Explainable Artificial Intelligence for Pharmacovigilance: What Features Are Important When Predicting Adverse Outcomes?

ISAAC RONALD WARD, School of Population & Global Health, and Department of Computer Science & Software Engineering, University of Western Australia, Australia
LING WANG, School of Population & Global Health, University of Western Australia, Australia
JUAN LU, School of Population & Global Health, University of Western Australia, Australia
MOHAMMED BENNAMOUN, Department of Computer Science & Software Engineering, University of Western Australia, Australia
GIrish DWivedI, Cardiology Department Fiona Stanley Hospital, Harry Perkins Institute of Medical Research, Medical School University of Western Australia, Australia
Frank M Sanfilippo*, School of Population & Global Health, University of Western Australia, Australia

Background and Objective
Explainable Artificial Intelligence (XAI) has been identified as a viable method for determining the importance of features when making predictions using Machine Learning (ML) models. In this study, we created models that take an individual’s health information (e.g. their drug history and comorbidities) as inputs, and predict the probability that the individual will have an Acute Coronary Syndrome (ACS) adverse outcome.

Methods
Using XAI, we quantified the contribution that specific drugs had on these ACS predictions, thus creating an XAI-based technique for pharmacovigilance monitoring, using ACS as an example of the adverse outcome to detect. Individuals aged over 65 who were supplied Musculo-skeletal system (anatomical therapeutic chemical (ATC) class M) or Cardiovascular system (ATC class C) drugs between 1993 and 2009 were identified, and their drug histories, comorbidities, and other key features were extracted from linked Western Australian datasets. Multiple ML models were trained to predict if these individuals would have an ACS related adverse outcome (i.e., death or hospitalisation with a discharge diagnosis of ACS), and a variety of ML and XAI techniques were used to calculate which features — specifically which drugs — led to these predictions.

Results
The drug dispensing features for rofecoxib and celecoxib were found to have a greater than zero contribution to ACS related adverse outcome predictions (on average), and it was found that ACS related adverse outcomes can be predicted with 72% accuracy. Furthermore, the XAI libraries LIME and SHAP were found to successfully identify both important and unimportant features, with SHAP slightly outperforming LIME.

Conclusions
ML models trained on linked administrative health datasets in tandem with XAI algorithms can successfully quantify feature importance, and with further development, could potentially be used as pharmacovigilance monitoring techniques.

1 INTRODUCTION
Artificial Intelligence (AI) and Machine Learning (ML) have recently demonstrated their utility in digital health applications regarding the prediction of outcome events [1, 2]. These techniques use models that learn patterns based on large quantities of data. These models typically demonstrate high predictive power, but they have been criticised as being ‘black box’ algorithms: their internal operations incomprehensible to human operators [3]. In critical decision making domains such as healthcare — the reason for a decision is often as equally important as the decision itself.

In response, Explainable Artificial Intelligence (XAI) has experienced a surge in interest and development. XAI is a field concerned with increasing the explainability and transparency of AI algorithms, by making their influencing variables, complex internal operations, and learned decision making paths interpretable [3, 4]. Although popular XAI methods such as ‘Local Interpretable Model-Agnostic Explanations’ (LIME) [5] and ‘SHapley Additive exPlanations’ (SHAP) [6] have proven to be useful in interpreting black box models [7, 8], more needs to be understood about these methods before they can be adopted in critical decision making domains [9].

In this work, we used pharmacovigilance monitoring as the critical decision making domain which demands transparency and trust. Pharmacovigilance systems aim to recognise pharmaceutical safety issues, and thus impact human well-being, public health, pharmaceutical companies [1], policy, and regulations at the highest scale. In one case in 2004, the non-steroidal anti-inflammatory drug (NSAID) rofecoxib was withdrawn from global markets based on evidence that it could double the risk of myocardial infarction and stroke if taken for 18 months or more [10]. At this time, celecoxib, another COX-2 selective inhibitor had its sales reduced by 50% [11]. Methods which could have quantified and exposed these risks earlier would have significantly improved public health and safety, and both statistical and AI-based methods have been suggested as candidate solutions [12–18].

The objective of this study was thus two-fold. Firstly, we used ML to predict the probability of an individual having an Acute Coronary Syndrome (ACS) based adverse outcome. We did this using only administrative health data (Section 2.3). Secondly, we analysed these ML models with XAI, and confirmed that expected patterns were identified correctly (Section 2.4). Among other patterns, we expected to find that taking rofecoxib and celecoxib would lead to

*The corresponding author.
a disproportionate increase in the predicted probability of an ACS related adverse outcome.

2 METHODS

2.1 Datasets

The study dataset was prepared by linking data from local core linked administrative datasets (Western Australian Department of Health) and pharmacy dispensing data from the Australian government. All data are de-identified so as to protect sensitive information. The exposure information is provided by the Pharmaceutical Benefits Scheme (PBS) data, which contains drug dispensing data from PBS-registered pharmacies (community and hospitals). Among other variables, rows in this dataset describe drug dispensing events by their date of supply, quantity, and ATC classification system codes. The outcome events from core datasets of the Western Australian Data Linkage System (WADLS): the Hospital Morbidity Data Collection (HMDC), Emergency Department Data Collection (EDDC), and Deaths Register [19]. These datasets contain records of all admissions to public and private hospitals in Western Australia (HMDC), emergency department presentations (EDDC) and deaths, and allow us to identify events by their codes from the International Classification of Diseases 9th revision clinical modification (ICD-9-CM) and 10th revision Australian Modification (ICD-10-AM).

