PrOCTOR: A Data-Driven Model for Predicting Drug Toxicity and Its Ethical Implications in Clinical Trials

Licza Lobo

Decemebr 12, 2024

Introduction

Over the past few decades, the rate at which drugs fail clinical trials has steadily increased (Sun). This high failure rate wastes resources, delays medical advancements, and heightens risks for trial participants. As the prevalence of chronic diseases and other health conditions continues to rise, contributing to the leading causes of death in the United States, the urgency to improve current clinical trial practices becomes more apparent. To put this issue into perspective, approximately 90% of all drug candidates fail, with 30% of these failures attributed to drug toxicity ("What Are Clinical Trials and Studies?"). This pressing challenge has prompted researchers to explore new data-driven approaches to enhance the efficiency and effectiveness of current methods (Benavidez).

In the paper, "A Data-Driven Approach to Predicting Successes and Failures of Clinical Trials," the authors applied a 'moneyball' strategy to evaluate current drug screening methods and discover overlooked factors in predicting drug toxicity. They developed a data-driven model, PrOCTOR, capable of predicting the likelihood of toxicity in clinical trials. This paper assesses the validity and adequacy of their methods and, through Immanuel Kant's philosophy of deontology, argues against its use due to violating patients' rights to autonomy and equitable treatment.

Analysis of Methods

Before attempting to critique the methods used by Gayvert et al., it is important to frame the approach they sought to replicate: the 'moneyball' approach. This approach, typically applied to the realm of sports, uses data to identify overlooked measures to guide decisions rather than relying on intuitive knowledge (Academy). In the case of drug toxicity, Gayvert et al. used this approach to reason that drug toxicity prediction could be improved by integrating properties beyond those currently in use, namely target-based properties. Therefore, their approach relies on the fact that the integration of said properties provides new information in predicting drug toxicity. To validate this claim, I will replicate the authors' methods by analyzing the initial 48 drug properties mentioned in the paper and attempting to reduce

them to the same features through their approach. Using their clinical trial dataset, I will compare my results to theirs, identify discrepancies, and critique their methodology.

Novel Analysis

When reading in the provided training dataset into RStudio, I noted the number of drugs as 846, which is 38 fewer than the 884 observations the authors claimed to use in their paper. Nevertheless, I continued with my analysis under the assumption that the paper misreported the number of observations. To begin, the authors assessed the correlation between all 48 initial features. These features include 10 chemical properties, 34 target-based properties, and 4 drug-likeness properties. Of the 34 target-based properties, 30 are gene-target expression values for 30 different tissues. Unsure of the specific correlation measure used in the paper, I computed Pearson correlations, as this method was later mentioned in their analysis. A correlation matrix plot revealed high correlation between features within the subset of gene-target expression properties and chemical properties, consistent with the claims in the paper. To further confirm this, I conducted a similar analysis on a second dataset provided by the authors, which contained the median gene-target expression values for over 54,000 genes in the Genotype-Tissue Expression Project. I chose this dataset because it was reportedly used to compute the expression values for the drugs in the training set. The visual confirmed high correlations between the tissue expression values. However, it revealed that two tissues, Pancreas and Pituitary, were not highly correlated with other tissue expression values.

