Scientific Reports

https://doi.org/10.1038/s41598-026-40267-2

Article in Press

A kNN based machine learning approach to automating causality assessment of adverse events

Received: 13 May 2025

Accepted: 11 February 2026

Published online: 15 February 2026

Cite this article as: Ren J., Carroll H., McCarthy K. *et al.* A kNN based machine learning approach to automating causality assessment of adverse events. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-40267-2

Jun Ren, Hua Carroll, Kerry McCarthy, Jeff Allen, Jeff Tam, Jeffrey Philip, Monica Mehta & Douglas Clark

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

A kNN Based Machine Learning Approach to Automating Causality Assessment of Adverse Events

Jun Ren[1,*], Hua Carroll[2], Kerry McCarthy[2], Jeff Allen[1], Jeff Tam[1], Jeffrey Philip[1], Monica Mehta[2] and Douglas Clark[1]

[1]Technology, Analytics & Data Insights (TADI), Biogen, Cambridge, MA 02142, USA
[2]Global PV, Reg Submissions Mgt, PV/Reg Quality, Biogen, Cambridge, MA 02142, USA
[*]**Corresponding author:** Jun Ren(jun.ren@biogen.com)

Abstract

This paper introduces a machine learning classifier designed to automate causality assessment in Individual Case Safety Reports (ICSRs), utilizing the principle of event similarity. The classifier's effectiveness was evaluated using adverse events from six marketed products. Furthermore, we incorporated an augmentation tool to efficiently manage the classification of 'unassessable' adverse events. To enhance medical review oversight, we developed a web-based application serving as a reliable decision-support tool for medical reviewers. The outcomes obtained from our model were highly encouraging, emphasizing the potential advantages of utilizing such a model in ICSR causality assessment.

Keywords: Pharmacovigilance, ICSR, Causality Assessment, Machine Learning, Similarity. ■

Statements and Declarations

Funding This research was funded by Biogen.

Conflict of Interest All authors are employees of Biogen.

Ethics Approval Not applicable. The research only utilized post-marketing AE data.

Consent to Participate Not applicable.

Consent for Publication Not applicable.

Code Availability Not applicable.

Availability of Data and Material The data is sensitive, proprietary, post-marketing adverse event data and is not available for sharing. However, inquiries regarding the data or the study can be directed to the corresponding author Jun Ren at jun.ren@biogen.com.

Author Contributions JR and HC led the model development effort and model augmentation activities. All authors contributed to the manuscript in accordance with the guidance. All authors read and approved the final version.

Competing Interest There are no Competing Interests.

1 Introduction

In the field of pharmacovigilance (PV), Individual Case Safety Reports (ICSRs) remain an important data source for monitoring the safety of pharmaceutical products after their market launch. These reports offer details about adverse events (AEs) experienced by individuals subsequent to the usage of a specific pharmaceutical product and are essential to ensure that approved products continue to be safe and effective when used in broader, uncontrolled, real-world settings. Medical evaluations of ICSRs, including causality assessment, form a critical component of the PV process, contributing to the determination of clinical relevance and identification of new safety concerns. ICSR causality assessment seeks to discern a potential causal relationship between the drug and the adverse event. This assessment aids in determining whether the event is a consequence of the usage of the medicinal product, or if it can be ascribed to other factors such as the patient's pre-existing medical conditions, concomitant usage of other drugs, or lifestyle factors.

Despite its significance, the task of causality assessment often proves to be knowledge- and labor-intensive due to the intricate and diverse factors that could potentially contribute to the occurrence of adverse events. This necessitates the involvement of trained medical professionals with specialized expertise. This complexity has spurred research into automated methods, such as Machine Learning (ML), to aid in causality adjudication. These innovative methods hold the potential to enhance efficiency, accuracy, and consistency of causality assessment.

The application of ML techniques in PV has seen considerable advancements in recent years. For instance, Zitu et al. [1] have illustrated how ML can be utilized to predict adverse drug reactions using electronic health records. Recent studies [2, 3, 4] demonstrated the efficacy of ML and Large Language Model (LLM) in assessing causality in ICSRs, providing evidence that AI/ML could significantly improve the precision and efficiency of causality assessments. Furthermore, a study conducted by Yan et al. [5] found that advanced ML techniques, such as deep learning, could be instrumental in identifying potential drug-drug interactions in ICSRs, further underlining the promising potential of ML in this field.

In this paper, we present a novel k-Nearest Neighbors (kNN)-based classification model designed to automate the causality assessment of adverse events. We employ the concept of event similarity or feature-space similarity between events, linking all adverse events when situated in a shared multivariate feature space. The degree of similarity between any two events is determined by calculating their spatial distance in relation to the selected event attributes. The expansive volume of adverse event reports allows for the identification of similar or closely related historical adverse events through the kNN algorithm. Given that similar events typically share the same causality relationship, our model classifies a new adverse event based on the causality classifications of its nearest events. This method allows us to leverage past assessments, significantly reducing assessment time and facilitating swift and consistent causality adjudication.

To further improve overall model performance, we introduce an augmentation tool specifically designed to effectively address the classification of 'unassessable' adverse events. This addition ensures appropriate handling of the unfavorable 'unassessable' assessment, improving the robustness of the classification process.

To enhance medical review oversight, we develop a web-based application that connects the ML model with the safety database, serving as a reliable decision-support tool for medical reviewers. This application enables the internal medical review team to efficiently identify specific case examples and emerging trends, thereby improving consistency and enhancing the overall quality of decision-making in medical review.

The remainder of this paper is structured as follows: Section 2 describes the data set used for model training and testing. Section 3 outlines the machine learning methodology, including the definition of event similarity and a concise overview of the kNN algorithm, detailing the classification process for adverse events. Section 4 presents empirical findings from the application of the model to marketed products, as well as the model augmentation efforts. Section 5 highlights the implementation of the machine learning model to support enhanced medical review oversight. Finally, Sections 6 and 7 consolidate the study's findings, providing a comprehensive discussion and summarizing the conclusions drawn.

2 Data set

Our experiment focused on post-marketing solicited reports from all reporter types (e.g., consumers, healthcare professionals, etc.) recorded in Biogen's safety database, encompassing both serious and non-serious adverse events. For each report, only the latest approved version was included. For each adverse event, company-assessed causality was categorized as related, probably related, possibly related, unlikely related, not related, or unassessable. Post-marketing spontaneous reports were excluded due to the implied causality required for regulatory reporting purposes. Clinical trial ICSRs were also omitted due to their lower volume and binary causality assessment (related or not related), with a predominant majority classified as 'not related', making the data set unsuitable for the applied machine learning method.

