Original Research

Bidirectional Encoder Representations from Transformers-like large language models in patient safety and pharmacovigilance: A comprehensive assessment of causal inference implications

Xingqiao Wang[1], Xiaowei Xu[1], Zhichao Liu[3] and Weida Tong[2]

[1]Department of Information Science, University of Arkansas at Little Rock, Little Rock, AR 72204, USA; [2]FDA/National Center for Toxicological Research, Jefferson, AR 72079, USA; [3]Nonclinical Drug Safety, Boehringer Ingelheim Pharmaceuticals, Inc., Ridgefield, CT 06877, USA
Corresponding authors: Zhichao Liu. Email: zhichao.liu@boehringer-ingelheim.com; Weida Tong. Email: weida.tong@fda.hhs.gov

Impact statement

In this study, we offered crucial insights into the application of Bidirectional Encoder Representations from Transformers (BERT)-like large language models (LLMs) in patient safety and pharmacovigilance (PSPV) causality assessment. Our findings reveal that while BERT-like LLMs maintain consistent performance across different data complexities, the model size is not a direct predictor of performance. Notably, domain-specific pre-trained LLMs, regardless of safety knowledge fine-tuning, outperform generic BERT models in causal inference. These revelations are pivotal in directing future LLM applications across a myriad of sectors, optimizing their deployment in PSPV.

Abstract

Causality assessment is vital in patient safety and pharmacovigilance (PSPV) for safety signal detection, adverse reaction management, and regulatory submission. Large language models (LLMs), especially those designed with transformer architecture, are revolutionizing various fields, including PSPV. While attempts to utilize Bidirectional Encoder Representations from Transformers (BERT)-like LLMs for causal inference in PSPV are underway, a detailed evaluation of "fit-for-purpose" BERT-like model selection to enhance causal inference performance within PSPV applications remains absent. This study conducts an in-depth exploration of BERT-like LLMs, including generic pre-trained BERT LLMs, domain-specific pre-trained LLMs, and domain-specific pre-trained LLMs with safety knowledge-specific fine-tuning, for causal inference in PSPV. Our investigation centers around (1) the influence of data complexity and model architecture, (2) the correlation between the BERT size and its impact, and (3) the role of domain-specific training and fine-tuning on three publicly accessible PSPV data sets. The findings suggest that (1) BERT-like LLMs deliver consistent predictive power across varied data complexity levels, (2) the predictive performance and causal inference results do not directly correspond to the BERT-like model size, and (3) domain-specific pre-trained LLMs, with or without safety knowledge-specific fine-tuning, surpass generic pre-trained BERT models in causal inference. The findings are valuable to guide the future application of LLMs in a broad range of application.

Keywords: Pharmacovigilance, patient safety, large language models, transformers, BERT

Experimental Biology and Medicine **2023; 248: 1908–1917. DOI: 10.1177/15353702231215895**

Introduction

Ensuring and promoting patient safety and pharmacovigilance (PSPV) is paramount in eliminating unexpected side effects and establishing reliable safety profiles throughout drug development.[1,2] As a pivotal component of PSPV, causality assessment is integral to identifying potential correlations between drug consumption and adverse events, thereby contributing to the detection and understanding of unforeseen risks and side effects.[3,4] Conventionally, the causal inference process has largely depended on controlled population studies, which are often time-consuming, costly, and, in some instances, impractical for certain PSPV challenges. Alternatively, targeted trials offer a viable solution for establishing causality based on observational data. These types of trials, such as decentralized clinical trials and those utilizing real-world data, inherently involve large volumes of data.[5,6] This necessitates the application of sophisticated statistical methods to effectively handle and interpret the information.[7–10]

The recent announcement of the Food and Drug Administration (FDA) Modernization Act 2.0 underscores

ISSN 1535-3702

the critical role of innovative non-animal-based methodologies like artificial intelligence and machine learning (AI/ML) in supporting drug development and its safety evaluation (https://www.congress.gov/bill/117th-congress/senate-bill/5002. By proposing these cutting-edge tools as alternatives, the Act aims to accelerate the drug development process, thereby expediting the availability of potentially life-saving medications. In concurrence with this legislative shift, the US FDA has unveiled a discussion paper that focuses on the application of AI/ML in the development of drugs and biological products (https://www.fda.gov/media/167973/download). The paper places a particular emphasis on ensuring the robustness and reliability of AI/ML solutions in causality assessments, affirming the importance of these technologies in shaping future healthcare landscapes.

Among the various AI/ML strategies, large language models (LLMs) are fast becoming the centerpiece of AI research.[11–13] Their capacity to understand and generate human-like text presents vast opportunities for innovation in drug development and PSPV.[14] By leveraging the predictive and analytical capabilities of LLMs, researchers can streamline the drug development process, analyze complex patient data, and predict potential adverse events more efficiently.[15] Furthermore, LLMs' ability to process and comprehend large volumes of text allows for a more comprehensive analysis of real-world data and clinical trials, contributing to a more holistic understanding of a drug's impact, thereby potentially transforming PSPV approaches, especially causality assessment.[16]

In our prior research, we explored the potential of two BERT-like models, namely, ALBERT and BioBERT, for performing causal inference in the realm of PSPV. In particular, InferBERT, a transformer-based causal inference framework, synergistically combines the powers of ALBERT and Judea Pearl's Do-calculus to establish potential causality in pharmacovigilance.[17] The effectiveness of this model is underscored by its demonstrated ability to accurately predict clinical events and infer their underlying causes. Similarly, DeepCausality[18] is another innovative, AI-driven causal inference framework. It uniquely amalgamates AI-powered language models (LMs), named entity recognition (NER) techniques, and Judea Pearl's Do-calculus into a comprehensive framework for causal inference.[18] The framework has been adeptly employed to estimate causative terms associated with idiosyncratic drug-induced liver injury (DILI), subsequently facilitating the generation of a knowledge-based causal tree. This causal tree serves as an invaluable tool for patient stratification in the context of idiosyncratic DILI.