Figure 1: Correlation Matrix Plot of Initial 48 Model Features

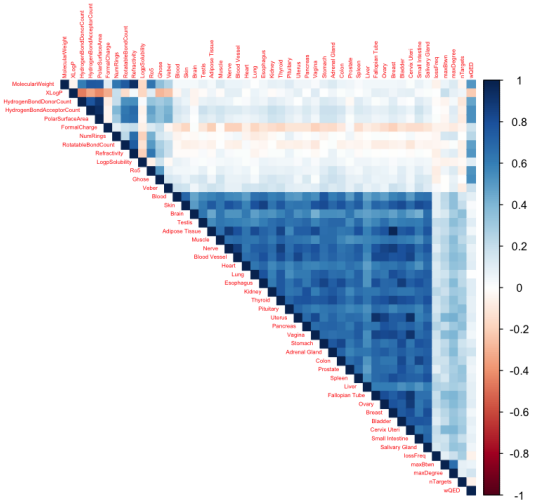
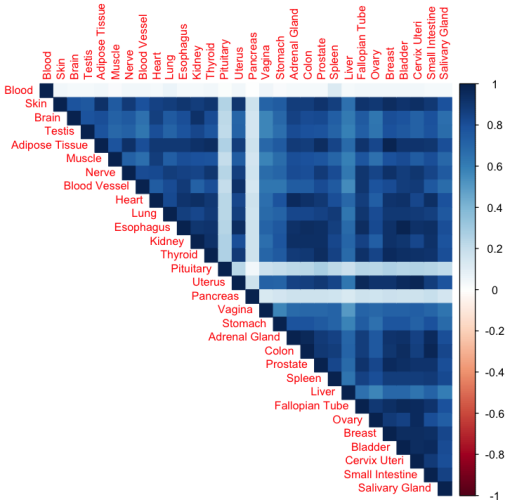


Figure 2: Correlation Matrix Plot of 30 Gene Target Expression Values

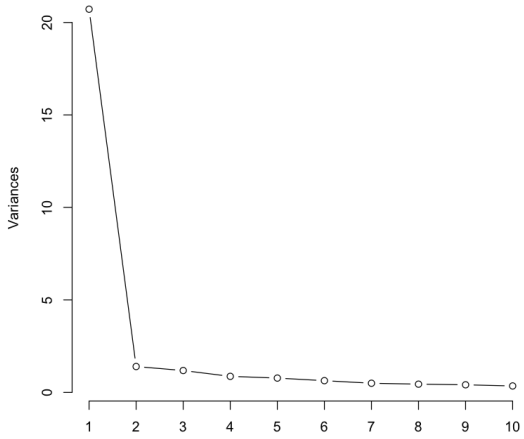


The authors then reported a maximum Pearson's correlation of r = 0.1942 between target-based expression features and chemical features, arguing that the addition of target-based features provides information independent of the chemical properties. I calculated a maximum Pearson's correlation coefficient of r = 0.2049, differing only slightly from the authors'

reported value and reaching similar conclusions. Additionally, I computed the maximum Pearson's r value across all target-based and chemical properties, resulting in r = 0.5396.

Next, the authors performed a principal component analysis (PCA) on all target expression values to reduce feature dimensionality, opting to use the first three principal components (PCs) in place of the raw expression values. However, they did not provide a clear rationale for this decision. To evaluate their approach, I conducted a similar PCA on the 30 gene target expression values in R, producing a Scree plot and calculating the cumulative variance explained by the first three PCs. My analysis showed a cumulative variance of 0.7765, indicating that the first three PCs capture most of the dataset's dimensionality. The Scree plot revealed that the majority of this variance is attributed to the first PC. Additionally, I computed the correlation between the three PCs and the chemical properties, finding a maximum Pearson's correlation of r = 0.1533, which supports the authors' claim that these components provide new information.

Figure 3: Scree Plot for PCA of 30 Gene Target Expression Values



At this stage in their methodology, the authors reduced the initial 48 features to 10 chemical features, 4 drug-likeness features, and 6 target-based features, which included the 3 principal components of gene-target expression and 3 of the 4 other computed target-based features. I followed the same process. However, one target-based feature, the variable representing the number of gene targets a drug is reported to have, was excluded by the authors without explanation. Since this variable was not heavily correlated with the expression values, its exclusion from the feature set appears unwarranted. In my comparative analysis of feature importance, I chose to retain this variable to assess its ranking relative to the authors' results and determine whether its omission was justified.

In R, I recreated this by sampling 100 approved drugs for the 100 failed drugs 30 times, creating balanced training sets. Using the random forest package in R, I trained trees with 50 sub-trees for each sample, calculating probabilities and PrOCTOR scores for each. To compare my model with the authors', I plotted the distribution of PrOCTOR scores and probabilities of approval by class. I then followed the authors' use of a Kolmogorov-Smirnov

test to determine the difference in the distributions of PrOCTOR's approval probability between the two class distributions. The authors' model yielded a D = 0.5343, $p < 2.2 \times 10^{-16}$ and my model resulted in D = 0.97051, $p < 2.2 \times 10^{-16}$. Although the D statistics differed, both models were significant and demonstrated that PrOCTOR's approval probability effectively separates the two drug classes.