In total, there were more than 800K solicited adverse events from six marketed drugs across two therapeutic areas collected over a span of six years from 2017 to 2022. For each drug, the adverse events were randomly divided into two distinct data sets: a training data set and a testing data set. The classifier was trained using approximately 550K adverse events from the training data sets and subsequently evaluated on the six testing data sets, which collectively comprised over 250K adverse events. All AEs in both the training and testing data sets represent distinct drug-AE pairs across the six drugs. In the training sets, the number of AEs per drug ranges from 17.8K (lowest) to 197.3K (highest), with a similar proportional distribution observed in the testing data sets.

3 ML method

3.1 Event similarity

In the field of machine learning, various classification methods exist to predict the category or class of a given input, including but not limited to Logistic Regression, Decision Trees, Support Vector Machines (SVM), and kNN. For our study, we selected the kNN method due to its simplicity in both implementation and interpretability. kNN is an instance-based learning algorithm that assigns the class of a sample based on the majority class among its k nearest neighbors. To assess the proximity between adverse events, we have introduced a measure of event similarity, which quantifies the closeness between two AEs.

Similarity serves as a fundamental building block for numerous artificial intelligence services such as recommendation systems, clustering, classification, and anomaly detection. For instance, in k-means clustering, data points are allocated to clusters through the calculation and comparison of similarities or distances to each cluster's center. Similarity is essentially a measure of the degree to which two or more objects resemble each other. It can be quantified by the distance in a feature space, where a smaller distance signifies a higher degree of similarity, and conversely, a larger distance indicates a lower degree of similarity.

To compute the degree of similarity, two key elements are necessary - a feature space and a distance metric. For instance, two songs may be deemed similar based on factors such as genre, artist, duration, or production date; likewise, two fruits might be considered similar due to their color, size, or taste. In these cases, the features (genre, artist, duration, date of production for songs; color, size, taste for fruits) constitute the feature space that describes the songs or fruits respectively. Once the feature space is given, a distance metric is required to quantify the distance between two objects within this multivariate feature space. The specifics of applying this concept to adverse events will be elaborated upon in the following subsections.

3.1.1 Feature space

Among all the attributes of an adverse event, we have selected the MedDRA *Preferred Term (PT)*, *dechallenge & rechallenge*, *expectedness*(per company core data sheet), *reporter causality*, *concomitant medications*, *medical history*, and *event seriousness* to constitute the feature space. This selection was made upon consultation with medical reviewers and considering the assessment process and company products. For the purposes of enhancing data quality and simplifying the classification model, we have chosen not to consider other attributes such as patient demographic information, AE onset/stop dates, and text-based event descriptions.

To digitally encapsulate an adverse event, we conduct feature engineering by transforming each event attribute into a vector, represented as

$$X = [x_1, x_2, x_3, \cdots, x_m].$$

Here, each x_i denotes the value of a unique attribute, while $m = 7$ indicates the dimension of the feature space that comprises the seven aforementioned AE features. As a result, the similarity or distance between any two adverse events can be depicted by the distance between their respective vectors within the 7-dimensional space.

3.1.2 Distance metric

A variety of distance metrics are available to measure the proximity between two vectors [9, 10]. Let's assume vectors X and Y, which represent two distinct AEs, can be expressed as

$$\begin{aligned} X &= [x_1, x_2, x_3, \cdots, x_m], \\ Y &= [y_1, y_2, y_3, \cdots, y_m]. \end{aligned} \tag{1}$$

Figure 1 illustrates four distance metrics in a 2-dimensional space (the points labeled A, B, C, D, E represent five different vectors or AEs). The choice of metric function can vary, depending on the nature of the data set and the problem to be solved. For data sets with specific properties, such as high dimensionality or correlated features, metrics like cosine similarity or Mahalanobis distance may be more suitable. On the other hand, metrics like Euclidean distance or Manhattan distance tend to perform better with low-dimensional data.

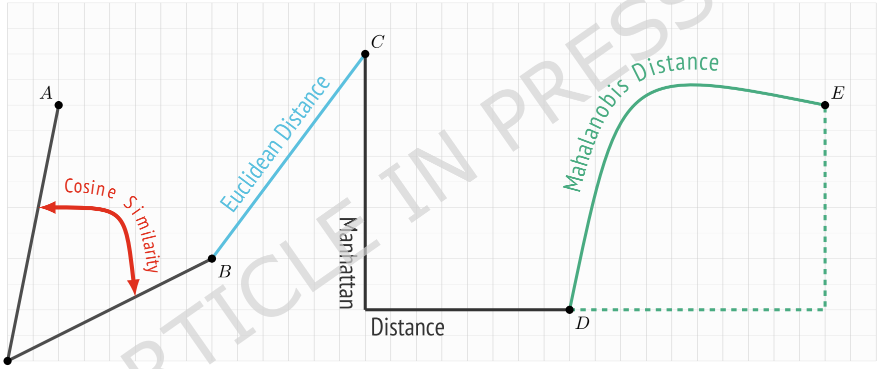


Figure 1: Example distance metrics in 2-dimensional space.

In our study, we opted for the Euclidean distance, a widely utilized distance function in many applications, to measure the distance between any two adverse events. We selected Euclidean distance because, after encoding and scaling, it measures straight line similarity across all features, giving equal opportunity for each to contribute to the similarity calculation. We also evaluated the Manhattan metric and found that Euclidean distance yielded similar or slightly better performance.

The Euclidean distance is defined as follows:

$$dist(X, Y) = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2}, \tag{2}$$

or

$$dist(X, Y) = \sqrt{\sum_{i=1}^{m} dist_i(x_i, y_i)^2}, \tag{3}$$

where $dist_i(\cdot, \cdot)$ represents the component distance between two AEs/vectors concerning the i-th attribute. The definition of these component distances will be discussed subsequently.

3.1.3 Component distance

The component distances $dist_i(\cdot, \cdot)$ are fundamental to the distance metric we have established above. In Appendix, we have explored these distances in greater detail for the seven event attributes we have selected: *event PT*, *dechallenge & rechallenge*, *expectedness*, *reporter causality*, *concomitant medications*, *medical history*, and *event seriousness*. For each event attribute, our methodology begins with feature engineering to transform its values into either categorical or numerical representation, followed by a thorough definition of the corresponding component distance. The resulting component distances are denoted as $disc_{pt}$, $dist_{dr}$, $dist_{ex}$, $dist_{rc}$, $dist_{cm}$, $dist_{mh}$, and $dist_{es}$.

