TABLE-TO-TEXT GENERATION FOR BIOMEDICAL CAUSAL INFERENCE

By

Andrew Bell

Supervised by

Dr. Gabriele Pergola

Department of Computer Science
University of Warwick

September 2023

Abstract

Adverse Drug Effects (ADEs) are harmful, unintended symptoms caused by the use of pharmacological substances. While clinical trials are an important measure to prove causal relations between drugs and adverse events, they cannot reliably detect rare ADEs. Causal inference tackles this problem by leveraging observational data to estimate causal relations between variables surrounding adverse events induced by drugs on the market. A promising approach in this domain is InferBERT, which uses natural language processing (NLP) techniques to harness the biomedical knowledge of a pre-trained language model (LM). However, the framework's applicability is constrained by relying on a template to convert tabular pharmacovigilance reports into sentences suitable for LM input.

To address this limitation, we introduce a table-to-text generation method to automatically generate sentences from pharmacovigilance reports. We performed experiments with two distinct methods: finetuning sequence-to-sequence models, and prompting large language models (LLMs). By evaluating them on a liver-failure case study used to test InferBERT, we demonstrated that both methods improve the generated sentences, enhancing the causal results. In addition, we applied the LLM prompting method to more clinical features than the original study; we identified clinical terms in each of these new features. This proves that our T2T method was able to extend InferBERT beyond its original scope.

Acknowledgements

I want to thank my supervisor, Gabriele Pergola, for inspiring this project. His guidance kept me on the right track and was always appreciated. I'd also like to thank Sanugi, who not only supported me throughout this year, but showed me how important it is to be dedicated to your work, and still make time for others.

Contents

List of Figures

List of Tables

1 | Introduction

1.1 Motivation

Adverse Drug Effects (ADEs) are harmful, unintended symptoms caused by the use of pharmacological substances. They pose a significant health risk to the public; approximately 5.3% of hospital admissions are associated with ADEs [2]. All approved medications must undergo clinical drug trials, which are able to identify common adverse events. However, due to the experimental nature of clinical trials, they are constrained by sample size, and therefore cannot reliably detect rare ADEs. For this reason agencies such as the World Health Organisation (WHO) and Food and Drug Administration (FDA) practice pharmacovigilance, the science of drug safety. As part of this initiative, the FDA maintains the Adverse Event Reporting System (FAERS), to which patients and medical professionals can submit reports of ADEs. The result of this is a database containing over 24 million reports, freely accessible for research purposes.

An important part of pharmacovigilance is the use of data mining techniques to find links between adverse events and their circumstances. However, to make informed decisions about drug safety we need to study causal relations. This is because a correlation between two variables can be spurious, where a third variable, called a confounder, affects both the perceived cause and effect. Causal inference is the study of causal relations between variables; statistical techniques are conventionally used, but often require constraints or assumptions. However, methods from the field of artificial intelligence have enabled causal inference techniques to be enhanced with machine learning, mitigating many limitations. By making improvements to causal inference methods, medical institutions make better decisions about drug safety and can reduce mortalities and healthcare costs associated with ADEs.

The interdisciplinary research field of AI-based causal inference is evolving rapidly, alongside innovations in deep learning such as the Transformer architecture [3], and pre-trained large language models. A promising approach, published by Wang et al. [1], involves finetuning ALBERT [4], a Transformer-based language model to make predictions about clinical terms. By using a model pre-trained on billions of tokens, the authors present a way to implicitly leverage biomedical knowledge. However, a language model alone can only tell us predictive probabilities, which represent correlation not causation. In order to make causal inferences, we need to model interventional probabilities (the effect of changing a single variable). In InferBERT, this is done using a do-calculus based causal inference algorithm and informing it with the predicted probabilities from ALBERT.

To apply their framework to pharmacovigilance case studies, Wang et al. [1] convert reports from FAERS into artificial sentences. This is necessary because a language model requires a textual input, but the reports are in a structured, tabular format. The method used to make this transformation from table to text will affect the input given to the language model for finetuning. This means that by improving the sentence generation method, ALBERT will make better predictions, positively influencing the causal results. This motivates our research into table-to-text (T2T) generation, to improve biomedical causal inference.

1.2 Objective

Our objective for this project is to establish a table-to-text generation method for pharmacovigilance reports, that fits the following criteria:

- Ability to infer implicit relations between report terms
- Ability to fluently express the clinical terms and the relations between them
- Applicability to arbitrary clinical terms

The input to our task is a list of input reports $R = (r_1, r_2, ..., r_N)$, each report r_i consisting of arbitrary clinical terms t_i. The desired output is a sentence s_i to describe each report r_i, by including its clinical terms t_i, as well as the implicit relations between them. By solving this task, we can replace the template approach in InferBERT with our suitable T2T generation method, and improve the quality of the artificial sentences. These sentences go on to affect the prediction results, and the final causal results.

In the two case studies used to evaluate InferBERT, they leverage seven clinical features from the case reports, but ignore useful features such as patient weight, administration route and secondary suspect drugs. While a new template could be made to extend the feature set, this requires manual input and is less generalisable to new types of data. By using T2T generation methods instead of a template, we generalise the framework, allowing it to be more readily applicable to new case studies or feature sets.

2 | Background

2.1 Pharmacovigilance

Adverse drug effects vary significantly in severity and rarity, from common, minor effects such as nausea to rare, but severe ones such as liver failure. The study of adverse drug effects is known as **Pharmacovigilance**, and encompasses the use of clinical trials as well as the analysis of post-market drug surveillance data. For most drugs, clinical trials have only a small number of participants and take place over a short period of time, so they can only detect common adverse effects. This necessitates the surveillance of adverse effects for drugs already on the market. Agencies such as the FDA maintain databases for reports of ADEs. The FAERS database is freely accessible and contains more than 24 million reports[1], one of which can be seen in example 1.

Example 1 (Pharmacovigilance Report)

Clinical Feature	Term
Patient Gender	Female
Patient Age	60 Years
Patient Weight	62kg
Primary Suspect Drugs (PSDs)	Morphine, Ephedrine, Dexamethasone
Indication	Anaesthesia, Gait Disturbance
Adverse Drug Effects (ADEs)	Hallucinations

An important aspect of pharmacovigilance is the detection of ADEs, and therefore the application of data mining on sources such as FAERS is especially important. Traditional techniques such as predictive modelling or clustering are useful for finding correlations between adverse events and their circumstances but are limited in their ability to infer causality [5]. Informed decisions about drug safety require knowledge of causality, which is why causal inference is an essential method for pharmacovigilance.

2.2 Causal Inference

An endpoint of a causal study is the proposed effect in a cause-effect relationship; **causal inference** aims to determine which factors cause the endpoint and to what extent. Causal inference encompasses methods which reason about both experimental and observational data. For example, the randomised controlled trial (RCT) is an experimental method that can be used to prove a causal relationship. This is done by randomly assigning participants between a control and intervention group, the latter of which is given a treatment (e.g a drug in a clinical trial). A statistically significant difference in the study's dependent variable (endpoint) between the two groups proves that the treatment causes the endpoint.

[1] FAERS

However, experimental methods such as the RCT are limited in sample size and therefore cannot be effectively used to monitor the safety of drugs after they are marketed. For this reason it is vital to make estimations of causality from observational data, such as pharmacovigilance databases. However, this type of data is neither controlled nor random, as confounding variables play a role in the distribution of its population. For this reason, we cannot rely on the same assumptions that can be made in an experimental study. To overcome this and leverage the abundance of observational data, causal inference methods can be applied to estimate causality. Traditionally, statistical techniques such as propensity score matching (PSM) or graph-based causal inference are used [6]. The purpose of PSM is to balance the confounder distribution between groups by subsampling the observed data, allowing an observational study to more closely match an experimental one [7].

2.2.1 Graphical Models of Causality

A Bayesian Network is a directed acyclic graph (DAG), where each node is a variable and each edge indicates a dependency between two variables [8]. Given a set of random variables (nodes) $\mathbf{X} = (X_1, X_2, \dots, X_n)$, a Bayesian network models the joint probability distribution $P(\mathbf{X})$ that assigns probability to each variable or node X_i. By the chain rule of probability, the joint probability distribution can be factored as

$$P(\mathbf{X}) = \prod_{i=1}^{n} P(X_i \mid Pa(X_i)),$$

where $Pa(X_i)$ denotes the set of parent nodes of X_i, or in other words, the variables on which X_i is conditioned. An example Bayesian Network is shown in fig. 2.1a; by observing its dependencies we can make the following example inferences:

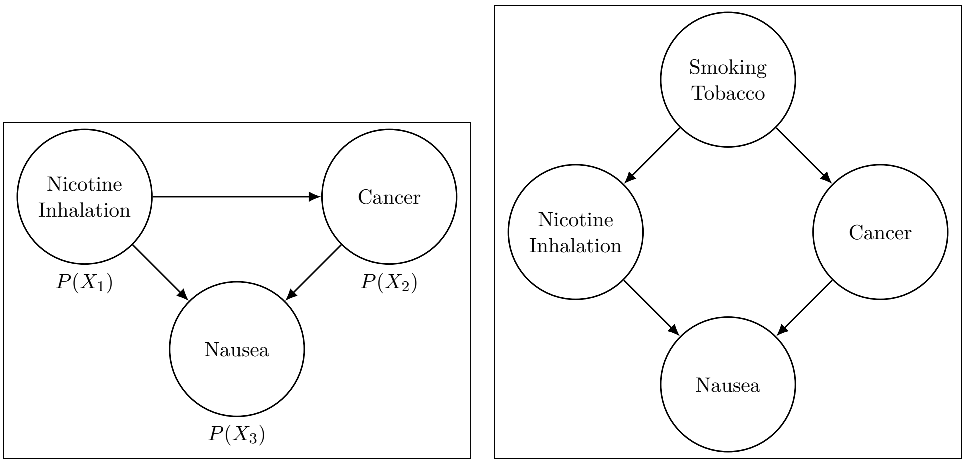
- Nicotine inhalation $P(X_1)$ is independent of nausea or cancer.
- Cancer depends on nicotine inhalation $P(X_2) = P(X_2 \mid X_1) \cdot P(X_1)$.
- Nausea depends on nicotine inhalation and cancer
  $P(X_3) = P(X_3 \mid X_2) \cdot P(X_2) + P(X_3 \mid X_1) \cdot P(X_1)$.

Causal (directed acyclic) graphs are a type of Bayesian network, where a directed edge (u, v) indicates that variable u has a causal effect on variable v [9]. An example causal graph is shown in fig. 2.1b, which includes an additional node to illustrate the confounding effect of smoking tobacco on nicotine inhalation and cancer. In this graph, there is no connection between nicotine and cancer, suggesting that neither one causes the other. Again, these are examples, and the dependencies between these variables are more complex in reality.

2.2.2 Do-Calculus

Definition 1 (Do-Calculus)**.** Let X, Y, Z, W be disjoint sets of nodes in a causal graph, and x, y, z, w be observations for the respective sets of nodes. Fix Y as the endpoint and let Z be the variables we want to transform or remove. Do-Calculus specifies the following three rules [10]:

1. (Deleting observation) $P(y \mid do(x), z, w) = P(y \mid do(x), w)$
   if set Z does not influence the outcome Y through any path
2. (Equating observation and intervention) $P(y \mid do(x), do(z), w) = P(y \mid do(x), z, w)$
   if set Z only influences the outcome Y through directed paths
3. (Deleting intervention) $P(y \mid do(x), do(z), w) = P(y \mid do(x), w)$
   if $do(z)$ does not influence the outcome Y through any path



(a) Bayesian Network - Edges are probabilistic dependencies

(b) Causal Graph - Edges are causal relations

Figure 2.1: Comparison of a Bayesian Network and a Causal Graph

Pearl's Do-Calculus (calculus of intervention) is a set of rules for reasoning about interventions in a causal graph [10]. An intervention is written as a function $do(X = x)$ or just $do(x)$, which hypothetically sets the variable X as a value x. Given an endpoint Y, we can express the effect of an intervention of a variable X as $P(Y \mid do(X))$. Such expressions can be manipulated according to the three rules of do-calculus in definition 1. By introducing hypothetical interventions, we can infer causality by computing the difference between $P(Y \mid do(x))$ and $P(Y \mid \text{not } do(x))$. A statistically significant difference implies that x causes Y.

2.3 AI and Causal Inference

Artificial Intelligence (AI), and its subfield machine learning (ML) have greatly impacted biomedical research, notably in the analysis of visual data such as pathology images and text such as healthcare records. ML typically follows a predictive paradigm in which a model predicts an outcome or label, given a set of feature variables. In contrast, causal inference aims to explain the factors that lead to an outcome, which is a challenging task to model with ML [6]. Some research uses ML to classify causality given a drug-ADE pair [11], but such methods are too constrained for the discovery of ADEs.

For this reason, we find more examples of machine learning to inform or enhance existing statistical methods such as PSM and graph-based methods. Monleuzun et al. [12] use deep learning to inform PSM in their case study of heart procedures, mitigating the difficulty of accounting for all confounders. In the field of pharmacovigilance, ML Graph-based causal inference methods are currently centred on premarketing studies, such as analysis of chemical structures in relation to adverse effects [6].

The relationship between natural language processing (NLP) and causality is more complex. In the context of NLP, causal relation extraction is a distinct task with the aim of identifying *linguistic* causal relations in text. This should not be confused with causal inference, which studies empirical relations. Furthermore, causal inference methods have been used to improve performance in NLP tasks in areas such as robustness [13]. These works are not directly relevant and will not be discussed. However, in the intersection of natural processing, machine learning and causal inference lies a promising approach introduced by Wang et al. [1], called InferBERT.

2.3.1 InferBERT

Wang et al. [1] introduced InferBERT, a recent AI-enhanced approach to biomedical causal inference. In their work, they finetune a transformer-based language model as an endpoint classifier, to inform causal inference. The authors evaluate their framework on two case studies: Analgesics Induced Liver Failure and Tramadol Related Mortalities, identifying more confirmed factors than traditional causal inference techniques. Fig. 2.2 gives an overview of the main steps of InferBERT; each step will now be explained in detail.

Step 1: Preprocessing

Each case study involves a fixed `endpoint` (e.g liver failure) and a dataset of pharmacovigilance reports $R = (r_1, r_2, \ldots, r_N)$, which will be used to finetune the language model. The dataset contains both positive and negative cases, i.e some that include the endpoint and others that do not. All the cases include further features about the case (patient age, administered drugs, dose, etc.). A key goal of the framework is to finetune a language model to predict the existence of the endpoint, given a list of clinical terms. To prepare the dataset for this goal, the following preprocessing steps are performed:

1. Let x_0 be the feature to which our endpoint belongs, and let $r_i(x_0)$ be the terms in report r_i belonging to feature x_0. For each report r_i, the terms $r_i(x_0)$ are removed, and used as a training label y_i by the following rule:

$$y_i = \begin{cases} 1, & \text{if } \texttt{endpoint} \in r_i(x_0) \\ 0, & \text{otherwise.} \end{cases}$$

2. All the other features $X = (x_1, x_2, \ldots, x_K)$ are cleaned, by removing uninformative terms such as ‘unknown’. Numerical features, such as age and dose are and normalised into standard units and sorted into categories. For example the term 16, in the age feature, becomes < 18.

Step 2: Artificial Sentence Generation

In a conventional machine learning approach, the features would directly be used to predict the outcome. However, by using a language model, the input must be first transformed into natural language. This is done in InferBERT by constructing artificial sentences from the clinical terms, using the following template:

“Patient ([gender] and [age]) takes [primary suspect drug] with [dose] to treat [indication], causing [adverse drug effects], leading to [outcomes].”

Generating artificial sentences is necessary because the goal of the finetuning is to leverage the (biomedical) knowledge of the pre-trained language model to make endpoint predictions. A language model like ALBERT is pre-trained on a huge dataset of natural language, and thus taking advantage of these parameters can be done best using a natural language input. Each report in R is converted into a sentence using the template, giving a list of sentences $S = (s_1, s_2, \ldots, s_N)$.

Step 3: Finetuning ALBERT as Classifier

From step 2 we have a list of artificial sentences $S = (s_1, s_2, \ldots, s_n)$, each one representing a pharmacovigilance report. From step 1, we have labels $Y = (y_1, y_2, \ldots, y_N)$ indicating the presence of the endpoint for each report. The goal of this step is to finetune a language model

(ALBERT), to predict labels from sentences. The dataset (S, Y) is split into a training, validation and test set.

Since ALBERT is a masked language model, it outputs contextual embeddings based on its given input. In order to use this model as a classifier, a linear layer is applied to its output which propagates to a single value between 0 and 1, which is the prediction for the label. Finetuning is performed using the implementation provided by ALBERT developers[2], with the hyperparameters shown in table 2.1. Using these, ALBERT is optimised over the training set. The model parameters giving the lowest cross entropy loss over the validation set are chosen for the final model.

The result of this is a language model, which is effectively a classifier. The classifier is used to predict a probability $p_i = P(\texttt{endpoint} \mid s_i)$. By performing a final prediction over the entire list of reports R, we get a list of conditional probabilities $\mathbf{P} = (p_1, p_2, \ldots, p_N)$. These probabilities are used to inform the causal inference algorithm in the next step.

Table 2.1: Hyperparameters used to finetune ALBERT for endpoint prediction

Parameter	Value
Model	ALBERT-Base
Loss Function	Cross Entropy
Optimiser	AdamW
Max Seq. Length	128
Learning Rate	2e-5
Training Steps	30k
Batch Size	32

Step 4: Do-Calculus Based Causal Inference

From the previous steps we have a list of reports $R = (r_1, r_2, \ldots, r_N)$, and for each report r_i, a conditional probability $p_i = P(\texttt{endpoint} \mid s_i)$, where s_i is the artificial sentence based on terms in r_i.