Figure 4: Distribution of PrOCTOR Scores

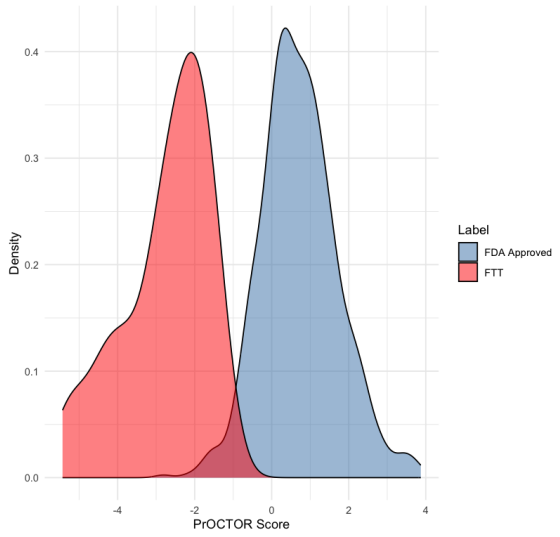
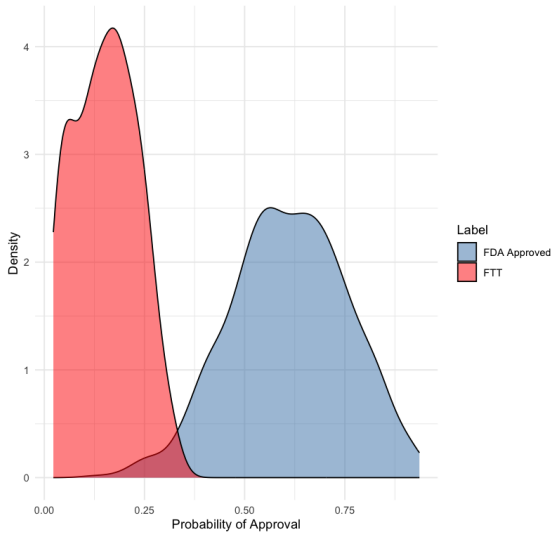


Figure 5: Distribution of Approval Probabilities



To further assess my model's predictive power and compare it to the authors' model, I evaluated its true positive rate (TPR) and true negative rate (TNR). Gayvert et al. reported a TPR of 0.7544 and a TNR of 0.7410. From my model's confusion matrix, I calculated a TPR of 0.72 and a TNR of 0.60. While differences in model development may account for these variations, both models demonstrated comparable accuracy and error rates in separating the two classes.

Having developed a comparable model to that of the authors, I conducted a feature importance analysis to evaluate the significance of target-based properties. Like the authors, I used the Mean Decrease Gini Index to rank feature importance, but my results differed substantially. While the authors claimed their rankings provided strong evidence that target-based features were highly discriminative, my analysis suggested otherwise. The top-ranked target-based feature in my results was 6th, compared to 1st in theirs, and I identified only one target-based feature in the top six, whereas the authors reported four. This discrepancy may be due to differences in sampling methods or model training, but it raises questions about the authors' claim regarding the importance of target-based features.

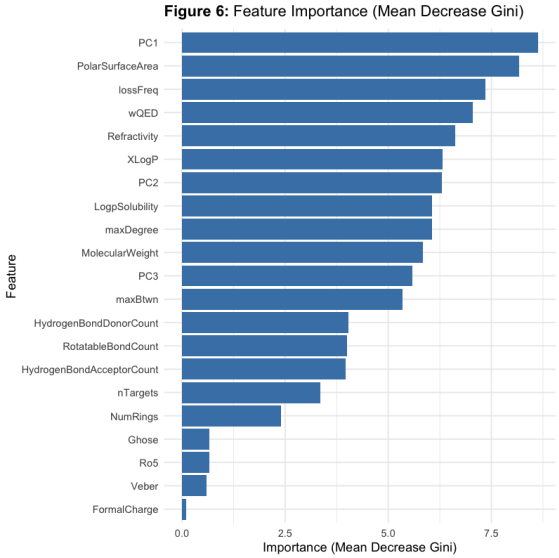


Figure 6: Feature Importance (Mean Decrease Gini)