3.1.4 Calculation of event similarity

We are now furnished with all the necessary tools to compute the event similarity between any two adverse events. The Euclidean distance (described in Equation 3) between any two adverse events X and Y can be revised as follows:

$$dist(X, Y) = \sqrt{\sum_f dist_f^2}\ , \tag{4}$$

where f spans all the seven features *pt, dr, ex, rc, cm, mh* and *es*.

Furthermore, we considered the varying importance of event attributes. Weights were assigned based on expert judgement from experienced medical reviewers, reflecting the relative importance of each attribute in causality assessment. Attributes with well established primary influence in regulatory and clinical practice (*PT*, *dechallenge & rechallenge*, and *expectedness*) were assigned higher weights, while other attributes with more supportive roles were assigned lower weights. This expert driven approach was chosen over formal MCDM methods such as AHP [6] to maintain interpretability and simplicity, and preliminary sensitivity analysis confirmed that results were robust to moderate weight variations. Specifically, we established the following weighting scheme:

$$\omega_f = \begin{cases} \frac{1}{5} & \text{if } f \text{ is } pt \text{ or } dr \text{ or } ex, \\ \frac{1}{10} & \text{if } f \text{ is } rc \text{ or } cm \text{ or } mh \text{ or } es. \end{cases} \tag{5}$$

For simplicity, the weighted Euclidean distance maintains its original notation and is defined as follows:

$$dist(X, Y) = \sqrt{\sum_f \omega_f * dist_f^2}. \tag{6}$$

Please note that the sum of all the weights amounts to 1, i.e. $\sum_f \omega_f = 1$.

The event similarity exhibits a negative correlation with the event distance. In essence, a smaller distance indicates a higher degree of similarity, while a larger distance suggests a lower degree of similarity.

3.2 k-Nearest Neighbors

The k-Nearest Neighbors (kNN) algorithm [8, 9] is among the most straightforward machine learning techniques, operating on a basic principle: for any given object, the algorithm identifies the 'k' most similar objects from a historical data set, ranked by similarity. The algorithm can be harnessed to make predictions or classifications by leveraging the characteristics of the identified nearest neighbors.

In this section, we will first use the kNN technique to identify the most similar adverse events for a given adverse event, based on the predefined event similarity metric. Following this, we will illustrate the process of classifying causality.

3.2.1 Identification of the most similar AEs

To identify the most similar adverse events for a target adverse event, it is essential to have access to a historical data set. For any particular drug, we assemble a collection of AEs, designated as $AE_1, AE_2, \cdots, AE_N$. Each of these AEs has undergone causality assessment by PV professionals. After applying the previously described feature selection and feature engineering processes on these AEs, we generate N vectors, represented as $X_1, X_2, \cdots, X_N$. Of note, each of these vectors exists within a 7-dimensional space.

$$AE_1, AE_2, \cdots, AE_N \xrightarrow[\text{feature engineering}]{\text{feature selection \&}} X_1, X_2, \cdots, X_N$$

In addition, we retain the company (vendor) causality findings, symbolized as $cau_1, cau_2, \cdots, cau_N$, which will later be employed for the causality classification of the target AE. We implement the same procedures on the target AE, resulting in the generation of a vector, denoted as T.

We then measure the distance between the target AE and the historical AEs. Specifically, we employ the distance function (Equation 6), applying it to both T and every X_i, where $1 \leqslant i \leqslant N$. This process yields N distances, denoted as $d_1, d_2, \cdots, d_N$, where

$$d_i = dist(T, X_i),\ 1 \leqslant i \leqslant N. \tag{7}$$

Following that, we arrange the historical AEs in order of their similarity to the target AE. The k AEs that demonstrate the greatest similarity or the smallest distances are selected. These k AEs then become the k nearest neighbors of the target adverse event.

3.2.2 Classifying causality

To classify the causality relationship of the target AE, we first determined the number and types of classifications. Six causality terms appear in our AEs: ‘related’, ‘probable’, ‘possible’, ‘not related’, ‘unlikely’, and ‘unassessable’. An AE can be categorized as ‘unassessable’ when a report suggests an adverse reaction that cannot be judged because information is insufficient or contradictory, and that cannot be supplemented or verified at the time of assessment [7].

We constructed a 3-class classification model, with each class defined as follows:

$$\begin{aligned} &\text{Class 1}: &&\text{‘related’ or ‘probable’ or ‘possible’},\\ &\text{Class 2}: &&\text{‘not related’ or ‘unlikely’},\\ &\text{Class 3}: &&\text{‘unassessable’}. \end{aligned} \tag{8}$$

In an ideal situation, we expect the classification model to be able to distinguish all six causality classifications. However, as the number of classes in the model increases, the performance of the model may decrease. Therefore, we have opted for a 3-class model to strike a balance between model performance and business requirements.

With the three classes clearly defined, we now examine the causality outcomes resulting from the identified k most similar AEs. For ease of reference, we label the k causality outcomes as $cau_1, cau_2, \cdots, cau_k$. For each cau_i, we determine its associated class based on the definitions of the three classes, and subsequently cast a vote for that class. In the end, the class receiving the most votes is chosen as the causality class for the target AE. For instance, if we select k to be 10, among the 10 most similar AEs, Class 1, Class 2, and Class 3 have 2, 3, and 5 AEs, respectively. In this scenario, the algorithm will categorize the target AE under Class 3. The procedure for this determination is depicted in Algorithm 1.

4 ML model performance

4.1 Mismatch rate/recall, precision, and F1 score

The mismatch rate denotes the ratio of adverse events predicted differently by a classifier within a specific class, thereby serving as an indicator of the accuracy of the classification model. It measures the proportion of misclassified positive instances out of all actual positive instances, mathematically represented as

$$\text{Mismatch rate} = \frac{\text{False Negatives}}{\text{True Positives} + \text{False Negatives}}. \tag{9}$$

Algorithm 1: Pseudocode for causality classification

```
Given a target $AE$ of interest
for every $AE_i$ in the collected historical data set do
    $d_i \leftarrow$ compute the distance between $AE$ and $AE_i$
end
Identify the $k$ AEs with the smallest distances
Class 1 $\leftarrow$ [related, probable, possible]
Class 2 $\leftarrow$ [not related, unlikely]
Class 3 $\leftarrow$ [unassessable]
for every $AE_i$ in the identified $k$ AEs do
    if causality($AE_i$) is in Class $j$ then
        count(Class j) $\leftarrow$ count(Class j) + 1
    end
end
return either Class 1, Class 2 or Class 3 whichever is the majority
```

Figure 2 illustrates the mismatch rates for the six drugs under examination. Notably, the classifier exhibits strong performance when predicting adverse events of Class 1 and Class 3 across all six drugs. However, it demonstrates a higher error rate when dealing with Class 2 adverse events. A likely reason is that,'not related' or 'unlikely' AEs (corresponding to Class 2) are less frequent than the other two classes in our data sets. ML models tend to learn patterns for the more common classes more effectively, while underperforming on underrepresented ones. In our testing data sets, Class 2 accounts for only 13.4% of all AEs (referred to Tabel 1). To address this imbalance, potential strategies include oversampling the 'not related/unlikely' class, undersampling the majority classes, or adjusting decision thresholds to be more conservative for the majority classes (as discussed in Section 4.3).

