Drug Safety (2022) 45:521–533
https://doi.org/10.1007/s40264-022-01160-9

ORIGINAL RESEARCH ARTICLE

Identifying and Mitigating Potential Biases in Predicting Drug Approvals

Qingyang Xu[1,4] · Elaheh Ahmadi[2,3] · Alexander Amini[2,3] · Daniela Rus[2,3] · Andrew W. Lo[1,2,3,4,5]

Accepted: 9 February 2022

Abstract

Introduction Machine learning models are increasingly applied to predict the drug development outcomes based on intermediary clinical trial results. A key challenge to this task is to address various forms of bias in the historical drug approval data.
Objective We aimed to identify and mitigate the bias in drug approval predictions and quantify the impacts of debiasing in terms of financial value and drug safety.
Methods We instantiated the Debiasing Variational Autoencoder, the state-of-the-art model for automated debiasing. We trained and evaluated the model on the Citeline dataset provided by Informa Pharma Intelligence to predict the final drug development outcome from phase II trial results.
Results The debiased Debiasing Variational Autoencoder model achieved better performance (measured by the F_1 score 0.48) in predicting the drug development outcomes than its un-debiased baseline (F_1 score 0.25). It had a much higher true-positive rate than baseline (60% vs 15%), while its true-negative rate was slightly lower (88% vs 99%). The Debiasing Variational Autoencoder distinguished between drugs developed by large pharmaceutical firms and those by small biotech companies. The model prediction is strongly influenced by multiple factors such as prior approval of the drug for another indication, whether the trial meets the positive/negative endpoints, and the year when the trial is completed. We estimate that the debiased model generates financial value for the drug developer in six major therapeutic areas, with a range of US$763–1,365 million.
Conclusions Our analysis shows that debiasing improves the financial efficiency of late-stage drug development. From the pharmacovigilance perspective, the debiased model is more likely to identify drugs that are both safe and effective. Meanwhile, it may predict a higher probability of success for drugs with potential adverse effects (because of its lower true-negative rate), thus it must be used with caution to predict the development outcomes of drug candidates currently in the pipeline.

1 Introduction

1.1 Financial Risks in Novel Drug Development

Despite groundbreaking advances in biomedical science and technology, the translational research and development of novel therapeutics has become more expensive and less likely to succeed over the last 5 decades. The number of new drugs approved by the US Food and Drug Administration (FDA) per billion US dollars (US$) spent on translational research and development has halved in inflation-adjusted terms about every 9 years since 1950 [1]. A recent analysis [2] estimates a median cost of US$985.3 million to bring a new drug to market in the period 2009–18, while the historical probability of success (PoS) of developing a novel drug from a phase I clinical trial to FDA approval is merely 10.8% in all therapeutic areas, and as low as 4.0% in oncology [3]. Novel scientific and business models are needed to reduce the financial risks and bridge the funding gap (also known as the "valley of death" [4]) in the translational research and development of novel therapeutics.

1.2 Machine Learning for Drug Approval Prediction

In recent years, machine learning and artificial intelligence have been increasingly applied to forecast the PoS (i.e., FDA

✉ Andrew W. Lo
alo-admin@mit.edu

1 MIT Laboratory for Financial Engineering, Cambridge, MA, USA

2 MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA, USA

3 MIT Department of Electrical Engineering and Computer Science, Cambridge, MA, USA

4 MIT Operations Research Center, Cambridge, MA, USA

5 Sante Fe Institute, Santa Fe, NM, USA

Key Points

There is significant bias present in machine learning models to predict drug development outcomes, largely because of the imbalance in the distributions of drug and clinical trial features and the imbalance in drug development outcomes.

The debiased model has better overall performance in predicting drug development outcomes and predicts safe and effective drug candidates more accurately than its un-debiased counterpart. Meanwhile, it may predict high probabilities of success for drugs with potential adverse effects and must be used with caution.

Debiasing significantly increases the financial value of novel drug development for the drug developer in six major therapeutic areas.

approval) for a novel drug candidate based on early or mid-stage (phase I or II) clinical trial results. Early works [5–8] revealed important factors that correlated with successes and failures. However, these conclusions have been limited by the relatively small size of the datasets, which consist of fewer than 100 drugs or 500 clinical trials. Beinse et al. [9] trained a regularized Cox model with 462 anti-neoplastic agents to predict drug approval from phase I results, although their model is affected by look-ahead bias, as the authors randomly split the data into training and testing sets, training their model with future data but evaluating it with past data. Lo et al. [10] were the first to train machine learning models on the Citeline dataset provided by Informa Pharma Intelligence [11], with more than 91,000 drugs and 374,000 clinical trials.[1] Using the Citeline dataset, Wong et al. [12–14] proposed a path-by-path approach to estimate the PoS of drug candidates in different therapeutic areas, which is robust against missing data. Recently, Siah et al. [15] organized a data science competition with over 50 participating teams to predict drug approvals using the Citeline dataset. The top-performing models in this competition used novel features that were highly predictive of drug development outcomes.

A key challenge to predicting the outcomes of drug development projects is to address the various forms of bias present in the dataset and prediction model. Two major sources of bias are data missingness and dataset imbalance. While data missingness can be addressed by imputation [10], dataset imbalance (i.e., over/underrepresentation in outcome labels and input features) remains a critical challenge that limits the predictive performance of machine learning models (Fig. 1). In its outcome labels, only 11.8% of all drugs in our dataset are approved by the FDA. Thus, the prediction model trained on the imbalanced dataset is incentivized to predict negative outcomes, and has a low true-positive (TP) rate. In the input feature space, imbalance occurs where many drugs have similar properties (e.g., "me too" [16] or repurposed drugs with biochemical features similar to previously approved drugs). As drugs with similar properties are not guaranteed to be effective to treat different diseases, the prediction model has a lower performance when predicting the approval outcomes of overrepresented drugs in the feature space.

1.3 Mitigating the Bias of Machine Learning Models

With the increasing reliance on machine learning models to automate the decision-making process in many applications, the issue of bias and fairness of machine learning models has become a central concern (see [17], for a systematic survey). Previous studies reported and addressed significant racial, gender, and socioeconomic biases in machine learning models applied to domains such as financial loans [18], criminal justice [19, 20], career advertisement [21], medical diagnosis [22, 23], and healthcare policy [24].

Bias in a machine learning model is often caused by biases within an imbalanced training dataset [17]. The biased model performs better on overrepresented subgroups in the dataset than on underrepresented minorities [25–27]. Common strategies used to mitigate algorithmic bias include rebalancing the training dataset by resampling the training data [28, 29], generating synthetic data [30, 31], or changing a subset of the class labels [20]. Each method, however, has its limitations. Resampling requires knowledge of the class imbalance, and is difficult when the class label is latent (e.g., gender or skin color in an image) and needs to be manually annotated. Zhou and Liu [28] showed that resampling methods that are effective for binary classification often do not generalize to multi-class classification. Generating synthetic data requires modeling the distribution of input features, which may be difficult for high-dimensional features, and may produce unrealistic synthetic samples. Recently, Bandi and Bertsimas [20] proposed an optimization framework that flips a subset of binary labels to achieve guaranteed demographic parity, although the model does not justify why certain labels are flipped, and it is evaluated on a testing set that contains samples with flipped labels.