2.2 Data preparation

To produce data features which can be used for ML, the PBS, HMDC, EDDC, and Deaths Register data were first cleaned, with null entries being dropped entirely. Sex and age are taken from the PBS data, and entries are linked across datasets by the patient’s anonymous encrypted identification code. ATC M (musculo-skeletal system) and C (cardiovascular system) class drugs were investigated for this study, as they capture the drugs of interest (rofecoxib and celecoxib) as drugs used by patients with cardiovascular diseases. We disregard uncommon drugs with less than 10,000 total dispensing events. ACS related hospitalisations and deaths were identified for each individual from the HMDC, EDDC, and Death Register. The study cohort included individuals who were classified as concessional beneficiaries, as their dispensing records within the PBS are more complete than individuals in the general category [20]. We additionally limit the study cohort to individuals aged over 65, as not all concessional beneficiaries are over 65. Finally, we used data from January 1st 1993 to December 31st 2009, which includes the exposure period of interest (2003-2004) with 10-year lookback for comorbidity history period, and ACS related hospital admissions or deaths were considered adverse outcomes if they were within the follow up period. These conditions were identified by ICD-9-CM and ICD-10-AM codes. Similarly, any PBS dispensing events for C or M class drugs were added to an individual’s drug history if they were within the drug history period. A full list of the conditions considered as adverse outcomes, and the ATC codes of considered drugs are shown in Appendix A.

Hospital admissions or emergency department presentations were considered comorbidities if they were within the 10 year comorbidity history period, and ACS related hospital admissions or deaths were considered adverse outcomes if they were within the follow up period. These conditions were identified by ICD-9-CM and ICD-10-AM codes. Similarly, any PBS dispensing events for C or M class drugs were added to an individual’s drug history if they were within the drug history period. A full list of the conditions considered as adverse outcomes, and the ATC codes of considered drugs are shown in Appendix A.

The 158-dimension, per patient, per supply feature vectors were constructed by concatenating an individual’s sex, age, drug history, comorbidities, and two additional random features. Drug history is represented by 59 counts of specific C class drug dispensings, and 19 counts of specific M class drug dispensings, where each element corresponds to a valid ATC drug code which was recorded in the PBS data. The value of each element corresponds to the number of dispensings events the individual had registered for that particular drug in the drug history period. Similarly, hospitalisation comorbidities were represented by 76 values, where each value corresponds to the number of hospitalisation events that the individual had for that specific condition in the comorbidity period. The random features were added because by definition they have no predictive power. This aids in quantifying the models’ tendency to overfit, and in quantifying the feature importance methods’ ability to detect truly important features. A random integer (in the range 0–158) and floating point number (in the range 0–1) is appended to each feature vector. Note that every feature in the feature vector is represented numerically, and that scikit-learn interprets all features as continuous numeric variables with threshold-based decision boundaries — thus preventing the possibility of bias due to data type.

The label for a feature vector was a 0 or 1 (one-hot encoding), denoting if an individual suffered an ACS related adverse outcome or death during the follow up period (1) or not (0). We removed all instances which had identical feature vectors but different labels, as these cases lacked the input information required to discriminate between the two possible outcomes. This limitation is inherent in administrative datasets; such datasets do not capture all the information about an individual that describes their full clinical history and presentation.
The resulting linked dataset is imbalanced: the number of instances which did not have ACS related adverse outcomes (the negative class) outnumbered those who did (the positive class) by a ratio 6.94:1. For the training set, we used 70% of the instances and randomly undersampled from the negative class population until the training dataset was exactly equal, totalling 278,608 instances [21]. Random undersampling approaches have been proven to be suitable for dealing with minority classes with relative class sizes that were even smaller than ours (our relative minority class size is 14.41%, and 0.1% minority classes have been trained with undersampling with minimal performance losses in [22]). Having a balanced training set will prevent the ML models from learning to predict the larger class in lieu of identifying meaningful patterns. We used 4-fold cross validation when training and testing, with a ratio of 30% of the total instances reserved for the testing set. Performance measures were averaged across folds.

2.3 Machine Learning
We used Decision Tree (DT) based classification models: Random Forests (RFs), Extra random Trees (ETs), and eXtreme Gradient Boosting machines (XGB) [23–25]. These tree based models benefit from being both widely accepted and used, as well as having accelerated feature importance value computations [6] (see Section 2.4). Moreover, using a variety of proven ML models allows us to investigate if our results are not model dependent.

A DT is a representation of an algorithm that follows a tree-like structure (see Figure 2). In ML, DTs can be generated from labelled datasets — during each training step, a condition is based on some selected feature (either randomly or based on some measure of optimality), and the data is split into subsets based on the outcome of this condition. Each splitting point creates new nodes, and this process continues recursively until some stopping criteria has been met (e.g., accuracy, node limits). One measure of optimality is Gini impurity, which measures the probability of an instance being classified incorrectly when classifying it based on the class distribution of the dataset.

It is also possible to identify the most important features by inspecting the DTs directly; the features that are used as splitting criteria in shallow nodes discriminate between feature vectors more effectively (i.e., these features have the highest information gain). In practice, we find that single DTs are prone to overfitting due to the small samples that occur near the tree’s leaf nodes. As such, all of our feature importance analyses are based on ensembles of trees, rather than single trees, which will ensure that the calculated feature importance scores are less variant and more indicative of the global patterns in the data.

An RF is an ensemble architecture which bags multiple DTs to produce more stable predictions. DTs with high complexity often overfit, whereas RFs create subsets of random features and build a greater number of smaller DTs using these subsets. This has a tendency to increase variance and thus reduce overfitting [23]. Furthermore, ETs further increase the variance by also randomly choosing the decision threshold at each node in each DT [24]. XGBs are another bagging approach to DTs, but in this case boosting is used to alter the evaluation criteria for each DT. These errors are minimised via gradient descent. XGBs have numerous performance optimisations which suit them for the purpose of this study [25, 26].

To optimally train these models we performed 128 rounds of Bayesian hyperparameter optimisation over a distribution of valid parameters (see Appendix B for the descriptions of these searches) [27]. In each round, a configuration of parameters is selected and the resulting model is trained and tested using 4-fold cross validation [28]. Performance measures were averaged across folds. The most performant model from the 128 trials is selected and its feature importance is analysed using both traditional methods and XAI (the results of which are averaged across the three models and presented in Figure 3).

