 10% each, it implies there are 100 non-selected patients and 90 of the remaining 900 have positive outcomes while the rest has negative outcomes.

in the performance among all the techniques, including Oracle, across all the settings. The smaller variation in the performance of MT-Net occurs because it has access to included as well as excluded patients and it employs shared learning between the tasks, which is helpful with limited data settings. Interestingly, the performance of KLIEP and Imputation fall below Naive technique which indicates that the bias handling negatively impacts the performance. Also, a naive multitasking, i.e., MT-Naive improves over Naive and performs better than the rest of the baselines. We study the effect of the non-selection rates, the event rates and dataset sizes on the performance of various SSB handling techniques in the following subsections.

5.2.1 Effect of the Non-selection Rates and the Event Rates

To study the effect of the non-selection rates and the event rates on the performance of different techniques to address SSB, we consider the event rates and the non-selection rates from [5, 10, 20, 30, 40]%, and present a summary of the performance averaged over dataset sizes from [1000, 2000, 3000, 4000, 5000]. For the sake of clarity, we consider the proposed techniques and two best baselines, i.e., DANN and KMM, along with the naive techniques, and present differences in their performance from Oracle as heatmaps in Fig. 5.

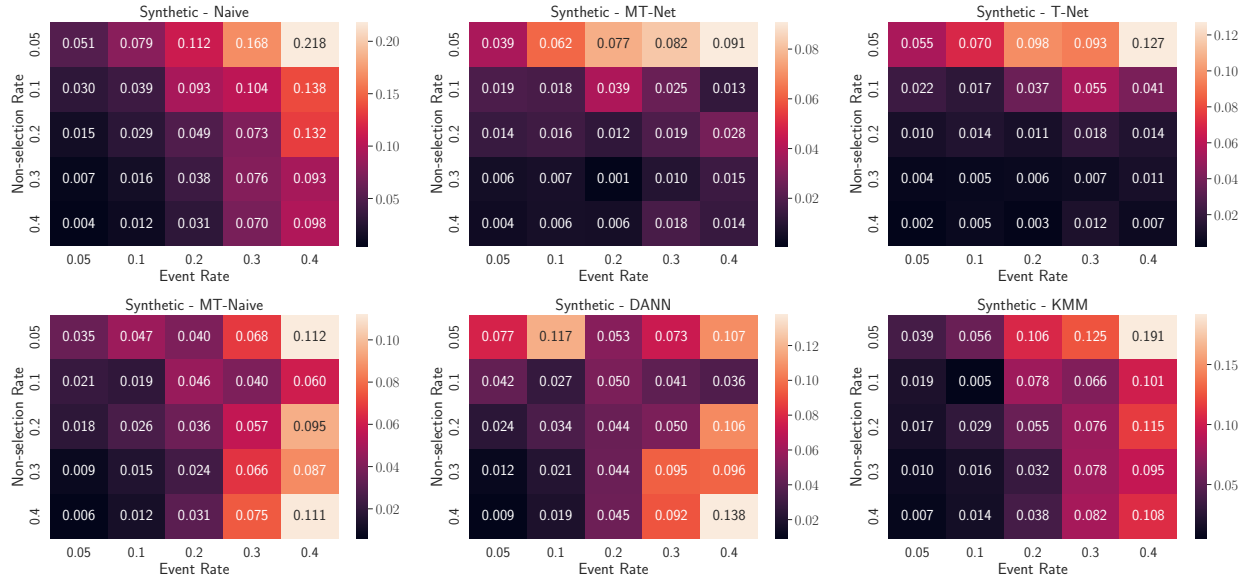


Figure 5: Effect of the event rates and the non-selection rates on the performance of different techniques to address SSB using the synthetic dataset, as averaged over dataset sizes ranging from 1000 to 5000.

From Fig. 5, it is clear that overall MT-Net and T-Net perform better than the baselines as the difference with Oracle is lesser than the baselines. The relative performance of Naive technique drops up to 22% for a low non-selection rate and a high event rate, while the lowest drop occurs at a high non-selection rate and a low event rate. This underlies the importance of addressing SSB in machine learning for healthcare as machine learning algorithms showing high performance may not show similar performance when deployed due to SSB. It is interesting to note that for some settings, like a high non-selection rate and a low event rate, all four techniques perform closely to Oracle and then the relative performance decreases as the non-selection rate decreases or the event rate increases, where the baselines show more drop in the performance as compared with the proposed techniques. The worst performance occurs at a low non-selection rate and a high event rate for all techniques, except DANN which shows a slightly different pattern at an event rate of 0.3 and 0.4. Since we keep the total number of data points fixed while we change one of the two rates, the pattern of a decrease in the relative performance with a decrease in the non-selection rate can be accounted for less available data to address SSB as well as an increase in imbalance for non-selection rate. On the other hand, the pattern of a decrease in the relative performance with an increase in the event rate, i.e., a decrease in class imbalance, for the event rate across all the techniques is a bit counter-intuitive, as generally, the performance improves with improvement in class imbalance. Further analysing, we observe that the performance of individual techniques does increase with an increase in the event rate, however, the relative performance presented in the heatmaps shows a decrease because Oracle shows less variability with event rate than the rest of the techniques resulting in a decrease in the performance with an increase in event rate. These results also show a consistency of the proposed techniques across different settings as compared with the baselines. Overall, Naive performs the worst and MT-Net performs the best to address SSB.

5.2.2 Effect of Dataset Size

To study the effect of dataset size on the performance of different techniques to address SSB with the synthetic dataset, we consider dataset sizes from [1000, 2000, 3000, 4000, 5000], and present a summary of the performance in Fig. 6 averaged over the event rates and the non-selection rates from [5, 10, 20, 30, 40] %.