Given these challenges, Amini et al. [27] proposed a novel framework called the Debiasing Variational Autoencoder (DB-VAE), which automatically identifies and mitigates the bias in large datasets without prior knowledge of the detailed

[1] As of 5 December 2021. The Citeline data is continually updated to include new clinical trials.

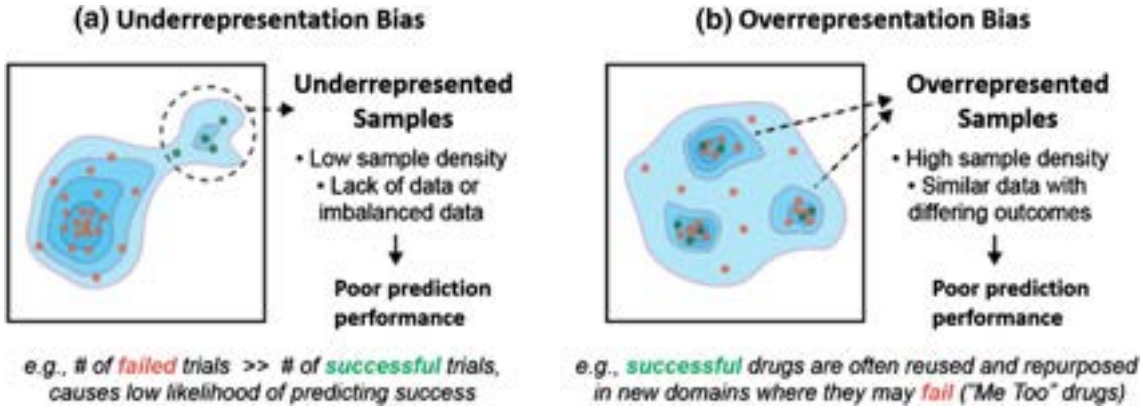

Fig. 1 Sources of bias in the Citeline dataset. (**A**) Underrepresentation in outcome labels with 11.8% of positive samples (green) and (**B**) overrepresentation in drug and clinical trial features (e.g., "me too" or repurposed drugs with similar drug features as previously approved drugs)

structure of bias (e.g., whether certain subgroups are over/underrepresented). The mathematical framework and debiasing mechanism of the DB-VAE are discussed in Sect. 3.2.

1.4 Contributions of this Work

In this work, we simultaneously identify and mitigate various forms of bias by instantiating the DB-VAE for drug approval predictions. The automatic debiasing feature of the DB-VAE is particularly useful in our application, as it is difficult to directly analyze the bias structure of complex drug and clinical trial data. The main contributions of our work are:

1. To the best of our knowledge, this work is the first to systematically address the significant bias present in the machine learning model for drug approval prediction using one of the largest datasets in this domain.
2. We show that the debiased model achieves better overall prediction performance and its prediction performance is more uniform across different subgroups of the dataset than its un-debiased counterpart.
3. We quantify the impact of different forms of bias on the prediction performance. Debiasing the imbalance of drug approval outcomes results in major improvements for all drugs, while debiasing the input feature distributions achieves improvements for drugs that are overrepresented in the input feature space, such as oncology and cardiovascular drugs.
4. We show that debiasing generates significant financial values of drug development in six major therapeutic areas.
5. From the pharmacovigilance perspective, we find that the debiased model predicts safe and effective drugs more accurately than its un-debiased counterpart.

2 Data

We query historical drug development data in the period from 2004 to 2020 (inclusive) in the Citeline data provided by Informa Pharma Intelligence [11]. As a drug developer may conduct multiple clinical trials for one drug to investigate its therapeutic efficacy for different diseases, we predict the binary outcome of whether the drug has been approved by the FDA to treat a particular indication, which we call a "drug-indication pair." For clarity of exposition, we shall refer to "drug-indication pair" simply as "drug" in the subsequent sections. We follow the data query and preprocessing procedures described in [10], and refer the reader to this work for additional details. We form the Phase 2 to Approval dataset, which consists of drug-indication pairs with known approval outcomes (success or failure) and their phase II clinical trial results. We train the machine learning model on the Phase 2 to Approval dataset to predict the binary approval outcome from drug features and phase II trial results. The summary statistics of the Phase 2 to Approval dataset are provided in Table 1. The raw data consist of both categorical and continuous features. A categorical feature may be single-labeled (e.g., whether the drug was previously approved for another indication) or multi-labeled (e.g., the drug developer may conduct clinical trials in different countries for a drug). We apply one-hot encoding to the multi-labeled features and create binary child features. Detailed descriptions of the drug and clinical trial features are summarized in Table S1 of the Electronic Supplementary Material (ESM). In addition, because of the different standards of post-study reporting of clinical trial results (especially before the 2007 FDA Amendments Act) [10], there is considerable missingness in certain drug and clinical trial features (Table S2 of the ESM) that needs to be imputed. We discuss the details of the imputation procedure in Sect. 3.3.

Table 1 Summary statistics of the phase 2 to approval (P2APP) dataset

	Drugs	Indications	Clinical trials	Drug-indication pairs
P2APP (2004–20)				
Approved	685	190	2,320	876
Failed	3,615	292	10,288	6,555
Total[a]	4,079	298	12,397	7,431
Training set (2004–18)				
Approved	595	182	2,060	752
Failed	3,200	275	8,090	5,630
Total	3,612	283	10,035	6,382
Testing set (2019–20)				
Approved	108	81	344	124
Failed	637	195	2,397	925
Total	740	212	2,711	1,049

The training set consists of drug approval outcomes from 2004 to 2018 (both inclusive) and the testing set consists of data from 2019 to 2020

[a]A drug may be approved to treat an indication but fail in other indications. Therefore, the sum of numbers of approved and failed drugs is not equal to the total number of unique drugs. However, the sum of numbers of approved and failed drug-indication pairs is equal to the total number of unique drug-indication pairs

3 Methods

3.1 Algorithm Fairness

We adopt the definition of algorithm fairness proposed in [27]. Given the training dataset $D_{\text{train}} = \left\{\left(x^{(i)}, y^{(i)}\right) : 1 \le i \le n\right\}$, where $x^{(i)} \in R^m$ denotes the m-dimensional feature vector and $y^{(i)} \in \{0, 1\}$ denotes the binary prediction target, our goal is to find a classifier $f : X \to Y$ that is fair with respect to sensitive features $z \in R^d$. The sensitive features, z, may either be observed in D_{train} (e.g., the outcome of drug development) or latent, which means that $z = z(x)$ are not directly observed, but can be represented as a function of the observed features, x. For example, in the computer vision task of face recognition, $x^{(i)}$ is the input image, $y^{(i)}$ denotes whether the image contains a human face, and the latent features $z^{(i)} = z\left(x^{(i)}\right)$ include skin color, gender, and the age of the subject in the image. We define an unbiased classifier f with respect to the sensitive features, z, if $f(x) = f(x, z)$, i.e., the classification decision is not affected by the additional sensitive features, and a biased classifier if $f(x) \neq f(x, z)$ for some z.

As highlighted in [27], in order to train an unbiased classifier, the training samples in D_{train} should be uniformly distributed across the latent feature space, Z. Furthermore, given a classifier f and testing dataset D_{test}, we can measure the bias of f by the variance of its prediction performance across different subgroups in Z. A larger variance indicates greater bias, as the classifier performs more poorly on certain subsets than others. For our purposes, the sensitive features of interest, z, are the drug approval outcomes (observed) and the degree of over/underrepresentation of a drug in the feature space (latent).