2.4 Feature importance
We used ML-based feature importance methods to understand the contribution of certain features to the model’s prediction. From a pharmacovigilance perspective, we want to know which drug(s) have the highest association with the outcome. The traditional methods that we employed were measures of Mean Decrease of Impurity (MDI), and Mean Decrease in Accuracy (MDA), and the XAI-based methods are LIME and SHAP.
MDI is based on the depth of certain nodes in a tree-based model’s DTs. If a node near the top of the DT uses some criterion based on a given feature, then that feature must contribute to the final decision for a larger fraction of samples than a feature which is used lower in the DT. This fraction can be used as an estimate of a feature’s relative importance. MDI is calculated by combining the decrease in Gini impurity with this relative importance [29]. We used scikit-learn’s implementation of MDI in this study.

MDA is defined as the decrease in model accuracy on the test set when a given feature is randomised or permuted [23, 30]. The drop in accuracy indicates how much the model depends on the given feature, and thus is an estimate of feature importance. MDA benefits from being model agnostic, unlike MDI measures which need to be performed on tree-based models. Additionally, MDA does not suffer from a bias towards high cardinality features (as MDI does). The method does still suffer from some bias however — when two or more features are highly correlated, permuting one feature does not restrict the model’s access to it, as it can still be accessed via the correlated feature(s).

Local Interpretable Model-agnostic Explanations (LIME) [5] generate local surrogate models to explain individual predictions produced from ML architectures. LIME takes a single instance, and locally perturbs the feature values to other valid values based on the dataset. These perturbed values are then fed back into the trained model. This creates a new dataset of input to output mappings, which an interpretable model is trained on. During the training of this local surrogate model, instances are weighted by the distance of the perturbed feature to the single feature of interest. The result is a ML model which is explainable and has a high local fidelity: it approximates the black box model locally in feature space, but not globally. LIME has been noted to increase model interpretability on tabular data [9], but relies on the correct definition of the local neighborhood and is thus highly dependent on kernel parameters [31].

SHapley Additive exPlanations (SHAP) [6] assigns each input feature an importance value for a given prediction, based on principles from cooperative game theory. For any given feature vector, SHAP takes a single feature’s value and replaces it with a sampled value from the dataset. The model of interest makes a prediction for this augmented feature vector and the difference in output values — the marginal contribution — is noted. This sampling process can be repeated to improve our estimates of marginal contribution. This is repeated for all possible coalitions, and the resulting average marginal contributions to each coalition are the Shapley values. SHAP can approximate these values accurately and quickly for tree-based models.

3 RESULTS

3.1 Predicting per-patient adverse outcomes

Table 1 presents the results from the hypertuning experiments for each of the model classes. The hypertuning procedure converges each model to similar levels of accuracy and Area Under the Receiver Operating Characteristic (AUC). The XGB classifier outperforms the RF classifier, which outperforms the ET classifier.

| Model  | AUC  | Macro-avg Precision (%) | Macro-avg Recall (%) |
|--------|------|-------------------------|----------------------|
| RF     | 0.70 | 71                      | 67                   | 70                   |
| ET     | 0.69 | 72                      | 66                   | 69                   |
| XGB    | 0.72 | 72                      | 68                   | 72                   |

Table 1. The test results after training the RF, ET and XGB models used in this study. For each model, a Bayesian hyperparameter optimisation search with 128 iterations was performed. Macro-average precision was used here, as this measure is insensitive to the imbalance of classes in the test dataset.

AUC, Area under the ROC Curve; RF, Random Forest; ET, Extra Trees classifier; XGB, Extreme Gradient Boosting.

3.2 Generating feature importance for pharmacovigilance

We analysed the models’ feature importance scores using the four measures described in Section 2.4. The per-feature importance scores as determined by MDI, MDA, LIME, and SHAP are plotted in Figure 3. There is an important distinction here: for MDI and MDA measures, we showed the importance of any feature when making either an adverse outcome or a non adverse outcome prediction, as these measures do not give per-instance feature contributions. For LIME and SHAP, we present results for specifically making adverse outcome predictions, and we note that it is possible for a feature to attain on-average negative contribution to an ACS related adverse outcome prediction (these features ‘protect’ a patient from having an ACS related prediction).

We observed that age and sex are almost always attributed with a high importance — in MDI and MDA especially, but less so under LIME analysis. There are repeating peaks and patterns of importance across MDI, MDA, LIME, and SHAP in C class drug history, M class drug history, and ICD-10-AM codes (see Figure 3 and Appendix A for more detail). Rofecoxib and celecoxib are the most important M class drug history features under MDI and MDA analysis, but not so under LIME analysis. Under SHAP analysis, rofecoxib and celecoxib are the second and third M class drug history features that on average contribute to an adverse outcome prediction respectively, being slightly behind M05BA04 (alendronic acid), with the feature importance scores of rofecoxib and alendronic acid being $1.16 \times 10^{-4}$ and $1.19 \times 10^{-4}$ respectively. We note the random features’ importances under MDI analysis, which is a result of the method’s bias towards high cardinality features [30].

We further analysed the feature importance rankings at the model level by displaying the key feature importance rankings in Table 2. The results largely reflect the average results, but the XGB classifier notably disfavors random features under MDI analysis, and favors the random floating point feature under MDA and SHAP analysis. We note that different model types may correlate to the features that are considered important — even across feature importance methods. This is potentially due to inductive bias in the learning algorithm’s design. Indeed, RFs and ETs are algorithmically similar, with the only key difference being the randomisation of selected thresholds and the sampling methods used in ETs.

Table 2. Feature importance rankings at the model level for each model class. The feature importance rankings are derived from the per-instance feature contributions as determined by MDI, MDA, LIME, and SHAP.

| Feature | MDI | MDA | LIME | SHAP |
|---------|-----|-----|------|------|
| age     | 0.01| 0.01| 0.01 | 0.01 |
| sex     | 0.02| 0.02| 0.02 | 0.02 |
| rofecoxib | 0.04| 0.04| 0.04 | 0.04 |
| celecoxib | 0.03| 0.03| 0.03 | 0.03 |

Table 2. Feature importance rankings at the model level for each model class. The feature importance rankings are derived from the per-instance feature contributions as determined by MDI, MDA, LIME, and SHAP.
We also provide the per-model rankings of the M class drug features in Table 3, and we observe that the XGB model disagrees with the RF and ET models under MDI analysis, but agrees with the RF and ET models under MDA analysis. This is potentially due to the shortcomings of MDI compared to MDA, which are discussed in the next section. LIME analysis suffers from high variance and inconsistency in the results, potentially due to the method’s high local fidelity. SHAP analysis results are largely consistent, apart from the RF not identifying celecoxib dispensing as an important feature whereas the ET and XGB models both do.