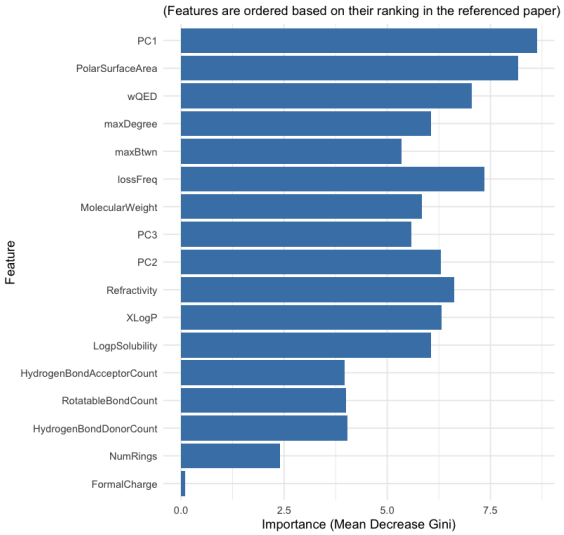


Figure 7: Feature Importance (Mean Decrease Gini)
(Features are ordered based on their ranking in the referenced paper)

To further validate these findings, I ranked feature importance using the Kolmogorov-Smirnov D statistic, a method consistently used throughout the paper, to assess univariate discriminatory power. When compared to their ranking by the Kolmogorov-Smirnov D statistic, **PC1**, **Hydrogen Bond Donor Count**, **qWED**, and **Molecular Weight** were consistently ranked within the top 6 in both my analysis and theirs, showing partial agreement with their conclusions. However, only one of their top six features, and only one of my top six features, namely **PC1**, is a target-based property.

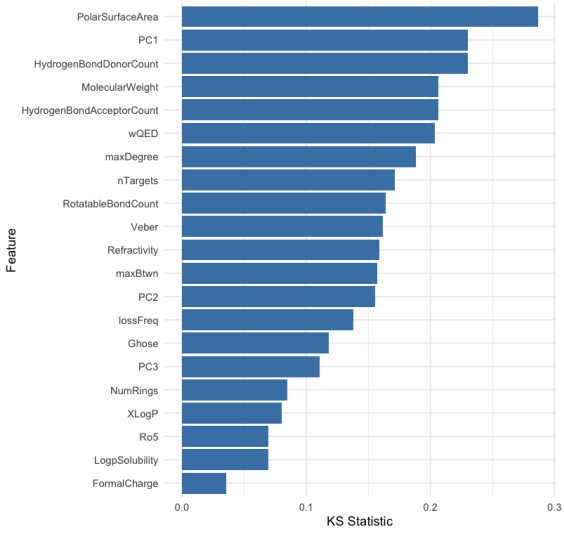


Figure 8: Features by KS Statistic

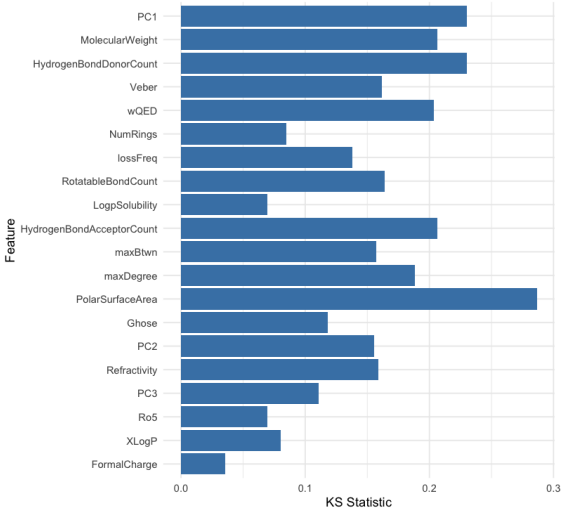


Figure 9: Features by KS Statistic
(Features are ordered based on their ranking in the referenced paper)

Critique of Methodology

One of the initial concerns with the paper lies in the inconsistency between the reported number of observations in the training set and the actual number provided. While this

could be a simple case of misreporting, it might also indicate missing critical information, making it challenging to replicate the authors' modeling approaches. This discrepancy raises questions about the reliability of their source data and the transparency of their methods.