3.2 Debiasing via the DB-VAE

We instantiate the DB-VAE [27] to train a debiased classifier for drug approval prediction. The DB-VAE consists of a Variational Autoencoder (VAE) [32], which learns the latent features $z(x)$ and predicts the approval outcome $\hat{y} = f(x)$, coupled with a feedback loop that debiases the training dataset by adaptively adjusting the resampling weights for over- and underrepresented samples. The model architecture of the DB-VAE is shown in Fig. 2. For clarity of presentation, we review the key components of the DB-VAE that are most relevant to our application, and refer the reader to the original work [27] for additional details.

We train the DB-VAE to simultaneously debias the imbalanced drug approval outcomes and the over/underrepresentation in the input feature space. To debias the outcome labels, we enforce that each training batch (of size 32) of the stochastic gradient descent contains an equal number of positive and negative training samples, which ensures that the model is not biased to predict negative outcomes due to overrepresentation in outcome labels.

To debias the input features, we utilize the low-dimensional latent representation $z(x) = \left[z_1(x), ..., z_d(x)\right]$ of the high-dimensional features $x \in R^m$ (with $d \ll m$) learned by the VAE with an encoder-decoder architecture. The encoder network produces the latent features, $z(x)$, as its output. The decoder network then reconstructs the input features, $\hat{x} = \hat{x}(z(x))$, from $z(x)$. To ensure that the reconstructed features are close to their original values, $\hat{x} \approx x$, we use L_1 loss function for reconstructing continuous features (denoted as $L_{\text{r,con}}$) and a weighted binary cross entropy loss (denoted as $L_{\text{r, cat}}$ in Eq. 1) for reconstructing categorical features:

$$L_{\text{r,cat}} = -\frac{1}{N}\sum_{\text{c}\in\text{cat}}\sum_{i=1}^{N}\frac{1-\bar{x}_{\text{c}}}{\bar{x}_{\text{c}}}x_{\text{c}}^{(i)}\log\hat{x}_{\text{c}}^{(i)} + \left(1-x_{\text{c}}^{(i)}\right)\log\left(1-\hat{x}_{\text{c}}^{(i)}\right). \tag{1}$$

Here, $x_{\text{c}}^{(i)}$ denotes the value of categorical feature c of the ith training sample, $\hat{x}_{\text{c}}^{(i)}$ denotes the reconstructed value by DB-VAE, and $\bar{x}_{\text{c}}$ denotes the average value of feature c in the training dataset. This weighted loss achieves a more accurate reconstruction for the subset of categorical features that are exceedingly sparse or dense. For sparse features (with $\bar{x}_{\text{c}}$ close to 0), we assign a higher weight to the positive samples ($x_{\text{c}}^{(i)} = 1$) and increase the TP rate of reconstruction.

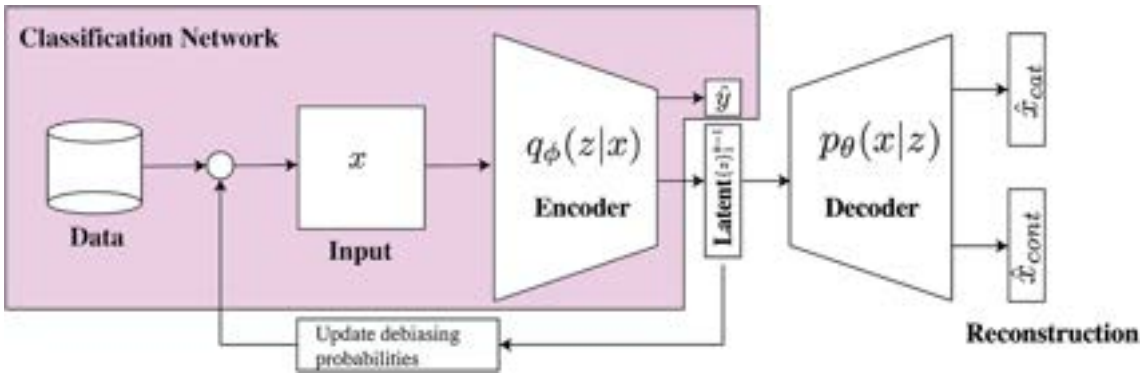

Fig. 2 Architecture of a Debiasing Variational Autoencoder instantiated for predicting drug development outcomes

For dense features (with $\bar{x}_c$ close to 1), we assign a higher weight to the negative samples ($x_c^{(i)} = 0$) and increase the true-negative (TN) rate of reconstruction.

In addition, VAE uses the Kullback–Leibler (KL) divergence [33] to regularize the latent space distribution and prevent overfitting. We denote this regularization loss by L_{KL}. As in [27], we use the encoder to predict the PoS of the drug approval outcome $\hat{y} = z_0(x) \in (0, 1)$, and denote the cross-entropy loss by L_c. The predicted PoS, $\hat{y}$, is not used in debiasing.

Given the latent representation $z(x)$ of every sample x in the training set, X_{train}, we compute the probability density function $Q_i(z_i(x)|X_{\mathrm{train}})$ of each latent dimension $i \in \{1, ..., d\}$ via a histogram with ten bins.[2] The joint probability density function in the d-dimensional latent space is $Q(z(x)|X_{\mathrm{train}}) = \prod_{i=1}^{d} Q_i(z_i(x)|X_{\mathrm{train}})$. Based on the latent space density, $Q(z(x)|X_{\mathrm{train}})$, Amini et al. [27] proposed a debiasing algorithm that assigns higher probabilities of resampling to training samples with lower $Q(z(x)|X_{\mathrm{train}})$ (i.e., samples that are underrepresented in the latent space) into the next training batch and assigns lower resampling probabilities to the overrepresented samples. Specifically, the debiasing resampling weight, $W(z(x)|X_{\mathrm{train}})$, for training sample x is given by:

$$W\left(z(x)|X_{\mathrm{train}}\right) \propto \prod_{i=1}^{d} \frac{1}{Q_i\left(z_i(x)|X_{\mathrm{train}}\right) + \alpha}. \tag{2}$$

The proportionality sign, $\propto$, indicates that the sum of $W(z(x)|X_{\mathrm{train}})$ is normalized to 1. The debiasing smoothing parameter, $\alpha > 0$, controls the degree of debiasing. A smaller value of α corresponds to more aggressive debiasing, as $W(z(x)|X_{\mathrm{train}})$ is mostly determined by the latent space density, $Q_i(z_i(x)|X_{\mathrm{train}})$. However, a large value of $\alpha \gg \max_{i,x} Q_i(z_i(x)|X_{\mathrm{train}})$ corresponds to uniform resampling, and does not debias the latent space distribution. The model parameters are trained by minimizing the loss function.

$$L_{\mathrm{DB-VAE}} = \lambda_1 L_c + \lambda_2 L_{\mathrm{r,\,cat}} + \lambda_3 L_{\mathrm{r,\,con}} + \lambda_4 L_{\mathrm{KL}}, \tag{3}$$

where the weights, $\lambda_i > 0$, are model hyperparameters. We choose the default values $\lambda_1 = 10$, $\lambda_2 = \lambda_3 = 1$, and $\lambda_4 = 0.001$ to reflect the relative importance of each term in the loss function $L_{\mathrm{DB-VAE}}$. A sensitivity analysis (Table S3 of the ESM) shows that the model performance is robust against a wide range of hyperparameter values.