Table 2. Rankings of features as measured by MDI, MDA, LIME, and SHAP. Key confounders sex and age are expected to be ranked highly, and random features to be ranked lower.

| Measure | Feature    | Overall ranking (out of 158) |
|---------|------------|------------------------------|
|         |            | RF  | ET  | XGB | Avg. |
| MDI     | sex        | 8   | 4   | 5   | 5    |
|         | age        | 1   | 1   | 20  | 1    |
|         | random float | 3  | 20  | 113 | 8    |
|         | random int | 4   | 21  | 116 | 10   |
| MDA     | sex        | 4   | 2   | 5   | 2    |
|         | age        | 1   | 1   | 1   | 1    |
|         | random float | 157  | 158 | 43  | 157  |
|         | random int | 158 | 157 | 158 | 158  |
| LIME    | sex        | 29  | 37  | 46  | 40   |
|         | age        | 43  | 53  | 63  | 57   |
|         | random float | 42  | 56  | 72  | 61   |
|         | random int | 75  | 101 | 84  | 86   |
| SHAP    | sex        | 7   | 2   | 7   | 5    |
|         | age        | 1   | 1   | 13  | 10   |
|         | random float | 155  | 128 | 20  | 35   |
|         | random int | 150 | 114 | 126 | 139  |

The average score is not the average rank, but the rank of the average importance for all models across that measure, as rendered in Figure 3. The feature with the highest importance is ranked as 1, and the lowest as 158. MDA, Mean Decrease in Impurity; MDA, Mean Decrease in Accuracy; LIME, Local Model Agnostic Explanations; SHAP, Shapley Additive Explanations; RF, Random Forest; ET, Extra Trees (classifier); XGB, Extreme Gradient Boosting.

Any dispensing of either drug decreased the likelihood. In other words: having rofecoxib and celecoxib dispensings on average increased the likelihood of an adverse outcome prediction, and having no such dispensings of these drugs on average decreased or had little effect on the likelihood (in both LIME and SHAP).

4 DISCUSSION

Our study showed that tree-based ML models trained on linked administrative datasets, in tandem with XAI techniques, have the potential to act as an ‘early warning system’ for per-patient ACS related adverse outcomes and for harmful drugs. To the best of our knowledge, this is the first study to demonstrate this using XAI.

The results of our feature importance analysis show that sex and age are ranked highly by MDI, MDA, and SHAP — which is expected, as these are the most important confounding variables in clinical and epidemiological studies. Certain C class drugs (e.g., digoxin, glyceryl trinitrate, isosorbide mononitrate, and furosemide) and comorbidities (see Figure 3 and Appendix A) are consistently reported as being highly important across all feature importance measures, which suggests some level of model independent feature importance. Importantly, the M class drugs celecoxib, rofecoxib, and alendronic acid repeatedly appear as important features across MDI, MDA, and SHAP analyses.

The LIME feature importance scores are inconsistent with the other measures for certain features (see Figure 3). A potential explanation is the variance across multiple consecutive runs in LIME [32] — features which were ranked highly in one run could be ranked lower in another [8]. Despite this, LIME has been shown to be at least as stable as SHAP for highly ranked features [9]. Our results agree
Fig. 3. Names of the features used to predict an ACS related adverse outcome, with feature importance scores (MDI, MDA, LIME, and SHAP) averaged across the RF, ET, and XGB models (to reduce cross-model variance). Sex and age are rendered in dark blue, C class drug history features in light blue, M class drug history features in magenta (with celecoxib and rofecoxib highlighted in red), comorbidities in green, and random features in black. In all importance measures other than LIME, the key confounders sex and age are identified as important features, and certain C class drugs, M class drugs, and ICD-10-AM codes are consistently identified as important in predicting an ACS related to the adverse outcome. Note the high importance of celecoxib (M01AH01) and rofecoxib (M01AH02) dispensings when compared to other M class drugs under MDI, MDA, and SHAP analysis (this is shown numerically in Table 3). All features are described in Appendix A. Image best viewed in colour. MDA, Mean Decrease in Impurity; MDA, Mean Decrease in Accuracy; LIME, Local Model Agnostic Explanations; SHAP, Shapley Additive Explanations; RF, Random Forest; ET, Extra Trees (classifier); XGB, Extreme Gradient Boosting; ACS, acute coronary syndrome; ICD, International Classification of Disease.
The Effect of Being Dispensed M01AH01 (celecoxib) and M01AH02 (rofecoxib) on the Likelihood of a Predicted ACS Related Adverse Outcome

Fig. 4. Dividing the test datasets into strata where an individual was a) dispensed celecoxib and rofecoxib, b) dispensed celecoxib and not rofecoxib, c) not dispensed celecoxib and dispensed rofecoxib, and d) where neither drug was dispensed. The contribution of M class drug history features under LIME analysis (left column) and SHAP analysis (right column) are presented. If individuals have been dispensed celecoxib or rofecoxib this on average resulted in an increased likelihood of a predicted ACS related adverse outcome. If an individual had not been dispensed rofecoxib or celecoxib, then this on average resulted in a near zero or decreased likelihood of a predicted ACS related adverse outcome. Note that, on average, the other M class drug dispensing events did not vary across these subsets, so we do not expect their feature scores to vary across rows a), b), c), and d). Celecoxib (M01AH01) and rofecoxib (M01AH02) are shown in red. M class drug codes are presented in Table 7.
with this: the highly ranked comorbidity features are consistently ranked highly across all feature importance measures. Random features were included because by definition they cannot have any predictive power — nevertheless, the MDI analysis ranked these features as important. MDI measures have a bias whereby variables with a large number of categories or potential values (high cardinality) are ranked as more important [30]. It is thereby expected that an impurity based approach will inflate the importance of random features. Moreover, MDI feature importance scores are calculated over the decision trees generated by the training dataset, so the importance of non-predictive features may be inflated. Under MDA, these high feature importance scores are reduced as the permutation importance scores are computed over a held out test set (Figure 3).