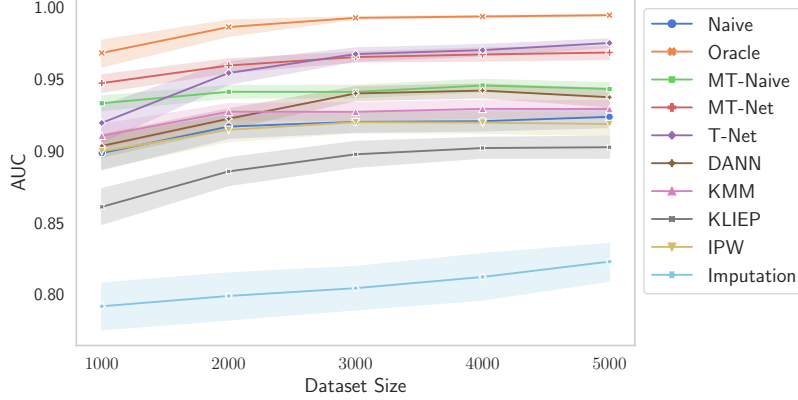


Figure 6: Effect of dataset size on the performance of different techniques to address SSB using the synthetic dataset, as averaged across event rates and non-selection rates ranging from 5% to 40% (shaded region represents one standard deviation over 10 runs).

From Fig. 6, we observe that Oracle is the best-performing technique to address SSB, as it is bias-free, while Imputation is the worst-performing technique. MT-Net and T-Net outperform the rest of the baselines across all dataset sizes, except for small datasets where MT-Naive performs better than T-Net, though MT-Net still outperforms MT-Naive. As observed earlier, MT-Net is more effective than T-Net for smaller datasets due to access to the included and the excluded patients and sharing of learning between the tasks. Interestingly, KLIEP and Imputation techniques perform worse than Naive technique which indicates that the bias correction fails to correct the bias and acts negatively resulting in a loss of performance. Another notable observation is that the naive multitasking, i.e., MT-Naive performs better than the rest of the baselines, including Naive technique. Overall, as expected, the performance of all the techniques for addressing SSB improves with an increase in the dataset size, however, once there are sufficient patients, there isn't much improvement. Similarly, deviation in the performance also improves with an increase in dataset size.

5.3 Analysis with the Semi-synthetic Data

Here, we present a comparative study of various techniques for handling SSB with two semi-synthetic datasets for COVID-19 and Diabetes. We need semi-synthetic datasets to evaluate the performance because in practice it is not possible to know the outcomes of the non-selected patients, as that's why they were dropped. For the semi-synthetic datasets, features and outcomes come from a real dataset but SSB is artificially introduced as discussed in Section 5.1. Similar to the synthetic dataset, first we present a summary of the performance across different settings using boxplots in Fig. 7. We consider the event rates and the non-selection rates from [5, 10, 20, 30, 40] %, and dataset sizes from [1000, 2000, 5000, 10,000, 15,000/25,000] for COVID-19/Diabetes datasets, where each setting is repeated and averaged over 10 seeds. We observe large variations in the performance of all the techniques due to diverse settings, where extreme cases like low event rates, low non-selection rates or small dataset sizes could lead to a large drop in the performance.

On analysing the performance of Naive technique as compared with Oracle, we observe an average drop of two percent and three percent for COVID-19 and Diabetes, respectively. Oracle performs the best in terms of median performance for COVID-19, however, for Diabetes dataset, KLIEP and DANN show performance similar to Oracle. Oracle is closely followed by T-Net and MT-Net, however, MT-Net shows the least variation in the performance among all the techniques across all the settings. The smaller variation in the performance of MT-Net occurs because it has access to included as well as excluded data points, resulting in shared learning between the tasks, which is helpful with limited data settings. Similar to the synthetic dataset, Imputation shows the worst performance both in terms of mean performance as well as variation in the performance. Interestingly, the performance of KMM and Imputation for COVID-19 and MT-Naive, DANN and Imputation for Diabetes falls below Naive method which indicates that the bias handling negatively impacts the performance. We study the effect of the non-selection rates, event rates and dataset sizes in the following subsections.

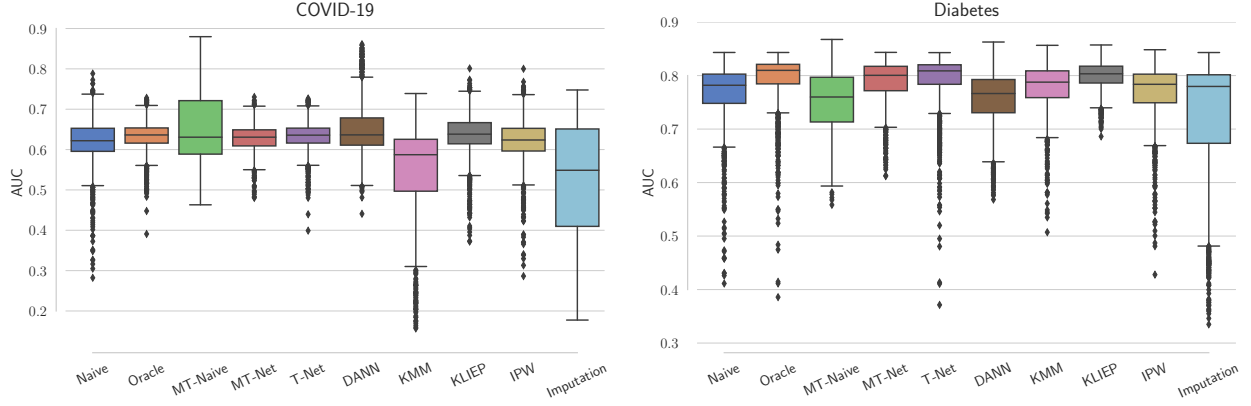


Figure 7: Comparative analysis of the average performance of various techniques for handling SSB using COVID-19 and Diabetes datasets across event rates and non-selection rates ranging from 5% to 40%, and dataset sizes ranging from 1000 to 15,000/25,000.