Despite the strides made in this field, a crucial question remains: what is the true impact of BERT-like LLMs on causal inference in PSPV? This study is designed to systematically explore and evaluate the performance of LLMs in this domain. In this endeavor, we compared two primary categories of LLMs: (1) common knowledge BERT-like LLMs, which are pre-trained using common corpora encompassing resources like webpages, books, and Wikipedia and (2) domain-specific BERT-like LLMs, which are not only pre-trained but also fine-tuned utilizing domain-specific corpora. The effectiveness of these two categories of LLMs was compared through their application in the causal inference of PSPV tasks. Our findings serve as an essential contribution to better position BERT-like LLMs in the appropriate context of PSPV applications.

Materials and methods

As illustrated in Figure 1, the primary objective of this study is to evaluate the influence of different BERT-like LLMs on causal inference performance in PSPV. This evaluation will be underpinned by an in-depth examination of three critical elements: the complexity of the data, the architecture of LLMs, and the specific domain of the LM. Through this comprehensive analysis, we aim to elucidate the interplay between these aspects and their collective impact on the efficacy and facilitate the "fit-for-purpose" causal inference approach selection in PSPV.

To assess the influence of data complexity, we leverage both structured and free-text data sets in the context of causal inference. For instance, our structured data set is a subset of the FDA Adverse Event Reporting System (FAERS) case reports related to tramadol-related deaths and acute liver failure, which includes free-text attributes, such as indications (https://www.fda.gov/drugs/questions-and-answers-fdas-adverse-event-reporting-system-faers/fda-adverse-event-reporting-system-faers-public-dashboard). Conversely, our unstructured data set, LiverTox, comprised entirely free text.[19]

Regarding the LM, we focus on two key factors: the architecture of the model and the training set used for pre-training, which also defines the model's domain. To examine the efficacy of domain-specific models, we utilize self-supervised learning to fine-tune the BERT model, functioning as our pre-trained LM. This is followed by employing a downstream task for the fine-tuning of all the candidate models. Subsequently, we evaluate all the fine-tuned models using the test set. We employ Judea Pearl's do-calculus mechanism to identify causal items and then utilize a one-tailed test to assess the significance of these enriched causal items. The overall performance of all candidate models is then compared across both structured and free-text data sets.

Data Set

FAERS data set

FAERS is a repository of adverse events and medication error reports submitted to the FDA. The design of this database aids the FDA's post-marketing safety surveillance program for drugs and therapeutic biologic products. In processing the FAERS data for our study, we adopted the same strategy as in our previous research,[17] which involved sentence extraction from each FAERS case report. These case reports encompass clinical features like gender, age, the primary suspected drug, dosage, indication, adverse events, and outcomes.

We used a specific template to transform the case reports into individual sentences, thereby generating a comprehensive sentence set. We selected two specific data sets for our study: one related to Analgesics-induced acute liver failure,

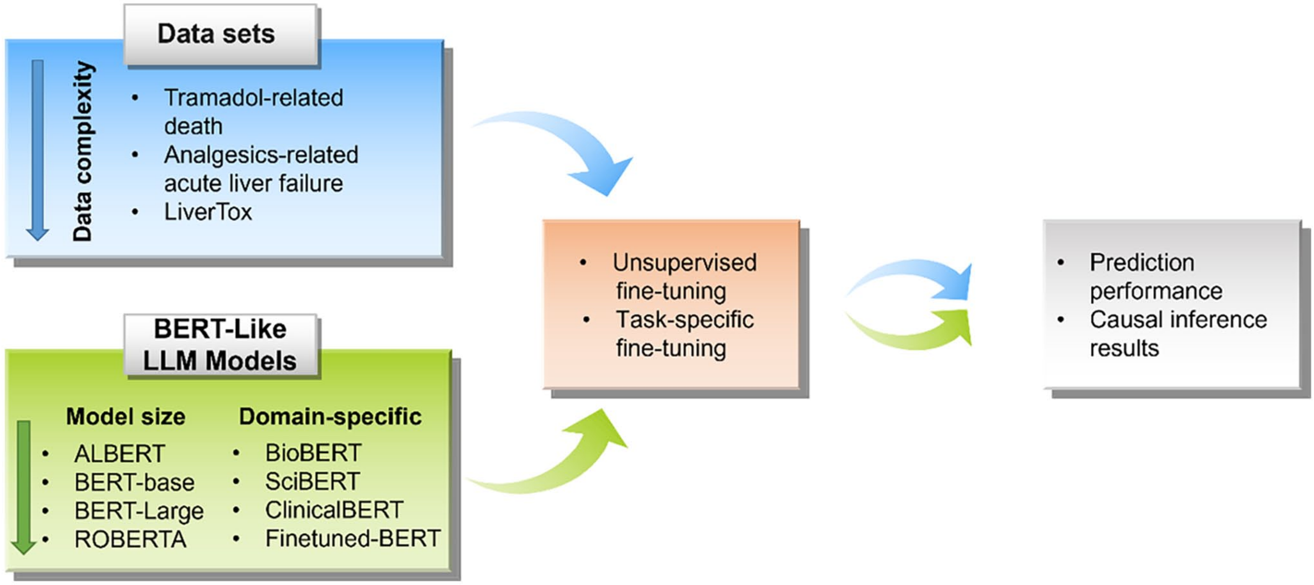

Figure 1. Workflow of this study.

and the other associated with tramadol-related deaths.[20] The data for analgesics-induced acute liver failure spans from November 3, 1997 to December 31, 2019. Similarly, the period for tramadol-related mortalities extends from November 3, 1997 to March 31, 2020. A keyword matching strategy was then employed to discern between positive and negative cases. Specifically, in the data set pertaining to analgesics-induced acute liver failure, "acute liver failure," as indicated within the clinical feature "adverse event," was designated as the endpoint. The cases mentioning "acute liver failure" were classified as positive, while the remaining cases were designated as negative. In contrast, for the tramadol-related death data, "outcomes" was the clinical feature utilized as the endpoint. Cases embedding the term "death" within the "outcomes" clinical feature were categorized as positive, whereas the rest were labeled as negative. Following this, the sentence set was segregated into training, development, and test subsets, utilizing a stratified splitting strategy that maintained a ratio of approximately 0.64:0.16:0.20, respectively.