Each of our feature importance analyses demonstrate that celecoxib and rofecoxib are important when predicting ACS based adverse outcomes, or more specifically, that they actually provide a positive probability contribution to ACS based adverse outcome predictions (see Figure 3). For MDI and MDA, celecoxib and rofecoxib are ranked as the most important M class drugs when making ACS based adverse outcomes predictions. LIME and SHAP analysis shows that both drugs contribute to ACS related adverse outcomes predictions, and SHAP analysis further ranks them as the second and third most contributing M class drugs. However, LIME analysis heavily mutes the importance of these features compared to SHAP, greatly reducing the magnitude of their on-average contributions.

Our stratified analysis (see Figure 4 and Table 3) show more detailed results: having any celecoxib dispensings in an individual’s drug history increases the likelihood of an ACS related adverse outcome prediction (M01AH01 / celecoxib’s contribution is positive in Figure 4 rows a and b), and not having any celecoxib dispensings decreases or does not impact the likelihood of an ACS related adverse outcome prediction (M01AH01 / celecoxib’s contribution is negative or near-zero in Figure 4 rows c and d). In other words, the presence of celecoxib dispensing events increases the likelihood of a positive ACS prediction independent of rofecoxib dispensing events.

This is the equivalent of the counterfactual in the epidemiological theory of causality. The counterfactual conditionals are of the form “if A had not occurred, then C would not have occurred”. This is exactly what is observed in Figure 4. These results provide more robust epidemiological evidence of the associations of celecoxib and rofecoxib with the outcomes.

We acknowledge the utility of existing statistical methods in pharmacovigilance. For example, Sequence Symmetry Analysis (SSA) has been determined to be a promising solution when compared to other quantitative methods such as Reporting Odds Ratio (ROR), Proportional Reporting Ratio (PRR), and Bayesian techniques [15, 16]. SSA considers the distribution of diseases and drugs, before and after initiation of treatment for an adverse event. Asymmetry in these distributions indicate that adverse events may be due to drug supply [13]. SSA’s time to signal detection for rofecoxib-induced myocardial infarction was within 1-3 years of market entry, whereas it took 5-7 years after market entry for trial results to lead to warnings and withdrawals of the drugs [16]. Despite this, there are limitations: in these investigations there are no adverse event signals which can be linked from a joint database; the signals are deduced from dispensing data only — e.g. furosemide initiation being used as an indicator of heart failure [18]. Using joint databases is more reliable as administrative data entries can validate if a comorbidity did actually occur. Moreover, XAI is better suited for domains which are already using ML-based approaches.

The administrative data used in this study has limitations. By definition, administrative data only includes information which is pertinent to administration. Clinical features which may impact the likelihood of a patient having an ACS related adverse outcome may not be present, and rich clinical information is reduced to ICD-10-AM codes. Also, there were assumptions which we had to apply — for example, we assumed that patients who are dispensed a given drug actually consume it. These shortcomings may have impacted the predictive performance of our models (Table 1). Moreover, the RF, ET, and XGB in this study converged to similar performances, suggesting that there may be a limit to the predictive performance which can be attained using these datasets. Indeed, other ML based predictors trained on administrative data have reached similar performance limits [2, 33].

Further potential research to progress the application of ML techniques in pharmacovigilance include the testing of more diverse ML models (and studies into their inductive biases), benchmarking these methods against other datasets that include more clinical data, investigating additional XAI algorithms [4], using the feature importance scores to iteratively reduce the full set of features, and handling multicollinear features whilst maintaining interpretability. Counterfactual examples may represent an interesting XAI approach to providing transparency in the pharmacovigilance domain, but calculating such explanations may be intractable for high-dimensionality datasets [34].

The strength of our study was the use of multiple linked administrative datasets that cover the whole-of-population. Another strength was the validation of feature importances scoring methods against features that have somewhat ‘known’ importances (e.g., it is known that rofecoxib increases the risk of ACS, and that random features are unimportant). Ultimately, we believe that this work highlights the utility of XAI in further analysing trained AI models in challenging domains, and in cases where performance may be constrained due to dataset limitations.

4.1 Conclusion
Tree-based ML models trained on linked administrative datasets, in tandem with XAI techniques, have the potential to act as an ‘early warning system’ for per-patient ACS related adverse outcomes and for harmful drugs in a pharmacovigilance monitoring system. MDA and SHAP methods exceed MDI and LIME methods in identifying known harmful drugs, key confounding variables, and random features. With the appropriate infrastructure and additional clinical data, these algorithms could provide an autonomous method of monitoring adverse outcomes from medications at the population level. This represents a valuable addition to the existing statistical techniques which are currently used and would be the next step in progressing towards a real-time pharmacovigilance monitoring system.
4.2 Author contributions
F.M.S., G.D. conceived the study and provided clinical interpretation. I.R.W. prepared the data based on a data preparation approach originally implemented by J.L. IRW conducted the experiments. M.B. contributed to the design of machine learning and the use of explainable artificial intelligence. J.L. and L.W. contributed to forthrightly discussions which were led by F.M.S., which shaped this work. I.R.W., F.M.S., J.L., and L.W have access to the underlying data and have verified it. F.M.S. contributed to the conception, study design, data acquisition, funding, analysis, and supervision of I.R.W.
All co-authors critically revised the work and gave final approval and agreed to be accountable for all aspects of the work ensuring integrity and accuracy.

4.3 Ethics statement
This study complies with the Declaration of Helsinki. Human Research Ethics Committee approvals were received from the Western Australian Department of Health (#2011/62); the Australian Department of Health (XI-16); and the University of Western Australia (RA/4/1/1130).