5.3.1 Effect of the Non-selection Rates and the Event Rates

To study the effect of the non-selection rates and the event rates on the performance of different techniques to address SSB, we consider the event rates and the non-selection rates from [5, 10, 20, 30, 40]%, and present a summary of the performance averaged over dataset sizes from [1000, 2000, 5000, 10,000, 15,000/25,000] for COVID-19/Diabetes datasets. For the sake of clarity, we consider the proposed techniques and two best baselines, i.e., DANN and KLIEP for COVID-19 and KMM and KLIEP for Diabetes datasets, along with the naive techniques, and present differences in their performance from Oracle as heatmaps in Fig. 8.

From Fig. 8, it is clear that overall MT-Net and T-Net perform better than the baselines as the difference with Oracle is lesser than the baselines. The relative performance of Naive technique drops more than 6% for COVID-19 and more than 10% for Diabetes datasets. As noted earlier, this underlies the importance of addressing SSB in machine learning for healthcare. Similar to the synthetic dataset, overall, the performance difference from Oracle decreases as the event rate increases or the non-selection rate decreases, however, the non-selection rate has more effect on the performance than the event rate. Interestingly, all the techniques show improvements over Oracle in certain settings, as indicated by negative values in heatmaps, and in those settings, MT-Naive performs the best. This happens when the non-selected patients lie on the correct side of the decision boundary and away from the decision boundary that they do not have any role in deciding the decision boundary. In such cases, the naive techniques will make correct predictions even if they have not seen those patients during the training, while all the bias handling techniques would lead to some loss of performance. Once again, these results show a consistency of the proposed techniques across different settings as compared with the baselines.

5.3.2 Effect of Dataset Size

To study the effect of dataset size on the performance of different techniques to address SSB, we consider dataset sizes from [1000, 2000, 3000, 4000, 5000], and present a summary of the performance in Fig. 9 averaged over the event rates and the non-selection rates from [5, 10, 20, 30, 40]%.

From the figure, we observe that MT-Net and T-Net techniques perform close to Oracle, which is the desired behaviour when we are addressing SSB in the dataset. Similar to the synthetic dataset, Imputation performs the worst for both the datasets, and Oracle performs the best on Diabetes dataset but for COVID-19 dataset MT-Naive performs the best. For COVID-19 dataset, KMM and Imputation perform worse than Naive, while for Diabetes dataset, MT-Naive, Imputation and DANN perform worse than Naive. This indicates the bias correction techniques fail and negatively impact the performance. As observed earlier, MT-Net is more effective than T-Net for smaller datasets due to access to included and excluded patients and from sharing of information. The proposed techniques perform consistently across settings and datasets, while the baselines show huge fluctuations, for example, MT-Naive outperforms the rest for COVID-19 but falls even below Naive for Diabetes. Similarly, KLIEP shows good performance for smaller datasets as observed for COVID-19 and Diabetes but lagged for the synthetic dataset. Overall, as expected, the performance of all the techniques for addressing SSB improves with an increase in the dataset size.

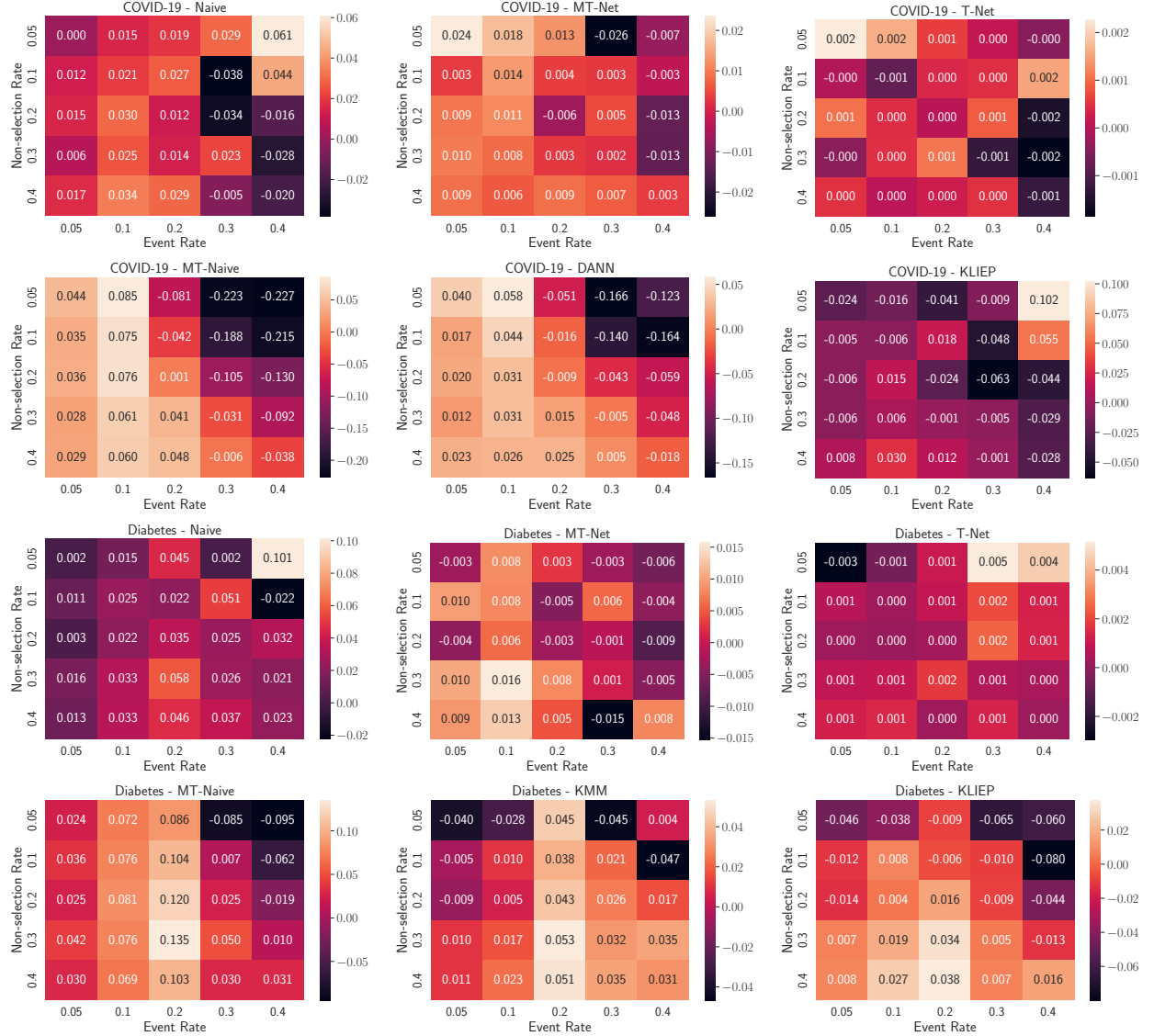


Figure 8: Effect of event rates and non-selection rates on the performance of different techniques to address SSB using COVID-19 and Diabetes datasets, as averaged over dataset sizes ranging from 1000 to 15,000/25,000.