LiverTox data set

LiverTox is a collaborative online resource curated by medical and scientific specialists that provides comprehensive, current, and easily accessible data on the diagnosis, cause, frequency, patterns, and management of liver injury attributable to prescription and nonprescription medications, herbals, and dietary supplements (www.livertox.nih.gov). Its purpose is to serve as a centralized repository of clinical information, supporting research on DILI. Our analysis utilized data available up to May 2021.

In processing the LiverTox data, we followed the same method as in our prior research.[18] We primarily focused on four sections of the data set: Introduction, Background, Hepatotoxicity, and Mechanism of Injury. For each drug, we collated the context from these sections and labeled each sentence in the Hepatotoxicity section with a "Likelihood score." The "Likelihood Score" within the LiverTox data set is devised to categorize medications according to the probability of their association with DILI, as detailed in the provided reference (https://www.ncbi.nlm.nih.gov/books/NBK548392/). A drug assigned to Category A is well-documented and thoroughly described to either directly cause or be associated with idiosyncratic liver injury, supported by evidence from more than 50 cases. Conversely, a drug in Category B is known or highly suspected to cause idiosyncratic liver injury, bearing a characteristic signature, with cases numbering between 12 and 50. In the context of this study, sentences scored as "A" or "B" were marked as positive. Sentences with any other score were labeled as negative. The data were then split into a training set (90%) and a test set (10%). Information on these two data sets can be found in Table 1.

Table 1. Data set information.

Data set	Number of positives instances	Number of total instances	Positive ratio
Acute liver failure	15,224	36,661	0.42
Tramadol-related death	9846	27,245	0.36
LiverTox data	3578	14,361	0.25

BERT-like LLMs

In this study, we deployed BERT-like LLMs to examine the effect of model selection on our causal inference task. The investigation encompassed two critical aspects: the models' relative sizes and the domain-specific corpus employed for pre- and fine-tuning.

To ascertain the influence of model size, we utilized ALBERT,[21] BERT,[22] and RoBERTa[23] LMs (size: RoBERTa > BERT > ALBERT). These models are trained on broad corpora, such as the Book Corpus and Wikipedia, neither of which are task specific. Furthermore, we contrasted these generic models with domain-specific ones pre- or fine-tuned using data sets germane to our task, including SciBERT, BioBERT, ClinicalBERT, and a task-specific fine-tuned BERT.

Common knowledge BERT-like LLMs

ALBERT, BERT base, BERT large, and RoBERTa base are all pre-existing models designed for natural language processing tasks. Each of them is rooted in the transformer architecture, known for effectively capturing contextual information in text.

ALBERT (A Lite BERT) is a streamlined version of BERT, designed for reduced memory usage and expedited training times.[21] This efficiency is achieved through a parameter-sharing technique that diminishes the number of trainable parameters while preserving high performance.

BERT base and BERT large are both BERT models that differ in size and computational requirements. The former, with 110 million parameters, is quicker to train, while the latter, with 340 million parameters, yields superior performance albeit at a higher computational cost.[22]

RoBERTa base, while similar to BERT, was pre-trained using a distinct training objective and larger batch sizes, resulting in improved performance on some downstream tasks.[23]

Domain-specific BERT-like LLMs

SciBERT,[24] BioBERT,[25] and ClinicalBERT[26] are task-specific variations of pre-trained LMs developed for scientific, biomedical, and clinical domains. They, too, are based on transformer architecture, proficient in capturing intricate relationships between words and their contexts, which makes them ideal for tasks like text classification, NER, and question answering.

What differentiates SciBERT, BioBERT, and ClinicalBERT from general-purpose models like BERT is their pre-training on domain-specific corpora. This specialization allows these models to better comprehend the unique characteristics and nuances of scientific, biomedical, and clinical language.

BioBERT and ClinicalBERT are pre-trained on large-scale biomedical and clinical text corpora, respectively, whereas SciBERT is pre-trained on a blend of general- and scientific-domain text corpora. Each model has demonstrated superior performance across a spectrum of scientific, biomedical, and clinical natural language processing tasks.

In summary, SciBERT, BioBERT, and ClinicalBERT are specialized pre-trained LMs designed to cater to the explicit needs of scientific, biomedical, and clinical applications. They leverage the strengths of transformer-based models and incorporate domain-specific knowledge and pre-training to deliver high performance on specialized tasks.

The task-specific data from this project were used to fine-tune the BERT base model through self-supervised learning using a masked language model (Mask LM). The fine-tuned models were named after their corresponding tasks: Trmol_DILI_BERT, Analgesics_DILI_BERT, and DILI_BERT.

Causal inference

Leveraging these LLMs for causal inference, we devised a task-specific downstream task consisting of a simple classification model, as outlined in our previous studies.[9,10] The procedure for conducting causal inference is contingent upon the organizational structure of the data set in use. For free-text data, we deployed an NER method to isolate pertinent named entities from the context. These entities were then perceived as potential causal candidates. On the other hand, with structured data, we regarded the values of each attribute as potential causal candidates. The do-calculus mechanism was applied in a manner consistent with our previous studies Ball and Dal Pan[9] and Wu *et al*.[10]

Performance metrics

We utilized standard classification metrics, including accuracy, recall, precision, and $F1$-score to assess the effectiveness of our downstream classification model.

Predictive positive rate (PPR) is a measure used in statistics and machine learning to evaluate the performance of a binary classification model. Specifically, PPR measures the proportion of true positive predictions among all positive predictions made by the model.

PPR can be calculated using the following formula:

$$\mathrm{PPR} = \mathrm{TP} / (\mathrm{TP} + \mathrm{FP})$$

where TP represents the number of true positive predictions and FP represents the number of false positive predictions.

PPR is often used in medical testing and diagnosis to evaluate the accuracy of a diagnostic test. In this context, PPR measures the proportion of correctly diagnosed positive cases among all cases that the test identified as positive. A higher PPR indicates a more accurate test, as it means that the test is correctly identifying a larger proportion of true positive cases among all cases that it identifies as positive.

Data and code availability

In this study, we conducted experiments using TensorFlow on a machine equipped with an NVIDIA V100 GPU. The data and code developed in this study could be accessed through GitHub.