Another significant issue lies in the authors' use of PCA. The authors claim to have used PCA because the gene target expressions were heavily correlated. However, they also report high correlations among the chemical properties. While my own analysis confirmed these correlations, it raises concerns about why PCA was applied to only one subset of properties. This inconsistency suggests a selective focus on gene expressions, potentially to support their claim. Additionally, the authors claim that target-based features provide new information, yet they fail to present measures of correlation between all target-based properties and chemical properties. Through my analysis, I found a moderate correlation between the two subsets ($r = 0.5396$). The omission of this relatively high correlation value, combined with the focus on expression values alone, raises doubt on the validity of their claim. This selective reporting calls into question the integrity of their methodology.

Furthermore, the authors fail to provide a rationale for retaining the top three PCs in place of the raw gene expression values, making it difficult to interpret their approach. To try and understand their decision, I computed the cumulative variance explained by the first three PCs and analyzed a Scree plot. My findings indicate that approximately 77% of the variance can be attributed to the first three PCs. The Scree plot also shows that most of the variance is explained by the first PC, with a leveling off after the third. Additionally, I extended the authors' approach by computing the maximum Pearson correlation between the three PC values and chemical properties, finding an $r = 0.1533$. While retaining the first three PCs appears appropriate, the authors should have justified this choice within the context of drug toxicity to ensure the validity of their claims.

When evaluating their model training approach, I developed a model with comparable power to the authors', even with the inclusion of the variable they excluded. This was assessed by replicating their use of the Kolmogorov-Smirnov (KS) test, which effectively evaluates the entire distribution of probabilities, providing insights into the model's ability to distinguish between the two classes across all thresholds. Using this model, I assessed feature importance in the same way as the authors, by calculating the Mean Decrease Gini. Regarding the excluded variable, my analysis showed it was no less important than other features retained by the authors, as demonstrated in Figure 6. This inconsistency in feature selection complicates efforts to replicate their training approach and raises concerns about transparency. The authors should have provided a clear rationale for excluding this feature to ensure their methodology's reproducibility and credibility. This discrepancy extends to the overall ranking of feature importance. Comparing Figures 6 and 7, target-based properties ranked much lower than chemical properties in my model, whereas the authors reported the opposite pattern, despite both models demonstrating comparable performance. This presents a significant challenge to their claim and raises questions about the validity of their approach. However, these differences may stem from variations in modeling techniques. Although I closely followed their methods, the lack of clarity on how class imbalances were addressed within the sub-sampling iterations made it difficult to precisely replicate their methodology.

To account for differences in modeling techniques, I also compared feature rankings using

KS statistics. Unlike the Mean Decrease Gini, which assesses feature importance based on the purity of splits within a random forest model, KS statistics measure the univariate discriminatory power of features by evaluating the maximum difference between the cumulative distributions of the two classes. This approach provides an additional layer of validation by focusing on the separation ability of individual features, independent of the PrOCTOR model as a whole. In this comparison, my results were more aligned with those of the authors, but both analyses ranked target-based properties lower compared to chemical properties. These findings were not explicitly discussed in the paper, raising suspicions that they may have been intentionally omitted since they contradict the authors' claim that target-based features have higher power than chemical properties. This highlights the importance of transparency in reporting results to build trust in the methodology and ensure clarity in the interpretation of results.