4.4 Data sharing statement
We will consider requests for data sharing on an individual basis, with the aim to share data whenever possible for appropriate research purposes. However, this research project uses data obtained from a third-party source under strict privacy and confidentiality agreements from the Western Australian Department of Health (State) and Australian Department of Health (Federal) databases, which are governed by their ethics committees and data custodians. The data were provided after approval was granted from their jurisdictional linked PBS data; the Victorian Department of Justice and Regulation for the cause of death data held in the National Coronial Information System. We also thank the University of WA for funding this work through the 2019 FHMS research grant.

REFERENCES

[1] Schneider, J. et al. Innovation in Pharmacovigilance: Use of Artificial Intelligence in Adverse Event Case Processing. Clinical Pharmacology & Therapeutics 105, 954–961 (2019). URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cpt.1255.
[2] Han, S. S., Azad, T. D., Suarez, P. A. & Ratliff, J. K. A machine learning approach for predictive models of adverse events following spine surgery. The Spine Journal 19, 1772–1781 (2019). URL http://www.sciencedirect.com/science/article/pii/S1529943019308356.
[3] Das, A. & Rad, P. Opportunities and challenges in explainable artificial intelligence (xai): A survey. ArXiv abs/2006.11371 (2020).
[4] Ribeiro, M. T., Singh, S. & Guestrin, C. Anchors: High-precision model-agnostic explanations. In AAAI Conference on Artificial Intelligence (AAAI) (2018).
[5] Ribeiro, M. T., Singh, S. & Guestrin, C., "why should I trust you?": Explaining the predictions of any classifier. CoRR abs/1602.04938 (2016). URL http://arxiv.org/abs/1602.04938.
[6] Lundberg, S. & Lee, S. A unified approach to interpreting model predictions. CoRR abs/1705.07874 (2017). URL http://arxiv.org/abs/1705.07874.
[7] Lai, V., Cai, J. Z. & Tan, C. Many faces of feature importance: Comparing built-in and post-hoc feature importance in text classification. CoRR abs/1910.08534 (2019). URL http://arxiv.org/abs/1910.08534.
[8] Man, X. & Chan, E. P. The best way to select features? comparing mda, lime, and shap. The Journal of Financial Data Science (2020).
[9] Dieber, J. & Kirrane, S. Why model why? assessing the strengths and limitations of lime (2020). 2012.00093.
[10] Singh, D. Merck withdraws arthritis drug worldwide. BMJ (Clinical research ed.) 329, 816 (2004).
[11] Painful lessons. Nature Structural & Molecular Biology 42, 205–205 (2005). URL https://doi.org/10.1038/nbt.1043.
[12] Lai, E. C-C. et al. Sequence symmetry analysis in pharmacovigilance and pharmacoepidemiologic studies. European Journal of Epidemiology 32 (2017).
[13] Tsioropoulos, I., Andersen, M. & Hallas, J. Adverse events with use of antiepileptic drugs: A prescription and event symmetry analysis. Pharmacoepidemiology and drug safety 19, 483–91 (2009).
[14] Hallas, J. Evidence of depression provoked by cardiovascular medication: A prescription sequence symmetry analysis. Epidemiology 7, 478–484 (1996). URL https://doi.org/10.1097/00001648-199609000-00005.
[15] Curtis, J. et al. Adaptation of bayesian data mining algorithms to longitudinal claims data. Medical care 46, 969–75 (2008).
[16] Wahab, I., Pratt, N., Kalisch Elett, L. & Roughhead, E. Comparing time to adverse drug reaction signals in a spontaneous reporting database and a claims database: A case study of rofecoxib-induced myocardial infarction and rosiglitazone-induced heart failure signals in australia. Drug safety: an international journal of medical toxicology and drug experience 37 (2013).
[17] Wahab, I., Pratt, N., Wiese, M., Kalisch Elett, L. & Roughhead, E. The validity of sequence symmetry analysis (ssa) for adverse drug reaction signal detection. Pharmacoepidemiology and drug safety 22 (2013).
[18] Wahab, I. A., Pratt, N. L., Elett, L. K. & Roughhead, E. E. Sequence symmetry analysis as a signal detection tool for potential heart failure adverse events in an administrative claims database. Drug safety 39, 347–354 (2016). URL https://doi.org/10.1007/s40264-015-0391-8.
[19] Holman, D., Bass, A., Rouse, J. & Hobbs, M. Population-based linkage of health records in western australia: development of a health services research linked database. Australian & New Zealand Journal of Public Health 23, 453–459 (1999).
[20] Paige, E., Kemp-Casey, A., Korda, P. & Banks, E. Using Australian pharmaceutical benefits scheme data for pharmacoepidemiological research: Challenges and approaches. Public Health Research and Practice 25 (2015).
[21] Lemaitre, G., Nogueira, F. & Aridas, C. K. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. The Journal of Machine Learning Research 18, 559–563 (2017).
[22] Hasanin, T. & Khoshgoftaar, T. The effects of random undersampling with simulated class imbalance for big data. In 2018 IEEE International Conference on Information Reuse and Integration (IRI), 70–79 (IEEE, 2018).
[23] Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001). URL https://doi.org/10.1023/A:1010933404324.
[24] Geurts, P., Ernst, D. & Wehenkel, L. Extremely randomized trees. Mach. Learn. 63, 3–42 (2006). URL https://doi.org/10.1007/s10994-006-6226-1.
[25] Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. CoRR abs/1603.02754 (2016). URL http://arxiv.org/abs/1603.02754.
[26] Shawatze-Ziv, R. & Armon, A. Tabular data: Deep learning is not all you need. arXiv preprint arXiv:2106.03253 (2021).
[27] Snoek, J., Larochelle, H. & Adams, R. P. Practical bayesian optimization of machine learning algorithms. In Pereira, F., Burges, C. J. C., Bottou, L. & Weinberger, K. Q. Advances in Neural Information Processing Systems, vol. 25, 2951–2959 (Curran Associates, Inc., 2012). URL https://proceedings.neurips.cc/paper/2012/file/05311655a15b75fab98956636e1819c9d-Paper.pdf.
A FEATURE LOOKUP TABLES AND CODE DESCRIPTIONS

Table 4. Descriptions of the features ‘sex’, ‘age’, ‘rand float’, and ‘rand int’.