3.3 Training the DB-VAE

Because drug development is a non-stationary process in which drugs approved in the past set higher standards for drug candidates in the future, there will be a significant look-ahead bias if we randomly split the dataset into training and testing sets, training the models on future data but evaluating on past data. To avoid this look-ahead bias, we train the models with historical data from 2004 to 2018, and evaluate the models on out-of-sample data from 2019 to 2020. The model parameters are optimized by minimizing the loss function (Equation 3) via a stochastic gradient descent with 200 training epochs, batch size 32, and learning rate 10^{-5}. To determine the training epochs and learning rate, we observe the evolution of the classification and reconstruction losses (L_c, $L_{\mathrm{r,\,cat}}$, $L_{\mathrm{r,\,con}}$) on a held-out validation set, formed by sampling 10% of the training set randomly without replacement, and terminate the stochastic gradient descent when the losses converge. We impute the missing entries using 5-nearest neighbor imputation [34] for training, testing, and validation sets separately. We apply a log-transform to the continuous features, and normalize each continuous feature to zero mean and unit variance. As most input features are

[2] We use the notation of conditional probability, $Q_i(\cdot|X_{\mathrm{train}})$, to emphasize that the latent space density is sensitive to the distribution of input features in the training dataset, X_{train}.

Table 2 Nomenclature of the Debiasing Variational Autoencoder models

Debiasing mechanism	Original outcome labels	Debiased outcome labels
Original latent space distribution	No-DB-Label, No-DB-Latent	DB-Label, No-DB-Latent
Debiased latent space distribution	No-DB-Label, DB-Latent	DB-Label, DB-Latent

sparse, we only use those features whose variance is above 0.2 before imputation. We implement the DB-VAE model in TensorFlow. The model configuration and hyperparameter values are summarized in Table S3 of the ESM. We perform a sensitivity analysis on the results against different values of model hyperparameters. The results are summarized in Table S4 of the ESM.

To quantify the contributions from the two forms of bias to the prediction performance, we train four instantiations of the DB-VAE, which differ by whether the model debiases the imbalance of outcome labels in each training batch (DB-Label) and whether it debiases the distribution of latent representations of input features (DB-Latent). The nomenclature is summarized in Table 2, and used in the subsequent sections.

4 Results

4.1 Prediction Performance

We evaluate the prediction performance of the trained models on drug development outcomes from 2019 to 2020. We compute the confusion matrix for binary classification, and report the TP rate, TN rate, and F_1 score of each classifier. To quantify the uncertainty from imputing the missing entries, we train 30 instances of a given set of model hyperparameters, each with randomly split training and validation sets, and report the average value of their performance metrics and associated standard errors. We also report the area under receiver operating characteristic curve [35] of each model, which needs to be interpreted with care because of the imbalance in drug approval outcomes (Fig. S1 of the ESM) [36].

The prediction performance is summarized in Table 3. Comparing each DB-Label model with its No-DB-Label counterpart, we find that debiasing the outcome labels significantly improves both the TP rate and F_1 score in all therapeutic areas. Debiasing the latent space distribution (DB-Latent) improves the TP rate in three therapeutic areas (oncology, cardiovascular, and central nervous system), with a slightly lower TN rate. The trade-off between a higher TP rate and a lower TN rate is consistent with the rationale of DB-Latent to achieve more uniform accuracy between different subgroups (approved vs failed drugs) of the dataset [27]. The performance of DB-Latent and No-DB-Latent are similar for autoimmune/inflammation, metabolic, and infectious disease. We conclude that the bias of imbalanced drug approval outcomes is a more severe issue that limits the prediction performance than the bias of over/underrepresentation in the input feature space.

To analyze the effect of debiasing on over/underrepresented drugs in the feature space, we evaluate the DB-VAE models on drugs in each quintile of $W(z(x)|X_{\text{test}})$. The results are shown in Fig. 3. For each DB-Latent model, we use three different values of smoothing parameter α and the results are robust against different values of α. We observe that DB-Latent improves the TP rate over its No-DB-Latent counterpart by 9.4%, 9.3%, 6.8%, and 1.8% in the lowest four quintiles, with the greatest improvement in the two lowest quintiles (i.e., the top 40% most overrepresented drugs in the test set). Debiasing the latent space distribution helps predict successful drugs that are overrepresented in the feature space (e.g., "me too" or repurposed drugs). This has major implications on the financial value of drug approval prediction, as will be shown in Sect. 4.4. The standard deviation of TP rates across the five quintiles is consistently lower for the DB-Label, DB-Latent models (5.1%, 5.8%, 6.5%) than their No-DB-Latent counterpart (8.2%), which confirms that DB-Latent reduces algorithmic bias across over/underrepresented subgroups.

4.2 Feature Importance

We use the saliency score [37] to identify the input features that are the most important for predicting drug approval. The saliency score of a feature is a real number whose magnitude reflects the sensitivity of the model predictions to changes in the feature value and is commonly used to measure feature importance in deep learning models. Table 4 lists the top ten features of the DB-Latent, DB-Label model that have the highest absolute saliency scores. Some of these features (prior approval of the drug for another indication, whether the phase II trial meets the positive/negative endpoints, whether the delivery medium is powder) were identified as important features in previous work [10]. Our debiased model also reveals previously unidentified therapeutic factors such as the two pharmacology families (inducing cancer cell apoptosis and insulin-like growth factor receptor antagonist) and one biological target (ion channel). The model prediction is also sensitive to the year when the phase II trial is completed and the year when drug approval outcome is known, which reflects the non-stationarity of the drug development process.

Table 3 Prediction performance (F_1 score, true-positive rate, and true-negative rate) for different instantiations of the Debiasing Variational Autoencoder

Therapeutic area	DB-label	DB-latent	F_1 score	SE	True positive	SE	True negative	SE
All (2019–20)	Yes	Yes	0.48	0.005	0.60	0.012	0.88	0.006
	Yes	No	0.49	0.004	0.57	0.005	0.90	0.002
	No	Yes	0.25	0.012	0.17	0.011	0.98	0.002
	No	No	0.25	0.007	0.15	0.005	0.99	< 0.001
Oncology	Yes	Yes	0.35	0.007	0.57	0.017	0.88	0.007
	Yes	No	0.36	0.006	0.45	0.009	0.93	0.002
	No	Yes	0.23	0.015	0.18	0.015	0.98	0.003
	No	No	0.22	0.012	0.13	0.008	1.00	< 0.001
Cardiovascular	Yes	Yes	0.42	0.031	0.52	0.041	0.87	0.010
	Yes	No	0.39	0.014	0.46	0.019	0.87	0.005
	No	Yes	0.17	0.047	0.12	0.033	0.99	0.003
	No	No	0.08	0.035	0.05	0.025	1.00	< 0.001
Central nervous system	Yes	Yes	0.60	0.007	0.63	0.013	0.90	0.005
	Yes	No	0.60	0.006	0.61	0.009	0.91	0.003
	No	Yes	0.20	0.021	0.12	0.015	0.99	0.002
	No	No	0.17	0.012	0.10	0.008	0.99	0.002
Autoimmune/inflammation	Yes	Yes	0.49	0.005	0.53	0.010	0.87	0.006
	Yes	No	0.49	0.005	0.52	0.007	0.88	0.003
	No	Yes	0.29	0.015	0.18	0.012	0.98	0.002
	No	No	0.33	0.011	0.21	0.008	0.99	0.001
Metabolic	Yes	Yes	0.56	0.011	0.66	0.019	0.81	0.012
	Yes	No	0.57	0.008	0.68	0.011	0.82	0.008
	No	Yes	0.26	0.017	0.17	0.015	0.97	0.005
	No	No	0.29	0.012	0.19	0.009	0.97	0.002
Infectious disease	Yes	Yes	0.53	0.009	0.65	0.011	0.82	0.012
	Yes	No	0.52	0.008	0.67	0.010	0.80	0.010
	No	Yes	0.31	0.018	0.21	0.015	0.98	0.002
	No	No	0.29	0.016	0.19	0.012	0.98	0.001