In the final step of InferBERT, a do-calculus based causal inference algorithm is used to weigh the factors of the chosen endpoint. Firstly, the set of all unique terms $T = \{t_1, t_2, \ldots, t_M\}$ is extracted from the reports, which are the potential causes of the study. For each term $t_j \in T$, let $L1_j = \{r \in R \mid t_j \in r\}$ be the set of all reports which contain the term, and $L2_j = \{r \in R \mid t_j \notin r\}$ be the set of reports that do not contain the term. We then calculate the mean of conditional probabilities for the reports in $L1$ and $L2$, respectively:

$$P(\texttt{endpoint} \mid \text{do}(t_j)) = \frac{1}{|L1_j|} \sum_{i \in L1_j} p_i$$

$$P(\texttt{endpoint} \mid \text{not do}(t_j)) = \frac{1}{|L2_j|} \sum_{i \in L2_j} p_i$$

We now have an empirical estimate for the interventional probabilities for each term t_j. The final step is to use a Z-test to determine if there is a statistically significant difference between $P(\texttt{endpoint} \mid \text{do}(t_j))$ and $P(\texttt{endpoint} \mid \text{not do}(t_j))$. The output of the causal inference is a subset of terms $T_{sig} \subseteq T$ which are statistically significant causal factors.

[2] https://github.com/google-research/albert

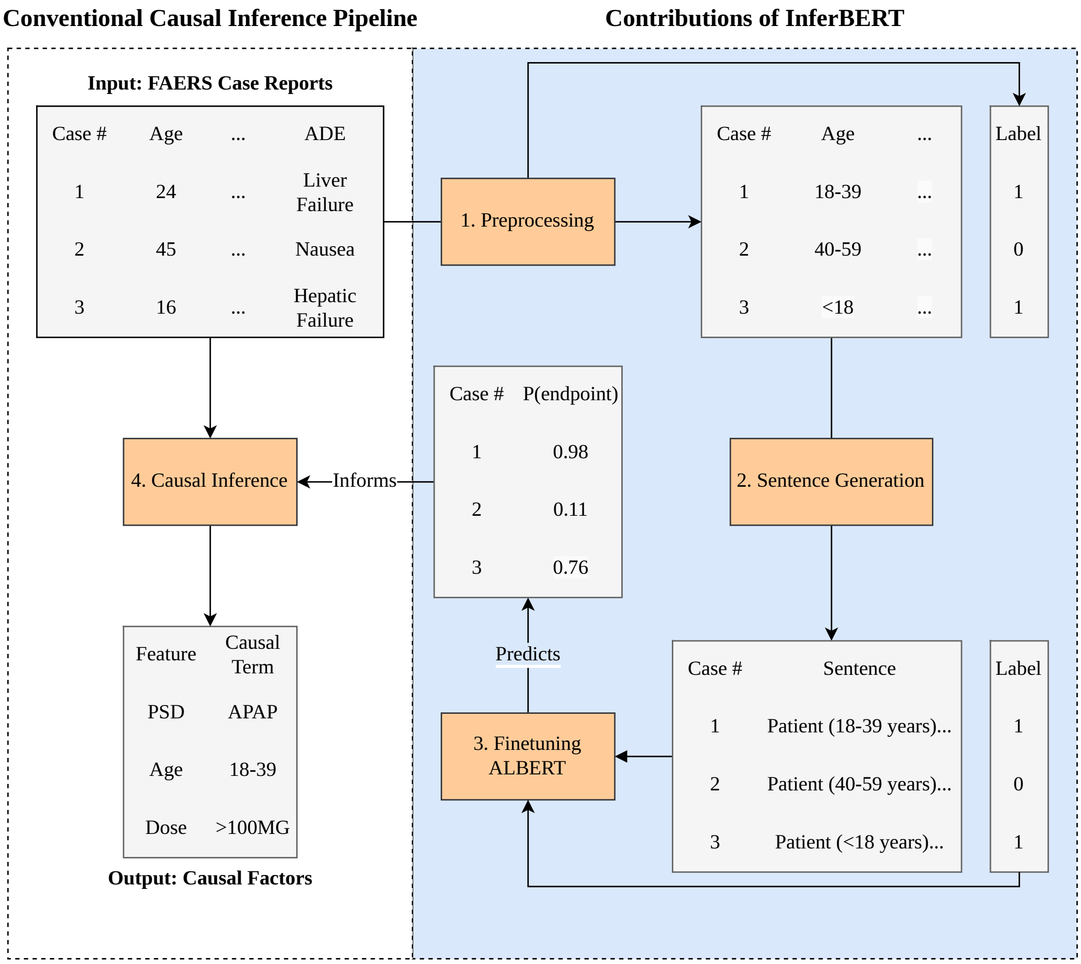


Figure 2.2: High level overview of the InferBERT framework, introduced by Wang et al. [1]

This procedure is used to infer the direct causes of the endpoint, but to further study conditional factors, the authors constrain to the algorithm to specific terms, and recursively apply it to generate a causal tree. These are shown in fig. 2.3; the root is the endpoint, and the leaves are conditional causes.

2.4 Table to Text Generation

Table-to-text (T2T) generation is the task of generating a description from tabular data, and is commonly applied to generate texts such as biographies or news. Traditionally, the task was separated into two problems: First, extracting the relevant information from the table (content selection) and second, generating a suitable descriptor of this information (sentence realisation). Early work in this field relied on rule-based approaches, such as templates or grammars [14].

Modern approaches use neural networks to learn the entire task (end-to-end), rather than using two distinct methods. Lebret et al. [15] introduced WikiBIO, a large dataset containing table-sentence pairs from Wikipedia biographies. They compare neural models, including a long short-term memory (LSTM) model, which is a recurrent neural network. When trained on sufficient data, neural models significantly outperform statistical language models. Building on this work, Liu et al. [16] utilised both word level and field level attention mechanisms. Finally, Yang et al. [17] achieve the current state of the art using the transformer architecture, outperforming the LSTM on the WikiBIO dataset. However, training a T2T transformer is

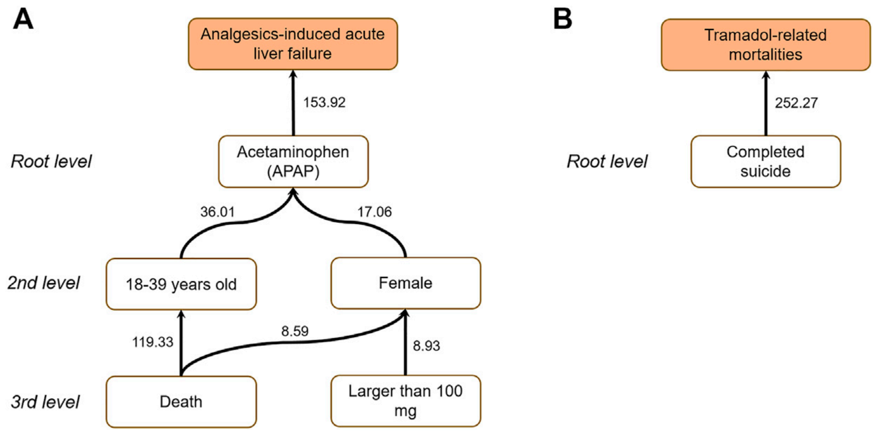


Figure 2.3: Causal Tree results produced by Wang et al. [1]

only suitable given a large dataset, and powerful computational resources.

2.4.1 Finetuning

Many applications of T2T are highly domain specific, including our task of generating sentences from ADE reports. Models trained on data such as Wikipedia infoboxes are unlikely to generalise well to applications in the biomedical domain. Furthermore, domain specific applications often lack large training sets on which to train a new model. For these reasons, we will now discuss methods suitable for few-shot settings, where only a few domain-specific examples are required.

Transfer learning encompasses machine learning methods that can be used to leverage the knowledge or capabilities of a model and apply it to a new domain [18]. A common approach to transfer learning involves updating the model's parameters by optimising over training examples in the new domain, which is known as finetuning. Such methods are often suited to few-shot learning, since a powerful but general model can be tuned towards the target task with a small number of training examples.

For table-to-text generation, Chen at al. [19] utilise a pre-trained language model (GPT-2 [20]) and finetune it in a number of different few-shot settings between 50 and 100 examples. They expand the original WikiBIO dataset to include tables on books and songs, and show significant improvement over baseline models in all few-shot settings. Gong et al. [21] further improve this method by using structure aware mechanisms as part of their architecture. Another type of pre-trained model is the sequence-to-sequence (seq2seq) model, often used in applications such as summarisation. Yermakov et al. [22] finetune seq2seq transformer based models including BART [23] and T5 [24], to generate drug labels from tabular data. They utilise a medium sized dataset (∼1,300 examples), but prove the efficacy of their methods in the biomedical domain.

2.4.2 Prompting

A promising new method is prompting, where we transform the input of our task into a prompt which is given as input to a language model. The input is formulated such that the language model is prompted to complete the text with the desired output to our task [25]. Prompting was proposed by Brown et al. [26] as an alternative to finetuning. They showed that a large language model (LLM) could learn tasks given just a few examples for context. When using powerful LLMs, only around 3-5 examples are necessary, which tends to be far less than is required for finetuning. Example 2 shows how a large language model can be prompted to give the desired output as a natural continuation.

Example 2 (In-Context Learning)

INPUT:

#	PSD	Dose	Age	
1	Tramadol	50mg	40-64	⇒ A 40-64 year old patient took 50mg Tramadol
2	Morphine	80mg	18-39	⇒ An 18-39 year old patient took 80mg Morphine
3	Naproxen	20mg	>64	⇒

OUTPUT: An older than 64 year old patient took 20mg Naproxen...

In the task of table-to-text generation, some works [27, 28] innovate methods to improve performance of existing seq2seq and language models, by utilising adaptive prompts. However, there is currently a gap in the literature regarding the use of prompting for T2T generation, without parameter finetuning. This gap is expected, given the only recent availability of high-performance, open-access LLMs. Furthermore, a major limitation of prompting is that the quality of the generations is entirely and unpredictably dependent on the prompt. This limitation has sparked research into prompt-based learning, which is the task of searching for optimal prompts for a given problem [29].

More recently, models such as ChatGPT[3], that are finetuned to respond to instructions in their input, were introduced. This makes a simpler type of prompting possible, where a task can be explicitly written as an instruction, rather than giving contextual examples. While generations are still volatile with respect to the prompt, instruction finetuning makes these changes much more interpretable. This means we can quickly find high quality prompts for tasks using prompt engineering, without the need for prompt-based learning.

[3] `https://openai.com/chatgpt`

3 | Methodology

In this chapter, we introduce our novel Table-to-Text (T2T) generation method for pharmacovigilance. Building on the work of Wang et al. [1], our T2T method replaces the existing sentence generation template used in the InferBERT. Fig. 3.1 shows how our work fits into the existing framework. Our method has three main advantages over using a template:

(1) ML-based T2T generation can be applied to arbitrary input data, containing different clinical features. This overcomes a significant limitation of the template approach, which must be manually crafted for a given feature set.

(2) T2T generation using pre-trained models can leverage learned biomedical/language knowledge to infer implicit relations between terms in the tables, and explicitly embed them in the artificial sentences.

(3) Large language models can produce faithful, complete and fluent sentences based on tabular pharmacovigilance reports. These sentences are closer in style to ALBERT's pre-training data.

The significance of items (2) and (3) is that by improving the artificial sentences that we generate, we allow the language model to make better predictions based on these sentences. The prediction $P(\texttt{endpoint} \mid s_i)$ for each sentence s_i is used as part of the do-calculus based causal inference to compute the interventional probabilities $P(\texttt{endpoint} \mid \mathrm{do}(t_j))$ for each term t_j. This means by using a better T2T generation method, we are able to produce better causal results.

However, as a task with domain-specific language, training data is very limited. Machine learning models trained solely on a small dataset are not able to generalise to new data. We overcome this limitation by using transfer learning. In addition, transfer learning can inject biomedical knowledge encoded in large pre-trained models into sentence generations. In our work, we used two distinct transfer learning approaches for T2T generation: Finetuning sequence-to-sequence (seq2seq) models, and prompting large language models. In the following sections, we describe in detail the methodology we used for each approach and the decisions we made.

3.1 Finetuning Sequence-to-Sequence Models

A natural approach to model this task is as a seq2seq problem where the input is the table encoded as a sequence, and the output is the descriptive sentence. In order to transform the input reports into sequence, we used the encoding scheme shown in example 3. This encoding scheme was inspired by Yermakov et al. [22], who use seq2seq finetuning for drug label generation.

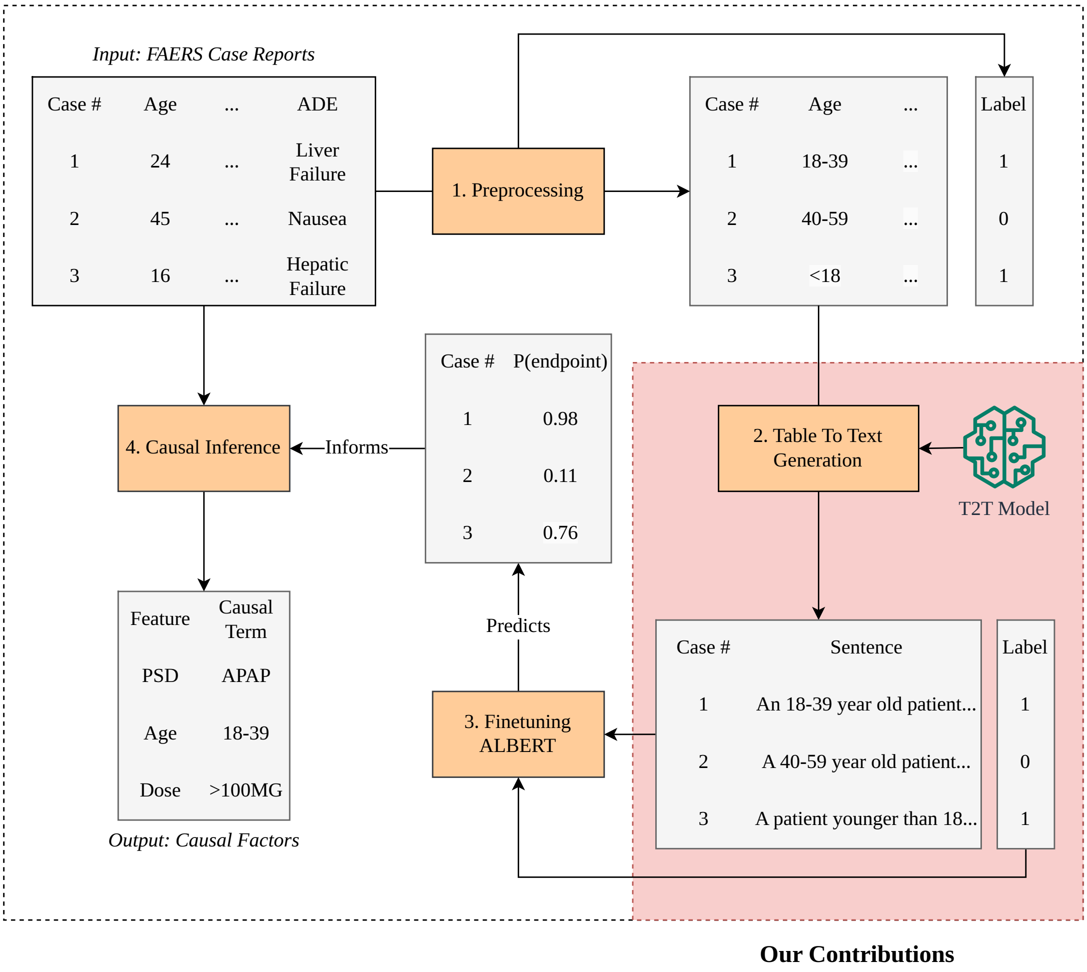


Figure 3.1: How our T2T generation method fits into InferBERT

Example 3 (Table Encoding)

Feature	Term
PSD	Tramadol
Dose	50-100mg
Age	40-59
ADE	Anxiety

⇒ "<psd>Tramadol</psd><dose>50-100MG</dose> <age>40-59</age><ade>Anxiety</ade>"

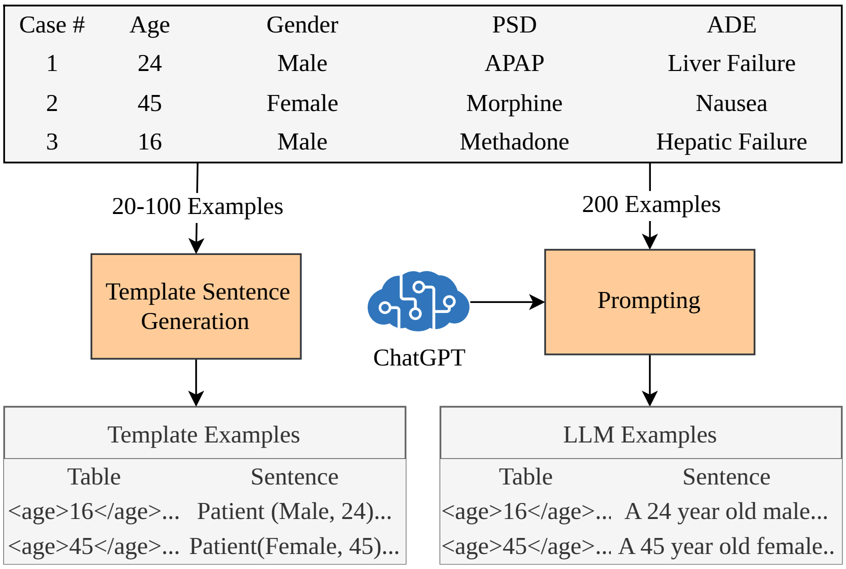
State-of-the-art Seq2seq models are based on the Transformer architecture, and are frequently used in tasks such as summarisation or machine translation. The most related task we found to our work was the use of T2T generation for medicine leaflet generation [22]. We took inspiration from their method and finetuned BART-Large [23], which is a recent Transformer-based seq2seq model. The model consists of an encoder (like BERT) and a decoder (like GPT), which makes it well suited to tasks whose input and output are arbitrary length sequences.

3.1.1 Training Data

To supervise the seq2seq finetuning, we needed an annotated dataset of reports. Fig. 3.2 shows an overview of our two annotation methods to get training examples for seq2seq finetuning.

We used the first method as part of our pilot study, and annotated each report with the template used in InferBERT; each encoded tabular report has a template-constructed sentence as annotation. We applied this method for different number of training examples to evaluate the model's ability to learn the task in a few-shot setting.

For our next method, we wanted to extend the annotations beyond the template approach. After testing zero-shot T2T generations using ChatGPT, we decided to use them as annotations. This is because ChatGPT has usage limits and could not be applied directly to the large datasets. Example 4 shows how we prompted ChatGPT to annotate the data.