5.4 Effect of SSB on the Non-selected Population

We analyse the performance of the best SSB handling baselines to study their performance for the selected, the non-selected and all the patients of the study with the synthetic and semi-synthetic datasets, and present results in Fig. 10. From the figure, we observe a large difference in the performance of the SSB handling techniques between the selected subpopulation compared to the non-selected subpopulation. We observe the largest difference in the performance for the synthetic dataset where the mean difference in the performance is 43%, 56% and 51% for DANN, KLIEP and KMM, respectively. For COVID-19, the mean performance difference is 5%, 8% and 7% for DANN, KLIEP and KMM, respectively, and for Diabetes, the mean performance difference is 33%, 15% and 16% for DANN, KLIEP and KMM, respectively. For the synthetic dataset, on one side the SSB handling techniques show near perfect performance for the patients selected in the study population but on the other side for the non-selected subpopulation, these techniques show very poor performance, close to 50% which show random predictions for the non-selected patient population. For the synthetic dataset, the performance for the non-selected subpopulation is always lesser than the included subpopulation, however for the semi-synthetic datasets, there is a larger variation in the performance compared to the synthetic dataset. We did not find any consistent pattern in the performance of the baselines which could show good performance for the non-selected subpopulation as well as the selected subpopulation. DANN shows the best performance for the synthetic

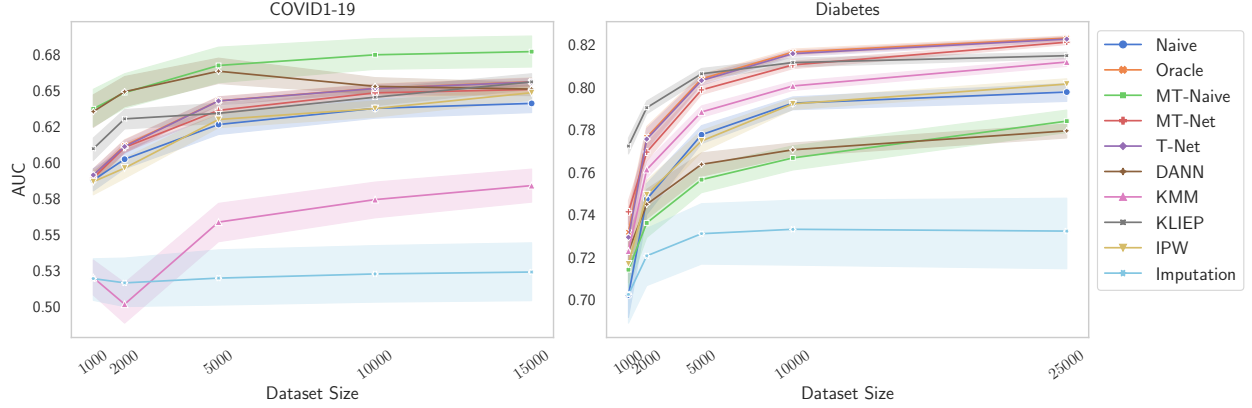


Figure 9: Effect of dataset size on the performance of different techniques to address SSB using COVID-19 and Diabetes datasets, as averaged across event rates and non-selection rates ranging from 5% to 40% (shaded region represents one standard deviation over ten runs).

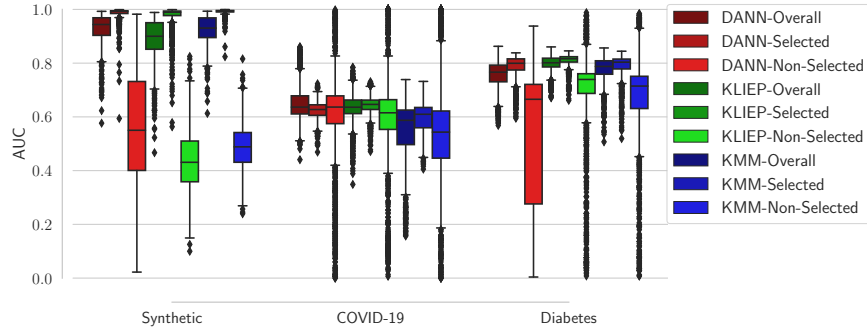


Figure 10: Analysis of the average performance for the selected, the non-selected and the overall population for the best baselines using the synthetic, COVID-19 and Diabetes datasets averaged over the event rates and the non-selection rates ranging from 5% to 40%, and dataset sizes ranging from 1000 to 5000/15,000/25,000.

and COVID-19 datasets but performs the worst for Diabetes dataset. Thus, the baselines lead to a poor performance for the non-selected subpopulation, resulting in a strong bias, for the non-selected subpopulation compared to the selected subpopulation. This would lead to harmful decisions for the non-selected subpopulation. Thus, these results highlight the dangers and the need to address SSB in machine learning algorithms for healthcare.