Both models with DB-Latent use smoothing parameter $\alpha = 10^{-7}$

SE standard error

4.3 Latent Space Clusters

The encoder of the DB-VAE learns a low-dimensional representation $z(x)$ that captures the structure of the high-dimensional distribution of input features x. The density of latent space distribution is then used in debiasing (Eq. 2). To interpret the latent space of the DB-VAE, we visualize the ten-dimensional latent representations of drugs in the testing set in two dimensions using t-SNE [38], and observe that the latent space of the DB-VAE consists of two distinct clusters (Fig. 4).

The drugs of the two clusters of DB-VAE latent space are separated by the value of track records of the clinical trial sponsor. If we measure the track record, T, by the number of phase I trials previously completed by the clinical trial sponsor, we find that the drugs in the top-left cluster (blue) have a normalized value of $T \geq 1$ while those in the bottom-right cluster (orange) have $T < 1$. The same separation holds if we use a different measure of sponsor track records (e.g., the number of phase II or phase III trials instead of phase I). We conclude that the encoder of the DB-VAE distinguishes drugs developed by sponsors with large track records, typically large multinational pharmaceutical companies, from those with limited track records, typically small biotech companies and academic medical centers.

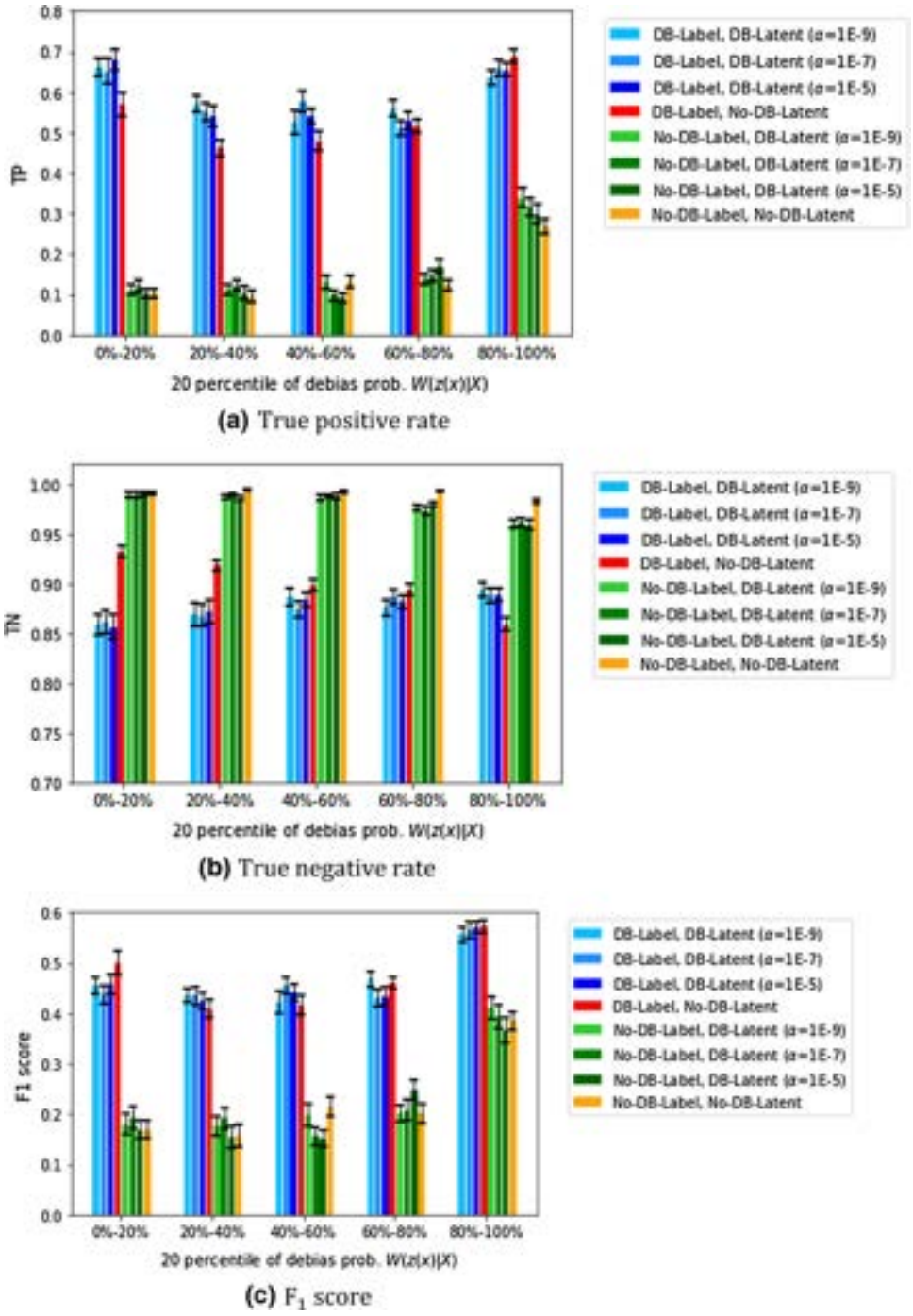

Fig. 3 Effects of debiasing the drug approval outcome labels (DB-Label) and latent space distributions (DB-Latent) on prediction performance, measured by (**a**) true-positive rate, (**b**) true-negative rate, and (**c**) F_1 score. For each DB-Latent model, we use three different values of smoothing parameter α and the results are robust against different values of α

4.4 Improving Financial Efficiency of Drug Development

As shown in Sect. 4.1, applying debiasing achieves higher TP rates and lower TN rates than the un-debiased counterpart. This trade-off between a higher TP rate and a lower TN rate leads to an overall improvement of the financial efficiency of drug development, as the revenue generated by correctly predicting a successful drug (the TP) far outweighs the costs saved by correctly predicting a failed drug (the TN). We use a simple financial model to illustrate this. Suppose the drug developer has completed a phase II clinical trial for a drug candidate and must decide whether to conduct a large-scale phase III clinical trial. We assume that the phase III trial costs US$100 million and takes 5 years to complete. If approved by the FDA, the drug will generate an annual profit of US$2 billion over a 10-year period of market exclusivity. Assuming 10% cost of capital per annum for

Table 4 Top ten drug and clinical trial features of the DB-label, DB-latent model with the highest magnitudes of saliency scores (measured in 10^{-5})