Results

Predictive performance of BERT-like models in PSPV

BERT-like models yield comparable predictive power for structured PSPV data sets. Table 2 reveals a consistently strong performance from all investigated BERT-like LLMs on the tramadol-related death task, boasting accuracy scores within the narrow range of 0.95–0.96. BERT-large and SciBERT rise to the top with accuracy scores of 0.96. In terms of precision, the LLMs achieve scores from 0.93 to 0.95, again with SciBERT and BERT-large recording the top precision scores of 0.95. Recall scores fall within 0.93–0.94, with

Table 2. Tramadol-related death task classification results on test set.

Model	Accuracy	Precision	Recall	*F*1-score
Common knowledge BERT-like LLMs				
ALBERT	0.95	0.94	0.93	0.94
BERT-base	0.95	0.94	0.93	0.94
BERT-large	0.96	0.95	0.94	0.94
ROBERTA	0.95	0.93	0.94	0.94
Domain-specific BERT-like LLMs				
SciBERT	0.96	0.95	0.93	0.94
BioBERT	0.96	0.94	0.94	0.94
ClinicalBERT	0.95	0.94	0.93	0.94
DILI_BERT	0.96	0.94	0.94	0.94

Table 3. Analgesics-related acute liver failure task classification result on test set.

Model	Accuracy	Precision	Recall	*F*1-score
Common knowledge BERT-like LLMs				
ALBERT	0.81	0.76	0.77	0.76
BERT-base	0.79	0.73	0.78	0.76
BERT-large	0.80	0.74	0.78	0.76
ROBERTA	0.80	0.74	0.78	0.76
Domain-specific BERT-like LLMs				
SciBERT	0.80	0.75	0.77	0.76
BioBERT	0.80	0.74	0.77	0.76
ClinicalBERT	0.80	0.75	0.79	0.77
DILI-BERT	0.80	0.74	0.78	0.76

SciBERT and BioBERT marking the highest recall scores of 0.93. Finally, all the domain-specific BERT-like LLMs secure an *F*1-score of 0.94, in a range of 0.94–0.94.

As per Table 3, the performance of all investigated BERT-like LLMs on the analgesics-related acute liver failure task is rather consistent, yielding accuracy scores from 0.79 to 0.81. Precision scores span from 0.73 to 0.76, with ClinicalBERT achieving the leading precision score of 0.75. Recall scores range between 0.77 and 0.79, with ClinicalBERT also topping the recall score with 0.79. Finally, the *F*1-scores range between 0.76 and 0.77, with ClinicalBERT yet again achieving the highest *F*1-score of 0.77. Taken together, these tables suggest that domain-specific LLMs and general-purpose LLMs exhibit a comparable performance on the structured data set in terms of precision, recall, and *F*1-score.

Predictive performance of lite version of BERT-like models is suboptimal for unstructured PSPV data set. Table 4 showcases the strong performance of all investigated BERT-like LLMs on the LiverTox task, with accuracy scores ranging from 0.89 to 0.91. BERT-base, BERT-large, RoBERTa, and the domain-specific LLMs all claim the highest accuracy score of 0.91, leaving ALBERT with the lowest accuracy score of 0.89. With respect to precision, recall, and *F*1-score, all LMs attain scores of 0.91, implying an equitable performance across these metrics. In essence, both general-purpose and domain-specific LLMs perform commendably on the task at hand, except for the ALBERT model.

Table 4. LiverTox task classification results on test set.

Model	Accuracy	Precision	Recall	*F*1-score
Common knowledge BERT-like LLMs				
ALBERT	0.89	0.89	0.89	0.89
BERT-base	0.91	0.91	0.91	0.91
BERT-large	0.91	0.91	0.91	0.91
ROBERTA	0.91	0.91	0.91	0.91
Domain-specific BERT-like LLMs				
SciBERT	0.91	0.91	0.91	0.91
BioBERT	0.91	0.91	0.91	0.91
ClinicalBERT	0.91	0.91	0.91	0.91
DILI-BERT	0.91	0.90	0.91	0.91

Causal inference

BERT-like models could well capture casual terms from structured PSPV data sets. Figure 2(A) showcases a comparative analysis of enriched causal terms as identified by various common knowledge BERT-like models, namely, ALBERT, BERT-base, BERT-large, and RoBERTa. Each algorithm successfully flagged "completed suicide" and "drug abuse" as key factors. Notably, the RoBERTa model further identified "sertraline hydrochloride," a known drug that interacts with tramadol. Similarly, as depicted in Figure 2(B), among the domain-specific BERT-like LLMs, Tramol_BERT, BioBERT, and SciBERT singled out "citalopram hydrobromide," another recognized tramadol-interacting drug.

In relation to the tramadol-related death task, we observed an overlap of 19 and 18 causal results among common knowledge and domain-specific BERT-like models, respectively. These overlaps, presented in ascending order of *p*-value in Figure 2(A) and (B), signify that all evaluated BERT-like LLMs aptly captured causal terms from the structured PSPV data set. Notably, a substantial overlap exists between the shared causal items identified by both common knowledge and domain-specific BERT-like LLMs, underscoring the effectiveness of these models in this task.

A similar pattern was found in another structured PSPV data set, specifically, the analgesics-related acute liver failure task. Figure 3 delineates the causal inference results generated by various BERT-like LLMs. Particularly among the common knowledge BERT-like LLMs, both ALBERT and BERT_base models identified "rivaroxaban" as a significant factor in the context of completed suicide and drug abuse. Furthermore, the term "breast cancer metastatic" surfaced in the ALBERT model's inference. BERT_large augmented the insights by flagging "morphine sulfate" as a relevant term (Figure 3(A)). Shifting focus to domain-specific models, both ClinicalBERT and SciBERT underscored the significance of "breast cancer metastatic." Simultaneously, "morphine sulfate" was noted by both Analgesics_DILI and ClinicalBERT models. The Analgesics_DILI model also spotlighted "rivaroxaban" in its findings (Figure 3(B)).