5.5 Analysis of the TPI Approach

Here, we analyse the performance of the proposed MT-Net and T-Net techniques, measured using AUC, to identify the target population with the synthetic, COVID-19 and Diabetes datasets, and present results in Fig. 11. We observe that overall MT-Net and T-Net techniques show good performance in identifying the target population. They show better performance on the semi-synthetic datasets as the performance drops from one to three percent from the perfect score, while they show a drop of three to 15% from the perfect score for the synthetic dataset. This is potentially due to the large dataset sizes of the semi-synthetic datasets and using different mechanisms for data generation. This figure also explains the overall pattern in the performance of the proposed techniques as their performance drops as their ability to identify the target population drops. In general, we observe that the performance improves with a decrease in the non-selection rate or an increase in the event rate. However, the effect of the event rate is more pronounced as compared to the non-selection rate. The effect of the non-selection rate is a bit unexpected, however, the effect of the event rate is as expected because, with an increase in the event rate, the class imbalance decreases. MT-Net and T-Net show similar performance for different studied settings.

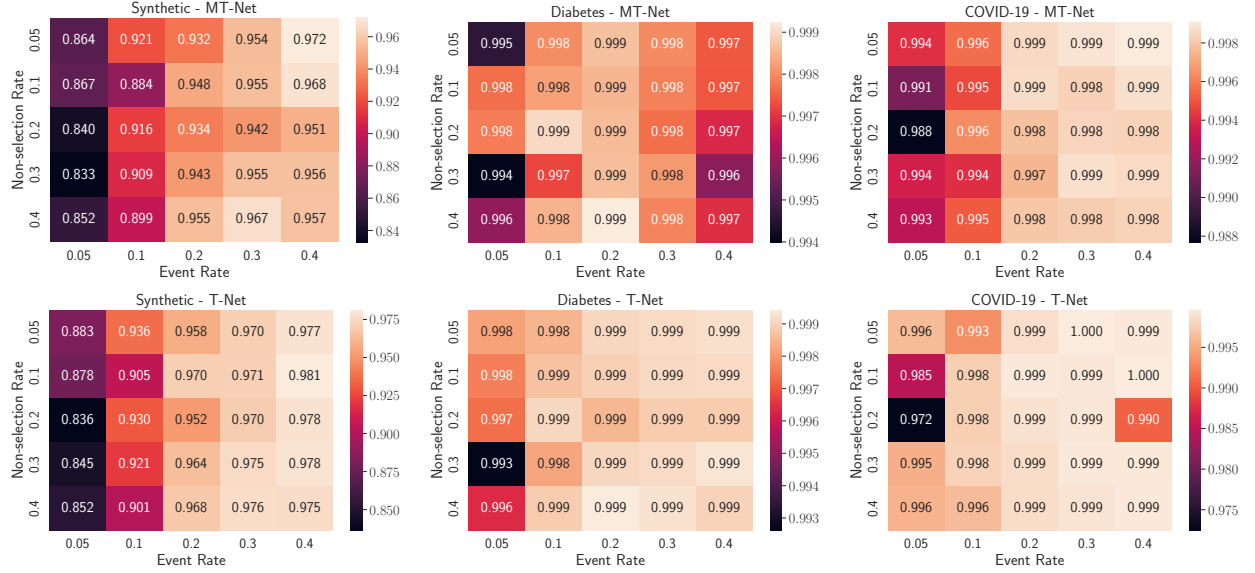


Figure 11: Analysis of the average performance of MT-Net and T-Net in AUC to identify the target population with the synthetic and the semi-synthetic datasets.

6 Discussion

SSB is a result of a non-random process to select patients in the study population while employing a random process to select patients in the target population, resulting in a distribution shift between the study population and the target population. The algorithms developed with the biased study population could lead to inaccurate predictions and harmful decisions for patients not selected in the study population. Despite being well known for decades and a threat to the validity of results, SSB is scarcely addressed in machine learning for healthcare. On the other hand, the existing techniques to address SSB, developed primarily in machine learning, are based on the bias correction approach where distributions of the study population and the target population are balanced to be similar. The existing balancing-based approach leads to the loss of predictive power of the algorithms, resulting in a loss of performance. Thus, the main objective of our work is to highlight the potential dangers of SSB, if unaddressed, on the performance of machine learning algorithms, and propose a new research direction for addressing the limitations of the existing techniques.

Our literature review and empirical study results converge on a crucial point that the presence of SSB in machine learning for healthcare poses significant risks to algorithmic performance and fairness. Several reviews have highlighted this issue, emphasising the potential harms associated with unaddressed SSB. For example, a meta-review of 50 reviews consisting of 2104 models by de Jong et al. [17] found that information on the risk of bias per domain was available for 1039 (47%) studies, of which 25% have a high risk of bias in participant selection. Our analysis of synthetic and semi-synthetic datasets aligns with these concerns. We observe that SSB could lead to a large overall performance drop of up to 22% in machine learning algorithms and a substantial performance gap of more than 40% between the non-selected subpopulation and the selected subpopulation, showing inaccurate and biased predictions for the non-selected subpopulation. These findings highlight the critical importance of addressing SSB in machine learning algorithms for healthcare.

The existing techniques, borrowed from other machine learning areas like domain-adaptation, are mostly based on the idea of bias correction and try to match distributions between the study population and the target population, leading to a loss of predictive performance of the machine learning algorithms. To address this issue, we propose a new research direction to address SSB based on the TPI, rather than bias correction. Here, an algorithm would focus first on identifying the target subpopulation representative of the study population and seen by the algorithm during the training time, and then making predictions for the identified subpopulation, while deferring the rest of the patients to clinicians. This approach bears resemblance to uncertainty quantification where predictions are generated only for those patients where an algorithm is confident otherwise the patients are referred to a clinician [38, 12]. Uncertainty quantification and TPI can also be regarded as subcategories within the broader framework of machine learning with rejection, where algorithms refrain from making predictions when they are likely to err [29]. The proposed approach is motivated by healthcare where certain patient populations fail to complete a study and the target population would contain any patients, including patient groups who failed to complete the study. This setting is unlike domain adaptation

where the target dataset is supposed to be different from the source and the algorithms are supposed to work with the target dataset. Additionally, compared to the naive algorithms, the proposed approach can utilise the non-selected patients for training machine learning algorithms which could be helpful for limited data settings.