Feature	Saliency
Year of phase II trial completion	−2.16
Trial outcome - completed, positive outcome/primary endpoint(s) met	1.72
Pharmacology - induce cancer cell apoptosis	−1.69
Year of drug approval outcome	1.43
Biological target - ion channel	−1.42
Trial outcome - terminated, lack of efficacy	−1.33
Medium - powder	1.15
Pharmacology - insulin-like growth factor receptor antagonist	1.09
Prior approval of drug for another indication	−1.00
Trial outcome - completed, negative outcome/primary endpoint(s) not met	−0.91

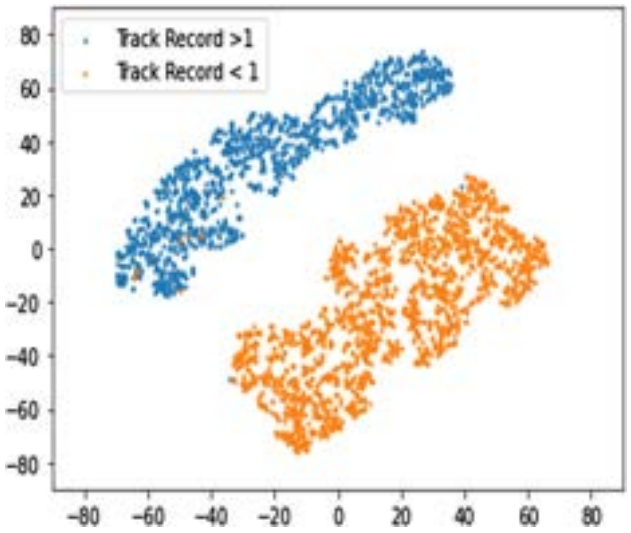

Fig. 4 t-SNE visualization of the latent representation of the DB-Label, DB-Latent model with smoothing parameter $\alpha = 10^{-7}$. Drugs in the two clusters are well separated by the values of track records for the clinical trial sponsors

cash flows of an approved drug and 15% cost of capital for cash flows of a phase III trial, the net present value (NPV) of an approved drug is $\text{NPV}_1 = \text{US\$6}$ billion, while that of a drug that fails a phase III trial is $\text{NPV}_0 = -\text{US\$100}$ million. The assumed values of the costs of capital are taken from the finance literature [39] and the calculation details are presented in Section S1 of the ESM. The drug developer uses a machine learning model (with TP and TN) to forecast the approval outcome of the drug candidate from its phase II results. The financial value, V, of the machine learning model is given by:

$$V = \text{PoS} \cdot \text{TP} \cdot \text{NPV}_1 + (1 - \text{PoS}) \cdot (1 - \text{TN}) \cdot \text{NPV}_0. \quad (4)$$

We evaluate the financial values in six major therapeutic areas with the most recent PoS estimates by Project ALPHA in Q2 2021 [3]. The results are summarized in Table 5. Compared with the un-debiased baseline (column 3), applying debiasing to the prediction model generates additional financial value in all six major therapeutic areas ranging from US$763 million to US$1,365. This illustrates the critical role of debiasing in improving financial efficiency and reducing the financial risks of late-stage drug development.

For the two DB-Label models that debias the outcome labels (columns 5 and 6 of Table 5), the additional financial value of DB-Latent is most significant in oncology (US$211 million) and cardiovascular diseases (US$170 million), and negative in metabolic (−US$41 million) and infectious diseases (−US$42 million). The differences in financial values for different therapeutic areas are correlated with their average debiasing resampling weights, $W(z(x)|X_{\text{test}})$, shown in Table 6. We find that oncology and cardiovascular drugs have the lowest $W(z(x)|X_{\text{test}})$ (i.e., they are overrepresented in the latent space distribution). As debiasing the latent space distribution leads to the greatest improvements in the TP rate for overrepresented drugs (Fig. 3), it makes sense that the increase in financial value is also highest in these two therapeutic areas. However, metabolic and infectious disease drugs are underrepresented (i.e., higher $W(z(x)|X_{\text{test}})$), and we do not observe the increase in financial value for these drugs by using DB-Latent.

5 Discussion

5.1 Implications of Debiasing Drug Approval Prediction

Debiasing improves the financial efficiency of drug development and the overall prediction performance (measured by F_1 score) by achieving a higher TP rate with a lower TN rate. This trade-off between a higher TP rate and a lower TN rate has important implications for pharmacovigilance. The debiased model is more likely to correctly identify drug candidates that are safe and effective (higher TP rate) than the un-debiased counterparts. Meanwhile, it may predict a

Table 5 Net present value to the drug developer by using debiased models to predict drug development outcomes

Therapeutic area	PoS [20] (%)	No-DB-label No-DB-latent	No-DB-label DB-latent	DB-label No-DB-latent	DB-label DB-latent	Financial value of debiasing
Oncology	29.7	238	314	790	1,001	763
Cardiovascular	47.2	151	340	1,298	1,468	1,317
Central nervous system	36.5	209	273	1,326	1,376	1,167
Autoimmune/inflammation	47.2	585	524	1,479	1,509	924
Metabolic	46.3	524	483	1,876	1,835	1,311
Infectious disease	49.6	576	626	1,983	1,941	1,365

Financial value of debiasing (last column) is estimated by the difference between the model that applies debiasing to both outcome labels and latent features (column 6) and the one with no debiasing (column 3). Net present value is measured in million US$. Both DB-Latent models use $\alpha = 10^{-7}$

PoS probability of success

Table 6 Average debiasing resampling weights, $W(z(x)|X_{\text{test}})$, of drugs in each therapeutic area

Therapeutic area	No-DB-label No-DB-latent	No-DB-label DB-latent	DB-label No-DB-latent	DB-label DB-latent
Oncology	2.94	2.83	2.93	2.88
Cardiovascular	3.61	3.43	3.56	3.21
Central nervous system	4.04	3.87	3.81	3.91
Autoimmune/inflammation	4.13	3.83	3.98	3.96
Metabolic	4.29	3.76	4.58	4.14
Infectious disease	3.83	3.72	3.84	3.91

Numerical values are measured in units of 0.01%. Oncology and cardiovascular drugs have the lowest $W(z(x)|X_{\text{test}})$ for all models

high PoS for drug candidates that have adverse effects (lower TN rate). Because of the potential risks of adverse effects, the predictions of the debiased model must be used with caution by the drug developer. To address the drug safety concerns, future work may use the debiased model to predict whether a drug will exhibit adverse effects on a particular patient population.

Additionally, it is somewhat surprising that the DB-VAE achieves greater improvements for the subset of drugs that are overrepresented in the latent space distribution with lower debiasing resampling weights, $W(z(x)|X_{\text{test}})$. This is contrary to the findings of [27] for image classification, where the DB-VAE improved the prediction for underrepresented minority subgroups. One possible explanation is that the overrepresented drugs correspond to the "me too" drugs [16] or repurposed drugs, which have similar drug properties as previously approved drugs. However, prior approval in one indication does not necessarily lead to a higher PoS in other indications, which makes the approval outcomes of overrepresented drugs more difficult to predict.

5.2 Limitations

Our analysis has several limitations that need to be addressed in future work. First, there are sources of bias in clinical trial development that are important in practice but not addressed by our paper. One example is the patient selection bias against certain demographic features such as race, gender, and socioeconomic status. A major difficulty in performing a statistical analysis on the bias in patient demographics is the significant under-reporting of the relevant information. As of 2 December 2021, the Citeline dataset contains 374,460 clinical trials in all phases of clinical development. Among these, only 1,589 trials (0.4%) recorded "white" or "Caucasian" in the "Patient Population" entry; 1,149 (0.3%) recorded "black" or "African American"; and 82 (0.02%) recorded "Latino." Future work should use natural language processing techniques to extract patient demographic information and apply debiasing on the patient demographic features.