Case #	Age	Gender	PSD	ADE
1	24	Male	APAP	Liver Failure
2	45	Female	Morphine	Nausea
3	16	Male	Methadone	Hepatic Failure

Template Examples

Table	Sentence
<age>16</age>...	Patient (Male, 24)...
<age>45</age>...	Patient(Female, 45)...

LLM Examples

Table	Sentence
<age>16</age>...	A 24 year old male...
<age>45</age>...	A 45 year old female..

Figure 3.2: Overview of our annotation methods for seq2seq finetuning

Example 4 (Zero-Shot Annotation using ChatGPT)

INPUT[1]: Convert this encoded ADE report into a single concise sentence:

Feature	Value
PSD	Tramadol
ADE	Vomiting
Outcome	Disability
Gender	Female
Age	57
Indications	Pain

OUTPUT: A 57-year-old female experienced disability after taking Tramadol for pain, reporting adverse events of vomiting.

Finally, given the annotated pharmacovigilance reports, we split the dataset, with 75% for training and 25% for validation. Note that there is no test set here, since evaluation will be performed downstream.

3.1.2 Training

In order to finetune BART for our task, we updated the model's parameters by optimising the cross entropy loss over the training set. We did not optimise the parameters in the encoder, because like BERT, the contextualised embeddings are effective without further tuning. Instead, we optimise the decoder, which is responsible for generating the output sequence given the input embeddings. The BART model was imported using HuggingFace Transformers, and the training procedure was implemented in PyTorch. We use the same hyperparameters for each of our seq2seq experiments, which can be seen in table 3.1. We use the validation set to ensure the model is generalising, and to determine when to stop training.

[1] table encoded as shown in example 3

Table 3.1: Hyperparameters used to finetune BART for table-to-text generation

Parameter	Value
Model	BART-Large
Loss Function	Cross Entropy
Optimiser	Adam
Learning Rate	0.001
Batch Size	16

3.2 Prompting Large Language Models

An alternate way to model this task is as a text-completion problem, where we transform the input sequence into a prompt which a large language model such as GPT-3 [26] attempts to complete. While prompting can be directly applied to a pre-trained language model, the quality of the results is volatile with respect to the prompt. We considered applying prompt-based learning to overcome this, but found that additional overhead and training examples were necessary.

Instead we decided to make use of instruction finetuned models. Prompting these models is a simpler process, since instructions can be given explicitly, rather than giving implicit context. While the same volatility is observed - it is easier to understand why certain prompts are more effective than others. For this reason we could rely on simple prompt engineering, instead of automated prompt-based learning. Furthermore, instruction finetuned models are able to perform generations in a zero-shot setting, which means we need little to no training examples to perform our task. In our seq2seq finetuning method, we already used ChatGPT in a zero-shot for annotation, but it was limited to 200 training examples, since ChatGPT has usage limitations. However, by leveraging the open-access Llama-2-Chat model, we were able to generate sentences for an entire case study.

3.2.1 Prompt Design

When designing a prompt to generate the sentences from the tabular reports, there were a number of important considerations.

- The prompt should be in the same format as the training prompts for Llama2Chat, including a system prompt with instructions.
- The instructions should be concise, to avoid giving a prompt which is too long to the model.
- The instructions should incite complete and faithful generations
- The instructions should cause the model to delimit the answer from its own verbiage

With this in mind, we conducted experiments with different prompts, and performed manual evaluation on the generations. Through our observations, we encountered a number of problems that needed addressing. The model could not deal well with empty fields in the input, and on occasion misunderstood the structure of the tabular report. To mitigate these issues and improve the generations, we first modified the input encoding, so that empty features were simply omitted, then we provided a single example (one-shot) to help the model understand the report structure. Our final prompt, as well as an example generation is given in example 5.

Example 5 (One-shot LLama2Chat Generation)

SYSTEM: Your role is a table-to-text generator for pharmacovigilance reports. You take as input a tabular report and give as output a single concise sentence which describes the report. Include every term provided in the output sentence. You must always use quotation marks around the summary sentence to delimit it. Here is an example: [example given]
USER: <age> 18-39 </age> <indication> back pain </indication>...
ASSISTANT: Sure! Here is the generated sentence “An 18-39 year old patient experienced back pain and took...”

3.3 Introducing Further Clinical Features

To test the generalisability of our T2T methods, we extended the preprocessing steps to include more features. We decided to include patient weight, administration route and secondary suspect drugs. While further features about administered drugs could be used from the dataset, the input lengths would become too long for models to handle given over around 5 different drugs. For this reason we chose secondary suspect drugs, which are more likely to be relevant to the case. This meant that there were three new types of terms which could be considered as causal factors in the InferBERT framework.

Patient weights are normalised into kilograms, and placed into one of four categories. This is necessary because the causal inference implementation can only consider identical terms, which is rare for continuous values. The administration route is truncated so that only the first word is used, e.g. ‘oral’ or ‘intravenous’. We only include a secondary suspect drug if it does not first appear in the primary suspect drug feature, and limit the total number of drugs to four. We normalise uninformative terms such as ’unknown’ into empty terms for all features.

Finally, we had to update the existing implementation of the causal inference algorithm. This is because by default, the procedure was only applied to the original 6 clinical features. We updated the implementation to apply the procedure to all features that are supplied by the preprocessing steps.

Aside from using additional features, the pipeline remained the same. We preprocessed the extended features, used them to generate sentences using a T2T method. The sentences are used to finetune ALBERT and finally we apply causal inference informed by the ALBERT-based classifier to output causal factors.

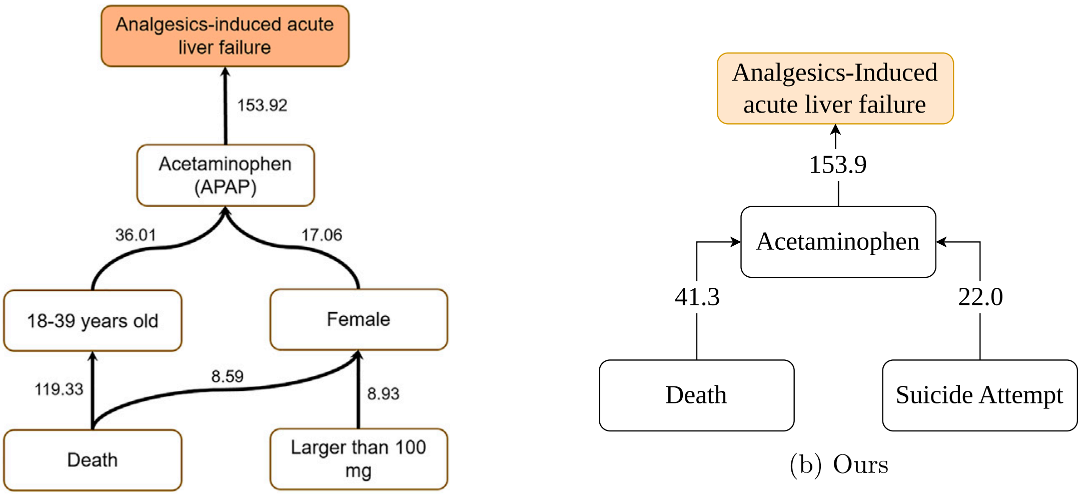
3.4 Evaluation

To evaluate the results of our experiments, we followed the standard set by Wang et al. [1] and report the top term for each feature, by Z-score. To remain consistent with the paper’s results, we ignore terms containing ‘hepatic’ as indication, and ‘oral’ as a dose. We also measure the proportion of shared terms between sets of results, to give a metric of similarity between results produced by different methods. Where appropriate, we perform intrinsic evaluation of the T2T generated sentences, by comparing them to a reference. This is done using metrics such as ROUGE [30].

3.4.1 Causal Tree Construction

To graphically compare the different results, we constructed causal trees for each set of results using Algorithm 1. In fig. 3.3, we compare the tree produced by our algorithm and the tree produced by Wang et al. [1]. Both trees were produced using the same causal results data reported in the original paper. While the trees are similar, there are some differences.

Because the authors' methods to produce their tree were undocumented, we do not know why these differences occur. We attempted to reverse engineer their method, but no consistent set of decisions could be found that could produce the tree. For example, they use the value of 119 for 'Death' in level 2, which comes from the Z-score for 'Death' as a root (level 0) cause. Furthermore, the value 8.59 appears in their causal tree, but could not be found in the raw results they report. For these reasons, we will compare the results of our experiments only to the trees produced by our own method.



(a) Wang et al. [1]'s method

(b) Ours

Figure 3.3: Comparison of tree generation methods using reported results from InferBERT

3.4.2 Interpretation of Causal Results

Causal trees are type of causal graph, with the additional constraint that each node may only have at most one parent. Technically speaking, the models we produce here are only trees if you reverse every arrow. However, this would have no impact on the interpretation of the graph, so we can still consider them trees for our analysis.

These trees allow us to make interpretations about the causal factors of the study's endpoint. The root term can be interpreted as a direct cause of the endpoint. The terms below the root are conditional causes, meaning that the causal relation depends on any terms they pass through to reach the endpoint. For example, in fig. 3.3a the term 'Female' is a cause of the endpoint if and only if Acetaminophen was administered. All the causal results we produce are empirical estimates of causality and cannot be considered proof of a causal relation between given terms. For instance the term 'death' appears as a conditional cause of liver failure, but we know that liver failure is not diagnosed posthumously, and therefore cannot be considered a cause. Terms like these may appear as artefacts of the dataset and therefore all results should only be interpreted as a causal hypotheses not as proofs.

Algorithm 1 Causal Tree Construction

```
function CONSTRUCT TREE(results,endpoint)
    $G$ = (V,E)
    $G$.add_vertex(endpoint)
    level_count = $[0, 0, 0]$
    used_terms = $\emptyset$
    sort results by Z-score
    for row in results do
        level, route, value, Z-score $\leftarrow$ row
        if term in used_terms or level_count[level]>level then
            skip iteration
        end if
        if ‘po’ in term or ‘hepatic’ in term then
            skip iteration
        end if
        if level=0 then
            $G$.add_vertex(value)
            $G$.add_edge(value, endpoint, weight=Z-score)
            level_count[level] $\leftarrow$ level_count[level]+1
            used_terms.add(value)
        else
            child_node $\leftarrow$ route[-1]
            if child_node in $G$ then
                $G$.add_vertex(value)
                $G$.add_edge(value,child_node,weight=Z-score)
                level_count[level]+=1
                used_terms.add(value)
            end if
        end if
    end for
     return $G$
end function
```

4 | Experiments

In this chapter, we discuss our experiments and results using the methodology described in chapter 3. We give our results for each experiment in the respective section, and the full raw results can be found in appendix A.

4.1 Control

To establish a control for our studies, we ran the unchanged InferBERT framework three times on the Analgesics-induced liver failure dataset. Table 4.1 shows the mean values for the top causal terms over the three runs. The mean number of statistically significant ($p<0.05$) causal terms per feature, across the three control runs can be seen in table 4.2. Note that we ignored 'malignant liver tumour' and 'oral' as a dose as terms, to be consistent with how the authors present the original results. In addition, we compute the proportion of shared terms between runs as 86.1%, which approximately concurs with Wang et al. [1], who report that 82.4% of terms are shared across three runs. We use the same finetuning parameters and implementation, adapted from the original ALBERT code[1]. Despite this, our control results achieve lower Z-scores across the board; this discrepancy likely comes from a difference in training hardware. For this reason, we will compare our experimental results to this control and not those reported in the paper.

We also produced a causal tree for the first control run, which can be seen in fig. 4.1b. To fairly compare this to the paper's results, we used our own function to generate a tree from the paper's raw results (fig. 4.1a). From these trees we can conclude that the results we reproduced for our control are different (as many terms have changed order of significance), but between our three control runs there were no changes to the structure of the generated trees. This is important since when we see changes in the tree in later runs we know this occurred due to changes we make to the framework, not random differences.

Table 4.1: Top causal terms and mean statistics for three control runs of Inferbert

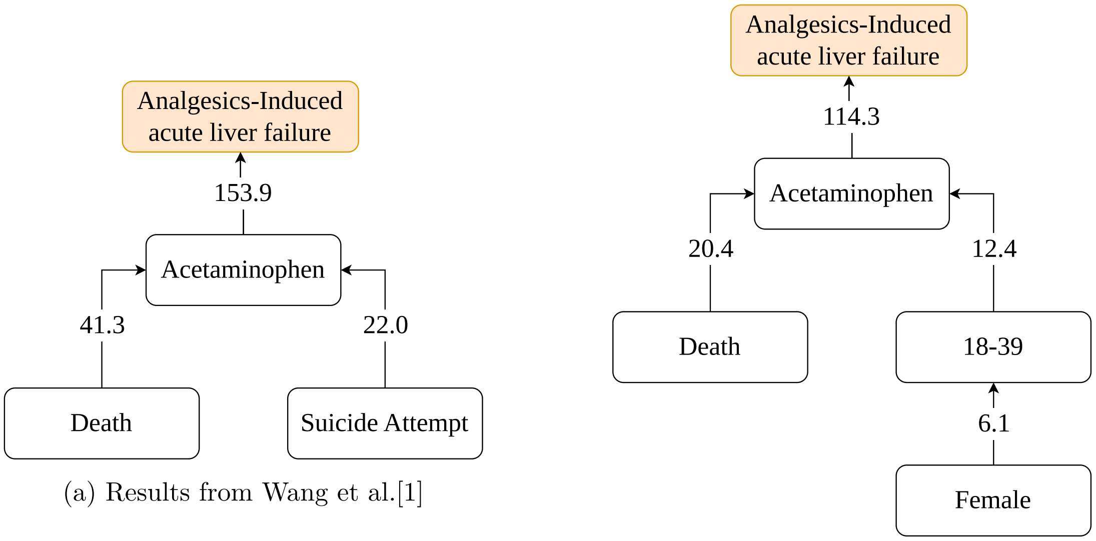
Feature	Top Term	mean P(do)	mean P(not do)	mean Z-score and (s.d)
PSD[2]	Acetaminophen	0.85	0.35	114.95 (4.46)
Outcome	Death	0.71	0.32	77.94 (0.59)
Age	18-39	0.56	0.37	27.29 (0.11)
Indication	Suicide Attempt	0.81	0.38	16.85 (0.12)
Gender	Female	0.43	0.38	10.26 (0.01)
Dose	>100 mg	0.40	0.37	5.01 (0.21)

[1] `https://github.com/google-research/albert`

[2] Primary Suspect Drug

Table 4.2: Mean causal terms per feature, over three control runs of InferBERT

Feature	Mean Number of Terms
PSD	14.00
Outcome	3.00
Age	2.00
Indication	11.67
Gender	1.00
Dose	2.00
Total	33.67



(a) Results from Wang et al.[1]

(b) Our first control run

Figure 4.1: Comparison of causal trees produced from the reported results, and our control

4.2 Pilot Study

Our first experiment's goal was to train a T2T model using the template-generated artificial sentences as training examples. We applied our seq2seq finetuning method (section 3.1) with 3000 template annotated training examples from the Analgesics-Induced Liver failure dataset. We compared the output of the finetuned model (template-BART) and the template itself and computed an average[3] Rouge score of 0.996. Rouge is a set of metrics comparing similarity between reference and generated text [30]. This suggests that using a finetuned seq2seq model can effectively mimic the template.

To verify that the sentences produced consistent results downstream, we replaced the default template with template-BART sentences in the InferBERT framework, producing a new set of causal results. We compared these to our control runs; the average proportion of shared terms between template-BART and the three control runs was 0.806. Furthermore, the main causal terms for each feature were the same and it identified 30 total causal terms. These results are highly consistent with our control runs, and thus we conclude that template-BART is able to reproduce the results with no less variation than the template itself.

[3] Mean of Rouge-1 Rouge-2 and Rouge-L

4.2.1 Few-shot Performance

To test whether similar results could be produced with fewer training examples, we reduced them to 20, 50 and 100. Table 4.3 shows the resulting ROUGE metrics comparing each model's outputs to the template generated sentences.

Results showed that 20 and 50 examples were insufficient for BART to adequately produce sentences from the reports, but 100 was sufficient. Interestingly the model trained on 50 examples performed worse in this metric than the one trained on 50. This unexpected result can be attributed to the 50-model was being more 'ambitious' than the 20-model, causing some drops in similarity score.

Table 4.3: Similarity metrics between sentences generated by the template, and BART finetuned using different numbers of examples

# Examples	Rouge-1	Rouge-2	Rouge-L	Mean
20	0.551	0.271	0.548	0.457
50	0.528	0.259	0.526	0.437
100	0.838	0.713	0.829	0.793
3000	0.997	0.994	0.997	0.996

4.3 T2T with LLM supervision

The template-BART model was able to produce similar results to the template. However, to improve this, we wanted to generalise the model to consider arbitrary features and generate fluent sentences. To do this we utilised the dataset from the paper's other case study (Tramadol Related Mortalities) and extracted 200 reports. From our few-shot experiments we found that around 100 training examples were sufficient to mimic the template. However because the goal of this model is to be generalisable to new features, we considered this a minimum number and increased this to 200 for this experiment.

We annotated the 200 reports using ChatGPT, as explained in section 3.1.1, and used them as training examples to finetune the BART-Large model. We plotted the loss curve for the finetuning over 20 epochs in fig. 4.2. Using this we determined that due to convergence of training and validation loss, it was sufficient to train to only 10 epochs, and so we retrained only up to this point. We applied the finetuned model to the Analgesics Induced Liver Failure dataset. Table 4.4 shows the results of applying the InferBERT framework using the sentences generated by the general T2T model. We plot the causal tree generated from this data in fig. 4.3.