Interestingly, a full overlap, consisting of 22 causal terms, was discovered among the shared items identified by both common knowledge and domain-specific BERT-like LLMs

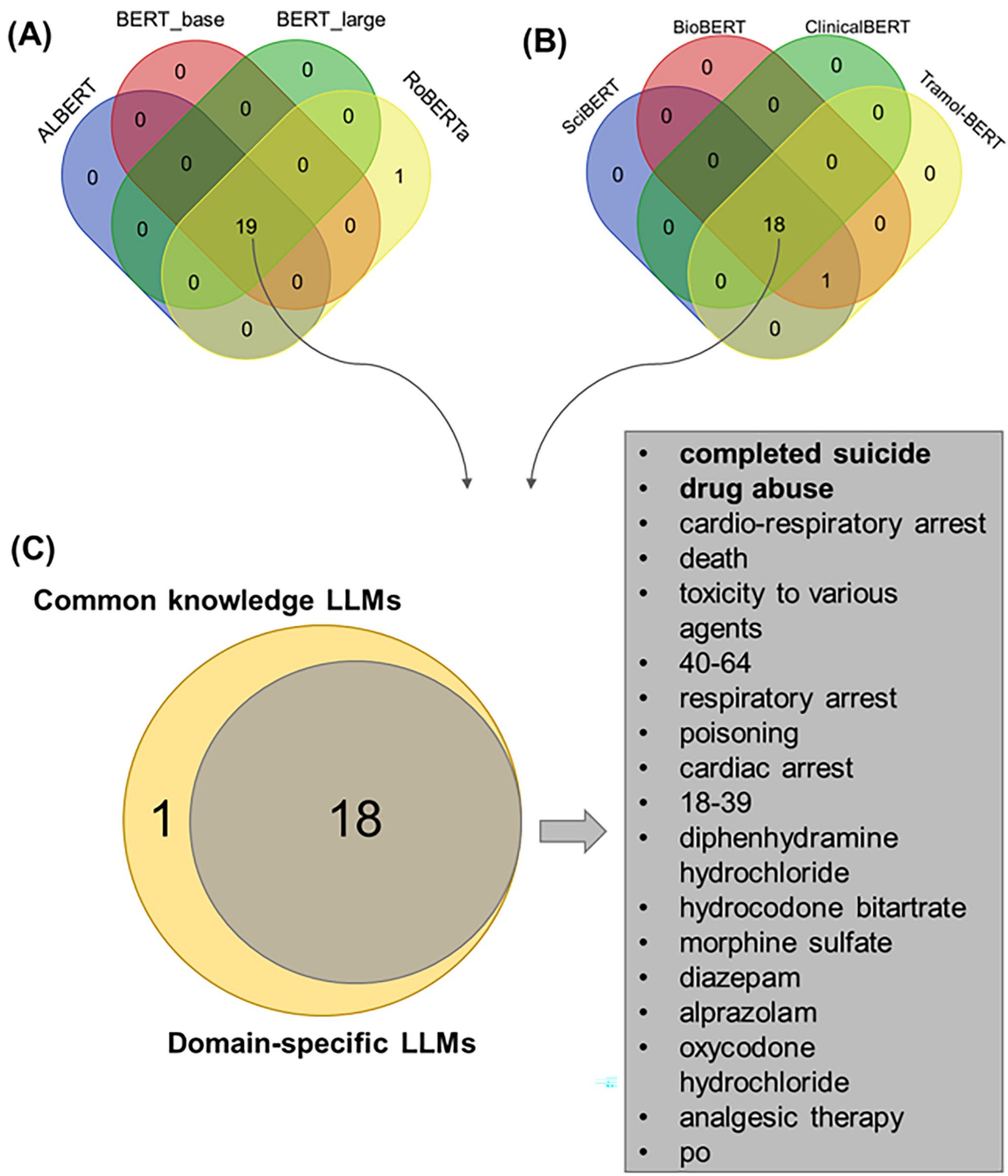

Figure 2. Causal inference results for tramadol-related death: (A) common knowledge BERT-like LLMs; (B) domain-specific BERT-like LLMs; (C) shared enriched causal terms between common knowledge BERT-like LLMs and domain-specific ones.

(Figure 3(C)). These shared causal terms are listed in Figure 3 in ascending order of *p*-value.

Domain-specific BERT-like LLMs with knowledge-based fine-tuning provided superior causal inference in PSPV. Figure 4(A) presents the enriched causal terms derived from various common knowledge, BERT-like LLMs. The most compact and extensive LLMs examined, specifically ALBERT and RoBERTa, yielded distinct enriched causal terms. Interestingly, most of these terms fell outside the liver injury domain. For instance, RoBERTa enriched more therapeutic information, including terms like benzodiazepines,