In addition, we observe that debiasing the latent space (DB-Latent) improves the TP rate for drugs that are overrepresented in the feature space (i.e., have similar features). However, our hypothesis that some of the overrepresented drugs are "me too" drugs needs to be confirmed, as the

Citeline dataset does not explicitly label the "me too" drugs. A potential solution is to use the drug novelty metric proposed in [40] based on the Tanimoto distance between the chemical structures of two drugs. Because the Tanimoto distance applies to small molecules but not necessarily to large biologics, future work needs to generalize the drug novelty metrics to all therapeutics (including combination therapies) and test whether debiasing improves the prediction for "me too" drugs.

Finally, the goal of our work is to illustrate the benefits of applying debiasing on a machine learning model with fixed structure. Despite its improved prediction performance, the debiased model has a Type I error (false-positive) rate of 12% and a Type II error (false-negative) rate of 40%, which should be further reduced by optimizing the model design. In our work, the debiasing and prediction tasks are simultaneously achieved with one neural network. While this is an efficient design, the capacity of the encoder network of the DB-VAE may be constrained by performing the dual tasks of reconstructing the input features and predicting the drug approval outcome. A natural extension is to implement debiasing resampling with other prediction models that may be better suited to learn tabular data.

6 Conclusions

By instantiating the DB-VAE to our purposes, we simultaneously identify and mitigate the bias from the imbalance of drug approval outcomes and the over/underrepresentation in the drug feature space. We find that debiasing the imbalance of drug approval outcomes results in major improvements in the TP rate and F_1 score for all drugs, and debiasing the imbalance in the feature space improves the TP rate for overrepresented drugs such as oncology and cardiovascular drugs. The debiased machine learning model predicts safe and effective drugs more accurately and generates financial value for the drug developer in six major therapeutic areas. Future work should address the patient selection bias based on demographic features, incorporate measures of drug novelty as an input feature, and optimize the design of the debiased model to further improve the prediction performance.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s40264-022-01160-9.

Acknowledgements We thank Informa Pharma Intelligence for providing us access to their Citeline data. We thank Kate Lyons and Jillian Ternullo for logistics support, and Jayna Cummings for editorial assistance. Research support from the MIT Laboratory for Financial Engineering is gratefully acknowledged. The views and opinions expressed in this article are those of the authors only, and do not necessarily represent the views and opinions of any institution or agency, any of their affiliates or employees acknowledged above.

Declarations

Funding Research support from MIT Laboratory for Financial Engineering is gratefully acknowledged. No direct funding was received for this study and no funding bodies had any role in the study design, data collection and analysis, decision to publish, or preparation of this article. The authors were personally salaried by their institutions during the period of writing (though no specific salary was set aside or given for the writing of this manuscript).

Conflict of interest Q.X. reports personal investments in publicly traded pharmaceutical companies. E.A. and A.A. are co-founders of Themis AI. D.R. reports personal investments in technology companies and mutual funds. D.R. is co-founder of Venti Technologies, ThemisAI, and The Routing Company. She is a member of the technology advisory board of British Telecom, Hyundai Motor Company, RobGlobal, Knowledge AI, Ten63, Venti, and RIIID. She is a member of the board of trustees of MBZUAI, a senior visiting fellow at The MITRE Corporation, and a member of Accenture's Luminary program. She was a member of PCAST and DIB and has given recent talks at GITEX, TransformAI, Stavros Niarchos Foundation, Purdue, Harvard Radcliff, UIUC, University of Pennsylvania, Johns Hopkins University, ETH, EPFL, KTH, and University of Cambridge. D.R.'s research is funded by the USA Air Force, NSF, ONR, DARPA, Toyota Research Institute, IBM, The Boeing Company, Amazon, JPMC, DSTA, DSO, GIST, the Israel Ministry of Defense, AMS, SMART, and a TED Audacious Prize. A.W.L. reports personal investments in private biotech companies, biotech venture capital funds, and mutual funds. A.W.L. is a co-founder and partner of QLS Advisors, a healthcare analytics and consulting company; an advisor to Apricity Health, Aracari Bio, BrightEdge Impact Fund, Enable Medicine, FINRA, Lazard, Quantile Health, SalioGen Therapeutics, the Swiss Finance Institute, Thalēs, and Think Therapeutics; a director of AbCellera, Atomwise, BridgeBio Pharma, Roivant Sciences, and Annual Reviews; and a member of the NIH's National Center for Advancing Translational Sciences Advisory Council. During the most recent 6-year period, A.W.L. has received speaking/consulting fees, honoraria, or other forms of compensation from: AlphaSimplex Group, Annual Reviews, the Bernstein Fabozzi Jacobs Levy Award, BIS, BridgeBio Pharma, Cambridge Associates, Chicago Mercantile Exchange, Financial Times, Harvard Kennedy School, IMF, JOIM, National Bank of Belgium, New Frontiers Advisors (for 2020 Harry M. Markowitz Prize), Q Group, Research Affiliates, Roivant Sciences, and the Swiss Finance Institute.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

Availability of data and material Citeline is a proprietary dataset provided by Informa Pharma Intelligence and can be accessed via commercial license. The probability of success estimates extracted from historical drug development data in Citeline can be accessed via Project ALPHA https://projectalpha.mit.edu/pos/.

Code availability The code for this work is available from the corresponding author on reasonable request.

Authors' contributions The study was first conceived by D.R. and A.W.L. Data curation and preprocessing were performed by Q.X. The code for training the deep learning models, data visualization, and hyperparameter sensitivity analysis was written by E.A., with help from Q.X. and A.A. The deep learning models were trained by Q.X. on

the MIT Engaging cluster. The code for formal analysis was primarily written by Q.X., with crucial inputs from E.A. and A.A. and supervision by D.R. and A.W.L. The first draft of the manuscript was written by Q.X. and E.A. All authors reviewed and commented on previous versions of the manuscript. All authors read and approved the final version of the manuscript.

