


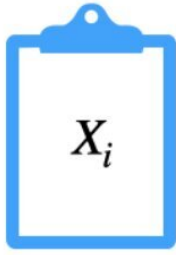



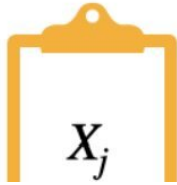


# Using large language models to accurately analyze doctors' notes

February 20 2024, by Jaimie Patterson

Time \ Patient	$T - 1$ (Progress)		$T$ (Discharge)	
	Caretaker	Note	Caretaker	Note
$i$				
$j$				

Generating counterfactual clinical notes for patients using auxiliary data with Algorithm 1(A). Credit: *arXiv* (2023). DOI: 10.48550/arxiv.2310.12803

The amount of digital data available is greater than ever before, including in health care, where doctors' notes are routinely entered into

electronic health record systems. Manually reviewing, analyzing, and sorting all these notes requires a vast amount of time and effort, which is exactly why computer scientists have developed artificial intelligence and machine learning techniques to infer medical conditions, demographic traits, and other key information from this written text.

However, safety concerns limit the deployment of such models in practice. One key challenge is that the medical notes used to train and validate these models may differ greatly across hospitals, providers, and time. As a result, models trained at one hospital may not perform reliably when they're deployed elsewhere.

Previous seminal works by Johns Hopkins University's Suchi Saria—an associate professor of computer science at the Whiting School of Engineering—and researchers from other top institutions recognize these "dataset shifts" as a major concern in the safety of AI deployment.

But a team of Johns Hopkins and Columbia University computer scientists has a plan to harness recent breakthroughs in [large language models](#) to combat the spurious correlations that may arise from AI-powered medical text analysis. They presented their new technique at the [37th Annual Conference on Neural Information Processing Systems](#) in December. The paper is also [published](#) on the *arXiv* preprint server.

"We found that we can greatly improve the robustness of these text models across different settings by making them less sensitive to changes in writing habits and styles observed between different caregivers," says Yoav Wald, a postdoctoral fellow working on the project with Saria.

For example, doctors often use specialized templates, such as headings or tables, in their notes. These templates have no inherent link to the patient's condition. However, AI systems can incorrectly deduce associations between certain templates and specific diagnoses, as the

same templates tend to be used by doctors treating certain subpopulations of patients, he explains. The same goes for doctors' writing styles, including word choice and grammar.

Though these style-related factors have nothing to do with the analysis being attempted, they can lead to poor results when a model is deployed, degrading its performance and resulting in inaccurate diagnoses.

One way of preventing models from learning these spurious correlations is to feed it the same medical note in many different writing styles. This way, the model learns to focus on the content rather than the writing style, the researchers say.

But rather than having each caregiver rewrite other physicians' notes—which would severely drain the already scarce resource of caregivers' time—the team used large language models to automate this process and create datasets that are resistant to the learning of faulty correlations based on writing style.

"Given a specific note that we wish to rewrite in the style of some caregiver—say, Dr. Beth—we instead ask an LLM, 'How would this note look had Dr. Beth written it?'" Wald explains.

By using LLMs to generate such counterfactual data—data that do not exist in the real world, but that can be used to negate spurious correlations in existing data—the researchers say they can reduce the likelihood that an ML model makes inaccurate predictions based on irrelevant details.

The team also proposes using available auxiliary data (like timestamps, document types, and patient demographics) associated with but not included in these medical notes to create better approximations of counterfactual data.

Through extensive experiments, the researchers demonstrate that using language models in a domain-informed manner improves an ML model's generalizability in challenging, safety-critical tasks like medical note analysis.

This project is part of an effort led by Saria with collaborators at regulating agencies such as the FDA toward the development of an AI safety framework for [health care](#) applications.

"As we increase our use of AI in real-world applications and learn about its strengths and weaknesses, it is important to develop tools that improve AI models' robustness and safety," says Saria. "This has been a key area of our focus over the last five years, and this new work takes an important step in that direction. The methods we've developed here are directly applicable across many important text classification tasks."

"Overall, we believe that causally motivated data augmentation methods like ours can help address challenges in developing robust and reliable ML systems, particularly in safety-critical applications," says Wald.

**More information:** Amir Feder et al, Data Augmentations for Improved (Large) Language Model Generalization, *arXiv* (2023). [DOI: 10.48550/arxiv.2310.12803](#)

Provided by Johns Hopkins University

Citation: Using large language models to accurately analyze doctors' notes (2024, February 20) retrieved 27 April 2024 from <https://medicalxpress.com/news/2024-02-large-language-accurately-doctors.html>

This document is subject to copyright. Apart from any fair dealing for the purpose of private

study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.