| Feature   | Description                                                                 |
|-----------|-----------------------------------------------------------------------------|
| sex       | The individual’s sex. The values ‘1’, and ‘2’ represent male and female. It was not possible to consider other values in this study, due to their low sample size. |
| age       | The individual’s age in years.                                             |
| rand float| A random floating point number in the range 0 — 1.                         |
| rand int  | A random integer in the range 0 — 157 inclusive (the number of features considered in the study). |

Table 5. Drug names of the ATC codes for the C class drugs C01AA05 — C09AA10. In this study, the features corresponding to these codes represent the number of dispensings that an individual had for these drugs in the drug history period.

| ATC code  | Drug name                        |
|-----------|----------------------------------|
| C01AA05   | digoxin                          |
| C01BC04   | flecaïnide                       |
| C01BD     | Antiarrhythmics, class III       |
| C01BD01   | amiodarone                       |
| C01DA02   | glyceryl trinitrate              |
| C01DA14   | isosorbid mononitrate            |
| C01DX16   | nicorandil                       |
| C02AB01   | β-methyldopa                     |
| C02AC01   | clonidine                        |
| C02CA01   | prazosin                         |
| C03AA01   | bendroflumethiazide              |
| C03AA03   | hydrochlorothiazide              |
| C03BA11   | indapamid                        |
| C03CA01   | furosemide                       |
| C03DA01   | spironolactone                   |
| C03DB01   | amiloride                        |
| C03EA01   | hydrochlorothiazide & potassium-sparing agents |
| C07AA03   | pindolol                         |
| C07AA05   | propranolol                      |
| C07AB02   | metoprolol                       |
| C07AB03   | atenolol                         |
| C07AB07   | bisoprolol                       |
| C07AG02   | carvedilol                       |
| C08CA01   | amiodipine                       |
| C08CA02   | felodipine                       |
| C08CA05   | nifedipine                       |
| C08CA13   | lercanidipine                    |
| C08DA01   | verapamil                        |
| C08DB01   | diltiazem                        |
| C09AA01   | captopril                        |
| C09AA02   | enalapril                        |
| C09AA03   | lisinopril                       |
| C09AA04   | perindopril                      |
| C09AA05   | ramipril                         |
| C09AA06   | quinapril                        |
| C09AA09   | fosinopril                       |
| C09AA10   | trandolapril                     |
Table 6. Drug names of the ATC codes for the C class drugs C09BA02 — C10BX03. In this study, the features corresponding to these codes represent the number of dispensings that an individual had for these drugs in the drug history period.

| ATC code  | Drug name                                      |
|-----------|------------------------------------------------|
| C09BA02   | enalapril and diuretics                        |
| C09BA04   | perindopril and diuretics                      |
| C09BA06   | quinapril and diuretics                        |
| C09BA09   | fosinopril and diuretics                       |
| C09CA02   | eprosartan                                     |
| C09CA04   | irbesartan                                     |
| C09CA06   | candesartan                                    |
| C09CA07   | telmisartan                                    |
| C09DA02   | eprosartan and diuretics                       |
| C09DA04   | irbesartan and diuretics                       |
| C09DA06   | candesartan and diuretics                      |
| C09DA07   | telmisartan and diuretics                      |
| C10AA01   | simvastatin                                    |
| C10AA03   | pravastatin                                    |
| C10AA04   | fluvastatin                                    |
| C10AA05   | atorvastatin                                   |
| C10AA07   | rosvastatin                                    |
| C10AB04   | gemfibrozil                                     |
| C10AB05   | fenofibrate                                     |
| C10AX09   | ezetimibe                                       |
| C10BA02   | simvastatin and ezetimibe                      |
| C10BX03   | atorvastatin andamlodipine                    |

Table 7. Drug names of the ATC codes for all the M class drugs considered in this study. In this study, the features corresponding to these codes represent the number of dispensings that an individual had for these drugs in the drug history period.

| ATC code  | Drug name                                      |
|-----------|------------------------------------------------|
| M01AB01   | indometacin                                    |
| M01AB05   | diclofenac                                     |
| M01AC01   | piroxicam                                      |
| M01AC06   | meloxicam                                      |
| M01AE01   | ibuprofen                                      |
| M01AE02   | naproxen                                       |
| M01AE03   | ketoprofen                                     |
| M01AH01   | celecoxib                                      |
| M01AH02   | rofecoxib                                      |
| M01AH06   | lumiracoxib                                    |
| M01CA     | Quinolines, antirheumatic drugs                |
| M03BX01   | baclefen                                       |
| M04AA01   | allopurinol                                    |
| M04AC01   | colchicine                                     |
| M05BA04   | alendronic acid                                 |
| M05BA07   | risedronic acid                                 |
| M05BB02   | risedronic acid and calcium, sequential         |
| M05BB03   | alendronic acid and colecalciferol             |
| M05BX     | Other drugs affecting bone structure and mineralization in ATC |

Table 8. Descriptions of the ICD-10-AM codes I20.0 — I50.9. In this study, the features corresponding to these codes represent the number of hospitalisations that an individual had due to these conditions in the comorbidity history period.