We propose two concrete techniques under the proposed TPI approach, called T-Net and MT-Net, which use two independent networks and a multitasking approach, respectively. They develop one network/task for predicting the selection of a patient in the study population and a second network/task for the risk prediction task. T-Net has the flexibility to be more expressive than MT-Net as it develops two independent neural networks for the selection and the risk prediction task. On the other hand, MT-Net has the advantage of information sharing across the two tasks due to shared learning which benefits from inductive learning. As our analysis finds, these properties of the proposed techniques make them suitable for certain settings, for example, MT-Net is better suited for limited data settings and settings with low non-selection rates, while T-Net is better suited for the rest of the settings.

Our analysis of the synthetic and the semi-synthetic datasets shows that the proposed SSB handling techniques are very effective across all settings and show consistent results for different selection rates, event rates, dataset sizes and datasets. The proposed techniques always improve over Naive technique while the baselines sometimes fall below Naive, indicating a negative effect of bias correction. For limited data settings, MT-Net is more effective than T-Net.

It is interesting to note that SSB might not always lead to inaccurate predictions for the non-selected population. This is because SSB impacts the performance of machine learning algorithms only when the non-selected patients play an active role in deciding the decision boundary and unequally favour different outcomes. Suppose, the non-selected patients lie away from the decision boundary but on the correct side then there will be no impact of SSB on the performance of algorithms, even when trained on the biased dataset. This means that by definition, Naive method will perform better than all the baselines because the predictions will be correct for the non-selected patients, resulting in better performance for Naive method. Moreover, all bias handling techniques would lead to a loss of performance. Thus, in such a setting, Naive method, i.e., ignoring the bias would be the best technique to deal with SSB. We can easily identify such a setting, and suggest the community to always develop Naive method, in addition to SSB handling techniques, and if Naive shows improvement in its performance on the target population as compared with the study population then that is an indicator that SSB does not affect the decision boundary and has no impact on algorithms.

Limitations and Future Work: The proposed TPI approach operates under the assumption, akin to IPW and imputation methods, that we can infer a patient’s selection or non-selection into a study population based on their attributes. Essentially, we assume a missing-at-random mechanism for outcomes of the non-selected patients, a common practice in statistical and clinical studies employing imputation techniques to address SSB. Moreover, similar to IPW, the proposed techniques work in two steps, where the first step identifies the target subpopulation representative of the study population, followed by predictions for the identified subpopulation. So, any misspecification in the first step would lead to errors in the second step. However, since the identification task predicts the selection of patients for inclusion in the study, it has access to patients included as well as excluded in the study, and hence relative to the prediction task it has more access to data that would reduce errors in learning.

The proposed techniques make predictions only for the target subpopulation representative of the study population, i.e., the selected patients. However, we want to argue that the non-random sample selection is inevitable in healthcare, for example, all the clinical studies have strict inclusion and exclusion criteria (say adult patients) implying that the results apply to the restricted population (only to adult patients). So, as long as the study and the target population are similar, the bias does not occur and the results/algorithms are still useful. It is better to make correct predictions for subpopulations than to make biased predictions for certain subpopulations when considering the entire population. Thus, the proposed SSB handling approach ensures robust and reliable predictions within the defined subpopulation, contributing to more accurate and effective healthcare decision-making.

Another limitation of the proposed work is that it is limited to setting where the target dataset is a superset of the source dataset, and may not work for domain adaptation where the entire target dataset could be different from the source. Hence, it would be interesting to explore settings across disciplines where the proposed approach could be beneficial and will be explored as part of our future work. Validating SSB handling techniques with real-world datasets poses a challenge, as outcomes for patients excluded from the study are unavailable. However, we can evaluate the effectiveness of these techniques by comparing their performance on the target population against the naive machine learning algorithm on the study population. A smaller performance gap indicates better handling of SSB. It’s worth noting that other biases or noise in the target population could contribute to performance differences. Nonetheless, this comparative analysis will offer insight into the relative effectiveness of different techniques. Future research will focus on exploring and validating our proposed methodology with real-world healthcare studies.

7 Conclusion

This paper underscores the critical need to address SSB in machine learning for healthcare, highlighting its potential to compromise algorithm performance and lead to inappropriate interventions for unrepresented subpopulations. To tackle this issue, we propose a novel research direction focusing on identifying the target subpopulation representative of the study population and making predictions specifically for these identified patients. In contrast to existing techniques that focus on correcting the bias, our proposed approach also prevents the loss of predictive accuracy caused by aligning distributions between the study and target populations. Furthermore, we introduce T-Net and MT-Net as implementations of the TPI approach. These models are trained to predict both the underlying task and the selection of patients into the study population. Our empirical results demonstrate that SSB leads to a large drop in performance for the target population compared to the study population and a substantial performance difference for the target subpopulations representative of the selected and the non-selected patients. Additionally, T-Net and MT-Net exhibit consistent performance in addressing SSB across various scenarios, including different non-selection rates, event rates, dataset sizes, and datasets.

Acknowledgements

This work was supported in part by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC) and in part by InnoHK Project Programme 3.2: Human Intelligence and AI Integration (HIAI) for the Prediction and Intervention of CVDs: Warning System at Hong Kong Centre for Cerebro-cardiovascular Health Engineering (COCHE). DAC was supported by an NIHR Research Professorship, an RAEng Research Chair, the InnoHK Hong Kong Centre for Cerebro-cardiovascular Health Engineering (COCHE), the NIHR Oxford Biomedical Research Centre (BRC), and the Pandemic Sciences Institute at the University of Oxford. GN is funded by NIHR (Grant number 302607) for a doctoral research fellowship. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, the Department of Health, the InnoHK – ITC, or the University of Oxford.

Competing Interests

The authors declare that they have no competing interests.

Availability of data and materials

The data used in the study are publicly available from the following links: Diabetes dataset is released by the Center for Disease Control and Prevention and available on Kaggle at <https://www.kaggle.com/datasets/cdc/behavioral-risk-factor-surveillance-system>. A cleaned, consolidated and balanced version is available at <https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset>. COVID-19 dataset is an anonymised dataset released by the Mexican government <https://www.gob.mx/salud/documentos/datos-abiertos-152127> and accessed from Kaggle <https://www.kaggle.com/datasets/tanmoymx/covid19-patient-precondition-dataset/data>.