Table 4.4: Top causal terms for results using LLM-supervised BART

Feature	Term	P(do)	P(not do)	Z-score
PSD	Acetaminophen	0.83	0.35	142.56
Outcome	Death	0.67	0.32	79.22
Age	18-39	0.56	0.37	30.16
Indication	Suicide Attempt	0.86	0.37	25.59
Gender	Female	0.43	0.38	10.14
Dose	>100 MG	0.41	0.37	5.56

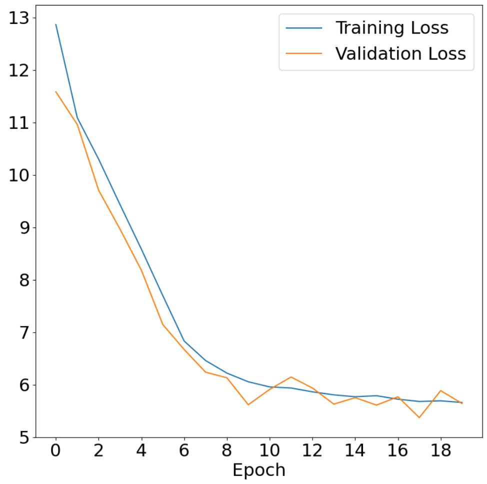


Figure 4.2: Loss curve for BART finetuning over 20 epochs

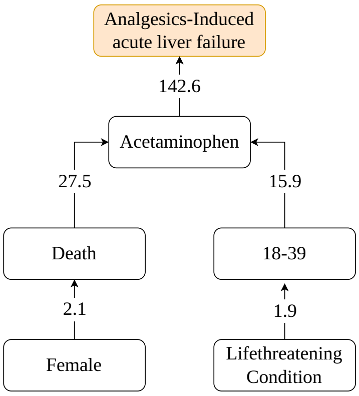

Analgesics-Induced acute liver failure
142.6
Acetaminophen
27.5
15.9
Death
18-39
2.1
1.9
Female
Lifethreatening Condition

Figure 4.3: Causal tree for results using LLM-supervised BART

4.4 Prompting LLMs

We applied our prompting method to Llama-2-Chat 13B, as explained in section 3.2, to the same case study and used the sentences as part of the InferBERT framework. The primary causal terms for each feature can be seen in table 4.5 and the causal graph constructed from this set of results can be seen in fig. 4.4

Table 4.5: Top causal terms for results using LLama-2-Chat

Feature	Term	P(do)	P(not do)	Z-score
PSD	Acetaminophen	0.84	0.35	109.95
Outcome	Death	0.71	0.32	78.59
Age	18-39	0.57	0.38	27.70
Indication	Suicide Attempt	0.81	0.38	15.64
Gender	Female	0.43	0.38	9.84
Dose	>100 MG	0.40	0.38	4.08

4.5 Feature Generalisability

We tested two of our T2T methods using the extended features we processed in section 3.3: the LLM-supervised BART model, and Prompting Llama2Chat. We quickly found that LLM-supervised BART did not succeed in generalising to more features; we observed from a small test that none of the reports included the extra features. For this reason we did not proceed the rest of the reports and moved on to our next method.

To adapt the Llama-2-Chat prompt for the extra features, we simply included them as part of the single example generation. Adapting the prompt is much faster than incorporating the additional features in a template. From early tests we could see that the model was able to adapt to the new features, and produced high quality sentences. We applied this method, with the modified example in the prompt to the liver failure case study, with the extended features. Table 4.6 shows the top terms for each feature by Z-score, and the causal tree for this set of

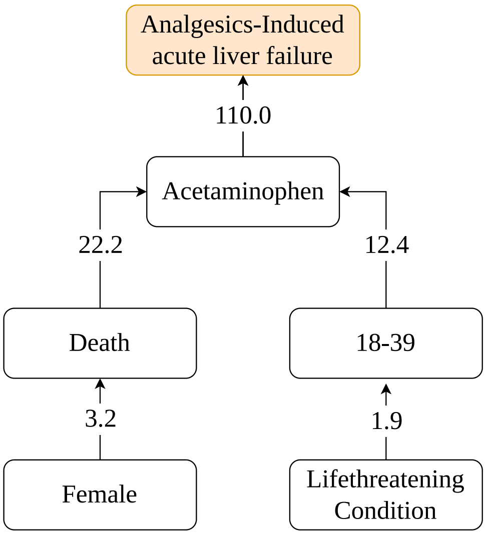

Analgesics-Induced acute liver failure
110.0
Acetaminophen
22.2
12.4
Death
18-39
3.2
1.9
Female
Lifethreatening Condition

Figure 4.4: Causal tree for results using Llama2Chat

results is plotted in fig. 4.5.

Table 4.6: Top causal terms for results using LLama2Chat, with extended features

Feature	Term	P(do)	P(not do)	Z-score
PSD	Sorafenib Tosylate	0.76	0.42	16.15
Outcome	Death	0.72	0.32	74.44
Age	18-39	0.58	0.37	27.69
Indication	Suicide Attempt	0.80	0.38	14.12
Gender	Female	0.43	0.38	10.39
Dose	>100 MG	0.41	0.37	5.57
Secondary suspect drug	Diphenhydramine Hydrochloride	0.82	0.36	15.55
Weight	50-70	0.44	0.39	5.75
Route	Intravenous	0.43	0.38	5.13

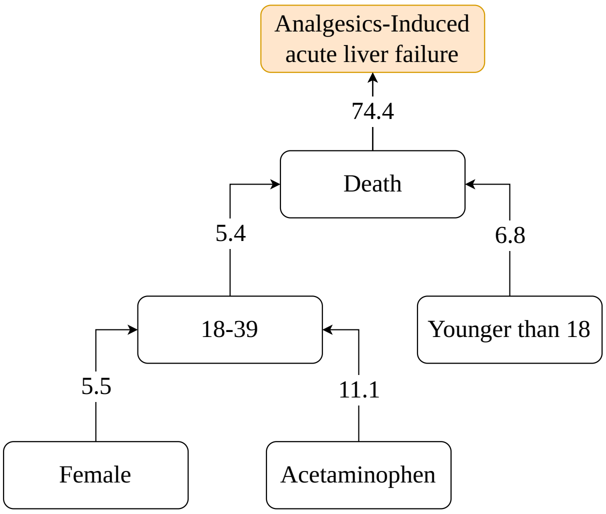


Figure 4.5: Causal tree for results using Llama-2-chat, with extended features

5 | Discussion

Our early experiments confirmed that we could reproduce the control results using BART finetuned for table-to-text generation instead of a template. By comparing generated sentences to the template as a reference, we found an estimate for the minimum number of training examples that were necessary to do so. Our results showed that with 100 training examples, we could achieve a mean ROUGE score of 0.793 between sentences from template-BART and the template itself.

Using 200 training examples, we were able to train GPT-supervised BART, which was able to produce better sentences than the template for the original feature set. The downstream results suggested greater confidence, with Z-scores well over the mean from our control runs. Furthermore, its causal tree was more extensive than our control runs. The total number of identified terms was 35, which is marginally higher than the control average. One unique term identified in these results was Paclixatel with a Z-score of 1.75. This PSD was not identified in the results of any prior experiment, nor the paper's reported results. There is at least one report of liver failure following a reaction to Paclixatel in literature, but there is no proven link [31]. Overall, we find that the results using the LLM-supervised model are on par with those using a template.

Similar results were achieved by prompting large language models on the original feature set. The causal tree had the same structure as GPT-Supervised BART, but Z-scores were more consistent with the control. However, by prompting the language model to generate sentences directly from the dataset, our method was able to extend InferBERT beyond its original scope. This is evidenced by its ability to identify statistically significant terms in the newly introduced features. While most causal terms remained consistent through the features, one feature changed significantly. Acetaminophen, which was the root cause in every causal tree so far, was no longer considered a significant causal term for the primary suspect drug. Instead, it appeared as a causal term for secondary suspect drugs. From our test with LLama on the original feature set, we know that our method is not responsible for this change, rather it is due to the introduction of the additional features. More specifically, without considering secondary suspect drugs, APAP was considered a very strong and primary cause; once secondary suspect drugs were introduced, APAP was considered to be a smaller but still significant causal term.

Furthermore, we see two new drugs as top causal terms in our extended features experiment. Sorafenib appears as the top PSD, with a Z-score of 16.25. According to LiverTox, Sorafenib is a likely cause of liver injury [32]. On the other hand, Diphenhydramine, which appears as the top SSD, with a Z-score of 15.55, is considered unlikely to be a cause of liver injury [33]. With regards to the other two features we added, a weight of 50-70 kg was considered a causal term, but is not especially informative. The intravenous route was also a causal term, which aligns with our intuition that administration directly into the bloodstream contributes to the risk of liver failure.

6 | Conclusions

6.1 Contributions

In our work, we addressed a limitation in a recent AI approach to causal inference, for pharmacovigilance. Our main results indicate that our LLM prompting method allows us to generalise InferBERT to arbitrary feature sets, and through our evaluation it finds causal terms within these new features. Furthermore, when applied to the same feature set, this method appears to produce more conditional causes, as shown in the generated causal trees. This is likely due to the improved sentences we generate, by leveraging pre-trained models' biomedical knowledge.

We met all the primary objectives we set out in the project proposal. These include collating research across numerous disciplines, generating improved sentences from pharmacovigilance reports, and finally expanding causal results using these improved sentences.

6.2 Limitations

The main limitation of our methods are the computational resources and time required to generate sentences, via transformer-based models. For example, generating sentences for the 36,000 reports used in the case study took around 2.5 hours for BART and 11 hours for LLama-2-chat. However, this limitation will become less significant in the future, since new innovations in LLMs and inference optimisations are rapidly being made.

While we have generalised the framework to be able to consider arbitrary clinical features, it is still limited in domain. This pipeline could be leveraged for other types of causal inference, but further work is needed to improve the framework for this purpose. InferBERT itself comes with many of its own limitations, including a lack of explainability. Implicitly encoding knowledge is an attractive concept for causal inference, but it makes it difficult to interpret how the results are achieved. Nevertheless, by constraining these classification results to a proven causal inference algorithm, the produced results remain robust.

6.3 Future Work

In the future we will apply the framework, with our additions, to new datasets and case studies, further proving its generalisability. Furthermore, we wish to study the relationship between the ALBERT model's final accuracy, and the downstream performance in causal inference. This work would help us make more informed decisions about feature selection and the number of features that is appropriate to include in a given study.

In addition, there are a number of ways our methods could be extended to improve the framework. In particular, while finetuning remains necessary in the framework, it might be possible to finetune a different model directly on the encoded reports, instead of artificial sentences. This would require a single model which is able to infer the information from the table, and make predictions using biomedical knowledge. This idea would require further experimentation to see if there is potential.

6.4 Climate Impact

By utilising large language models, we used GPU accelerated inference code. For the seq2seq finetuning and inference we used 1 Nvidia RTX 6000 GPU, with an approximate usage of 80 GPU hours. For inference with Llama-2-chat (13B) we used 2 Nvidia RTX 6000 GPUs with an approximate usage of 97 GPU hours. Given the maximum power consumption of each GPU is 300W, we get an upper bound energy usage of approximately 53.1kWh. On average, in the United Kingdom, this energy usage is equivalent to the emission of around 14kg CO_2eq.

6.5 Reproducibility

The supplementary data contains all the code necessary to reproduce the experiments detailed in this report. ALBERT and BART finetuning is not deterministic and will produce slightly different results each time, as explained in section 4.1. Furthermore differences in training hardware may cause small discrepancies in the results. Results can be reproduced by following the instructions in `readme.md` and including the necessary model files specified. For convenience we uploaded the raw dataset and two finetuned models to Figshare: `https://doi.org/10.6084/m9.figshare.24077526`.

6.6 Data Privacy

All data used in this project originates from FAERS, an open access database for pharmacovigilance reports. The database does not contain any personal information, and we do not make use of database features such as location, reporter occupation or date of event. The specific dataset curated from the FAERS data was downloaded from `https://github.com/XingqiaoWang/DeepCausalPV-master` which is the repository containing the code and data for InferBERT [1].

Bibliography

[1] X. Wang, X. Xu, W. Tong, R. Roberts, and Z. Liu, "InferBERT: A Transformer-Based Causal Inference Framework for Enhancing Pharmacovigilance," *Frontiers in Artificial Intelligence*, vol. 4, 2021.

[2] C. Kongkaew, P. R. Noyce, and D. M. Ashcroft, "Hospital admissions associated with adverse drug reactions: A systematic review of prospective observational studies," *Annals of Pharmacotherapy*, vol. 42, no. 7-8, pp. 1017–1025, 2008.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

[4] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations," in *International Conference on Learning Representations*, Mar. 2020.

[5] A. M. Wilson, L. Thabane, and A. Holbrook, "Application of data mining techniques in pharmacovigilance," *British Journal of Clinical Pharmacology*, vol. 57, no. 2, pp. 127–134, Feb. 2004.

[6] Y. Zhao, Y. Yu, H. Wang, Y. Li, Y. Deng, G. Jiang, and Y. Luo, "Machine Learning in Causal Inference: Application in Pharmacovigilance," *Drug Safety*, vol. 45, no. 5, pp. 459–476, May 2022.

[7] P. C. Austin, "An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies," *Multivariate Behavioral Research*, vol. 46, no. 3, pp. 399–424, May 2011.

[8] J. Pearl, "Fusion, propagation, and structuring in belief networks," *Artificial Intelligence*, vol. 29, no. 3, pp. 241–288, Sep. 1986.

[9] J. Pearl, "Causal Diagrams for Empirical Research," *Biometrika*, vol. 82, no. 4, pp. 669–688, 1995.

[10] J. Pearl, "The do-calculus revisited," in *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, ser. UAI'12. Arlington, Virginia, USA: AUAI Press, Aug. 2012, pp. 3–11.

[11] K. Kreimeyer, O. Dang, J. Spiker, M. A. Muñoz, G. Rosner, R. Ball, and T. Botsis, "Feature engineering and machine learning for causality assessment in pharmacovigilance: Lessons learned from application to the FDA Adverse Event Reporting System," *Computers in Biology and Medicine*, vol. 135, p. 104517, Aug. 2021.

[12] D. J. Monlezun, S. Lawless, N. Palaskas, S. Peerbhai, K. Charitakis, K. Marmagkiolis, J. Lopez-Mattei, M. Mamas, and C. Iliescu, "Machine Learning-Augmented Propensity

Score Analysis of Percutaneous Coronary Intervention in Over 30 Million Cancer and Noncancer Patients," *Frontiers in Cardiovascular Medicine*, vol. 8, 2021.

[13] A. Feder, K. A. Keith, E. Manzoor, R. Pryzant, D. Sridhar, Z. Wood-Doughty, J. Eisenstein, J. Grimmer, R. Reichart, M. E. Roberts, B. M. Stewart, V. Veitch, and D. Yang, "Causal Inference in Natural Language Processing: Estimation, Prediction, Interpretation and Beyond," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 1138–1158, 2022.

[14] F. Mairesse, R. Prasad, A. Stent, and M. A. Walker, "Individual and Domain Adaptation in Sentence Planning for Dialogue," *Journal of Artificial Intelligence Research*, vol. 30, pp. 413–456, Nov. 2007.

[15] R. Lebret, D. Grangier, and M. Auli, "Neural Text Generation from Structured Data with Application to the Biography Domain," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing.* Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 1203–1213.

[16] T. Liu, K. Wang, L. Sha, B. Chang, and Z. Sui, "Table-to-Text Generation by Structure-Aware Seq2seq Learning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, Apr. 2018.

[17] Y. Yang, J. Cao, Y. Wen, and P. Zhang, "Table to text generation with accurate content copying," *Scientific Reports*, vol. 11, no. 1, p. 22750, Nov. 2021.

[18] S. J. Pan and Q. Yang, "A Survey on Transfer Learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[19] Z. Chen, H. Eavani, W. Chen, Y. Liu, and W. Y. Wang, "Few-Shot NLG with Pre-Trained Language Model," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.* Online: Association for Computational Linguistics, Jul. 2020, pp. 183–190.

[20] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[21] H. Gong, Y. Sun, X. Feng, B. Qin, W. Bi, X. Liu, and T. Liu, "TableGPT: Few-shot Table-to-Text Generation with Table Structure Reconstruction and Content Matching," in *Proceedings of the 28th International Conference on Computational Linguistics.* Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 1978–1988.

[22] R. Yermakov, N. Drago, and A. Ziletti, "Biomedical Data-to-Text Generation via Fine-Tuning Transformers," in *Proceedings of the 14th International Conference on Natural Language Generation.* Aberdeen, Scotland, UK: Association for Computational Linguistics, Aug. 2021, pp. 364–370.

[23] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.* Online: Association for Computational Linguistics, Jul. 2020, pp. 7871–7880.

[24] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 140:5485–140:5551, Jan. 2020.

[25] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, “Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing,” *ACM Computing Surveys*, vol. 55, no. 9, pp. 195:1–195:35, Jan. 2023.

[26] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, “Language Models are Few-Shot Learners,” in *Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901.

[27] Y. Luo, M. Lu, G. Liu, and S. Wang, “Few-shot Table-to-text Generation with Prefix-Controlled Generator,” in *Proceedings of the 29th International Conference on Computational Linguistics*. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, Oct. 2022, pp. 6493–6504.

[28] Z. Guo, M. Yan, J. Qi, J. Zhou, Z. He, Z. Lin, G. Zheng, and X. Wang, “Adapting Prompt for Few-shot Table-to-Text Generation,” Aug. 2023.

[29] T. Gao, A. Fisch, and D. Chen, “Making Pre-trained Language Models Better Few-shot Learners,” in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 3816–3830.

[30] C.-Y. Lin, “ROUGE: A Package for Automatic Evaluation of Summaries,” in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81.

[31] “Paclitaxel,” in *LiverTox: Clinical and Research Information on Drug-Induced Liver Injury*. Bethesda (MD): National Institute of Diabetes and Digestive and Kidney Diseases, 2012.

[32] “Sorafenib,” in *LiverTox: Clinical and Research Information on Drug-Induced Liver Injury*. Bethesda (MD): National Institute of Diabetes and Digestive and Kidney Diseases, 2012.

[33] “Diphenhydramine,” in *LiverTox: Clinical and Research Information on Drug-Induced Liver Injury*. Bethesda (MD): National Institute of Diabetes and Digestive and Kidney Diseases, 2012.