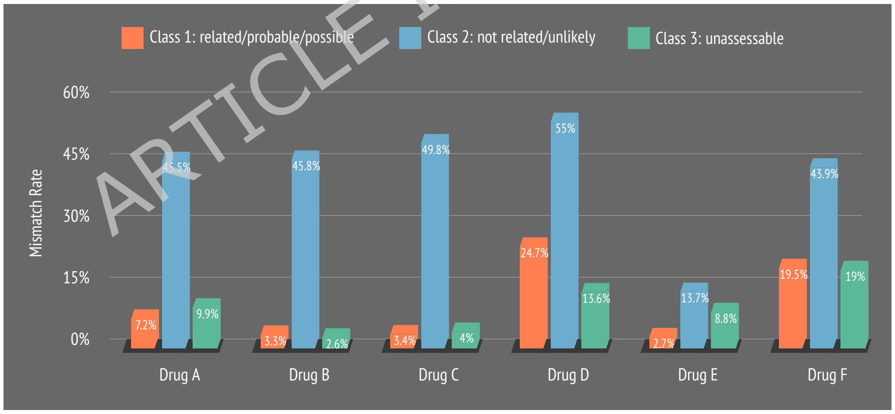


Figure 2: Mismatch rates of the model for the six drugs.

To evaluate the overall model performance across all six drugs, we utilized the confusion matrix presented in Table 1. The term 'ML i' indicates that the causality was classified as Class i by the ML classifier. The testing data set includes 88,722, 34,934, and 136,185 adverse events, which were labeled as Class 1, Class 2, and Class 3, respectively, by medical reviewers. The distribution of false positives and false negatives for each class is summarized in Table 2.

Class	ML 1	ML 2	ML 3	Total
Class 1	82,515	254	5,953	88,722
Class 2	1,070	20,625	13,239	34,934
Class 3	6,286	5,449	124,450	136,185
Total	89,871	26,328	143,642	259,841

Table 1: The confusion matrix.

Class	False Positives	False Negatives
Class 1	7,356	6,207
Class 2	5,703	14,309
Class 3	19,192	11,735

Table 2: False positives and false negatives.

The mismatch rates for these three classes are 7.0%, 41.0%, and 8.6%, respectively, referred in Table 3. The elevated mismatch rate for Class 2 is consistent with the patterns observed in the per-drug breakdown discussed earlier.

Precision is defined as the proportion of true positive predictions relative to the total number of positive predictions made by the model, mathematically represented as

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}. \tag{10}$$

Essentially, this metric gauges the model's proficiency in correctly identifying positive instances from all instances it classifies as positive. A model demonstrating high precision indicates a strong capability of minimizing false positive errors. The precision values for these three classes are 0.918, 0.783, and 0.866, respectively, referred in Table 3.

Class	Mismatch rate	Precision	Recall	F1 score
Class 1	7.0%	0.918	0.93	0.924
Class 2	41.0%	0.783	0.59	0.673
Class 3	8.6%	0.866	0.914	0.889
Macro-average F1				0.829
Weighted-average F1				0.872

Table 3: Mismatch rate, precision, recall, and F1 score.

Recall (also referred to as sensitivity or the true positive rate) is the complement of the mismatch rate and quantifies the proportion of actual positive instances that are correctly identified by the model. It is defined as:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}. \tag{11}$$

A high recall value reflects strong detection capability, with few false negatives (low mismatch rate), whereas a low recall value indicates that a substantial share of true positives remain undetected. The recall values for the three classes are 0.93, 0.59, and 0.914, respectively (see Table 3).

The F1 score combines precision and recall into a single metric by taking their harmonic mean:

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{12}$$

A high F1 score indicates that the model is achieving both high precision and high recall, whereas a low F1 score suggests weaknesses in at least one of these dimensions. In the present results, the disparity between precision and recall is most evident for Class 2, which exhibits a relatively high precision (0.783) but substantially lower recall (0.59), resulting in a reduced F1 score 0.673 (see Table 3). This pattern reflects the model's tendency to be conservative when predicting Class 2 events (producing fewer false positives), but at the cost of missing a substantial number of actual positives.

The two average F1 scores computed across classes, the macro-average F1 score and weighted-average F1 score (as defined in Equation 13), are 0.829 and 0.872, respectively. The lower macro-average reflects the weaker performance on Class 2, which receives equal consideration despite its smaller representation in the dataset. By contrast, the weighted-average is higher because the stronger performance on the more prevalent Classes 1 and 3 exerts a greater influence on the overall score.

$$\begin{gathered}\text{Macro-average F1} = \frac{1}{3}\sum_{i=1}^{3} \text{F1}_i, \quad \text{where } \text{F1}_i \text{ is the F1 score of Class } i. \\ \text{Weighted-average F1} = \frac{\sum_{i=1}^{3} n_i \cdot \text{F1}_i}{\sum_{i=1}^{3} n_i}, \quad \text{where } n_i \text{ is the number of AEs in Class } i.\end{gathered} \tag{13}$$

4.2 Proportions of causality classes

The distribution of the proportions of causality classes can offer valuable insights into the frequency of various types of causal relationships between a drug and adverse events. The proportions may influence the course of action by medical reviewers, such as prompting further investigations. Additionally, if a machine learning model underrepresents a particular causality class, it may become necessary to gather more data for that class to enhance the model's learning capability.

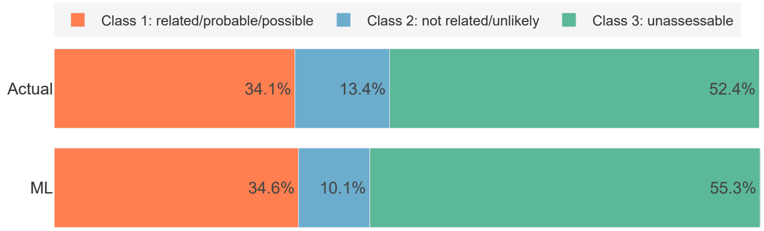


Figure 3: Proportions of causality classes.

Figure 3 illustrates a comparison between the two distributions, one derived from medical reviewers and the other from the machine learning classifier, across all six drugs. It is observed that the proportions of all three causality classes align closely between the actual assessment results and the outcomes generated by the machine learning model. This observation suggests that the model assigns causality results in a manner that is remarkably consistent with the medical reviewers.