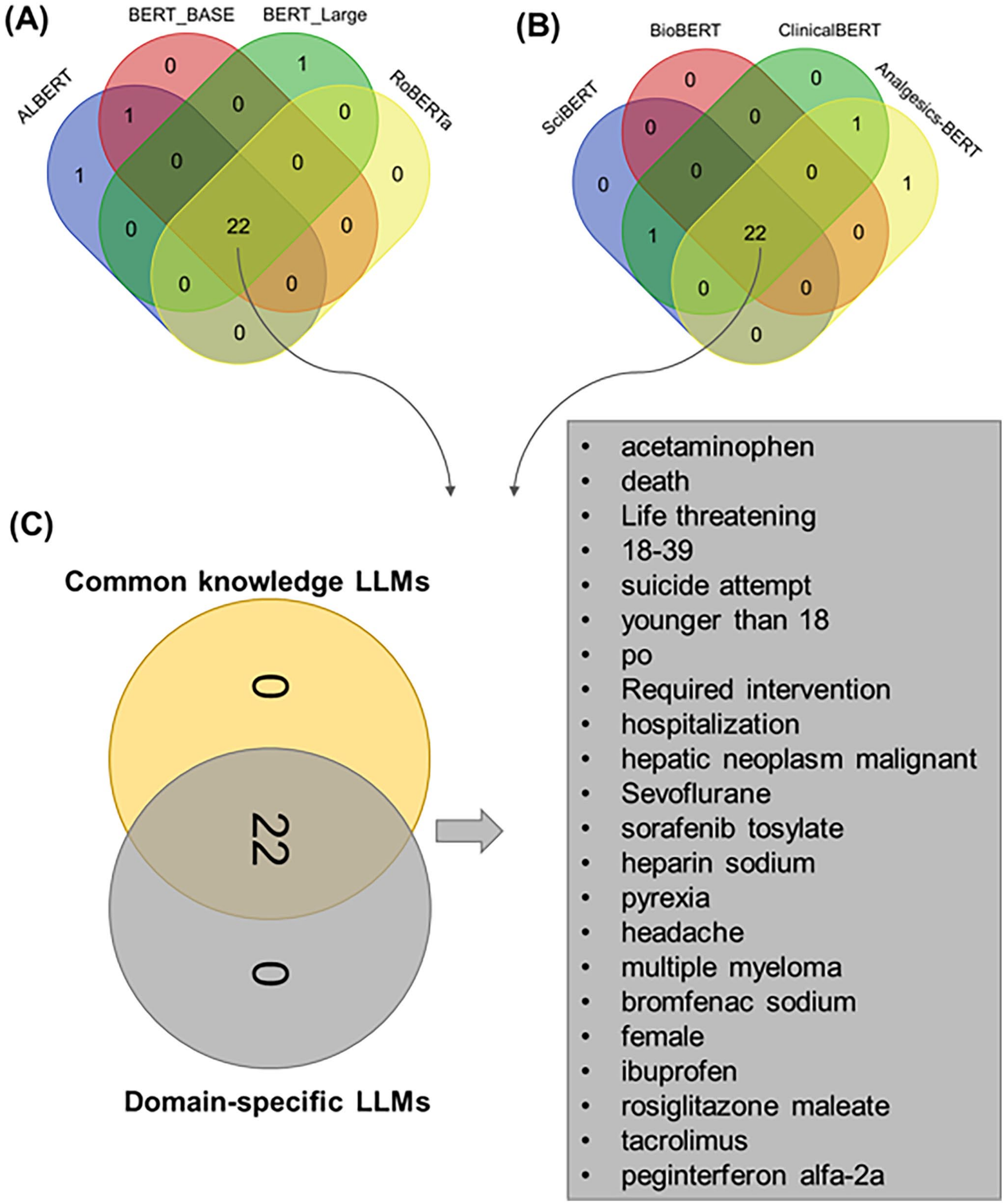

Figure 3. Causal inference results for Analgesics-related acute liver failure: (A) common knowledge BERT-like LLMs; (B) domain-specific BERT-like LLMs; (C) shared enriched causal terms between common knowledge BERT-like LLMs and domain-specific ones.

dopamine, and tyrosine. In comparison, BERT_base and BERT_large produced remarkably similar causal terms, predominantly pertaining to liver injury.

Figure 4(B) showcases the results from three domain-specific BERT-like LLMs: BioBERT, ClinicalBERT, and DILI-BERT. These models generated fairly comparable causal terms. SciBERT, however, yielded a unique, smaller set of causal terms, differing considerably from the other three LLMs. We extended our study by comparing the 18 commonly enriched causal terms from BERT_base

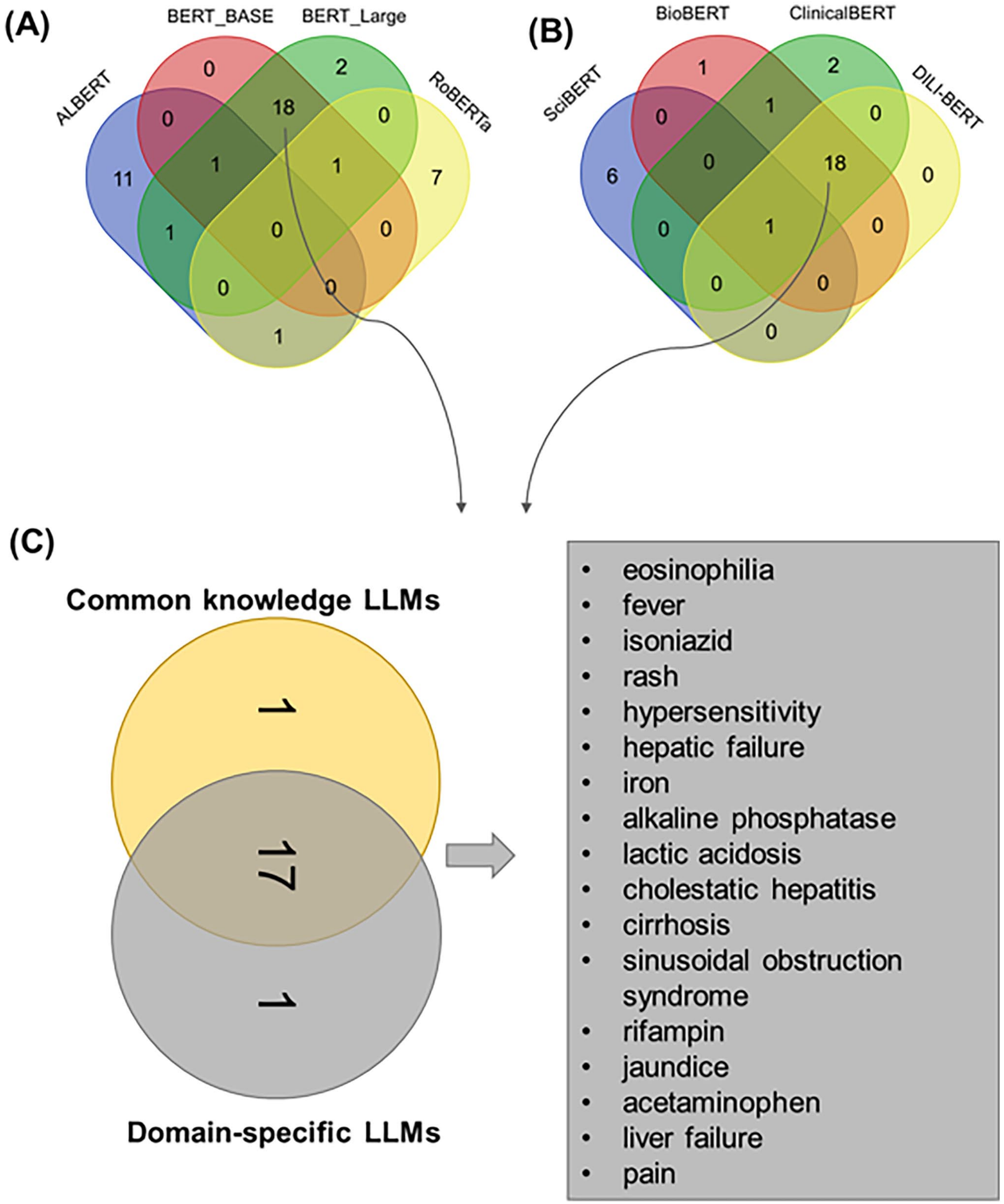

Figure 4. Causal inference results for LiverTox: (A) common knowledge BERT-like LLMs; (B) domain-specific BERT-like LLMs; (C) shared enriched causal terms between common knowledge BERT-like LLMs and domain-specific ones.

and BERT_large to the 18 shared terms among BioBERT, ClinicalBERT, and DILI-BERT. We found an overlap of 17 common terms, with the domain-specific LLMs contributing an additional causal term: cholestasis, a well-established DILI pattern, as displayed in Figure 4(C).