A | Raw Results

A.1 First Control Results

A.1.1 Root Causal Terms

feature	value	z score	P(do)	P(not do)	delta	p value	support
psd	Acetaminophen	114.310347	0.855975	0.344849	0.511125	0.000000e+00	4990
outcome	Death	76.882889	0.709555	0.317756	0.391799	0.000000e+00	11311
outcome	LifeThreatening	31.068673	0.638773	0.423499	0.215274	0.000000e+00	5488
age	18-39	26.847784	0.561084	0.371934	0.189151	0.000000e+00	5350
indication	Hepatoma	19.754968	0.719683	0.374743	0.344940	0.000000e+00	271
indication	SUICIDE ATTEMPT	17.234751	0.826860	0.375502	0.451358	0.000000e+00	170
age	younger than 18	15.555547	0.579967	0.395188	0.184779	0.000000e+00	1687
psd	Sorafenib Tosylate	15.524659	0.689824	0.411688	0.278135	0.000000e+00	360
dose	PO	14.604209	0.672018	0.371813	0.300204	0.000000e+00	457
gender	Female	10.429742	0.428632	0.376399	0.052233	0.000000e+00	17843
indication	PYREXIA	9.571886	0.624649	0.375628	0.249021	0.000000e+00	297
indication	HEADACHE	8.984391	0.702705	0.376927	0.325778	0.000000e+00	139
psd	Sevoflurane	8.726726	0.751948	0.413302	0.338646	0.000000e+00	121
psd	Heparin Sodium	7.477228	0.684502	0.413273	0.271229	3.796963e-14	155
psd	Bromfenac Sodium	6.007936	0.626746	0.413664	0.213082	9.395005e-10	130
outcome	RequiredIntervention	5.054285	0.535111	0.459276	0.075835	2.160030e-07	1078
dose	larger than 100 MG	4.868486	0.400621	0.366883	0.033738	5.622834e-07	6919
indication	ATRIAL FIBRILLATION	4.717715	0.523072	0.377368	0.145704	1.192544e-06	244
indication	BCa	4.413204	0.549163	0.377721	0.171442	5.092604e-06	162
indication	MULTIPLE MYELOMA	3.541390	0.506920	0.377986	0.128933	1.990126e-04	170
psd	Ibuprofen	3.363931	0.459928	0.413108	0.046820	3.842044e-04	1027
psd	Tacrolimus	3.280608	0.556439	0.413918	0.142521	5.179180e-04	129
psd	Moxifloxacin Hydrochloride	2.989044	0.551184	0.414004	0.137180	1.399259e-03	111
indication	BACK PAIN	2.961555	0.453230	0.377720	0.075509	1.530450e-03	368
psd	Rosiglitazone Maleate	2.721393	0.517636	0.413976	0.103661	3.250368e-03	157

Continued on next page

feature	value	z score	P(do)	P(not do)	delta	p value	support
indication	PULMONARY HYPERTENSION	2.537566	0.494534	0.378411	0.116122	5.581322e-03	108
psd	Troglitazone	2.448274	0.471239	0.413916	0.057323	7.177132e-03	322
indication	ACUTE MYELOID LEUKAEMIA	2.093418	0.474966	0.378464	0.096502	1.815594e-02	118
indication	PLASMA CELL MYELOMA	2.065549	0.452538	0.378392	0.074146	1.943555e-02	175
psd	Trovafloxacin Mesylate	2.006618	0.483989	0.414061	0.069928	2.239515e-02	188
psd	Peginterferon Alfa-2b	1.975796	0.496523	0.414139	0.082384	2.408896e-02	125
psd	Sofosbuvir	1.865826	0.476315	0.414174	0.062141	3.103283e-02	145
psd	Peginterferon Alfa-2a	1.744499	0.488975	0.414193	0.074782	4.053609e-02	111
psd	Rivaroxaban	1.719440	0.490042	0.414180	0.075863	4.276711e-02	116
indication	NON-SMALL CELL LUNG CANCER	1.702522	0.446767	0.378522	0.068245	4.432882e-02	148

A.1.2 Causal Tree Terms

level	route	feature	value	z score	P(do)	P(not do)	delta	p value
0	NaN	psd	Acetaminophen	114.310347	0.855975	0.344849	0.511125	0.000000e+00
0	NaN	outcome	Death	76.882889	0.709555	0.317756	0.391799	0.000000e+00
0	NaN	outcome	LifeThreatening	31.068673	0.638773	0.423499	0.215274	0.000000e+00
0	NaN	age	18-39	26.847784	0.561084	0.371934	0.189151	0.000000e+00
0	NaN	indication	Hepatoma	19.754968	0.719683	0.374743	0.344940	0.000000e+00
0	NaN	indication	SUICIDE ATTEMPT	17.234751	0.826860	0.375502	0.451358	0.000000e+00
0	NaN	age	younger than 18	15.555547	0.579967	0.395188	0.184779	0.000000e+00
0	NaN	psd	Sorafenib Tosylate	15.524659	0.689824	0.411688	0.278135	0.000000e+00
0	NaN	dose	PO	14.604209	0.672018	0.371813	0.300204	0.000000e+00
0	NaN	indication	PYREXIA	9.571886	0.624649	0.375628	0.249021	0.000000e+00
0	NaN	indication	HEADACHE	8.984391	0.702705	0.376927	0.325778	0.000000e+00
0	NaN	psd	Sevoflurane	8.726726	0.751948	0.413302	0.338646	0.000000e+00
0	NaN	psd	Heparin Sodium	7.477228	0.684502	0.413273	0.271229	3.796963e-14
0	NaN	psd	Bromfenac Sodium	6.007936	0.626746	0.413664	0.213082	9.395005e-10

Continued on next page

level	route	feature	value	z score	P(do)	P(not do)	delta	p value
0	NaN	outcome	RequiredIntervention	5.054285	0.535111	0.459276	0.075835	2.160030e-07
0	NaN	indication	ATRIAL FIBRILLATION	4.717715	0.523072	0.377368	0.145704	1.192544e-06
0	NaN	indication	BCa	4.413204	0.549163	0.377721	0.171442	5.092604e-06
0	NaN	indication	MULTIPLE MYELOMA	3.541390	0.506920	0.377986	0.128933	1.990126e-04
0	NaN	psd	Tacrolimus	3.280608	0.556439	0.413918	0.142521	5.179180e-04
0	NaN	psd	Moxifloxacin Hydrochloride	2.989044	0.551184	0.414004	0.137180	1.399259e-03
0	NaN	psd	Rosiglitazone Maleate	2.721393	0.517636	0.413976	0.103661	3.250368e-03
1	Acetaminophen	outcome	Death	20.396541	0.954967	0.822811	0.132155	0.000000e+00
1	Acetaminophen	dose	PO	12.695141	0.946943	0.764757	0.182186	0.000000e+00
1	Acetaminophen	age	18-39	12.391853	0.907122	0.795752	0.111370	0.000000e+00
1	Acetaminophen	gender	Female	7.493808	0.871821	0.800122	0.071699	3.341771e-14
1	Acetaminophen	outcome	LifeThreatening	6.446461	0.932435	0.871266	0.061169	5.724610e-11
1	Acetaminophen	indication	SUICIDE ATTEMPT	5.002852	0.873468	0.727618	0.145850	2.824413e-07
1	Acetaminophen	age	younger than 18	3.136228	0.878775	0.838672	0.040103	8.556792e-04
1	Death	psd	Acetaminophen	29.303276	0.943154	0.639641	0.303513	0.000000e+00
1	Death	age	younger than 18	9.977210	0.899651	0.645397	0.254254	0.000000e+00
1	Death	age	18-39	6.949458	0.782602	0.636895	0.145707	1.833422e-12
1	Death	gender	Female	4.453764	0.714198	0.645047	0.069151	4.218894e-06
1	Death	dose	larger than 100 MG	4.312924	0.738300	0.641443	0.096858	8.055464e-06
1	LifeThreatening	age	40-64	2.810226	0.516963	0.408877	0.108087	2.475335e-03
1	LifeThreatening	gender	Female	2.790064	0.517413	0.422117	0.095296	2.634885e-03
1	18-39	psd	Acetaminophen	53.097512	0.907692	0.417123	0.490569	0.000000e+00
1	18-39	outcome	Death	30.070328	0.846254	0.490688	0.355566	0.000000e+00
1	18-39	dose	PO	11.677179	0.849634	0.478346	0.371289	0.000000e+00
1	18-39	gender	Female	8.036785	0.602275	0.496129	0.106147	4.440892e-16
1	18-39	outcome	LifeThreatening	4.487831	0.666910	0.594937	0.071973	3.597591e-06
1	18-39	outcome	RequiredIntervention	2.888269	0.682833	0.605990	0.076842	1.936844e-03
1	18-39	psd	Ibuprofen	2.017771	0.608942	0.558211	0.050730	2.180754e-02
1	18-39	indication	MULTIPLE SCLEROSIS	1.679679	0.553461	0.472809	0.080652	4.650990e-02
1	younger than 18	psd	Acetaminophen	24.732389	0.879603	0.421320	0.458283	0.000000e+00

Continued on next page

level	route	feature	value	z score	P(do)	P(not do)	delta	p value
1	younger than 18	outcome	Death	13.994175	0.853633	0.537710	0.315923	0.000000e+00
1	younger than 18	indication	PYREXIA	8.203980	0.772748	0.480413	0.292335	1.110223e-16
1	younger than 18	outcome	LifeThreatening	3.073445	0.687027	0.602812	0.084216	1.058012e-03
1	Sorafenib Tosylate	indication	Hepatoma	3.199321	0.732424	0.580069	0.152355	6.887580e-04
1	PO	psd	Acetaminophen	17.342918	0.949232	0.393557	0.555675	0.000000e+00
1	PO	outcome	Death	10.539077	0.873212	0.436908	0.436304	0.000000e+00
1	PYREXIA	age	younger than 18	5.373019	0.804554	0.510643	0.293911	3.871458e-08
1	PYREXIA	psd	Acetaminophen	3.719662	0.746513	0.553368	0.193144	9.974491e-05
1	ATRIAL FIBRILLATION	gender	Female	2.281914	0.588147	0.444919	0.143228	1.124722e-02
2	Acetaminophen->Death	age	18-39	2.900583	0.961106	0.907017	0.054089	1.862347e-03
2	Acetaminophen->18-39	outcome	Death	9.789695	0.968427	0.883569	0.084857	0.000000e+00
2	Acetaminophen->18-39	gender	Female	6.109159	0.930140	0.853473	0.076667	5.007886e-10
2	Acetaminophen->Female	outcome	Death	11.685221	0.956839	0.859011	0.097828	0.000000e+00
2	Acetaminophen->Female	age	18-39	11.157157	0.930140	0.808263	0.121878	0.000000e+00
2	Acetaminophen->Female	dose	PO	7.977480	0.951636	0.802248	0.149388	7.771561e-16
2	Acetaminophen->younger than 18	outcome	Death	6.675412	0.980947	0.862740	0.118207	1.232692e-11
2	Death->Acetaminophen	age	18-39	2.900583	0.961106	0.907017	0.054089	1.862347e-03
2	Death->18-39	psd	Acetaminophen	9.289103	0.961106	0.703471	0.257635	0.000000e+00
2	Death->18-39	gender	Female	5.118445	0.870308	0.672822	0.197486	1.540326e-07
2	Death->Female	psd	Acetaminophen	18.786674	0.945277	0.667229	0.278049	0.000000e+00
2	Death->Female	age	18-39	8.463106	0.870308	0.665160	0.205148	0.000000e+00
2	Death->Female	dose	larger than 100 MG	3.502675	0.772343	0.655788	0.116555	2.303058e-04
2	18-39->Acetaminophen	outcome	Death	9.789695	0.968427	0.883569	0.084857	0.000000e+00
2	18-39->Acetaminophen	gender	Female	6.109159	0.930140	0.853473	0.076667	5.007886e-10
2	18-39->Death	psd	Acetaminophen	9.289103	0.961106	0.703471	0.257635	0.000000e+00
2	18-39->Death	gender	Female	5.118445	0.870308	0.672822	0.197486	1.540326e-07
2	18-39->Female	psd	Acetaminophen	42.157719	0.930512	0.444123	0.486389	0.000000e+00
2	18-39->Female	outcome	Death	23.844872	0.887399	0.547685	0.339714	0.000000e+00
2	18-39->Female	outcome	LifeThreatening	6.359281	0.760655	0.634738	0.125917	1.013505e-10
2	18-39->Female	outcome	RequiredIntervention	3.238516	0.758590	0.653586	0.105004	6.007664e-04

Continued on next page

level	route	feature	value	z score	P(do)	P(not do)	delta	p value
2	18-39->Female	dose	larger than 100 MG	2.489852	0.565537	0.502179	0.063358	6.389806e-03
2	younger than 18->Acetaminophen	outcome	Death	6.675412	0.980947	0.862740	0.118207	1.232692e-11

A.2 Template-Supervised BART results

A.2.1 Root Causal Terms

feature	value	z score	P(do)	P(not do)	delta	p value	support
psd	Acetaminophen	115.208506	0.851006	0.346936	0.504070	0.000000e+00	4990
outcome	Death	76.429732	0.707054	0.318770	0.388284	0.000000e+00	11311
outcome	LifeThreatening	29.374121	0.629398	0.425196	0.204202	0.000000e+00	5488
age	18-39	28.305645	0.568861	0.371384	0.197476	0.000000e+00	5350
age	younger than 18	16.798436	0.591774	0.395472	0.196302	0.000000e+00	1687
indication	SUICIDE ATTEMPT	16.792601	0.826332	0.377651	0.448681	0.000000e+00	170
dose	PO	14.488178	0.666958	0.371849	0.295109	0.000000e+00	457
psd	Sorafenib Tosylate	11.961930	0.662287	0.413099	0.249189	0.000000e+00	360
indication	Hepatoma	11.890637	0.651840	0.377741	0.274099	0.000000e+00	271
gender	Female	10.703474	0.430165	0.376872	0.053293	0.000000e+00	17843
indication	PYREXIA	10.124118	0.641001	0.377562	0.263439	0.000000e+00	297
psd	Sevoflurane	9.919259	0.782693	0.414330	0.368363	0.000000e+00	121
indication	HEADACHE	9.292295	0.705415	0.379052	0.326363	0.000000e+00	139
psd	Heparin Sodium	8.750057	0.717361	0.414264	0.303097	0.000000e+00	155
indication	ATRIAL FIBRILLATION	6.303168	0.562876	0.379075	0.183801	1.458111e-10	244
indication	BCa	5.340529	0.582191	0.379621	0.202571	4.633789e-08	162
psd	Ibuprofen	4.942285	0.483268	0.413594	0.069674	3.860619e-07	1027
outcome	RequiredIntervention	4.271528	0.522130	0.459458	0.062672	9.706901e-06	1078
dose	larger than 100 MG	4.138795	0.397121	0.368675	0.028446	1.745671e-05	6919
psd	Bevacizumab	3.963032	0.565539	0.414896	0.150643	3.700191e-05	158
indication	MULTIPLE MYELOMA	3.882370	0.523684	0.380001	0.143683	5.172163e-05	170
psd	Fentanyl	3.507393	0.538531	0.414979	0.123552	2.262600e-04	168
psd	Bromfenac Sodium	3.340132	0.528059	0.415145	0.112914	4.186928e-04	130
indication	BACK PAIN	3.069658	0.456730	0.379825	0.076905	1.071521e-03	368
psd	Peginterferon Alfa-2a	2.781290	0.532936	0.415189	0.117747	2.707164e-03	111

Continued on next page

feature	value	z score	P(do)	P(not do)	delta	p value	support
indication	MULTIPLE SCLEROSIS	2.674449	0.442470	0.380133	0.062336	3.742613e-03	345
indication	COLORECTAL CANCER	2.274461	0.487144	0.380582	0.106563	1.146915e-02	109
psd	Trovafloxacin Mesylate	1.993870	0.483228	0.415197	0.068031	2.308312e-02	188
psd	Amiodarone Hydrochloride	1.866001	0.479312	0.415219	0.064093	3.102064e-02	187
psd	Tacrolimus	1.786842	0.491170	0.415279	0.075892	3.698151e-02	129

A.2.2 Causal Tree Terms

level	route	feature	value	z score	P(do)	P(not do)	delta	p value
0	NaN	psd	Acetaminophen	115.208506	0.851006	0.346936	0.504070	0.000000e+00
0	NaN	outcome	Death	76.429732	0.707054	0.318770	0.388284	0.000000e+00
0	NaN	outcome	LifeThreatening	29.374121	0.629398	0.425196	0.204202	0.000000e+00
0	NaN	age	18-39	28.305645	0.568861	0.371384	0.197476	0.000000e+00
0	NaN	age	younger than 18	16.798436	0.591774	0.395472	0.196302	0.000000e+00
0	NaN	indication	SUICIDE ATTEMPT	16.792601	0.826332	0.377651	0.448681	0.000000e+00
0	NaN	dose	PO	14.488178	0.666958	0.371849	0.295109	0.000000e+00
0	NaN	psd	Sorafenib Tosylate	11.961930	0.662287	0.413099	0.249189	0.000000e+00
0	NaN	indication	Hepatoma	11.890637	0.651840	0.377741	0.274099	0.000000e+00
0	NaN	indication	PYREXIA	10.124118	0.641001	0.377562	0.263439	0.000000e+00
0	NaN	psd	Sevoflurane	9.919259	0.782693	0.414330	0.368363	0.000000e+00
0	NaN	indication	HEADACHE	9.292295	0.705415	0.379052	0.326363	0.000000e+00
0	NaN	psd	Heparin Sodium	8.750057	0.717361	0.414264	0.303097	0.000000e+00
0	NaN	indication	ATRIAL FIBRILLATION	6.303168	0.562876	0.379075	0.183801	1.458111e-10
0	NaN	indication	BCa	5.340529	0.582191	0.379621	0.202571	4.633789e-08
0	NaN	outcome	RequiredIntervention	4.271528	0.522130	0.459458	0.062672	9.706901e-06
0	NaN	psd	Bevacizumab	3.963032	0.565539	0.414896	0.150643	3.700191e-05
0	NaN	indication	MULTIPLE MYELOMA	3.882370	0.523684	0.380001	0.143683	5.172163e-05
0	NaN	psd	Fentanyl	3.507393	0.538531	0.414979	0.123552	2.262600e-04