4.3 Model augmentation

One of the limitations of post-marketing solicited reports is incomplete and/or ambiguous information, which renders causality assessment for some AEs impossible. Nevertheless, over utilization of 'unassessable' as causality assessment hinders safety data evaluation at individual case level and aggregate level. The ML model has proven its proficiency in predicting ICSR causality assessment. Can this model be augmented to help decrease the reliance on 'unassessable' causality assessment? To answer this question, we tested the implementation of a threshold for predicting Class 3 ('unassessable') adverse events.

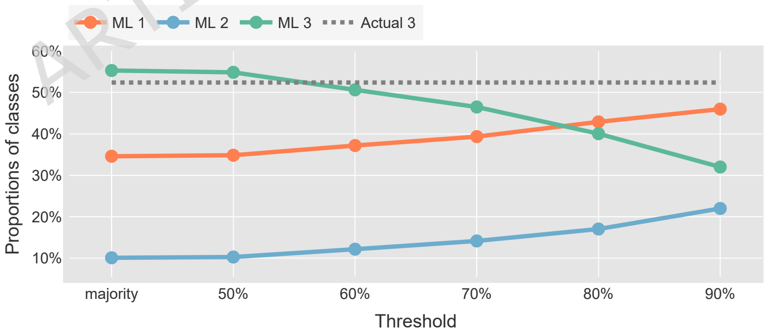


Figure 4: Proportions of causality classes against thresholds.

In the present setup of the kNN algorithm (referred in Algorithm 1), class determination is simply made based on its most frequent occurrence among the nearest neighbors. Applying this rule to Class 1 and Class 2 AEs is suitable, considering there is no compelling reason to dismiss the prediction in these two scenarios. However, a more stringent rule can be adopted when categorizing Class 3 AEs, given that 'unassessable' often serves as a surrogate category for various uncertainties. Consequently, we suggest incorporating a threshold percentage for Class 3, rather than solely relying on frequency of occurrence (i.e., simple majority). Specifically, an adverse event is predicted as Class 3 only if the occurrence of Class 3 AEs surpasses a pre-set threshold among its nearest neighbors. If this criterion is not met, the predicted class defaults to either Class 1 or Class 2, depending on which is more prevalent among the neighbors. The higher the threshold, the more stringent the criteria for an adverse event to be classified as Class 3. This targeted augmentation not only reduces the over-representation of ‘unassessable’ cases, but also addresses class imbalance effects by making minority class assignment more deliberate. While this approach improves robustness in the current proof-of-concept, future work will investigate advanced resampling and cost-sensitive learning strategies (e.g., SMOTE, Tomek links) to further enhance minority-class performance.

The augmentation tool was implemented as a post-processing module within the kNN classifier workflow. For a given threshold, the tool calculates the overall proportion of events classified as Class 3. If this proportion exceeds a level deemed too high by medical reviewers for a specific product, the threshold can be adjusted interactively in the web application, and the classification process is re-run to observe the changes in predicted class distribution and performance metrics. The threshold is a user-configurable parameter in the web application and can be set globally or per-product, allowing reviewers to calibrate the stringency of ‘unassessable’ predictions in accordance with product-specific AE profiles and review priorities.

In our study, we tested a variety of thresholds (50%, 60%, 70%, 80%, 90%) to assess their impact on the model performance across all six drugs. It was observed in Figure 4 that the percentage of the AEs predicted as Class 3 (denoted as ML 3) declined from 54.9% to 32.0% as the threshold increases from 50% to 90%. Consequently, the percentages of AEs categorized as ML 1 and ML 2 both increased, as a larger number of AEs were reclassified into these two categories. Furthermore, the proportion of ML 3 AEs started to fall below the proportion of actual Class 3 AEs when the threshold reached 60%. Note that ‘majority’ (i.e., simple majority) was included in the figure for comparison purposes.

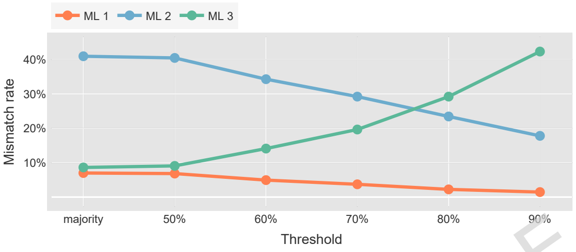


Figure 5: Mismatch rate against thresholds.

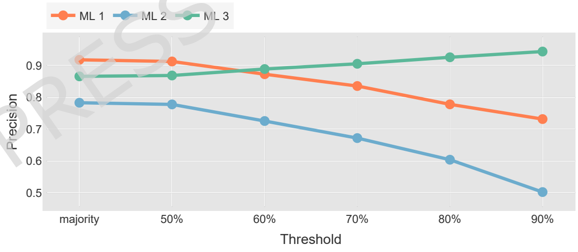


Figure 6: Precision against thresholds.

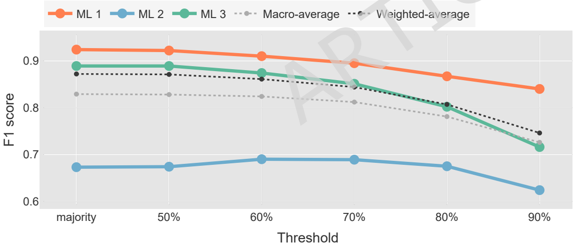


Figure 7: F1 score against thresholds.

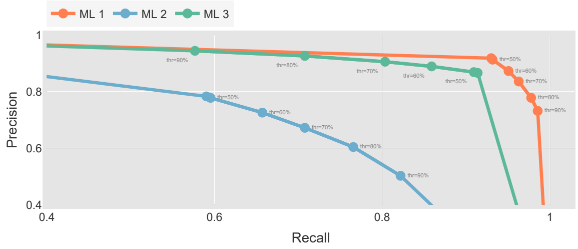


Figure 8: Precision-Recall curve.

