Extending ADE detection with Explainable Counterfactuals and BioBERT

Shrey Patel and Abhishek Jani and Mustafa Adil
Rutgers University
New Brunswick, NJ, USA

Abstract

Adverse Drug Events (ADEs) are a significant clinical safety concern, making their effective detection essential. Although state-of-the-art natural language processing (NLP) models like BioBERT have yielded strong performance in ADE classification from clinical text, their lack of interpretability can hinder real-world application. This paper bridges this gap by demonstrating an end-to-end pipeline integrating BioBERT fine-tuning with counterfactual explanations to achieve both precise and explainable performance. Empirical results on the ADE-Corpus V2 demonstrate high classification performance (F1: 95.6%, Precision: 95.3%, Recall: 96.2%) and reveal the model's sensitivity to drug name masking (76.5% flip rate for ADE sentences vs. 1.9% for non-ADE sentences) and brand name swapping (8.6% overall flip rate, 13.1% for ADE sentences) through counterfactual analysis. Our approach paves the way for explainable clinical AI that can meet practitioners' requirements for transparency and trust.

1 Motivation

Automated Adverse Drug Event (ADE) detection from clinical text is increasingly used to reduce manual effort and strengthen drug safety. Identifying ADEs promptly is crucial for patient safety and effective healthcare delivery. While deep learning models offer powerful tools for this task, their adoption in clinical settings faces hurdles. Clinicians and physician experts require not only accurate predictions but also understandable reasons behind those predictions to justify clinical decision-making and build trust in AI systems (4). Addressing this need for transparency alongside high performance is essential for the successful integration of NLP tools into clinical workflows.

2 Problem Statement

Despite the state-of-the-art accuracy produced by deep learning models like transformers in ADE classification, their lack of explainability often limits its clinical acceptance. Traditional ADE detection models, while proficient in classification, often fail to convey **why** a given sentence is tagged as ADE-related. Furthermore, high accuracy on benchmark datasets does not guarantee robustness to the variations commonly found in real-world clinical text, such as the use of brand names versus generic names or minor lexical differences. This work addresses this dual challenge of performance versus explainability and robustness by investigating a fine-tuned BioBERT model. Specifically, we aim to understand the model's reliance on specific drug entities and its sensitivity to lexical variations through the integration of explainable counterfactuals (8) and systematic brand-generic substitution tests within the ADE classification task.

3 Related Work

Early approaches to ADE detection involved traditional sequence tagging models such as Conditional Random Fields (CRFs) and Support Vector Machines (SVMs) . More recently, deep learning models, including Long Short-Term Memory networks (LSTMs) and variants of BERT (1), have become prominent. BioBERT (5), pre-trained on PubMed abstracts and PMC full-text articles, generally outperforms standard domain transformers on biomedical NLP tasks.

Recent explanation techniques for NLP models include local attribution methods like LIME and SHAP , as well as counterfactual reasoning for explaining predictions. However, fewer studies have employed counterfactual analysis specifically for biomedical NLP, including ADE detection. The present work contributes by applying BioBERT with systematic masking and drug substitution-

based counterfactual experiments to evaluate robustness and gain insights into model behavior.

4 Data

We utilized the ADE-Corpus V2 (2), a gold-standard corpus comprising three core file types: (i) drug-ADE relation pairs, (ii) ADE classification examples (presence/absence labels), and (iii) negative examples, sourced from Hugging Face Datasets (3). The primary dataset for our classification task (`Ade_corpus_v2_classification`) contains sentences annotated with a binary label indicating the presence (1) or absence (0) of an ADE mention. Key attributes are the raw sentence `text` and the corresponding `label`. This classification set comprises over 23,000 clinical sentences. From this set, we focused on 6,821 ADE-positive sentences for the drug masking analysis. We also analyzed 5,093 non-ADE sentences where drug masking was applied. For the brand substitution tests, 8,448 sentences were ultimately modified and analyzed.

To identify drug names for masking and substitution, we used the related `Ade_corpus_v2_drug_ade_relation` dataset. This dataset contains attributes such as the sentence `text`, the identified `drug` entity string, and the `ade` string, along with their character indices within the text. From this relation corpus component, we extracted 1,049 unique drug names. We used the RxNorm API to find corresponding brand names for these generic drugs, successfully mapping 541 of them.

5 Methodology

We fine-tuned the `dmis-lab/biobert-base-cased-v1.1` model using the Hugging Face Transformers library (9). The model was trained for 10 epochs with a batch size of 16 and a maximum sequence length of 128 tokens. Standard procedures using the Trainer and Datasets APIs were followed for tokenization, dataset conversion, and evaluation. The model achieved **96.3% accuracy**, **95.3% precision** (macro), **96.2% recall** (macro), and **95.6% F1-score** (macro), based on evaluation metrics reported at epoch 10.