Continued on next page

level	route	feature	value	z score	P(do)	P(not do)	delta	p value
0	NaN	psd	Bromfenac Sodium	3.340132	0.528059	0.415145	0.112914	4.186928e-04
0	NaN	psd	Peginterferon Alfa-2a	2.781290	0.532936	0.415189	0.117747	2.707164e-03
1	Acetaminophen	outcome	Death	20.145927	0.949380	0.822385	0.126995	0.000000e+00
1	Acetaminophen	age	18-39	15.128441	0.915460	0.789431	0.126028	0.000000e+00
1	Acetaminophen	dose	PO	13.075302	0.937192	0.756827	0.180365	0.000000e+00
1	Acetaminophen	gender	Female	7.292012	0.867211	0.799701	0.067511	1.526557e-13
1	Acetaminophen	outcome	LifeThreatening	6.744972	0.930105	0.868526	0.061579	7.652878e-12
1	Acetaminophen	indication	SUICIDE ATTEMPT	5.215545	0.884115	0.735331	0.148785	9.163891e-08
1	Acetaminophen	age	younger than 18	2.836389	0.874301	0.839800	0.034501	2.281341e-03
1	Death	psd	Acetaminophen	30.465265	0.941258	0.634036	0.307222	0.000000e+00
1	Death	age	younger than 18	11.868730	0.917094	0.638734	0.278360	0.000000e+00
1	Death	gender	Female	4.859862	0.710363	0.635309	0.075054	5.873393e-07
1	Death	age	18-39	4.423275	0.735422	0.638864	0.096558	4.860787e-06
1	Death	dose	larger than 100 MG	3.998162	0.726889	0.636931	0.089958	3.191818e-05
1	LifeThreatening	gender	Female	3.912135	0.541867	0.408952	0.132915	4.574187e-05
1	LifeThreatening	age	40-64	2.571919	0.513827	0.415572	0.098256	5.056833e-03
1	18-39	psd	Acetaminophen	55.528924	0.914617	0.425253	0.489364	0.000000e+00
1	18-39	outcome	Death	26.996557	0.830905	0.509175	0.321729	0.000000e+00
1	18-39	dose	PO	10.053301	0.819952	0.489466	0.330486	0.000000e+00
1	18-39	gender	Female	6.712568	0.601870	0.513831	0.088039	9.561463e-12
1	18-39	outcome	LifeThreatening	5.001930	0.679828	0.600307	0.079521	2.837959e-07
1	18-39	psd	Ibuprofen	3.828240	0.656909	0.563575	0.093334	6.453156e-05
1	18-39	outcome	Hospitalization	2.395921	0.625109	0.580435	0.044674	8.289326e-03
1	18-39	outcome	RequiredIntervention	2.002933	0.666290	0.614654	0.051636	2.259226e-02
1	younger than 18	psd	Acetaminophen	23.778722	0.874668	0.441992	0.432677	0.000000e+00
1	younger than 18	outcome	Death	13.543928	0.854601	0.554575	0.300026	0.000000e+00
1	younger than 18	indication	PYREXIA	8.420824	0.784947	0.487343	0.297604	0.000000e+00
1	younger than 18	outcome	LifeThreatening	2.523975	0.687934	0.619331	0.068603	5.801804e-03
1	PO	psd	Acetaminophen	17.659231	0.947199	0.393077	0.554121	0.000000e+00
1	PO	outcome	Death	10.349360	0.861578	0.434672	0.426906	0.000000e+00

Continued on next page

level	route	feature	value	z score	P(do)	P(not do)	delta	p value
1	PO	gender	Female	2.241772	0.674376	0.567385	0.106991	1.248806e-02
1	PYREXIA	age	younger than 18	5.867936	0.827532	0.512305	0.315227	2.206263e-09
1	PYREXIA	psd	Acetaminophen	3.058347	0.743790	0.583445	0.160345	1.112810e-03
1	PYREXIA	gender	Female	1.971348	0.693664	0.580872	0.112792	2.434207e-02
2	Acetaminophen->18-39	outcome	Death	8.321982	0.959726	0.900962	0.058763	0.000000e+00
2	Acetaminophen->18-39	gender	Female	4.757417	0.929509	0.880978	0.048531	9.804314e-07
2	Acetaminophen->18-39	outcome	LifeThreatening	2.198844	0.949882	0.922357	0.027525	1.394452e-02
2	Acetaminophen->18-39	outcome	Hospitalization	1.662171	0.929705	0.902219	0.027486	4.823929e-02
2	Acetaminophen->Female	outcome	Death	13.556750	0.958179	0.854157	0.104022	0.000000e+00
2	Acetaminophen->Female	age	18-39	11.573187	0.929509	0.813827	0.115682	0.000000e+00
2	Acetaminophen->Female	dose	PO	7.945445	0.940809	0.794618	0.146191	9.992007e-16
2	Acetaminophen->Female	outcome	LifeThreatening	2.340180	0.926434	0.894808	0.031627	9.637226e-03
2	Acetaminophen->younger than 18	outcome	Death	6.956498	0.968089	0.849084	0.119005	1.744160e-12
2	Acetaminophen->younger than 18	gender	Female	3.618721	0.890536	0.788002	0.102534	1.480312e-04
2	Death->Female	psd	Acetaminophen	19.565060	0.929270	0.665868	0.263401	0.000000e+00
2	Death->Female	age	18-39	5.046682	0.806850	0.674435	0.132414	2.247740e-07
2	Death->Female	dose	larger than 100 MG	2.047886	0.739381	0.669586	0.069796	2.028558e-02
2	Death->18-39	psd	Acetaminophen	9.681176	0.926950	0.650518	0.276432	0.000000e+00
2	Death->18-39	gender	Female	4.062608	0.806850	0.640356	0.166494	2.426374e-05
2	18-39->Acetaminophen	outcome	Death	8.321982	0.959726	0.900962	0.058763	0.000000e+00
2	18-39->Acetaminophen	gender	Female	4.757417	0.929509	0.880978	0.048531	9.804314e-07
2	18-39->Acetaminophen	outcome	LifeThreatening	2.198844	0.949882	0.922357	0.027525	1.394452e-02
2	18-39->Acetaminophen	outcome	Hospitalization	1.662171	0.929705	0.902219	0.027486	4.823929e-02
2	18-39->Death	psd	Acetaminophen	9.681176	0.926950	0.650518	0.276432	0.000000e+00
2	18-39->Death	gender	Female	4.062608	0.806850	0.640356	0.166494	2.426374e-05
2	18-39->Female	psd	Acetaminophen	42.988923	0.927920	0.444772	0.483147	0.000000e+00
2	18-39->Female	outcome	Death	21.518924	0.867765	0.555370	0.312395	0.000000e+00
2	18-39->Female	outcome	LifeThreatening	7.765462	0.779086	0.628386	0.150700	4.107825e-15
2	18-39->Female	psd	Ibuprofen	3.273311	0.693112	0.596942	0.096170	5.314764e-04
2	18-39->Female	dose	larger than 100 MG	2.930958	0.579230	0.506089	0.073141	1.689593e-03

Continued on next page

level	route	feature	value	z score	P(do)	P(not do)	delta	p value
2	18-39->Female	outcome	RequiredIntervention	2.579384	0.735854	0.653647	0.082207	4.948831e-03
2	18-39->Ibuprofen	gender	Female	1.816856	0.693112	0.605151	0.087961	3.461956e-02
2	18-39->Hospitalization	psd	Acetaminophen	47.344068	0.897633	0.301308	0.596324	0.000000e+00
2	18-39->Hospitalization	psd	Ibuprofen	1.871443	0.522416	0.462040	0.060375	3.064185e-02
2	younger than 18->Acetaminophen	outcome	Death	6.956498	0.968089	0.849084	0.119005	1.744160e-12
2	younger than 18->Acetaminophen	gender	Female	3.618721	0.890536	0.788002	0.102534	1.480312e-04

A.3 GPT-Supervised BART results

A.3.1 Root Causal Terms

feature	value	z score	P(do)	P(not do)	delta	p value	support
psd	Acetaminophen	142.555254	0.831005	0.347795	0.483210	0.000000e+00	4990
outcome	Death	79.217076	0.673835	0.322225	0.351610	0.000000e+00	11311
age	18-39	30.164009	0.561430	0.372565	0.188865	0.000000e+00	5350
indication	SUICIDE ATTEMPT	25.585835	0.864948	0.373748	0.491200	0.000000e+00	170
outcome	LifeThreatening	24.739699	0.579840	0.423745	0.156095	0.000000e+00	5488
indication	Hepatoma	21.345459	0.719311	0.373283	0.346028	0.000000e+00	271
psd	Sorafenib Tosylate	18.267121	0.685804	0.410866	0.274938	0.000000e+00	360
age	younger than 18	16.182208	0.569651	0.396415	0.173236	0.000000e+00	1687
dose	PO	13.995601	0.632102	0.380083	0.252019	0.000000e+00	457
gender	Female	10.135883	0.425959	0.381348	0.044611	0.000000e+00	17843
psd	Heparin Sodium	9.948791	0.706356	0.412323	0.294033	0.000000e+00	155
indication	PYREXIA	8.290012	0.573317	0.374861	0.198456	1.110223e-16	297
indication	HEADACHE	7.779816	0.658375	0.375752	0.282622	3.663736e-15	139
indication	MULTIPLE MYELOMA	7.310623	0.557822	0.376133	0.181689	1.330047e-13	170
psd	Bromfenac Sodium	7.305064	0.646777	0.412736	0.234042	1.385558e-13	130
psd	Sevoflurane	6.642788	0.667432	0.412725	0.254707	1.539024e-11	121
psd	Rosiglitazone Maleate	5.711784	0.568718	0.412898	0.155819	5.589905e-09	157
dose	larger than 100 MG	5.557569	0.408077	0.373852	0.034225	1.367793e-08	6919
outcome	RequiredIntervention	5.253085	0.518766	0.449174	0.069591	7.478624e-08	1078
indication	BACK PAIN	5.069647	0.487783	0.375663	0.112120	1.992766e-07	368
indication	ATRIAL FIBRILLATION	5.057290	0.512542	0.376023	0.136519	2.126276e-07	244
psd	Ibuprofen	4.640391	0.464234	0.412105	0.052129	1.738749e-06	1027
indication	BCa	4.052558	0.514564	0.376519	0.138045	2.533035e-05	162
psd	Peginterferon Alfa-2a	4.033445	0.551611	0.413146	0.138464	2.748256e-05	111
psd	Bevacizumab	3.788906	0.551690	0.412968	0.138722	7.565615e-05	158

Continued on next page

feature	value	z score	P(do)	P(not do)	delta	p value	support
psd	Troglitazone	3.546303	0.484142	0.412940	0.071202	1.953381e-04	322
indication	PLASMA CELL MYELOMA	3.465906	0.483926	0.376682	0.107244	2.642241e-04	175
psd	Tacrolimus	2.956535	0.530040	0.413154	0.116885	1.555585e-03	129
indication	METASTATIC RENAL CELL CARCINOMA	2.765789	0.489148	0.376902	0.112246	2.839263e-03	124
indication	NON-SMALL CELL LUNG CANCER	2.719544	0.479082	0.376847	0.102235	3.268603e-03	148
psd	Fentanyl	2.672209	0.486611	0.413229	0.073381	3.767684e-03	168
indication	PULMONARY HYPERTENSION	2.477089	0.456074	0.377146	0.078928	6.622945e-03	108
dose	ORAL	2.306130	0.423856	0.384724	0.039132	1.055167e-02	628
psd	Peginterferon Alfa-2b	1.916386	0.488533	0.413309	0.075223	2.765796e-02	125
psd	Paclitaxel	1.748104	0.480443	0.413361	0.067082	4.022305e-02	112

A.3.2 Casual Tree Terms

level	route	feature	value	z score	P(do)	P(not do)	delta	p value
0	NaN	psd	Acetaminophen	142.555254	0.831005	0.347795	0.483210	0.000000e+00
0	NaN	outcome	Death	79.217076	0.673835	0.322225	0.351610	0.000000e+00
0	NaN	age	18-39	30.164009	0.561430	0.372565	0.188865	0.000000e+00
0	NaN	indication	SUICIDE ATTEMPT	25.585835	0.864948	0.373748	0.491200	0.000000e+00
0	NaN	outcome	LifeThreatening	24.739699	0.579840	0.423745	0.156095	0.000000e+00
0	NaN	indication	Hepatoma	21.345459	0.719311	0.373283	0.346028	0.000000e+00
0	NaN	psd	Sorafenib Tosylate	18.267121	0.685804	0.410866	0.274938	0.000000e+00
0	NaN	age	younger than 18	16.182208	0.569651	0.396415	0.173236	0.000000e+00
0	NaN	dose	PO	13.995601	0.632102	0.380083	0.252019	0.000000e+00
0	NaN	psd	Heparin Sodium	9.948791	0.706356	0.412323	0.294033	0.000000e+00
0	NaN	indication	PYREXIA	8.290012	0.573317	0.374861	0.198456	1.110223e-16
0	NaN	indication	HEADACHE	7.779816	0.658375	0.375752	0.282622	3.663736e-15
0	NaN	indication	MULTIPLE MYELOMA	7.310623	0.557822	0.376133	0.181689	1.330047e-13
0	NaN	psd	Bromfenac Sodium	7.305064	0.646777	0.412736	0.234042	1.385558e-13

Continued on next page

level	route	feature	value	z score	P(do)	P(not do)	delta	p value
0	NaN	psd	Sevoflurane	6.642788	0.667432	0.412725	0.254707	1.539024e-11
0	NaN	psd	Rosiglitazone Maleate	5.711784	0.568718	0.412898	0.155819	5.589905e-09
0	NaN	outcome	RequiredIntervention	5.253085	0.518766	0.449174	0.069591	7.478624e-08
0	NaN	indication	ATRIAL FIBRILLATION	5.057290	0.512542	0.376023	0.136519	2.126276e-07
0	NaN	indication	BCa	4.052558	0.514564	0.376519	0.138045	2.533035e-05
0	NaN	psd	Peginterferon Alfa-2a	4.033445	0.551611	0.413146	0.138464	2.748256e-05
0	NaN	psd	Bevacizumab	3.788906	0.551690	0.412968	0.138722	7.565615e-05
0	NaN	psd	Tacrolimus	2.956535	0.530040	0.413154	0.116885	1.555585e-03
1	Acetaminophen	outcome	Death	27.459026	0.917797	0.796617	0.121181	0.000000e+00
1	Acetaminophen	age	18-39	15.926088	0.882172	0.789019	0.093153	0.000000e+00
1	Acetaminophen	indication	SUICIDE ATTEMPT	14.295520	0.926391	0.754302	0.172088	0.000000e+00
1	Acetaminophen	gender	Female	10.539733	0.854760	0.785275	0.069486	0.000000e+00
1	Acetaminophen	dose	PO	7.654595	0.887421	0.801493	0.085928	9.658940e-15
1	Acetaminophen	age	younger than 18	6.728609	0.870643	0.822504	0.048139	8.564593e-12
1	Acetaminophen	outcome	LifeThreatening	5.906041	0.882802	0.843589	0.039213	1.752139e-09
1	Acetaminophen	indication	PYREXIA	3.192595	0.828598	0.761955	0.066642	7.050036e-04
1	Death	psd	Acetaminophen	38.254684	0.920117	0.606681	0.313437	0.000000e+00
1	Death	age	younger than 18	9.012908	0.836413	0.614140	0.222273	0.000000e+00
1	Death	age	18-39	7.517256	0.746713	0.604482	0.142232	2.797762e-14
1	Death	gender	Female	3.998909	0.673506	0.620547	0.052959	3.181757e-05
1	Death	dose	larger than 100 MG	2.801817	0.651900	0.596233	0.055666	2.540785e-03
1	18-39	psd	Acetaminophen	59.879881	0.882606	0.428031	0.454575	0.000000e+00
1	18-39	outcome	Death	30.353551	0.814578	0.493753	0.320825	0.000000e+00
1	18-39	dose	PO	10.327096	0.801603	0.502463	0.299140	0.000000e+00
1	18-39	gender	Female	6.107163	0.589016	0.516767	0.072249	5.070870e-10
1	18-39	psd	Ibuprofen	2.892417	0.615079	0.558209	0.056870	1.911452e-03
1	18-39	outcome	LifeThreatening	2.854422	0.634946	0.592902	0.042044	2.155760e-03
1	18-39	outcome	RequiredIntervention	2.847298	0.665371	0.597910	0.067461	2.204603e-03
1	18-39	indication	MULTIPLE SCLEROSIS	2.443774	0.576823	0.494533	0.082290	7.267267e-03
1	LifeThreatening	gender	Female	4.175783	0.530922	0.399148	0.131774	1.484811e-05

Continued on next page

level	route	feature	value	z score	P(do)	P(not do)	delta	p value
1	LifeThreatening	age	40-64	2.865223	0.508178	0.406502	0.101676	2.083581e-03
1	Hepatoma	age	40-64	3.037211	0.765899	0.670965	0.094934	1.193892e-03
1	Sorafenib Tosylate	indication	Hepatoma	4.571157	0.735944	0.566813	0.169131	2.425194e-06
1	Sorafenib Tosylate	age	40-64	4.108913	0.748772	0.630458	0.118315	1.987627e-05
1	younger than 18	psd	Acetaminophen	31.468967	0.871431	0.409870	0.461561	0.000000e+00
1	younger than 18	outcome	Death	13.583765	0.806700	0.524231	0.282469	0.000000e+00
1	younger than 18	outcome	LifeThreatening	3.996985	0.675372	0.576321	0.099051	3.207714e-05
1	younger than 18	dose	larger than 100 MG	2.621620	0.577589	0.498727	0.078863	4.375645e-03
1	younger than 18	indication	PYREXIA	2.251979	0.573388	0.482473	0.090915	1.216181e-02
1	PO	psd	Acetaminophen	19.447919	0.886260	0.368834	0.517426	0.000000e+00
1	PO	outcome	Death	13.121477	0.828227	0.387164	0.441063	0.000000e+00
1	PO	gender	Female	2.556675	0.645921	0.540194	0.105727	5.283889e-03
1	PYREXIA	psd	Acetaminophen	11.735561	0.829473	0.372518	0.456955	0.000000e+00
1	PYREXIA	gender	Female	1.919445	0.611610	0.510279	0.101330	2.746400e-02
2	Acetaminophen->Death	age	18-39	3.680038	0.931799	0.888612	0.043186	1.165997e-04
2	Acetaminophen->Death	gender	Female	2.140917	0.921242	0.894679	0.026564	1.614035e-02
2	Acetaminophen->18-39	outcome	Death	16.339714	0.933903	0.857790	0.076113	0.000000e+00
2	Acetaminophen->18-39	gender	Female	5.913708	0.895541	0.849187	0.046354	1.672454e-09
2	Acetaminophen->18-39	outcome	LifeThreatening	1.930118	0.905477	0.888278	0.017198	2.679608e-02
2	Acetaminophen->Female	outcome	Death	14.963706	0.920152	0.845654	0.074498	0.000000e+00
2	Acetaminophen->Female	age	18-39	12.498368	0.895541	0.813221	0.082320	0.000000e+00
2	Acetaminophen->Female	age	younger than 18	4.437328	0.887729	0.847937	0.039793	4.554113e-06
2	Acetaminophen->Female	dose	PO	3.563889	0.891251	0.842024	0.049227	1.827002e-04
2	Acetaminophen->Female	outcome	LifeThreatening	2.317544	0.893674	0.875436	0.018238	1.023705e-02
2	Acetaminophen->younger than 18	outcome	Death	6.189436	0.921661	0.853922	0.067740	3.018985e-10
2	Acetaminophen->younger than 18	gender	Female	5.001201	0.887729	0.810411	0.077318	2.848716e-07
2	Death->Acetaminophen	age	18-39	3.680038	0.931799	0.888612	0.043186	1.165997e-04
2	Death->Acetaminophen	gender	Female	2.140917	0.921242	0.894679	0.026564	1.614035e-02
2	Death->18-39	psd	Acetaminophen	11.393663	0.931799	0.664665	0.267134	0.000000e+00
2	Death->18-39	gender	Female	4.340079	0.815580	0.661815	0.153765	7.121573e-06