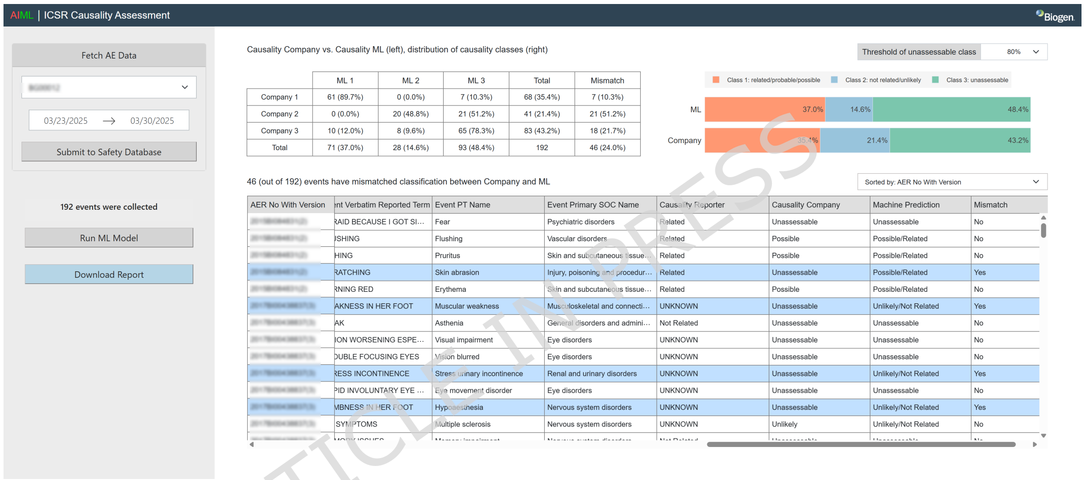
We also observed a notable improvement in mismatch rates, with Class 1 decreasing from 6.8% to 1.5%, and Class 2 experiencing a significant reduction from 41.5% to 17.8%, as illustrated in Figure 5. Naturally, this improvement occurred at the expense of a higher mismatch rate for Class 3, which rose from 9.1% to 42.4%. In term of precision, values for Class 1 and Class 2 both fell to 0.731 and 0.502 respectively, as shown in Figure 6, while the precision value for Class 3 saw an increase to 0.944.

The F1 score trends largely reflect the interplay between these precision and recall shifts. As shown in Figure 7, Class 1’s F1 score decreased gradually from 0.924 to 0.84, driven primarily by the drop in precision despite relatively stable recall. Class 3 also saw a decline in F1 score, from 0.889 to 0.716, due to a sharp reduction in recall that outweighed the gains in precision. By contrast, Class 2 showed an initial improvement in F1 score, peaking at 0.69 when the threshold reached 60-70%, indicating a more balanced precision–recall trade-off at this point. However, further increases in threshold led to a decline to 0.624, suggesting that the loss in precision at higher thresholds could no longer be offset by gains in recall. Overall, both the macro-average and weighted-average F1 scores showed a consistent downward trend as the threshold increases.

The Precision–Recall curves in Figure 8 show clear trade-offs as the threshold increases. For Class 1, precision fell from 0.918 to 0.732 while recall rose slightly from 0.93 to 0.985, indicating modest recall gains at the cost of precision. Class 2 showed a sharper precision drop (0.783 to 0.502) alongside consistent recall gains (0.590 to 0.822), whereas Class 3 displayed the reverse pattern, with precision increasing (0.866 to 0.944) but recall declining sharply (0.914 to 0.578). These patterns underscore the class-specific nature of the precision–recall trade-off.

5 ML model for medical review oversight

The machine learning model provides a systematic and objective approach to causality determination, effectively minimizing variability and mitigating potential human bias inherent in manual assessments. As a result, the model can serve as a robust decision-support tool for medical reviewers and pharmacovigilance professionals, enabling enhanced medical review oversight and training of staff and vendors.



AIML | ICSR Causality Assessment

[redacted]

03/23/2025 → 03/30/2025

192 events were collected

Causality Company vs. Causality ML (left), distribution of causality classes (right)

Threshold of unassessable class: 80%

	ML 1	ML 2	ML 3	Total	Mismatch
Company 1	61 (89.7%)	0 (0.0%)	7 (10.3%)	68 (35.4%)	7 (10.3%)
Company 2	0 (0.0%)	20 (48.8%)	21 (51.2%)	41 (21.4%)	21 (51.2%)
Company 3	10 (12.0%)	8 (9.6%)	65 (78.3%)	83 (43.2%)	18 (21.7%)
Total	71 (37.0%)	28 (14.6%)	93 (48.4%)	192	46 (24.0%)

Class 1: related/probable/possible
Class 2: not related/unlikely
Class 3: unassessable
ML
37.0%
14.6%
48.4%
Company
35.4%
21.4%
43.2%

46 (out of 192) events have mismatched classification between Company and ML

Sorted by: AER No With Version

AER No With Version	ent Verbatim Reported Term	Event PT Name	Event Primary SOC Name	Causality Reporter	Causality Company	Machine Prediction	Mismatch
[redacted]	RAID BECAUSE I GOT SI…	Fear	Psychiatric disorders	Related	Unassessable	Unassessable	No
[redacted]	JSHING	Flushing	Vascular disorders	Related	Possible	Possible/Related	No
[redacted]	HING	Pruritus	Skin and subcutaneous tissue…	Related	Possible	Possible/Related	No
[redacted]	RATCHING	Skin abrasion	Injury, poisoning and procedur…	Related	Unassessable	Possible/Related	Yes
[redacted]	RNING RED	Erythema	Skin and subcutaneous tissue…	Related	Possible	Possible/Related	No
[redacted]	AKNESS IN HER FOOT	Muscular weakness	Musculoskeletal and connecti…	UNKNOWN	Unassessable	Unlikely/Not Related	Yes
[redacted]	AK	Asthenia	General disorders and admini…	Not Related	Unassessable	Unassessable	No
[redacted]	ION WORSENING ESPE…	Visual impairment	Eye disorders	UNKNOWN	Unassessable	Unassessable	No
[redacted]	OUBLE FOCUSING EYES	Vision blurred	Eye disorders	UNKNOWN	Unassessable	Unassessable	No
[redacted]	RESS INCONTINENCE	Stress urinary incontinence	Renal and urinary disorders	UNKNOWN	Unassessable	Unlikely/Not Related	Yes
[redacted]	PID INVOLUNTARY EYE …	Eye movement disorder	Eye disorders	UNKNOWN	Unassessable	Unassessable	No
[redacted]	MBNESS IN HER FOOT	Hypoaesthesia	Nervous system disorders	UNKNOWN	Unassessable	Unlikely/Not Related	Yes
[redacted]	SYMPTOMS	Multiple sclerosis	Nervous system disorders	UNKNOWN	Unlikely	Unlikely/Not Related	No

Figure 9: A screenshot of the web-based application developed.

We developed a web-based application that integrates the ML model into the medical review process. A screenshot of the application is shown in Figure 9. This application connects the trained ML model directly to the safety database, allowing the company's internal medical reviewers to interact with the model in real-time. Specifically, reviewers can efficiently select solicited post-marketing ICSRs of interest by filtering cases based on products and case completion time periods. This interactive functionality ensures a user-friendly experience, improving the overall efficiency of the review process.