When providing evidence for PrOCTOR's superiority over current practices, the authors did not compare its performance with standard clinical practices. While they evaluated the model against individual rule-based methods, they failed to assess it in the context of the common clinical approach, which combines all multiple rule-based methods with expert judgment. Instead, the authors focused on validation through internal and external analyses, such as 10-fold cross-validation, application to external drug datasets, and comparisons of PrOCTOR scores with side effect scores. While these analyses demonstrated the model's ability to predict drug toxicity effectively, they did not show clear improvements over existing practices, leaving little evidence to substantiate the authors' claims.

Overall, my analysis supported PrOCTOR's predictive and discriminatory ability but produced contradictory results regarding feature importance. These findings challenge the validity of the authors' main claim that target-based features add significant value to the model. Despite contradictions regarding feature importance, the methodology of using random forests paired with sub-sampling was appropriate in this case. Statistically, addressing class imbalance is critical when there is a significant disparity between classes, in this case 100 approved drugs versus 746 failed drugs. Sub-sampling ensures that the model does not become biased toward the majority class, allowing for better representation of the minority class during training. However, given the large pool of clinical trial data publicly available, it raises the question of why the authors did not seek out a more balanced dataset to address this issue without relying on sub-sampling.

Analysis of Normative Consideration

Immanuel Kant's framework of Deontology emphasizes duties, principles, and the inherent rights of individuals, focusing on the intentions behind actions rather than their outcomes. The critique of toxicity prediction models can be framed through the lens of deontological reasoning, using Kant's two formulations of the categorical imperative, acting according to actions that can be universalized, and treating individuals as ends in themselves, never merely as a means to an end. Integrating a model like PrOCTOR into drug screening practices violates patients' rights through issues of patient autonomy, lack of transparency, and equity.

Patients have the right to make informed decisions about their medical treatment, yet when decisions rely solely on algorithmic predictions, this autonomy is undermined. They are denied the ability to participate in their medical choices. From the perspective of the first formulation of the categorical imperative, denying patients autonomy based on opaque model predictions could not be universalized. If universally applied, this practice would reduce all patients to passive recipients of care, denying them of their right to self-determination in medical decisions. Similarly, under the second formulation, using machine learning predictions to dictate treatment options treats patients as mere instruments of efficiency, failing to respect their inherent dignity and ability to make informed choices.

Ethical transparency is another critical concern. Machine learning models are often considered "black boxes," with their decision-making processes opaque, oftentimes even to clinicians. This lack of transparency deprives patients of a clear understanding about how decisions regarding their care are made, violating ethical duties of clarity and honesty. According to the first formulation of the categorical imperative, a practice where decisions are made without providing comprehensible reasons to those affected could not be universally justified. Universalizing such an approach would decrease trust in medical systems. Moreover, by denying patients and clinicians a clear understanding of the reasoning behind decisions, the model treats them as tools for system efficiency, rather than respecting their need for informed involvement in care. These conditions create an environment where clinical judgment is overridden by data-driven models, with no transparent explanation provided to either clinicians or patients.

While the sole use of PrOCTOR poses risks to patient autonomy and transparency, some may argue that its implementation offers significant benefits. Models like PrOCTOR have the potential to improve efficiency, reduce human bias, and accelerate drug screening processes. By processing vast amounts of data consistently, it can identify high-risk candidates earlier, saving resources and potentially preventing harm to patients in clinical trials. Additionally, removing subjective clinical judgment ensures decisions are grounded in empirical evidence, reducing variability and error.

However, these arguments overlook ethical concerns. Efficiency cannot justify practices that undermine patient rights, perpetuate inequities, and decrease trust in medical systems. The lack of transparency in how predictions are made deprives patients and clinicians of the ability to understand and challenge decisions that impact care. Furthermore, reliance on algorithmic models risks institutionalizing systemic biases, disproportionately affecting underrepresented groups. More specifically, it is inherently unfair to deny patients suffering from underrepresented or severe diseases access to treatments based on model predictions. According to the categorical imperative, a system where treatment is denied to specific groups based on statistical predictions could not be universalized as a standard, as it perpetuates systemic inequities. Under the second formulation, denying treatments based on predictions reduces individuals, particularly those in underrepresented groups, to statistical probabilities. This instrumentalizes patients as mere data points in a broader efficiency-driven model, rather than valuing them as individuals deserving of treatment and healthcare access. Therefore, while PrOCTOR may offer practical advantages, its sole use without regulations compromises ethical principles and reduces patients to data points rather than

respecting them as autonomous individuals deserving equitable and transparent care.