We further explored the relevance of these enriched causal terms to liver injury through a domain-expert manual review. Figure 5 depicts the predictive positive value (PPV) of the investigated BERT-like LLMs, measuring the enrichment rate of liver injury-related causal terms. Overall,

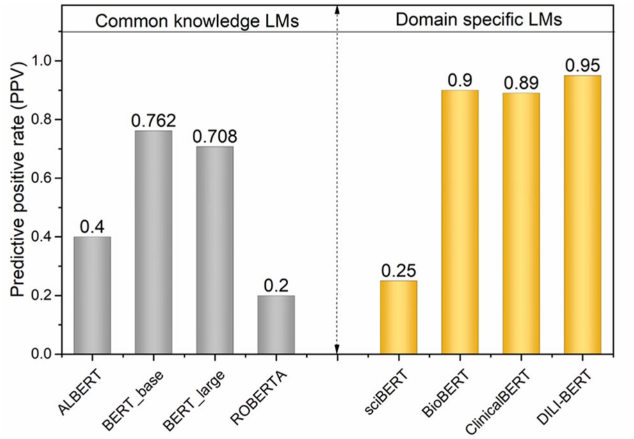

Figure 5. Predictive positive values (PPVs) from BERT-like LLMs for LiverTox.

domain-specific LLMs (excluding SciBERT) exhibited substantially higher PPVs compared to the common knowledge models. RoBERTa, despite its extensive model size, returned the lowest PPV (0.2), indicating no direct correlation between causal inference performance and the size of BERT-like LLMs in the unstructured PSPV data set. On the other hand, DILI-BERT yielded the highest PPV (0.95), underscoring the critical role of domain-specific, knowledge-based fine-tuning. It is worth noting that, unlike BioBERT, ClinicalBERT, and DILI-BERT, which are built on the foundation of BERT models, SciBERT is trained from scratch using scientific literature. This difference in development approach may partly explain its lower PPV (0.25).

Discussion

Causality assessment is an indispensable facet of PSPV, as it assists in establishing potential correlations between drug administration and the incidence of adverse events.[27] Robust causality assessments bolster the overall efficacy of pharmacovigilance systems, promoting superior adverse event reporting and empowering regulatory authorities to make timely and informed decisions. Upon the initial success in merging BERT-like models with advanced statistical methodologies for causal inference in PSPV,[17,18] we have undertaken a comprehensive evaluation of the influence exerted by various BERT-like LLMs to guide "fit-for-purpose" applications.

Distinctly, SciBERT, in contrast to BioBERT, ClinicalBERT, and DILI-BERT – which are extensions of foundational BERT models – is initialized from the ground up, exclusively utilizing scientific literature for its training. This foundational distinction may factor into its relatively lower PPV of 0.25. While SciBERT's comprehensive training furnishes it with a broad knowledge base, apt for various scientific contexts, domain-specific models like BioBERT deftly leverage the foundational weights of BERT – pre-trained on general-domain texts – and further refine them with biomedical content. This nuanced approach enables such models to smoothly amalgamate general linguistic patterns with specific biomedical expertise. Therefore, although SciBERT proves proficient in handling a diverse range of scientific literature, its precision may not parallel that of models specifically honed for the biomedical domain, as reflected in the observed PPV disparities.

Our study has revealed several significant insights that could shape the "best practice" of employing BERT-like LLMs for causal inference in PSPV:

1. For well-structured data, such as the FAERS database, common knowledge BERT-like LLMs can offer comparable predictive power and causal inference capabilities.
2. Interestingly, the size of BERT-like models does not correlate with their causal inference performance.
3. For unstructured free text, domain-specific training and knowledge-based fine-tuning can ensure reliable and robust causal inference results.
4. The pre-training strategy significantly affects causal inference performance, especially in the context of unstructured free text. Our current investigation suggests that models pre-trained on the basis of common knowledge BERT-like LLMs are superior to those trained from scratch.

Several directions warrant further exploration to fully unleash the potential of LLMs in PSPV-related causal inference. Although our current study has focused on the potential of BERT-like LLMs for causal inference, the exploration of generative LLMs, such as ChatGPT/GPT4, Cluade 2, and Bard in PSPV-related causal inference holds great promise. Early endeavors leveraging ChatGPT for causal inference have already been initiated.[16] In addition, incorporating more PSPV data can help consolidate the findings of our current study. Given the complexity of PSPV data, such as electronic medical records and patient narratives from clinical trials, the development of more sophisticated fine-tuning strategies for BERT-like LLMs may be necessary. Furthermore, it would be valuable to explore the integration of LLM-based causality assessment with traditional rule-based tools like the Roussel Uclaf Causality Assessment Method (RUCAM) to enhance performance.[28] Meanwhile, it is highly recommended to further verify the conclusions of current study on more sets of data to further increase its credibility. Data privacy is another critical factor to consider while developing secure AI solutions. Technologies like Lang chain, which pair well with generative LLMs, such as GPT4, offer new ways to access custom data. However, these also pose significant data security challenges. Therefore, a secure data governance infrastructure should be established to facilitate the safe integration of LLMs in PSPV.

In conclusion, BERT-like LLMs hold substantial potential for PSPV-related causal inference, an application that could significantly enhance public health and expedite safe drug development. The comprehensive assessment provided by our current study can guide the development of "fit-for-purpose" causal inference solutions in PSPV.