Availability of code

The complete code used in the study will be released after acceptance at https://github.com/jmdvinodjmd/SSB_MLHC.

References

- [1] Alaimo Di Loro, P., Scacciatelli, D., and Tagliaferri, G. (2023). 2-step gradient boosting approach to selectivity bias correction in tax audit: an application to the vat gap in Italy. *Statistical Methods & Applications*, 32(1):237–270.
- [2] Banack, H. R., Kaufman, J. S., Wactawski-Wende, J., Troen, B. R., and Stovitz, S. D. (2019). Investigating and remediating selection bias in geriatrics research: the selection bias toolkit. *Journal of the American Geriatrics Society*, 67(9):1970–1976.
- [3] Berk, R. A. (1983). An introduction to sample selection bias in sociological data. *American sociological review*, pages 386–398.
- [4] Bia, M., Huber, M., and Laff ers, L. (2023). Double machine learning for sample selection models. *Journal of Business & Economic Statistics*, pages 1–12.

- [5] Bishop, C. (1995). Neural networks for pattern recognition. *Clarendon Press google schola*, 2:223–228.
- [6] Bonita, R., Beaglehole, R., and Kjellström, T. (2006). *Basic epidemiology*. World Health Organization.
- [7] Boonstra, P. S., Little, R. J., West, B. T., Andridge, R. R., and Alvarado-Leiton, F. (2021). A simulation study of diagnostics for selection bias. *Journal of official statistics*, 37(3):751–769.
- [8] Bradley, V. and Nichols, T. E. (2022). Addressing selection bias in the uk biobank neurological imaging cohort. *MedRxiv*, pages 2022–01.
- [9] Breen, R., Choi, S., and Holm, A. (2015). Heterogeneous causal effects and sample selection bias. *Sociological Science*, 2(17):351–369.
- [10] Brewer, D., Carlson, A., et al. (2021). Addressing sample selection bias for machine learning methods. Technical report, Department of Economics, University of Missouri.
- [11] Chauhan, V. K., Molaei, S., Tania, M. H., Thakur, A., Zhu, T., and Clifton, D. A. (2023a). Adversarial de-confounding in individualised treatment effects estimation. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206, pages 837–849. PMLR.
- [12] Chauhan, V. K., Thakur, A., O’Donoghue, O., Rohanian, O., Molaei, S., and Clifton, D. A. (2024). Continuous patient state attention model for addressing irregularity in electronic health records. *BMC Medical Informatics and Decision Making*, 24(1):117.
- [13] Chauhan, V. K., Zhou, J., Lu, P., Molaei, S., and Clifton, D. A. (2023b). A brief review of hypernetworks in deep learning. *arXiv preprint arXiv:2306.06955*.
- [14] Christiansen, L. B., Koch, S., Bauman, A., Toftager, M., Petersen, C. B., and Schipperijn, J. (2023). Device-based physical activity measures for population surveillance—issues of selection bias and reactivity. *Frontiers in Sports and Active Living*, 5.
- [15] Christopher M. Bishop, H. B. (2023). *Deep Learning: Foundations and Concepts*. Springer Cham, Switzerland.
- [16] Cortes, C., Mohri, M., Riley, M., and Rostamizadeh, A. (2008). Sample selection bias correction theory. In *International conference on algorithmic learning theory*, pages 38–53. Springer.
- [17] de Jong, Y., Ramspek, C. L., Zoccali, C., Jager, K. J., Dekker, F. W., and van Diepen, M. (2021). Appraising prediction research: a guide and meta-review on bias and applicability assessment using the prediction model risk of bias assessment tool (probast). *Nephrology*, 26(12):939–947.
- [18] de Mathelin, A., Deheeger, F., Richard, G., Mougeot, M., and Vayatis, N. (2021). Adapt: Awesome domain adaptation python toolbox. *arXiv preprint arXiv:2107.03049*.
- [19] Dost, K. (2022). *Selection Bias Identification and Mitigation With No Ground Truth Information*. PhD thesis, ResearchSpace@ Auckland.
- [20] Du, W., Wu, X., and Tong, H. (2022). Fair regression under sample selection bias. In *IEEE International Conference on Big Data*, pages 1435–1444.
- [21] Elkan, C. (2001). The foundations of cost-sensitive learning. In *International joint conference on artificial intelligence*, volume 17(1), pages 973–978. Lawrence Erlbaum Associates Ltd.
- [22] Ellen, J. G., Matos, J., Viola, M., Gallifant, J., Quion, J., Anthony Celi, L., and Abu Hussein, N. S. (2024). Participant flow diagrams for health equity in ai. *Journal of Biomedical Informatics*, page 104631.
- [23] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. (2016). Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030.
- [24] Greenland, S., Pearl, J., and Robins, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology*, pages 37–48.
- [25] Gronau, R. (1974). Wage comparisons—a selectivity bias. *Journal of political Economy*, 82(6):1119–1143.
- [26] Heckman, J. (1990). Varieties of selection bias. *The American Economic Review*, 80(2):313–318.
- [27] Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. In *Annals of economic and social measurement*, volume 5, number 4, pages 475–492. NBER.
- [28] Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, pages 153–161.
- [29] Hendrickx, K., Perini, L., Van der Plas, D., Meert, W., and Davis, J. (2024). Machine learning with a reject option: A survey. *Machine Learning*, pages 1–38.