| ICD-10   | Description                                                                 |
|----------|-----------------------------------------------------------------------------|
| I20.0    | Unstable angina                                                             |
| I20.1    | Angina pectoris with documented spasm                                       |
| I20.8    | Other forms of angina pectoris                                              |
| I20.9    | Angina pectoris, unspecified                                                |
| I21.0    | ST elevation myocardial infarction of anterior wall                         |
| I21.1    | ST elevation myocardial infarction of inferior wall                         |
| I21.2    | ST elevation myocardial infarction of other sites                           |
| I21.3    | ST elevation myocardial infarction of unspecified site                      |
| I21.4    | Non-ST elevation myocardial infarction                                      |
| I21.9    | Acute myocardial infarction, unspecified                                    |
| I22.0    | Subsequent ST elevation and non-ST elevation myocardial infarction          |
| I22.1    | Subsequent ST elevation myocardial infarction of inferior wall              |
| I22.8    | Subsequent ST elevation myocardial infarction of other sites               |
| I22.9    | Subsequent ST elevation myocardial infarction of unspecified site           |
| I23.0    | Certain current complications following ST elevation and non-ST elevation myocardial infarction (within the 28 day period) |
| I23.1    | Atrial septal defect as current complication following acute myocardial infarction |
| I23.2    | Ventricular septal defect as current complication following acute myocardial infarction |
| I23.3    | Rupture of cardiac wall without hemopericardium as current complication following acute myocardial infarction |
| I23.5    | Rupture of papillary muscle as current complication following acute myocardial infarction |
| I23.6    | Thrombosis of atrium, auricular appendage, and ventricle as current complications following acute myocardial infarction |
| I23.8    | Other current complications following acute myocardial infarction          |
| I24.0    | Other acute ischemic heart diseases                                         |
| I24.1    | Dressler’s syndrome                                                          |
| I24.8    | Other forms of acute ischemic heart disease                                  |
| I24.9    | Acute ischemic heart disease, unspecified                                   |
| I25.0    | Chronic ischemic heart disease                                              |
| I25.1    | Atherosclerotic heart disease of native coronary artery                      |
| I25.2    | Old myocardial infarction                                                    |
| I25.3    | Aneurysm of heart                                                            |
| I25.4    | Coronary artery aneurysm and dissection                                      |
| I25.5    | Ischemic cardiomyopathy                                                     |
| I25.6    | Silent myocardial ischemia                                                   |
| I25.8    | Other forms of chronic ischemic heart disease                                |
| I25.9    | Chronic ischemic heart disease, unspecified                                 |
| I50.0    | Heart failure                                                                |
| I50.1    | Left ventricular failure, unspecified                                        |
| I50.9    | Heart failure, unspecified                                                   |
Table 9. Descriptions of the ICD-10-AM codes I60.0 — I64, K92.0 – K92.2, and R58. In this study, the features corresponding to these codes represent the number of hospitalisations that an individual had due to these conditions in the comorbidity history period.

| ICD-10  | Description                                                                 |
|---------|-----------------------------------------------------------------------------|
| I60.0   | Nontraumatic subarachnoid hemorrhage from unspecified carotid siphon and bifurcation |
| I60.1   | Nontraumatic subarachnoid hemorrhage from unspecified middle cerebral artery   |
| I60.2   | Nontraumatic subarachnoid hemorrhage from anterior communicating artery       |
| I60.3   | Nontraumatic subarachnoid hemorrhage from unspecified posterior communicating artery |
| I60.4   | Nontraumatic subarachnoid hemorrhage from basilar artery                      |
| I60.5   | Nontraumatic subarachnoid hemorrhage from unspecified vertebral artery        |
| I60.6   | Nontraumatic subarachnoid hemorrhage from other intracranial arteries          |
| I60.7   | Nontraumatic subarachnoid hemorrhage from unspecified intracranial artery     |
| I60.8   | Other nontraumatic subarachnoid hemorrhage                                   |
| I60.9   | Nontraumatic subarachnoid hemorrhage, unspecified                            |
| I61.0   | Nontraumatic intracerebral hemorrhage in hemisphere, subcortical              |
| I61.1   | Nontraumatic intracerebral hemorrhage in hemisphere, cortical                 |
| I61.2   | Nontraumatic intracerebral hemorrhage in hemisphere, unspecified             |
| I61.3   | Nontraumatic intracerebral hemorrhage in brain stem                           |
| I61.4   | Nontraumatic intracerebral hemorrhage in cerebellum                          |
| I61.5   | Nontraumatic intracerebral hemorrhage, intraventricular                      |
| I61.6   | Nontraumatic intracerebral hemorrhage, multiple localized                    |
| I61.8   | Other nontraumatic intracerebral hemorrhage                                  |
| I61.9   | Nontraumatic intracerebral hemorrhage, unspecified                           |
| I62.0   | Nontraumatic subdural hemorrhage, unspecified                                |
| I62.1   | Nontraumatic extradural hemorrhage                                            |
| I62.9   | Nontraumatic intracranial hemorrhage, unspecified                            |
| I63.0   | Cerebral infarction due to thrombosis of unspecified precerebral artery      |
| I63.1   | Cerebral infarction due to embolism of unspecified precerebral artery        |
| I63.2   | Cerebral infarction due to unspecified occlusion or stenosis of unspecified precerebral arteries |
| I63.3   | Cerebral infarction due to thrombosis of unspecified cerebral artery          |
| I63.4   | Cerebral infarction due to embolism of unspecified cerebral artery           |
| I63.5   | Cerebral infarction due to unspecified occlusion or stenosis of unspecified cerebral artery |
| I63.6   | Cerebral infarction due to cerebral venous thrombosis, nonpyogenic           |
| I63.8   | Other cerebral infarction                                                    |
| I63.9   | Cerebral infarction, unspecified                                             |
| I64     | Stroke, not specified as haemorrhage or infarction                           |
| K92.0   | Hematemesis                                                                  |
| K92.1   | Melena                                                                       |
| K92.2   | Gastrointestinal hemorrhage, unspecified                                     |
| R58     | Hemorrhage, not elsewhere classified                                         |

B HYPERPARAMETER SEARCHES

Table 10. The range of values that the hyperparameters could take during our RF and ET hyperparameter optimisation.

| Hyperparameter       | Search distribution |
|----------------------|---------------------|
| n_estimators         | 8 — 32              |
| max_features         | ’auto’, ’sqrt’       |
| max_depths           | 8 — 64, no limit    |
| min_samples_split    | 2 — 12              |
| min_samples_leaf     | 2 — 8               |
| bootstrap            | True, False         |

Table 11. The range of values that the hyperparameters could take during our XGB hyperparameter optimisation.

| Hyperparameter        | Search distribution            |
|-----------------------|-------------------------------|
| booster               | ‘gbtree’, ’gblinear’           |
| max_depth             | 2 — 32, no limit              |
| sampling_method       | ’uniform’, ’gradient_based’   |
| alpha                 | 0, 0.1, 0.5                   |
| lambda                | 0.8, 1, 1.2                   |
| grow_policy           | ’depthwise’, ’lossguide’       |