The application provides several key outputs, including a confusion matrix that compares model predictions against the actual causality assessment, the distribution of causality classes, and the classification results, as illustrated in the right section of Figure 9. Within the classification results table, events with mismatched causality assessments between medical reviewers and the ML model are highlighted in blue, enabling reviewers to efficiently identify and evaluate discrepancies. Additionally, the application offers the flexibility to adjust the threshold for Class 3 ('unassessable') predictions (as discussed in Section 4.3), allowing medical reviewers to pre-define Class 3 threshold based on specific considerations such as product and its AE profile.

To demonstrate the adoption of the ML model in medical review oversight, we applied it to the adjudication of 'unassessable' adverse events. Specifically, the model was tested on AEs collected in February 2025 from four products, totaling 3,712 AEs. The numbers of distinct MedDRA PTs for these products were 161, 135, 184, and 610, respectively, totaling 1,090 drug-PTs across the dataset. The adjudication results are summarized in

Oversight adjudication	Agree with ML predication (%)	Agree with medical reviewers (%)	New classification (%)	Total mismatched unassessable AEs
drug A	18 (69.2%)	1 (3.8%)	7 (26.9%)	26
drug B	24 (85.7%)	2 (7.1%)	2 (7.1%)	28
drug C	30 (68.2%)	6 (13.6%)	8 (18.2%)	44
drug D	377 (67.1%)	60 (10.7%)	125 (22.2%)	562
Total	449 (68.0%)	69 (10.5%)	142 (21.5%)	660

Table 4: Oversight adjudication of mismatched ‘unassessable’ AEs.

Table 4. Using a default prediction threshold of 80%, the ML model identified 660 mismatched events wherein medical reviewers classified them as Class 3 (‘unassessable’), but the model predicted them as either Class 1 or Class 2.

Among these 660 mismatched events, internal medical reviewers who are experienced in evaluating post-marketing ICSRs, agreed with the model’s causality prediction in 449 events (68.0%), aligned with the actual medical reviewer assessment in 69 events (10.5%), and assigned a new causality class in 142 events (21.5%). These findings underscore the effectiveness of the ML model in enhancing medical review oversight. The ML model’s capability of identifying mismatched causality assessment in minutes enables detailed and focused oversight review of causality assessment in a much larger volume of ICSRs. Furthermore, the high degree of alignment between internal medical reviewers and the model prediction provides confidence and verification in the ML model’s performance.

6 Discussion

Comparative analysis - this proof-of-concept study explored the feasibility of leveraging structured ICSR attributes within a simple, transparent, and interpretable framework. While the primary focus was on kNN, we acknowledge that other baseline models such as Random Forest, XGBoost, and Logistic Regression are widely used for classification tasks and could potentially be applied here. These methods can offer higher predictive accuracy in some settings, but they generally require more complex training, extensive hyperparameter tuning, and may produce decision boundaries that are less intuitive for end-users in pharmacovigilance. In contrast, kNN provides a straightforward, instance-based approach that is easy to explain, requires minimal training, and allows rapid, interactive parameter adjustment in collaboration with medical reviewers. Preliminary exploratory testing with Random Forest and Logistic Regression on a subset of the dataset yielded performance metrics in a similar range to kNN, supporting its suitability for this application. Given that the intended application is to support case review prioritization rather than replace medical judgment, transparency and explainability were prioritized over maximal predictive accuracy. Future work will incorporate benchmarking against alternative models to assess trade-offs between accuracy, interpretability, and operational applicability.

Historical AE data - the causality classifier we have developed relies on the causality outcomes of historical adverse events. Initially, we utilized eight years of data (2014-2021) for the selected products; however, the model performance was mixed, with a higher mismatch rate observed for certain products. We hypothesized that the implementation of a new safety database in 2015 and changes in company causality assessment conventions may have negatively impacted the model performance. To address this, we adjusted the data set time frame to six years, starting from 2017 and ending in 2022. This adjustment resulted in a significant improvement in model performance. From a life cycle management viewpoint, it is recommended to refresh the historical data set for re-training after a certain time interval to ensure model performance based on accurate and up to date data. Furthermore, if the quality of case processing and medical review improve over time, periodic re-training may improve model performance even further. The optimal time interval for re-training may depend on various factors, including maturity of the product, evolving safety profiles, and company convention changes, etc.

Feature selection - feature selection in this study was guided by domain expertise to ensure transparency and alignment with regulatory review practices. While this supports interpretability and clinical relevance, it does not guarantee statistical optimality. Future work will incorporate empirical feature selection methods (e.g., permutation importance, SHAP), correlation diagnostics, and exploration of additional variables such as demographics and temporal attributes to evaluate their contribution to model performance.

Dimension of feature space - when constructing the feature space, it's essential to ensure that the number of features isn't excessively high to avoid the 'curse of dimensionality' [11]. The 'curse of dimensionality', referring to various phenomena that arise when working with high-dimensional data, particularly in the context of the kNN algorithm, can present several challenges. One such issue arises in high-dimensional spaces where the distances between data points tend to equalize, thereby complicating the task of distinguishing between 'close' and 'far'. Consequently, this uniformity of distances hampers the kNN algorithm's ability to accurately identify the k nearest neighbors, which could potentially result in sub-optimal performance.

Feature normalization - it should be noted that all seven component distances have been standardized, exhibiting values that range from 0 to 1. Such standardization proves to be vital in machine learning models, as it guarantees that each feature is represented on a comparable scale within the feature space. Any deviation from this uniformity could result in distorted outcomes or misinterpretations. This is primarily because features with significantly diverse scales can dominate the model, diminishing the impact of other features. Therefore, preserving equivalent scales is indispensable to ensure an unbiased and precise representation in the machine learning model.

Selection of k - in our study, the k value in the kNN algorithm, the number of nearest AE neighbors, was selected as 10, after extensive testing. Generally, there is no universally optimal k value. The choice of k is highly dependent on the specific problem, and it is always necessary to validate the performance using the specific data set with which we are working. In our experiment, we have tested a range of k values, from 5 to 30, and the performance of the model does not exhibit a significant variation.

Threshold - the threshold used in the model augmentation functions as a supplementary mechanism for handling the classification of Class 3 adverse events. In general, elevating the threshold reduces the proportion of 'unassessable' adverse events while improving the model's accuracy for Class 1 and Class 2 adverse events. To determine the range of threshold, we engaged internal medical reviewers to provide causality assessments for a randomized set of 2,000 solicited cases received in 2023 for four out of the previously selected products. Among the 5,052 adverse events reviewed, the medical reviewers deemed 329 events (approximately 6.5%) as 'unassessable'. Based on this finding, we tested thresholds ranging from 50% to 90%. Please note that each drug could have a distinct threshold, rather than applying a universal threshold across all drugs, allowing for customization based on its specific AE data.