After fine-tuning, we performed two primary counterfactual analyses:

Drug Name Masking For identified ADE-positive sentences (6,821) and non-ADE sentences (5,093) containing known drugs, we replaced all mentions of those drugs with the `[MASK]` token. We then measured the prediction flip rate (original prediction vs. prediction on masked sentence) separately for each group to analyze the model's dependence on specific drug tokens in different contexts.

Brand Name Substitution The 541 generic drug names mapped via RxNorm were used to create perturbed sentences. In the original classification dataset, occurrences of these generic names were replaced with their corresponding brand name equivalents, affecting 8,448 sentences. We calculated the overall prediction flip rate on these sentences and also calculated the flip rate specifically for the subset of these sentences that were originally labeled as ADE-positive (5,155 sentences).

Intermediate results, including token substitutions and model predictions for original and perturbed inputs, were stored for analysis.

6 Results

Classification Performance The fine-tuned BioBERT model achieved a strong baseline performance with **96.3% accuracy**, **95.3% precision** (macro), **96.2% recall** (macro), and **95.6% F1-score** (macro) on the held-out evaluation data (reported at epoch 10).

Metric	Value (%)
Accuracy	96.3
Precision (macro)	95.3
Recall (macro)	96.2
F1-score (macro)	95.6

Table 1: Performance metrics of the fine-tuned BioBERT model on the ADE-Corpus V2 classification task (evaluation set, epoch 10).

Drug Masking Flip Rate Analyzing the 6,821 ADE-positive sentences subjected to masking revealed that **5,217 sentences (76.5%)** had their predictions flipped (typically from ADE to Not ADE). In contrast, analyzing 5,093 non-ADE sentences subjected to masking showed that only **94 sentences (1.9%)** had their predictions flipped (typically from Not ADE to ADE). This stark difference highlights a significantly higher reliance on the drug token itself for positive ADE classification compared to non-ADE classification.

Analysis Type	Processed	Flipped	Flip Rate (%)
Masking (ADE Only)	6,821	5,217	76.5
Masking (Non-ADE Only)	5,093	94	1.9
Brand Subst. (All)	8,448	726	8.6
Brand Subst. (ADE Only)	5,155	675	13.1

Table 2: Prediction flip rates under drug masking and brand name substitution perturbations.

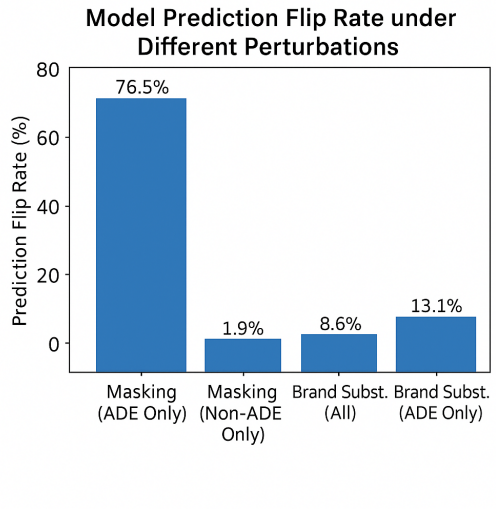

Figure 1: Visual comparison of model prediction flip rates under drug name masking and brand name substitution perturbations. Masking drug names in ADE sentences causes a significantly higher flip rate compared to other conditions.

Brand Name Flip Rate For the 8,448 sentences where a generic name was substituted with a brand name, **726 predictions flipped**, resulting in an overall flip rate of **8.6%**. Focusing on the 5,155 ADE-positive sentences within this set, **675 predictions flipped**, yielding a higher flip rate for ADE-only sentences of **13.1%**.

Brand Mapping Yield The RxNorm API search yielded brand name mappings for 541 of the 1,049 unique generic drugs extracted from the corpus (approx. 51.6% yield). These mappings enabled the modification of 8,448 sentences in the dataset for the substitution analysis.

These results highlight the model's dependence on drug-specific lexical features and its varying sensitivity to different types of textual perturbations. Table 2 summarizes the flip rate findings numerically, and Figure 1 provides a visual comparison.

7 Discussion

Our results demonstrate a nuanced dependency phenomenon in the fine-tuned BioBERT model. The very high flip rate observed after masking drug mentions in ADE sentences (**76.5%**) confirms that these drug names act as powerful predictors, heavily influencing the model's classification decision. This aligns with clinical intuition, where the mention of specific medications known to cause certain side effects is a strong signal for a potential ADE. The extremely low flip rate when masking drugs in non-ADE sentences (**1.9%**) further reinforces this; the absence of an ADE is likely determined more strongly by the surrounding context (or lack of adverse context) rather than the specific drug mentioned. This contrast, clearly visible in Figure 1, strongly suggests the model learns that certain drug mentions are necessary, though not sufficient, signals for ADE classification.