AUTHOR CONTRIBUTIONS

ZL, XX, and WT conceived and designed the study and the utilization of the PSPV. XW and ZL performed data analysis. ZL,

XW, and XX wrote the manuscript. ZL and WT revised the manuscript. All authors read and approved the final manuscript.

DECLARATION OF CONFLICTING INTERESTS

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

FUNDING

The author(s) received no financial support for the research, authorship, and/or publication of this article.

DISCLAIMER

This manuscript reflects the views of the authors and does not necessarily reflect those of the Food and Drug Administration. Any mention of commercial products is for clarification only and is not intended as approval, endorsement, or recommendation.

ORCID ID

Weida Tong https://orcid.org/0000-0003-3488-6148

REFERENCES

1. Downing NS, Shah ND, Aminawung JA, Pease AM, Zeitoun JD, Krumholz HM, Ross JS. Postmarket safety events among novel therapeutics approved by the US Food and Drug Administration between 2001 and 2010. *JAMA* 2017;**317**:1854–63
2. Wysowski DK, Swartz L. Adverse drug event surveillance and drug withdrawals in the United States, 1969-2002: the importance of reporting suspected reactions. *Arch Intern Med* 2005;**165**:1363–9
3. Zheng C, Dai R, Gale RP, Zhang MJ. Causal inference in randomized clinical trials. *Bone Marrow Transplant* 2020;**55**:4–8
4. Singh S, Loke YK. Drug safety assessment in clinical trials: methodological challenges and opportunities. *Trials* 2012;**13**:138
5. Lavertu A, Vora B, Giacomini KM, Altman R, Rensi S. A new era in pharmacovigilance: toward real-world data and digital monitoring. *Clin Pharmacol Ther* 2021;**109**:1197–202
6. Kompa B, Hakim JB, Palepu A, Kompa KG, Smith M, Bain PA, Woloszynek S, Painter JL, Bate A, Beam AL. Artificial intelligence based on machine learning in pharmacovigilance: a scoping review. *Drug Saf* 2022;**45**:477–91
7. Bates DW, Levine D, Syrowatka A, Kuznetsova M, Craig KJT, Rui A, Jackson GP, Rhee K. The potential of artificial intelligence to improve patient safety: a scoping review. *NPJ Digit Med* 2021;**4**:54
8. Cherkas Y, Ide J, van Stekelenborg J. Leveraging machine learning to facilitate individual case causality assessment of adverse drug reactions. *Drug Saf* 2022;**45**:571–82
9. Ball R, Dal Pan G. "Artificial intelligence" for pharmacovigilance: ready for prime time? *Drug Saf* 2022;**45**:429–38
10. Wu JQ, Horeweg N, de Bruyn M, Nout RA, Jürgenliemk-Schulz IM, Lutgens LCHW, Jobsen JJ, van der Steen-Banasik EM, Nijman HW, Smit VTHBM, Bosse T, Creutzberg CL, Koelzer VH. Automated causal inference in application to randomized controlled clinical trials. *Nat Mach Intell* 2022;**4**:436–44
11. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med* 2023;**29**:1930–40
12. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *npj Digit Med* 2023;**6**:120
13. Liu Z, Roberts RA, Lal-Nag M, Chen X, Huang R, Tong W. AI-based language models powering drug discovery and development. *Drug Discov Today* 2021;**26**:2593–607
14. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, Scales N, Tanwani A, Cole-Lewis H, Pfohl S, Payne P, Seneviratne M, Gamble P, Kelly C, Babiker A, Schärli N, Chowdhery A, Mansfield P, Demner-Fushman D, Arcas BA, Webster D, Corrado GS, Matias Y, Chou K, Gottweis J, Tomasev N, Liu Y, Rajkomar A, Barral J, Semturs C, Karthikesalingam A, Natarajan V. Large language models encode clinical knowledge. *Nature* 2023;**620**:172–80
15. Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ, Rajpurkar P. Foundation models for generalist medical artificial intelligence. *Nature* 2023;**616**:259–65
16. Kıcıman E, Ness R, Sharma A, Tan C. Causal reasoning and large language models: opening a new frontier for causality. https://arxiv.org/abs/2305.00050
17. Wang X, Xu X, Tong W, Roberts R, Liu Z. InferBERT: a transformer-based causal inference framework for enhancing pharmacovigilance. *Front Artif Intell* 2021;**4**:659622
18. Wang X, Xu X, Tong W, Liu Q, Liu Z. DeepCausality: a general AI-powered causal inference framework for free text: a case study of LiverTox. *Front Artif Intell* 2022;**5**:999289
19. Hoofnagle JH, Serrano J, Knoben JE, Navarro VJ. LiverTox: a website on drug-induced liver injury. *Hepatology* 2013;**57**:873–4
20. Xie J, Strauss VY, Martinez-Laguna D, Carbonell-Abella C, Diez-Perez A, Nogues X, Collins GS, Khalid S, Delmestri A, Turkiewicz A, Englund M, Tadrous M, Reyes C, Prieto-Alhambra D. Association of tramadol vs codeine prescription dispensation with mortality and other adverse clinical outcomes. *JAMA* 2021;**326**:1504–15
21. Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R. ALBERT: a lite BERT for self-supervised learning of language representations. https://arxiv.org/abs/1909.11942
22. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. https://arxiv.org/abs/1810.04805
23. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V. RoBERTa: a robustly optimized BERT pretraining approach. https://arxiv.org/abs/1907.11692
24. Beltagy I, Lo K, Cohan A. SciBERT: a pretrained language model for scientific text. https://arxiv.org/abs/1903.10676
25. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 2020;**36**:1234–40
26. Huang K, Altosaar J, Ranganath R. ClinicalBERT: modeling clinical notes and predicting hospital readmission. https://arxiv.org/abs/1904.05342
27. Agbabiaka TB, Savović J, Ernst E. Methods for causality assessment of adverse drug reactions: a systematic review. *Drug Saf* 2008;**31**:21–37
28. Rockey DC, Seeff LB, Rochon J, Freston J, Chalasani N, Bonacini M, Fontana RJ, Hayashi PH; US Drug-Induced Liver Injury Network. Causality assessment in drug-induced liver injury using a structured expert opinion process: comparison to the Roussel-Uclaf causality assessment method. *Hepatology* 2010;**51**:2117–26

(Received August 31, 2023, Accepted October 24, 2023)