References

1. Scannell J, Blanckley A, Boldon H, Warrington B. Diagnosing the decline in pharmaceutical R&D efficiency. Nat Rev Drug Discov. 2012;11:191–200. https://doi.org/10.1038/nrd3681.
2. Wouters OJ, McKee M, Luyten J. Estimated research and development investment needed to bring a new medicine to market, 2009–2018. JAMA. 2020;323(9):844–53. https://doi.org/10.1001/jama.2020.1166.
3. Project ALPHA. MIT Laboratory for Financial Engineering. 2021. https://projectalpha.mit.edu/pos/. Accessed 15 Jul 2021.
4. Butler D. Translational research: crossing the valley of death. Nature. 2008;453:840–2. https://doi.org/10.1038/453840a.
5. DiMasi JA, Hermann JC, Twyman K, Kondru RK, Stergiopoulos S, Getz KA, et al. A tool for predicting regulatory approval after phase II testing of new oncology compounds. Clin Pharmacol Ther. 2015;98(5):506–13. https://doi.org/10.1002/cpt.194.
6. Goffin J, Baral S, Tu D, Nomikos D, Seymour L. Objective responses in patients with malignant melanoma or renal cell cancer in early clinical studies do not predict regulatory approval. Clin Cancer Res. 2005;11(16):5928–34. https://doi.org/10.1158/1078-0432.CCR-05-0130.
7. El-Maraghi RH, Eisenhauer EA. Review of phase II trial designs used in studies of molecular targeted agents: outcomes and predictors of success in phase III. J Clin Oncol. 2008;26(8):1346–54. https://doi.org/10.1200/JCO.2007.13.5913.
8. Malik L, Mejia A, Parsons H, Ehler B, Mahalingam D, Brenner A, et al. Predicting success in regulatory approval from phase I results. Cancer Chemother Pharmacol. 2014;74:1099–103. https://doi.org/10.1007/s00280-014-2596-4.
9. Beinse G, Tellier V, Charvet V, Deutsch E, Borget I, Massard C, et al. Prediction of drug approval after phase I clinical trials in oncology: RESOLVED2. JCO Clin Cancer Inform. 2019;3:1–10. https://doi.org/10.1200/CCI.19.00023.
10. Lo AW, Siah KW, Wong CH. Machine learning with statistical imputation for predicting drug approvals. Harv Data Sci Rev. 2019. https://doi.org/10.1162/99608f92.5c5f0525.
11. Informa Pharma Intelligence. Citeline. n.d. https://pharmaintelligence.informa.com/products-and-services/clinical-planning/citeline. Accessed 5 Dec 2021.
12. Wong CH, Siah KW, Lo AW. Estimation of clinical trial success rates and related parameters. Biostatistics. 2019;20(2):273–86. https://doi.org/10.1093/biostatistics/kxx069.
13. Wong CH, Siah KW, Lo AW. Estimating clinical trial success rates and related parameters in oncology. SSRN preprint. 2019. https://doi.org/10.2139/ssrn.3355022. Accessed 5 Dec 2021.
14. Lo AW, Siah KW, Wong CH. Estimating probabilities of success of vaccine and other anti-infective therapeutic development programs. Harv Data Sci Rev. 2020. https://doi.org/10.1162/99608f92.e0c150e8.
15. Siah KW, Kelley NW, Ballerstedt S, Holzhauer B, Lyu T, Mettler D, et al. Predicting drug approvals: the Novartis data science and artificial intelligence challenge. Patterns. 2021;2(8):100312. https://doi.org/10.1016/j.patter.2021.100312.
16. Aronson JK, Green AR. Me-too pharmaceutical products: history, definitions, examples, and relevance to drug shortages and essential medicines lists. Br J Clin Pharmacol. 2020;86:2114–22. https://doi.org/10.1111/bcp.14327.
17. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. a survey on bias and fairness in machine learning. ACM Comput Surv. 2021;54(6):1–35. https://doi.org/10.1145/3457607.
18. Weber M, Yurochkin M, Botros S, Markov V. Black loans matter: distributionally robust fairness for fighting subgroup discrimination. NeurIPS Fair AI in Finance Workshop 2020. https://arxiv.org/abs/2012.01193. Accessed 5 Dec 2021.
19. Yapo A, Weiss JW. Ethical implications of bias in machine learning. Proceedings of 2018 Hawaii International Conference on System Sciences. 2018. https://doi.org/10.24251/HICSS.2018.668.
20. Bandi H, Bertsimas D. The price of diversity. Arxiv preprint. 2021. https://arxiv.org/abs/2107.03900. Accessed 5 Dec 2021.
21. Lambrecht A, Tucker C. Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads. Manag Sci. 2019;65(7):2966–81. https://doi.org/10.1287/mnsc.2018.3093.
22. Mazurowski MA, Habas PA, Zurada JM, Lo JY, Baker JA, Tourassi GD. Training neural network classifiers for medical decision making: the effects of imbalanced datasets on classification performance. Neural Netw. 2008;21(2–3):427–36. https://doi.org/10.1016/j.neunet.2007.12.031.
23. Seyyed-Kalantari L, Liu G, McDermott M, Chen IY, Ghassemi M. CheXclusion: fairness gaps in deep chest X-ray classifiers. Pac Symp Biocomput. 2021;26:232–43. https://doi.org/10.1142/9789811232701_0022.
24. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447–53. https://doi.org/10.1126/science.aax2342.
25. Bauder RA, Khoshgoftaar TM, Hasanin T. An empirical study on class rarity in big data. 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA). 2018. p. 785–790 https://doi.org/10.1109/ICMLA.2018.00125.
26. Bauder RA, Khoshgoftaar TM. The effects of varying class distribution on learner behavior for Medicare fraud detection with imbalanced big data. Health Inf Sci Syst. 2018;6(1):9. https://doi.org/10.1007/s13755-018-0051-3.
27. Amini A, Soleimany AP, Schwarting W, Bhatia SN, Rus D. Uncovering and mitigating algorithmic bias through learned latent structure. Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. 2019. https://doi.org/10.1145/3306618.3314243.
28. Zhou ZH, Liu XY. Training cost-sensitive neural networks with methods addressing the class imbalance problem. IEEE Trans Knowl Data Eng. 2006;18(1):63–77. https://doi.org/10.1109/TKDE.2006.17.
29. More A. Survey of resampling techniques for improving classification performance in unbalanced datasets. Arxiv preprint. 2016. https://arxiv.org/abs/1608.06048. Accessed 5 Dec 2021.
30. Sattigeri P, Hoffman SC, Chenthamarakshan V, Varshney KR. Fairness GAN: generating datasets with fairness properties using a generative adversarial network. IBM J Res Dev. 2019;63(4–5):3:1-3:9. https://doi.org/10.1147/JRD.2019.2945519.
31. Calmon FP, Wei D, Vinzamuri B, Ramamurthy KN, Varshney KR. Optimized pre-processing for discrimination prevention. Adv Neural Inf Process Syst. 2017. https://doi.org/10.5555/3294996.3295155.
32. Kingma DP, Welling M. Auto-encoding variational bayes. Arxiv preprint. 2013. https://arxiv.org/abs/1312.6114. Accessed 5 Dec 2021.
33. Kullback S, Leibler RA. On information and sufficiency. Ann Math Stat. 1951;22(1):79–86. https://doi.org/10.1214/aoms/1177729694.

34. Cover TM, Hart PE. Nearest neighbor pattern classification. IEEE Trans Inf Theory. 1967;13(1):21–7. https://doi.org/10.1109/TIT.1967.1053964.
35. Fawcett T. An introduction to ROC analysis. Pattern Recognit Lett. 2006;27(8):861–74. https://doi.org/10.1016/j.patrec.2005.10.010.
36. Brabec J, Machlica L. Bad practices in evaluation methodology relevant to class-imbalanced problems. Adv Neural Inf Process Syst. 2018. https://arxiv.org/abs/1812.01388. Accessed 5 Dec 2021.
37. Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: visualising image classification models and saliency maps. ICLR 2014. https://arxiv.org/abs/1312.6034. Accessed 5 Dec 2021.
38. Van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res. 2008;9(86):2579–605.
39. Harrington SE. Cost of capital for pharmaceutical, biotechnology, and medical device firms. In: Danzon PM, Nicholson S, editors. The oxford handbook of the economics of the biopharmaceutical industry. New York:Oxford University Press, Inc.; 2012.
40. Krieger J, Li D, Papanikolaou D. Missing novelty in drug development. Rev Financ Stud. 2022;35(2):636–79. https://doi.org/10.1093/rfs/hhab024.