However, the lower, yet significant, flip rate observed during brand name substitution (**8.6%** overall, **13.1%** for ADE-only sentences) suggests a degree of lexical inflexibility. The model appears less robust when encountering brand names that might have been less frequent or absent during pre-training or fine-tuning compared to their generic equivalents. This indicates a potential direction for improvement: incorporating more comprehensive drug synonym knowledge (including brand names, aliases, common misspellings) during model training, perhaps through data augmentation or knowledge-enhanced pre-training strategies.

These findings not only underscore the utility of counterfactual analysis for explainability—showing **what** the model relies on by observing changes when features are altered—but also serve as practical robustness checks for deploying clinical NLP models. Future work could explore finer-grained word-level attribution by integrating methods like LIME and SHAP , investigating the impact of masking other entities like symptoms, or testing robustness against multi-token substitutions and paraphrasing.

Limitations

This study has several limitations inherent in its scope and methodology. Firstly, our analysis relies on a single model architecture, BioBERT; other transformer models or architectures might exhibit different sensitivities to the tested perturbations. Secondly, the findings are based on the ADE-Corpus V2 dataset, and generalization to other clinical text sources or ADE definitions may vary. Thirdly, the drug name extraction and subsequent

brand mapping via RxNorm were incomplete (approx. 51.6% yield), potentially limiting the coverage of our perturbation analysis. Fourthly, the types of counterfactuals explored (masking, brand substitution) represent only a subset of possible linguistic variations encountered in real-world text. Lastly, our primary metric was the prediction flip rate, which captures classification changes but not subtle shifts in model confidence.

8 Conclusion

In this work, we presented an approach for ADE detection that integrates a high-performing BioBERT model with explainability derived from counterfactual analysis. By systematically masking drug names and substituting generic names with brand equivalents, we made the model's decision boundaries and its reliance on drug mentions more transparent. Our results reinforce the need for explainable clinical NLP pipelines where performance, transparency, and trust are interconnected. The flip-rate analysis, particularly the contrast between masking effects on ADE versus non-ADE sentences, serves as a valuable tool for robustness testing, illustrating both the strengths and limitations of the model's ability to generalize across lexical variations. This direction fosters the development of safer, more interpretable AI applications suitable for clinical environments.

9 Future Work

We propose to integrate complementary explainability methods to gain deeper insights. In particular, we will utilize LIME (7) to achieve finer-grained, token-level attributions, complementing the instance-level explanations provided by our current counterfactual analysis. This will allow for a more detailed examination of which specific parts of the text, beyond just the drug name, contribute most significantly to the model's predictions. Additionally, we will investigate more complex counterfactuals, such as masking symptom entities or testing against paraphrased sentences, to further probe model understanding and reliability in clinically relevant scenarios. Exploring data augmentation strategies to enhance robustness against lexical variations like brand names also remains a key direction. Alongside these methodological explorations, we will concurrently work towards addressing the limitations identified in Section 7, such as expanding the evaluation to different model architectures and datasets to ensure broader applicability and robustness.

Acknowledgments

We would like to acknowledge the developers and maintainers of the open-source libraries used in this work, including Hugging Face Transformers, PyTorch, and scikit-learn.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics. DOI:10.18653/v1/N19-1423.

[2] H. Gurulingappa, A.M. Rajput, A. Roberts, J. Fluck, M. Hofmann-Apitius, and L. Toldo. 2012. Development of a benchmark corpus to support ADE identification between drugs and adverse effects from clinical notes. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, pages 319–328. ACM. DOI:10.1145/2110363.2110404.

[3] ADE Benchmark Corpus Contributors. 2023. ADE Corpus V2. Hugging Face Dataset. Accessed: [Insert Date You Accessed It, e.g., May 5, 2025]. `https://huggingface.co/datasets/ade-benchmark-corpus/ade_corpus_v2`.

[4] Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. 2019. Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(4):e1312. DOI:10.1002/widm.1312.

[5] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240. DOI:10.1093/bioinformatics/btz682.

[6] Roisin M. Murphy, Johanna E. Klopotowska, Nicolette F. de Keizer, Kitty J. Jager, Janna H. Leopold, Dave A. Dongelmans, Ameen Abu-Hanna, and Menno C. Schut. 2023. Adverse drug event detection using natural language processing: A scoping review of supervised learning methods. *PLoS One*, 18(1):e0279842. DOI:10.1371/journal.pone.0279842. PMID: 36595517; PMCID: PMC9810201.

[7] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings*

of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 1135–1144. ACM. DOI:10.1145/2939672.2939778.

[8] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2):841–887. https://jolt.law.harvard.edu/assets/articlePDFs/v31/31HarvJLTech841.pdf.

[9] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics. https://aclanthology.org/2020.emnlp-demos.6.