Conclusion

Gayvert et al.'s paper, "A Data-Driven Approach to Predicting Successes and Failures of Clinical Trials," introduces an innovative approach to using machine learning in drug development through the PrOCTOR model. While the methodology offers potential in addressing increasing drug failure rates, my analysis highlights significant issues that challenge the authors' claims and raise ethical concerns about the broader implications of integrating such models into clinical practice.

My comparative analysis of the PrOCTOR model demonstrated its capability to classify drugs into FDA-approved and failed categories with comparable performance to the authors' findings. However, discrepancies in the transparency of the methods, the rationale behind feature selection, and the claimed importance of target-based properties undermine the reliability of the paper's conclusions. Additionally, the omission of a critical variable without justification, the failure to compare the model against existing practices, and selective reporting further weaken the paper's validity and illustrate the potential challenges of integrating machine learning tools into healthcare decisions.

Additionally, using Kant's deontological framework, I have argued that incorporating a model like PrOCTOR into drug screening practices violates principles of medical ethics. By relying on algorithmic predictions rather than transparent, patient-centered decision-making, such practices fail to respect individuals as ends in themselves, instead treating them as mere means to achieve efficiency. This approach disregards patients' inherent rights and autonomy, undermining the trust that is fundamental to medical care. It perpetuates inequities and shifts the focus of healthcare away from its moral obligation to value and prioritize each patient as an individual, not as a data point within a system.

As machine learning continues to play a larger role in clinical trials and drug development, the issues highlighted in Gayvert et al.'s paper show how important it is to prioritize thorough evaluation, transparency, and ethical responsibility when building these models. While their work demonstrates the potential of data-driven approaches to improve drug screening efficiency, it also exposes risks. The impact of this paper extends beyond its methodology, serving as both a step forward in applying machine learning to the medical field and an example of what can go wrong when ethical considerations are overlooked. Ensuring these tools align with fairness, patient involvement, and medical ethics is essential if we want them to truly advance our healthcare systems (Husnain).

References

Academy, U. S. Sports. "An Examination of the Moneyball Theory: A Baseball Statistical Analysis." The Sport Journal, 2 Jan. 2005, thesportjournal.org/article/an-examination-of-the-moneyball-theory-a-baseball-statistical-analysis/.

Benavidez, Gabriel A., et al. "Chronic Disease Prevalence in the US: Sociodemographic and Geographic Variations by Zip Code Tabulation Area." Preventing Chronic Disease, vol. 21, no. 21, 29 Feb. 2024, www.cdc.gov/pcd/issues/2024/23_0267.htm, https://doi.org/10.5888/pcd21.230267.

Gayvert, Kaitlyn M., et al. "A Data-Driven Approach to Predicting Successes and Failures of Clinical Trials." Cell Chemical Biology, vol. 23, no. 10, Oct. 2016, pp. 1294–1301, https://doi.org/10.1016/j.chembiol.2016.07.023.

Husnain, Ali, et al. "View of Revolutionizing Pharmaceutical Research: Harnessing Machine Learning for a Paradigm Shift in Drug Discovery." International Journal of Multidisciplinary Sciences and Arts, vol. 2, no. 2, Dec. 2023, jurnal.itscience.org/index.php/ijmdsa/article/view/2897/2219. Accessed 25 Oct. 2024.

Sun, Duxin. "90% of Drugs Fail Clinical Trials." ASBMB, 12 Mar. 2022, www.asbmb.org/asbmb-today/opinions/031222/90-of-drugs-fail-clinical-trials. Accessed 25 Oct. 2024.

"What Are Clinical Trials and Studies?" National Institute on Aging, 22 Mar. 2023, www.nia.nih.gov/health/clinical-trials-and-studies/what-are-clinical-trials-and-studies.