Continued on next page

level	route	feature	value	z score	P(do)	P(not do)	delta	p value
2	Death->Female	psd	Acetaminophen	24.854778	0.921242	0.623151	0.298091	0.000000e+00
2	Death->Female	age	18-39	8.744956	0.815580	0.617453	0.198127	0.000000e+00
2	18-39->Acetaminophen	outcome	Death	16.339714	0.933903	0.857790	0.076113	0.000000e+00
2	18-39->Acetaminophen	gender	Female	5.913708	0.895541	0.849187	0.046354	1.672454e-09
2	18-39->Acetaminophen	outcome	LifeThreatening	1.930118	0.905477	0.888278	0.017198	2.679608e-02
2	18-39->Death	psd	Acetaminophen	11.393663	0.931799	0.664665	0.267134	0.000000e+00
2	18-39->Death	gender	Female	4.340079	0.815580	0.661815	0.153765	7.121573e-06
2	18-39->Female	psd	Acetaminophen	47.143449	0.895791	0.441205	0.454586	0.000000e+00
2	18-39->Female	outcome	Death	24.750227	0.850214	0.534621	0.315593	0.000000e+00
2	18-39->Female	outcome	LifeThreatening	5.874977	0.724224	0.617574	0.106650	2.114513e-09
2	18-39->Female	dose	larger than 100 MG	2.961638	0.584838	0.517557	0.067281	1.530037e-03
2	18-39->Female	outcome	RequiredIntervention	1.861668	0.692028	0.635558	0.056471	3.132496e-02
2	Sorafenib Tosylate->Hepatoma	age	40-64	2.350430	0.768465	0.700995	0.067470	9.375857e-03
2	younger than 18->Acetaminophen	outcome	Death	6.189436	0.921661	0.853922	0.067740	3.018985e-10
2	younger than 18->Acetaminophen	gender	Female	5.001201	0.887729	0.810411	0.077318	2.848716e-07
2	younger than 18->larger than 100 MG	psd	Acetaminophen	13.494615	0.853602	0.389268	0.464334	0.000000e+00
2	younger than 18->larger than 100 MG	gender	Female	3.468625	0.632825	0.467762	0.165063	2.615645e-04

A.4 Llama T2T results

A.4.1 Root Causal Terms

feature	value	z score	P(do)	P(not do)	delta	p value	support
psd	Acetaminophen	109.952385	0.843057	0.351320	0.491737	0.000000e+00	4990
outcome	Death	78.586902	0.714103	0.319593	0.394510	0.000000e+00	11311
outcome	LifeThreatening	31.786216	0.643855	0.425845	0.218010	0.000000e+00	5488
age	18-39	27.697822	0.569967	0.376684	0.193284	0.000000e+00	5350
indication	Hepatoma	20.833077	0.745387	0.379221	0.366166	0.000000e+00	271
psd	Sorafenib Tosylate	18.569843	0.733491	0.415125	0.318366	0.000000e+00	360
age	younger than 18	15.865773	0.586678	0.400599	0.186079	0.000000e+00	1687
indication	SUICIDE ATTEMPT	15.641200	0.814972	0.380369	0.434602	0.000000e+00	170
dose	PO	13.748480	0.666717	0.378772	0.287945	0.000000e+00	457
gender	Female	9.837794	0.431426	0.382472	0.048954	0.000000e+00	17843
psd	Sevoflurane	9.288299	0.756995	0.417129	0.339866	0.000000e+00	121
psd	Heparin Sodium	8.992280	0.721102	0.416965	0.304137	0.000000e+00	155
indication	PYREXIA	8.702539	0.611096	0.380615	0.230481	0.000000e+00	297
indication	HEADACHE	8.535302	0.696896	0.381732	0.315164	0.000000e+00	139
indication	MULTIPLE MYELOMA	5.072737	0.565151	0.382309	0.182842	1.960672e-07	170
psd	Ibuprofen	4.808919	0.484100	0.416353	0.067747	7.587436e-07	1027
outcome	RequiredIntervention	4.691210	0.530871	0.462368	0.068503	1.357972e-06	1078
indication	BCa	4.679434	0.562003	0.382399	0.179603	1.438340e-06	162
indication	ATRIAL FIBRILLATION	4.584499	0.521577	0.382176	0.139401	2.275384e-06	244
dose	larger than 100 MG	4.084605	0.403702	0.375526	0.028176	2.207596e-05	6919
psd	Bromfenac Sodium	3.898813	0.550953	0.417779	0.133174	4.833273e-05	130
indication	PLASMA CELL MYELOMA	3.573574	0.506373	0.382737	0.123636	1.760708e-04	175
indication	BACK PAIN	3.317154	0.467367	0.382299	0.085068	4.546968e-04	368
psd	Bevacizumab	3.191665	0.538693	0.417730	0.120963	7.072771e-04	158
indication	PULMONARY HYPERTENSION	2.511288	0.493296	0.383179	0.110117	6.014572e-03	108

Continued on next page

feature	value	z score	P(do)	P(not do)	delta	p value	support
psd	Fentanyl	2.483843	0.508024	0.417838	0.090187	6.498656e-03	168
psd	Troglitazone	2.364345	0.473295	0.417763	0.055532	9.030994e-03	322
psd	Tacrolimus	2.326156	0.518834	0.417896	0.100938	1.000512e-02	129
psd	Ribavirin	2.282419	0.501360	0.417916	0.083443	1.123230e-02	147
psd	Moxifloxacin Hydrochloride	2.044682	0.513134	0.417963	0.095171	2.044311e-02	111
psd	Peginterferon Alfa-2a	1.861280	0.501151	0.417999	0.083152	3.135235e-02	111

A.4.2 Causal Tree Terms

level	route	feature	value	z score	P(do)	P(not do)	delta	p value
0	NaN	psd	Acetaminophen	109.952385	0.843057	0.351320	0.491737	0.000000e+00
0	NaN	outcome	Death	78.586902	0.714103	0.319593	0.394510	0.000000e+00
0	NaN	outcome	LifeThreatening	31.786216	0.643855	0.425845	0.218010	0.000000e+00
0	NaN	age	18-39	27.697822	0.569967	0.376684	0.193284	0.000000e+00
0	NaN	indication	Hepatoma	20.833077	0.745387	0.379221	0.366166	0.000000e+00
0	NaN	psd	Sorafenib Tosylate	18.569843	0.733491	0.415125	0.318366	0.000000e+00
0	NaN	age	younger than 18	15.865773	0.586678	0.400599	0.186079	0.000000e+00
0	NaN	indication	SUICIDE ATTEMPT	15.641200	0.814972	0.380369	0.434602	0.000000e+00
0	NaN	dose	PO	13.748480	0.666717	0.378772	0.287945	0.000000e+00
0	NaN	psd	Sevoflurane	9.288299	0.756995	0.417129	0.339866	0.000000e+00
0	NaN	psd	Heparin Sodium	8.992280	0.721102	0.416965	0.304137	0.000000e+00
0	NaN	indication	PYREXIA	8.702539	0.611096	0.380615	0.230481	0.000000e+00
0	NaN	indication	HEADACHE	8.535302	0.696896	0.381732	0.315164	0.000000e+00
0	NaN	indication	MULTIPLE MYELOMA	5.072737	0.565151	0.382309	0.182842	1.960672e-07
0	NaN	outcome	RequiredIntervention	4.691210	0.530871	0.462368	0.068503	1.357972e-06
0	NaN	indication	BCa	4.679434	0.562003	0.382399	0.179603	1.438340e-06
0	NaN	indication	ATRIAL FIBRILLATION	4.584499	0.521577	0.382176	0.139401	2.275384e-06
0	NaN	psd	Bromfenac Sodium	3.898813	0.550953	0.417779	0.133174	4.833273e-05

Continued on next page

level	route	feature	value	z score	P(do)	P(not do)	delta	p value
0	NaN	indication	PLASMA CELL MYELOMA	3.573574	0.506373	0.382737	0.123636	1.760708e-04
0	NaN	psd	Bevacizumab	3.191665	0.538693	0.417730	0.120963	7.072771e-04
0	NaN	psd	Fentanyl	2.483843	0.508024	0.417838	0.090187	6.498656e-03
0	NaN	psd	Tacrolimus	2.326156	0.518834	0.417896	0.100938	1.000512e-02
0	NaN	psd	Ribavirin	2.282419	0.501360	0.417916	0.083443	1.123230e-02
0	NaN	psd	Moxifloxacin Hydrochloride	2.044682	0.513134	0.417963	0.095171	2.044311e-02
0	NaN	psd	Peginterferon Alfa-2a	1.861280	0.501151	0.417999	0.083152	3.135235e-02
1	Acetaminophen	outcome	Death	22.244307	0.947359	0.802236	0.145123	0.000000e+00
1	Acetaminophen	age	18-39	12.400980	0.896862	0.787609	0.109254	0.000000e+00
1	Acetaminophen	dose	PO	12.083208	0.939923	0.761961	0.177962	0.000000e+00
1	Acetaminophen	gender	Female	8.167440	0.862923	0.785038	0.077885	1.110223e-16
1	Acetaminophen	outcome	LifeThreatening	7.175243	0.924059	0.855190	0.068869	3.609335e-13
1	Acetaminophen	indication	SUICIDE ATTEMPT	3.626359	0.849500	0.727185	0.122315	1.437227e-04
1	Acetaminophen	age	younger than 18	3.383765	0.870437	0.829439	0.040998	3.574963e-04
1	Death	psd	Acetaminophen	31.143091	0.950154	0.638516	0.311638	0.000000e+00
1	Death	age	younger than 18	9.550135	0.894029	0.651608	0.242420	0.000000e+00
1	Death	age	18-39	5.576646	0.766855	0.646248	0.120607	1.225999e-08
1	Death	gender	Female	4.335144	0.712261	0.645552	0.066709	7.283228e-06
1	Death	dose	larger than 100 MG	3.348666	0.724683	0.649679	0.075004	4.060075e-04
1	LifeThreatening	gender	Female	6.442372	0.622386	0.407217	0.215169	5.881029e-11
1	18-39	psd	Acetaminophen	50.619347	0.896574	0.434313	0.462260	0.000000e+00
1	18-39	outcome	Death	28.740611	0.842180	0.503710	0.338470	0.000000e+00
1	18-39	dose	PO	10.107138	0.826822	0.486738	0.340084	0.000000e+00
1	18-39	outcome	LifeThreatening	7.052340	0.703502	0.593798	0.109704	8.796297e-13
1	18-39	gender	Female	6.766981	0.603800	0.515327	0.088473	6.574852e-12
1	18-39	psd	Ibuprofen	3.178864	0.647794	0.565295	0.082500	7.392668e-04
1	Sorafenib Tosylate	indication	Hepatoma	1.714253	0.759037	0.681764	0.077272	4.324116e-02
1	younger than 18	psd	Acetaminophen	23.960196	0.871289	0.435987	0.435302	0.000000e+00
1	younger than 18	outcome	Death	14.298249	0.861023	0.547187	0.313836	0.000000e+00
1	younger than 18	indication	PYREXIA	6.784631	0.746150	0.486771	0.259379	5.819123e-12

Continued on next page

level	route	feature	value	z score	P(do)	P(not do)	delta	p value
1	younger than 18	outcome	LifeThreatening	3.966963	0.711858	0.606191	0.105667	3.639715e-05
1	PO	psd	Acetaminophen	17.687622	0.954664	0.389923	0.564741	0.000000e+00
1	PO	outcome	Death	9.850769	0.871450	0.460842	0.410608	0.000000e+00
1	PO	gender	Female	2.858914	0.703747	0.567709	0.136038	2.125468e-03
1	PYREXIA	age	younger than 18	5.283432	0.772747	0.473254	0.299493	6.339288e-08
1	PYREXIA	psd	Acetaminophen	4.093294	0.734495	0.520114	0.214381	2.126439e-05
1	ATRIAL FIBRILLATION	gender	Female	1.654457	0.570563	0.467542	0.103021	4.901731e-02
2	Acetaminophen->Death	gender	Female	3.167332	0.961885	0.917120	0.044764	7.692231e-04
2	Acetaminophen->Death	age	18-39	2.003098	0.973713	0.939500	0.034213	2.258340e-02
2	Acetaminophen->18-39	outcome	Death	10.516641	0.957750	0.872160	0.085590	0.000000e+00
2	Acetaminophen->18-39	gender	Female	7.015725	0.921533	0.839504	0.082029	1.143752e-12
2	Acetaminophen->18-39	outcome	LifeThreatening	1.930901	0.933172	0.905136	0.028037	2.674766e-02
2	Acetaminophen->Female	outcome	Death	14.686810	0.959463	0.841126	0.118338	0.000000e+00
2	Acetaminophen->Female	age	18-39	10.684067	0.921533	0.809817	0.111716	0.000000e+00
2	Acetaminophen->Female	dose	PO	6.422089	0.937874	0.814910	0.122964	6.720824e-11
2	Acetaminophen->Female	outcome	LifeThreatening	1.973645	0.916763	0.888548	0.028214	2.421103e-02
2	Acetaminophen->younger than 18	outcome	Death	8.317891	0.976448	0.842530	0.133918	0.000000e+00
2	Acetaminophen->younger than 18	gender	Female	2.851051	0.876406	0.797688	0.078718	2.178748e-03
2	Death->Acetaminophen	gender	Female	3.167332	0.961885	0.917120	0.044764	7.692231e-04
2	Death->Acetaminophen	age	18-39	2.003098	0.973713	0.939500	0.034213	2.258340e-02
2	Death->18-39	psd	Acetaminophen	10.778413	0.973713	0.675155	0.298558	0.000000e+00
2	Death->18-39	gender	Female	4.030381	0.836740	0.675018	0.161722	2.784324e-05
2	Death->Female	psd	Acetaminophen	21.909567	0.961885	0.661523	0.300362	0.000000e+00
2	Death->Female	age	18-39	6.071396	0.836740	0.673600	0.163140	6.340144e-10
2	18-39->Acetaminophen	outcome	Death	10.516641	0.957750	0.872160	0.085590	0.000000e+00
2	18-39->Acetaminophen	gender	Female	7.015725	0.921533	0.839504	0.082029	1.143752e-12
2	18-39->Acetaminophen	outcome	LifeThreatening	1.930901	0.933172	0.905136	0.028037	2.674766e-02
2	18-39->Death	psd	Acetaminophen	10.778413	0.973713	0.675155	0.298558	0.000000e+00
2	18-39->Death	gender	Female	4.030381	0.836740	0.675018	0.161722	2.784324e-05
2	18-39->Female	psd	Acetaminophen	40.922891	0.920576	0.451171	0.469405	0.000000e+00

Continued on next page

level	route	feature	value	z score	P(do)	P(not do)	delta	p value
2	18-39->Female	outcome	Death	23.697463	0.885832	0.546349	0.339482	0.000000e+00
2	18-39->Female	outcome	LifeThreatening	9.256527	0.798204	0.623489	0.174715	0.000000e+00
2	18-39->Female	dose	larger than 100 MG	2.370148	0.568742	0.508977	0.059766	8.890480e-03
2	18-39->Female	outcome	RequiredIntervention	2.347938	0.728041	0.654106	0.073935	9.438831e-03
2	Sorafenib Tosylate->Hepatoma	age	older than 65	1.646785	0.783820	0.731382	0.052438	4.980112e-02
2	younger than 18->Acetaminophen	outcome	Death	8.317891	0.976448	0.842530	0.133918	0.000000e+00
2	younger than 18->Acetaminophen	gender	Female	2.851051	0.876406	0.797688	0.078718	2.178748e-03

A.5 LLama Results with Extended Features

A.5.1 Root Causal Terms

feature	value	z score	P(do)	P(not do)	delta	p value	support
outcome	Death	74.440307	0.715568	0.319010	0.396558	0.000000e+00	11311
outcome	LifeThreatening	30.105857	0.641076	0.426657	0.214418	0.000000e+00	5488
patient age	18-39	27.685173	0.575937	0.372583	0.203354	0.000000e+00	5350
indication	Hepatoma	19.092432	0.800617	0.378591	0.422026	0.000000e+00	271
dose	PO	16.620928	0.711481	0.375686	0.335795	0.000000e+00	503
psd	Sorafenib Tosylate	16.152675	0.756602	0.415895	0.340706	0.000000e+00	360
ssd	Diphenhydramine	15.549527	0.816578	0.363357	0.453222	0.000000e+00	180
patient age	younger than 18	14.849305	0.580270	0.398531	0.181739	0.000000e+00	1687
indication	SUICIDE ATTEMPT	14.118277	0.802179	0.380527	0.421652	0.000000e+00	170
ssd	Hydrocodone Bitartrate	13.542467	0.752468	0.362978	0.389490	0.000000e+00	226
ssd	Alcohol	12.407953	0.659109	0.361220	0.297889	0.000000e+00	396
patient gender	Female	10.389699	0.433251	0.378891	0.054361	0.000000e+00	17843
indication	HEADACHE	9.860729	0.746329	0.381478	0.364850	0.000000e+00	139
psd	Sevoflurane	9.246430	0.771095	0.418076	0.353020	0.000000e+00	121
ssd	Propofol	8.325379	0.724785	0.365849	0.358936	0.000000e+00	109
indication	PYREXIA	8.076194	0.609379	0.380698	0.228681	3.330669e-16	297
psd	Heparin Sodium	7.322564	0.688936	0.418096	0.270840	1.216804e-13	155
ssd	Ibuprofen	7.185829	0.501900	0.362576	0.139324	3.339551e-13	681
ssd	Metamizole	6.060700	0.584041	0.365711	0.218330	6.776524e-10	190
patient weight	50-70	5.752971	0.438426	0.387536	0.050890	4.384424e-09	5060
dose	larger than 100 MG	5.573317	0.408989	0.369859	0.039130	1.249668e-08	7377
ssd	Acetaminophen	5.269953	0.408182	0.358471	0.049711	6.822953e-08	3496
route	intravenous	5.126232	0.429980	0.382844	0.047136	1.477994e-07	3252
route	respiratory	4.269811	0.548142	0.387697	0.160445	9.781930e-06	175
outcome	RequiredIntervention	4.217994	0.527303	0.462674	0.064629	1.232428e-05	1078