- [30] Hernan, M. and Robins, J. (2023). *Causal Inference: What If*. Chapman & Hall/CRC monographs on statistics & applied probability. Taylor & Francis.
- [31] Hernán, M. A., Hernández-Díaz, S., and Robins, J. M. (2004). A structural approach to selection bias. *Epidemiology*, pages 615–625.
- [32] Hong, C., Liu, M., Wojdyla, D. M., Hickey, J., Pencina, M., and Henao, R. (2024). Trans-balance: Reducing demographic disparity for prediction models in the presence of class imbalance. *Journal of Biomedical Informatics*, 149:104532.
- [33] Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H., and Aerts, H. J. (2018). Artificial intelligence in radiology. *Nature Reviews Cancer*, 18(8):500–510.
- [34] Huang, J., Gretton, A., Borgwardt, K., Schölkopf, B., and Smola, A. (2006). Correcting sample selection bias by unlabeled data. *Advances in neural information processing systems*, 19.
- [35] J Banasik, J. C. and Thomas, L. (2003). Sample selection bias in credit scoring models. *Journal of the Operational Research Society*, 54(8):822–832.
- [36] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589.
- [37] Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- [38] Kompa, B., Snoek, J., and Beam, A. L. (2021). Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digital Medicine*, 4(1):4.
- [39] Kouw, W. M. and Loog, M. (2019). A review of domain adaptation without target labels. *IEEE transactions on pattern analysis and machine intelligence*, 43(3):766–785.
- [40] Kundu, R., Shi, X., Morrison, J., and Mukherjee, B. (2023). A framework for understanding selection bias in real-world healthcare data. *arXiv preprint arXiv:2304.04652*.
- [41] Liu, X., Morelli, D., Littlejohns, T. J., Clifton, D. A., and Clifton, L. (2023). Combining machine learning with cox models to identify predictors for incident post-menopausal breast cancer in the uk biobank. *Scientific Reports*, 13(1):9221.
- [42] Long, J. P. and Ha, M. J. (2022). Sample selection bias in evaluation of prediction performance of causal models. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 15(1):5–14.
- [43] Mei, J. and Xia, E. (2019). Knowledge learning symbiosis for developing risk prediction models from regional ehr repositories. In *MEDINFO 2019: Health and Wellbeing e-Networks for All*, pages 258–262. IOS Press.
- [44] Navarro, C. L. A., Damen, J. A. A., Takada, T., Nijman, S. W. J., Dhiman, P., Ma, J., Collins, G. S., Bajpai, R., Riley, R. D., Moons, K. G. M., and Hooft, L. (2021). Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. *BMJ*, 375.
- [45] Parsons, R. E., Liu, X., Collister, J. A., Clifton, D. A., Cairns, B. J., and Clifton, L. (2023). Independent external validation of the qrisk3 cardiovascular disease risk prediction model using uk biobank. *Heart*, 109(22):1690–1697.
- [46] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.
- [47] Pearl, J. and Mackenzie, D. (2018). *The book of why: the new science of cause and effect*. Basic books.
- [48] Porta, M. (2014). *A dictionary of epidemiology*. Oxford university press.
- [49] Rajkomar, A., Dean, J., and Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine*, 380(14):1347–1358.
- [50] Rojas-Saunero, L. P., Glymour, M. M., and Mayeda, E. R. (2023). Selection bias in health research: Quantifying, eliminating, or exacerbating health disparities? *Current Epidemiology Reports*, pages 1–10.
- [51] Salaun, A., Knight, S., Wingfield, L. R., and Zhu, T. (2023). Interpretable machine learning in kidney offering: Multiple outcome prediction for accepted offers. *medRxiv*, pages 2023–08.
- [52] Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244.
- [53] Smith, L. H. (2020). Selection mechanisms and their consequences: understanding and addressing selection bias. *Current Epidemiology Reports*, 7:179–189.

- [54] Stekhoven, D. J. and Bühlmann, P. (2012). Missforest - non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118.
- [55] Sugiyama, M., Nakajima, S., Kashima, H., Buenau, P., and Kawanabe, M. (2007). Direct importance estimation with model selection and its application to covariate shift adaptation. *Advances in neural information processing systems*, 20.
- [56] Swanson, J. M. (2012). The uk biobank and selection bias. *The Lancet*, 380(9837):110.
- [57] Thakur, S. S., Li, H., Chan, A. M., Tudor, R., Bigras, G., Morris, D., Enwere, E. K., and Yang, H. (2018). The use of automated ki67 analysis to predict oncotype dx risk-of-recurrence categories in early-stage breast cancer. *PLoS One*, 13(1):e0188983.
- [58] Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 25(1):44–56.
- [59] Vella, F. (1998). Estimating models with sample selection bias: a survey. *Journal of Human Resources*, pages 127–169.
- [60] Vogel, R., Achab, M., Cléménçon, S., and Tillier, C. (2020). Weighted empirical risk minimization: Sample selection bias correction based on importance sampling.
- [61] Vokinger, K. N., Feuerriegel, S., and Kesselheim, A. S. (2021). Mitigating bias in machine learning for medicine. *Communications medicine*, 1(1):25.
- [62] Wagenaar, D., Hermawan, T., van den Homberg, M. J., Aerts, J. C., Kreibich, H., de Moel, H., and Bouwer, L. M. (2021). Improved transferability of data-driven damage models through sample selection bias correction. *Risk analysis*, 41(1):37–55.
- [63] Weuve, J., Tchetgen, E. J. T., Glymour, M. M., Beck, T. L., Aggarwal, N. T., Wilson, R. S., Evans, D. A., and de Leon, C. F. M. (2012). Accounting for bias due to selective attrition: the example of smoking and cognitive decline. *Epidemiology (Cambridge, Mass.)*, 23(1):119.
- [64] Wolff, R. F., Moons, K. G., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S., Reitsma, J. B., Kleijnen, J., Mallett, S., and Group†, P. (2019). Probast: a tool to assess the risk of bias and applicability of prediction model studies. *Annals of internal medicine*, 170(1):51–58.
- [65] Yu, A. C. and Eng, J. (2020). One algorithm may not fit all: how selection bias affects machine learning performance. *Radiographics*, 40(7):1932–1937.
- [66] Zadrozny, B. (2004). Learning and evaluating classifiers under sample selection bias. In *Proceedings of the twenty-first international conference on Machine learning*, page 114.
- [67] Zhelonkin, M., Genton, M. G., and Ronchetti, E. (2016). Robust inference in sample selection models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(4):805–827.