Continued on next page

feature	value	z score	P(do)	P(not do)	delta	p value	support
ssd	Oxycodone Hydrochloride	3.894731	0.506301	0.366538	0.139763	4.915384e-05	200
indication	ATRIAL FIBRILLATION	3.650350	0.500041	0.382478	0.117564	1.309414e-04	244
indication	BCa	3.519942	0.521245	0.382761	0.138484	2.158209e-04	162
ssd	Codeine Phosphate	3.476279	0.454540	0.366122	0.088419	2.542114e-04	485
ssd	Fluorouracil	3.442080	0.497388	0.366805	0.130583	2.886293e-04	175
ssd	Tacrolimus	2.906737	0.509417	0.367262	0.142155	1.826102e-03	106
indication	BACK PAIN	2.849608	0.456223	0.382549	0.073674	2.188660e-03	368
indication	MULTIPLE MYELOMA	2.815883	0.488168	0.382968	0.105200	2.432172e-03	170
psd	Bromfenac Sodium	2.792816	0.537695	0.418820	0.118876	2.612574e-03	130
route	oral	2.752520	0.393520	0.374435	0.019084	2.956924e-03	19399
ssd	Cefazolin Sodium	2.631113	0.493818	0.367344	0.126474	4.255291e-03	108
psd	Rosiglitazone Maleate	2.391166	0.508392	0.418833	0.089559	8.397479e-03	167
ssd	Methylprednisolone	2.350201	0.485186	0.367448	0.117738	9.381645e-03	101
ssd	Clarithromycin	2.330696	0.468059	0.367378	0.100681	9.884688e-03	130
psd	Moxifloxacin Hydrochloride	2.166318	0.520987	0.418932	0.102055	1.514346e-02	111
ssd	Prednisolone	2.054599	0.469637	0.367535	0.102102	1.995885e-02	102
psd	Infliximab	2.050013	0.487300	0.418828	0.068471	2.018160e-02	221
indication	MULTIPLE SCLEROSIS	2.045269	0.437308	0.382928	0.054380	2.041416e-02	345
indication	PNEUMONIA	2.020595	0.480140	0.383300	0.096840	2.166085e-02	109
dose	ORAL	1.987008	0.420851	0.382666	0.038185	2.346077e-02	666
ssd	Aspirin	1.689134	0.393545	0.366589	0.026956	4.559685e-02	984

A.5.2 Causal Tree Terms

level	route	feature	value	z score	P(do)	P(not do)	delta	p value
0	NaN	outcome	Death	74.440307	0.715568	0.319010	0.396558	0.000000e+00
0	NaN	outcome	LifeThreatening	30.105857	0.641076	0.426657	0.214418	0.000000e+00
0	NaN	patient age	18-39	27.685173	0.575937	0.372583	0.203354	0.000000e+00

Continued on next page

level	route	feature	value	z score	P(do)	P(not do)	delta	p value
0	NaN	indication	Hepatoma	19.092432	0.800617	0.378591	0.422026	0.000000e+00
0	NaN	dose	PO	16.620928	0.711481	0.375686	0.335795	0.000000e+00
0	NaN	psd	Sorafenib Tosylate	16.152675	0.756602	0.415895	0.340706	0.000000e+00
0	NaN	ssd	Diphenhydramine	15.549527	0.816578	0.363357	0.453222	0.000000e+00
0	NaN	patient age	younger than 18	14.849305	0.580270	0.398531	0.181739	0.000000e+00
0	NaN	indication	SUICIDE ATTEMPT	14.118277	0.802179	0.380527	0.421652	0.000000e+00
0	NaN	ssd	Hydrocodone Bitartrate	13.542467	0.752468	0.362978	0.389490	0.000000e+00
0	NaN	ssd	Alcohol	12.407953	0.659109	0.361220	0.297889	0.000000e+00
0	NaN	indication	HEADACHE	9.860729	0.746329	0.381478	0.364850	0.000000e+00
0	NaN	psd	Sevoflurane	9.246430	0.771095	0.418076	0.353020	0.000000e+00
0	NaN	ssd	Propofol	8.325379	0.724785	0.365849	0.358936	0.000000e+00
0	NaN	indication	PYREXIA	8.076194	0.609379	0.380698	0.228681	3.330669e-16
0	NaN	psd	Heparin Sodium	7.322564	0.688936	0.418096	0.270840	1.216804e-13
0	NaN	ssd	Ibuprofen	7.185829	0.501900	0.362576	0.139324	3.339551e-13
0	NaN	ssd	Metamizole	6.060700	0.584041	0.365711	0.218330	6.776524e-10
0	NaN	route	respiratory	4.269811	0.548142	0.387697	0.160445	9.781930e-06
0	NaN	outcome	RequiredIntervention	4.217994	0.527303	0.462674	0.064629	1.232428e-05
0	NaN	ssd	Oxycodone Hydrochloride	3.894731	0.506301	0.366538	0.139763	4.915384e-05
0	NaN	indication	ATRIAL FIBRILLATION	3.650350	0.500041	0.382478	0.117564	1.309414e-04
0	NaN	indication	BCa	3.519942	0.521245	0.382761	0.138484	2.158209e-04
0	NaN	ssd	Tacrolimus	2.906737	0.509417	0.367262	0.142155	1.826102e-03
0	NaN	psd	Bromfenac Sodium	2.792816	0.537695	0.418820	0.118876	2.612574e-03
0	NaN	psd	Rosiglitazone Maleate	2.391166	0.508392	0.418833	0.089559	8.397479e-03
0	NaN	psd	Moxifloxacin Hydrochloride	2.166318	0.520987	0.418932	0.102055	1.514346e-02
1	Death	patient age	younger than 18	6.785020	0.850141	0.640814	0.209327	5.803469e-12
1	Death	patient age	18-39	5.353730	0.756468	0.633340	0.123128	4.307979e-08
1	Death	patient gender	Female	4.572395	0.704629	0.629983	0.074646	2.410906e-06
1	Death	dose	larger than 100 MG	4.205922	0.741422	0.646438	0.094983	1.300102e-05
1	Death	ssd	Aspirin	2.754603	0.735571	0.621055	0.114516	2.938165e-03
1	Death	route	oral	2.743450	0.702624	0.643759	0.058865	3.039869e-03

Continued on next page

level	route	feature	value	z score	P(do)	P(not do)	delta	p value
1	Death	patient weight	50-70	2.433214	0.708085	0.638503	0.069583	7.482735e-03
1	LifeThreatening	psd	Acetaminophen	17.466727	0.940435	0.426393	0.514042	0.000000e+00
1	LifeThreatening	patient gender	Female	4.685269	0.571861	0.410200	0.161660	1.397960e-06
1	LifeThreatening	patient age	40-64	2.180137	0.524061	0.438269	0.085792	1.462367e-02
1	LifeThreatening	dose	larger than 100 MG	1.877775	0.521997	0.424432	0.097565	3.020599e-02
1	18-39	outcome	Death	26.813514	0.848799	0.512620	0.336178	0.000000e+00
1	18-39	dose	PO	10.071869	0.837317	0.504294	0.333023	0.000000e+00
1	18-39	ssd	Acetaminophen	8.660281	0.656551	0.471099	0.185452	0.000000e+00
1	18-39	patient gender	Female	8.451468	0.620752	0.504106	0.116645	0.000000e+00
1	18-39	route	oral	6.871446	0.566119	0.423958	0.142162	3.177680e-12
1	18-39	patient weight	50-70	6.564497	0.616860	0.456015	0.160845	2.610445e-11
1	18-39	ssd	Ibuprofen	6.422252	0.707611	0.503691	0.203920	6.713663e-11
1	18-39	outcome	LifeThreatening	5.126056	0.691303	0.607739	0.083564	1.479377e-07
1	18-39	ssd	Alcohol	4.339225	0.682209	0.511541	0.170668	7.149289e-06
1	18-39	ssd	Aspirin	3.493567	0.676415	0.514888	0.161527	2.383067e-04
1	18-39	indication	MULTIPLE SCLEROSIS	3.208423	0.647141	0.493876	0.153265	6.673262e-04
1	18-39	dose	larger than 100 MG	2.092612	0.543164	0.503134	0.040030	1.819188e-02
1	PO	psd	Acetaminophen	14.468120	0.942538	0.412379	0.530159	0.000000e+00
1	PO	outcome	Death	10.063301	0.899856	0.473601	0.426254	0.000000e+00
1	PO	patient age	18-39	4.165734	0.810971	0.610070	0.200901	1.551763e-05
1	PO	patient gender	Female	2.005982	0.712458	0.615586	0.096872	2.242908e-02
1	Sorafenib Tosylate	indication	Hepatoma	3.687556	0.803245	0.617145	0.186100	1.132092e-04
1	Sorafenib Tosylate	patient age	40-64	2.179041	0.793106	0.701932	0.091174	1.466429e-02
1	younger than 18	outcome	Death	12.587780	0.839086	0.542888	0.296198	0.000000e+00
1	younger than 18	indication	PYREXIA	5.551811	0.718760	0.484065	0.234695	1.413628e-08
1	younger than 18	route	oral	4.788880	0.571035	0.414522	0.156513	8.385732e-07
1	younger than 18	ssd	Acetaminophen	3.768648	0.600157	0.447780	0.152377	8.206704e-05
1	younger than 18	outcome	LifeThreatening	3.425673	0.694886	0.599761	0.095125	3.066389e-04
1	younger than 18	patient weight	50-70	3.241118	0.620528	0.481490	0.139037	5.953101e-04
1	PYREXIA	patient age	younger than 18	3.699897	0.741942	0.510037	0.231905	1.078435e-04

Continued on next page

level	route	feature	value	z score	P(do)	P(not do)	delta	p value
1	PYREXIA	psd	Acetaminophen	3.670316	0.730059	0.518803	0.211256	1.211256e-04
1	Ibuprofen	outcome	Death	15.718421	0.839616	0.401192	0.438424	0.000000e+00
1	Ibuprofen	patient age	18-39	10.668286	0.684620	0.367401	0.317219	0.000000e+00
1	Ibuprofen	ssd	Acetaminophen	6.558068	0.659955	0.423946	0.236010	2.725464e-11
1	Ibuprofen	patient gender	Male	2.625454	0.530582	0.451566	0.079016	4.326680e-03
1	Ibuprofen	indication	PYREXIA	2.534569	0.571477	0.437069	0.134408	5.629282e-03
1	Ibuprofen	outcome	LifeThreatening	1.906797	0.567248	0.496475	0.070774	2.827345e-02
1	ATRIAL FIBRILLATION	patient gender	Female	1.678365	0.548064	0.437579	0.110485	4.663792e-02
2	Death->18-39	psd	Acetaminophen	11.090087	0.979737	0.639980	0.339757	0.000000e+00
2	Death->18-39	patient gender	Female	5.525178	0.857636	0.621321	0.236315	1.645762e-08
2	Death->Female	patient age	18-39	7.295851	0.857636	0.662003	0.195633	1.484368e-13
2	Death->Female	dose	larger than 100 MG	1.942858	0.728780	0.661489	0.067291	2.601664e-02
2	Death->larger than 100 MG	route	oral	1.757043	0.774364	0.684610	0.089754	3.945524e-02
2	Death->oral	patient age	18-39	5.320751	0.838197	0.671521	0.166676	5.166986e-08
2	Death->oral	dose	larger than 100 MG	4.366948	0.774364	0.647794	0.126570	6.299744e-06
2	18-39->Death	psd	Acetaminophen	11.090087	0.979737	0.639980	0.339757	0.000000e+00
2	18-39->Death	patient gender	Female	5.525178	0.857636	0.621321	0.236315	1.645762e-08
2	18-39->Acetaminophen	outcome	Death	9.289193	0.979265	0.873918	0.105347	0.000000e+00
2	18-39->Acetaminophen	patient gender	Female	4.614967	0.926213	0.852700	0.073513	1.965787e-06
2	18-39->Acetaminophen	dose	PO	4.174770	0.947527	0.842505	0.105022	1.491433e-05
2	18-39->Female	outcome	Death	22.137177	0.898362	0.567789	0.330573	0.000000e+00
2	18-39->Female	outcome	LifeThreatening	9.668775	0.822913	0.640411	0.182502	0.000000e+00
2	18-39->Female	ssd	Acetaminophen	6.691598	0.716054	0.533741	0.182313	1.103728e-11
2	18-39->Female	route	oral	6.272006	0.610820	0.444234	0.166587	1.782124e-10
2	18-39->Female	patient weight	50-70	6.070173	0.623571	0.437812	0.185759	6.388623e-10
2	18-39->Female	dose	larger than 100 MG	3.089998	0.608401	0.531601	0.076800	1.000790e-03
2	18-39->Female	ssd	Ibuprofen	3.000461	0.697982	0.573712	0.124270	1.347856e-03
2	18-39->Female	outcome	RequiredIntervention	2.905000	0.766904	0.671246	0.095658	1.836265e-03
2	18-39->oral	outcome	Death	27.388677	0.912739	0.476234	0.436505	0.000000e+00
2	18-39->oral	dose	PO	9.425973	0.837317	0.515987	0.321329	0.000000e+00

Continued on next page

level	route	feature	value	z score	P(do)	P(not do)	delta	p value
2	18-39->oral	patient gender	Female	5.887581	0.610820	0.495757	0.115063	1.959450e-09
2	18-39->oral	patient weight	50-70	4.599318	0.618508	0.458719	0.159789	2.119380e-06
2	18-39->oral	ssd	Ibuprofen	3.532164	0.680945	0.514897	0.166048	2.060867e-04
2	18-39->oral	outcome	LifeThreatening	3.456174	0.683706	0.601207	0.082499	2.739509e-04
2	18-39->oral	ssd	Acetaminophen	3.001200	0.597274	0.503491	0.093782	1.344590e-03
2	18-39->oral	outcome	RequiredIntervention	1.694524	0.669168	0.612995	0.056173	4.508289e-02
2	18-39->50-70	outcome	Death	8.293320	0.858102	0.593730	0.264372	0.000000e+00
2	18-39->50-70	outcome	LifeThreatening	8.244788	0.856310	0.592970	0.263340	1.110223e-16
2	18-39->50-70	route	oral	3.491384	0.618508	0.450299	0.168210	2.402629e-04
2	18-39->50-70	dose	larger than 100 MG	1.970986	0.635830	0.536381	0.099449	2.436272e-02
2	18-39->Ibuprofen	outcome	Death	8.535590	0.925866	0.587580	0.338286	0.000000e+00
2	18-39->Ibuprofen	ssd	Acetaminophen	3.240273	0.796815	0.629571	0.167244	5.970760e-04
2	18-39->Ibuprofen	ssd	Ibuprofen	3.160399	0.805231	0.640580	0.164651	7.877652e-04
2	18-39->Ibuprofen	patient gender	Female	1.832224	0.718593	0.629556	0.089037	3.345904e-02
2	18-39->larger than 100 MG	outcome	Death	13.176973	0.852937	0.497451	0.355486	0.000000e+00
2	18-39->larger than 100 MG	outcome	LifeThreatening	7.097173	0.783356	0.553616	0.229740	6.367129e-13
2	18-39->larger than 100 MG	patient gender	Female	5.104027	0.608401	0.460173	0.148228	1.662505e-07
2	18-39->larger than 100 MG	route	oral	3.329503	0.542817	0.396620	0.146197	4.350053e-04
2	18-39->larger than 100 MG	patient weight	50-70	3.304492	0.635830	0.450353	0.185477	4.757434e-04
2	18-39->larger than 100 MG	outcome	RequiredIntervention	2.402930	0.694925	0.591103	0.103822	8.132155e-03
2	18-39->larger than 100 MG	ssd	Acetaminophen	2.008586	0.553299	0.445448	0.107851	2.229051e-02
2	younger than 18->oral	psd	Acetaminophen	12.906641	0.827499	0.409338	0.418162	0.000000e+00
2	younger than 18->oral	outcome	Death	11.155791	0.900667	0.535021	0.365646	0.000000e+00
2	younger than 18->oral	patient weight	50-70	2.230282	0.641437	0.511680	0.129758	1.286435e-02
2	younger than 18->oral	patient gender	Female	2.179208	0.579989	0.493289	0.086700	1.465810e-02
2	younger than 18->Acetaminophen	outcome	Death	5.789510	0.955717	0.824155	0.131562	3.529600e-09
2	Ibuprofen->18-39	outcome	Death	8.535590	0.925866	0.587580	0.338286	0.000000e+00
2	Ibuprofen->18-39	ssd	Acetaminophen	3.240273	0.796815	0.629571	0.167244	5.970760e-04
2	Ibuprofen->18-39	ssd	Ibuprofen	3.160399	0.805231	0.640580	0.164651	7.877652e-04
2	Ibuprofen->18-39	patient gender	Female	1.832224	0.718593	0.629556	0.089037	3.345904e-02

Continued on next page

level	route	feature	value	z score	P(do)	P(not do)	delta	p value
2	Ibuprofen->Acetaminophen	ssd	Ibuprofen	2.138524	0.885727	0.776708	0.109019	1.623713e-02
2	Ibuprofen->Male	patient age	18-39	2.729271	0.629556	0.492336	0.137220	3.173725e-03