Computing time - when the historical data set for a drug is exceptionally large (e.g., containing hundreds of thousands of events), identifying similar AEs can become computationally intensive, with the processing time scaling proportionally to the size of the training data set. To mitigate this issue, we recommend leveraging parallel computation to optimize performance and fully utilize the multi-core architecture of modern computing processors. By partitioning the training data set and distributing the computational workload across multiple cores, significant reductions in processing time can be achieved. To assess scalability, we benchmarked the pipeline on the data set containing over 550K AEs, using a server with 64 GB RAM and varying CPU allocations (1, 16, 32, and 64 cores). Processing time decreased from approximately 22 hours on a single core to ∼2.5 hours on 16 cores, ∼1.5 hours on 32 cores, and ∼1 hour on 64 cores. Although the speed-up was considerable, it was not directly proportional to the number of cores, likely due to parallelization overhead, non-parallelizable components of the workflow, and memory bandwidth limitations. These results demonstrate that parallel processing can greatly accelerate analysis and enable efficient handling of large-scale datasets without compromising performance.

Limitations

- A limitation concerns the scope of information provided to the model. As a data-driven model, like other machine learning models, it processes a finite set of inputs, which is considerably narrower compared to the breadth of data humans can evaluate. For example, in constructing the event similarity metric, we relied on a selected subset of adverse event attributes, excluding potentially valuable data such as patient demographics and unstructured text-based information. Moreover, it is fundamentally impractical for a machine learning model to fully emulate the depth of knowledge and nuanced judgment of medical reviewers.

- A limitation of the present study is that our analysis relies solely on structured variables, without incorporating unstructured text such as narrative descriptions in adverse event reports. These narratives often contain additional clinical details, temporal information, and causal cues that could enhance event similarity calculations or causality assessment. Future work could leverage advances in natural language processing (NLP) or large language models (LLMs) to extract structured representations, such as causal relationships and sentiment, from free-text narratives. These derived features could then be integrated with existing numerical and categorical variables to improve similarity assessment and downstream analyses. We consider such integration represents an important direction for future research to enhance both the accuracy and interpretability of our system.

- Another limitation relates to the timing of model development. When dealing with new drugs or drugs with a limited volume of historical data, it is advisable to allow the model sufficient time to accumulate a robust data set of AEs for effective learning. Without adequate data volume, the model's performance may become inconsistent and exhibit significant variability when conducting causality assessments.

- Researchers have explored different ways to improve ML model performance [2]. We believe using concomitant medication labels and concurrent disease attributes can facilitate causality assessment. The challenge is the availability of drug and disease databases with structured data, such as MedDRA coding of adverse drug reactions and manifestations of diseases and WHODrug coding of products. Improvement to databases such as Up-To-Date and DrugBank may make this a reality in the future.

- The component distance values in this study were defined heuristically to maintain interpretability for end-users. While this approach ensures transparency, it does not capture continuous gradations of semantic similarity and may not scale optimally to all edge cases. Similarly, the substitution of a midpoint value for missing components is a simplifying assumption that may introduce bias. Future work will investigate ontology-based continuous similarity metrics and more rigorous methods for handling missing data to improve the robustness of the distance computation.

- While we recognize the importance of validating the model with independent datasets, this was not feasible in the current study due to the absence of company assessment results (target labels) in publicly available sources such as FAERS, as well as differences in data structure and variable coding. Future work will focus on developing data harmonization methods and pursuing collaborations to access external datasets with compatible input variables and causality outcomes to enable robust validation.

Practical use - this ML model was developed in close consultation with the company's ICSR Medical Review team. During its implementation, the Medical Review team was thoroughly trained on the model's usage and provided with detailed information about its methodology, training process, performance, and limitations. Feedback from the Medical Review team, such as incorporating medical reviewers' selection of new causality classification categories, will be considered for future model refinements. Human intervention is built in when using this type of ML model whether as retrospective oversight and training tool or when integrating it into the safety database to suggest ICSR causality assessments to medical reviewers. As AI technology continues to advance and reliance on human intervention diminishes, maintaining human oversight will remain critical to ensure the responsible and ethical application of AI in pharmacovigilance.

Generalizability - although the present study was conducted using Biogen's post-marketing solicited ICSR data, the underlying methodology is not specific to any one therapeutic area or sponsor. The kNN classifier and augmentation tool operate on structured AE features, including MedDRA terms, seriousness, and other standard pharmacovigilance variables, which are consistently collected across the industry. As such, the framework could be applied to other therapeutic areas or external datasets, provided that equivalent data elements are available. Nevertheless, product-specific AE profiles, differences in case completeness, and variations in medical review conventions may influence model performance. In such contexts, retraining the model and calibrating the augmentation threshold would likely be necessary to achieve optimal results.

7 Conclusion

In this study, we leveraged the principle of event similarity to develop a 3-class classification model aiming at automating ICSR causality assessments. The model was subsequently evaluated using adverse events from six marketed drugs. Results demonstrated impressive low mismatch rates, high precision values, and strong F1 scores, highlighting the model's effectiveness. Furthermore, we incorporated an augmentation tool to effectively address the classification of 'unassessable' adverse events. This innovative addition notably enhanced the overall performance of the machine learning classifier. Additionally, we demonstrated the practical application of the model in enhancing medical review oversight and facilitating vendor training. In conclusion, our findings validate the potential of artificial intelligence and machine learning in causality assessments while laying the foundation for future advancements and applications in this exciting field.

References

1. Zitu MM, Zhang S, Owen DH, Chiang C, Li L. *Generalizability of machine learning methods in detecting adverse drug events from clinical narratives in electronic medical records.* Front Pharmacol. 2023 Jul 12;14:1218679. doi: 10.3389/fphar.2023.1218679. PMID: 37502211; PMCID: PMC10368879

2. Cherkas Y, Ide J, van Stekelenborg J. *Leveraging machine learning to facilitate individual case causality assessment of adverse drug reactions.* Drug Saf. 2022; 45:571–82

3. Zhao Y., Yu Y., Wang H., Li Y., Deng Y., Jiang G., et al. (2022). *Machine learning in causal inference: Application in pharmacovigilance.* Drug Saf. 45 (5), 459–476. 10.1007/s40264-022-01155-6

4. Kıcıman E, Ness R, Sharma A, Tan C. *Causal reasoning and large language models: opening a new frontier for causality.* https://arxiv.org/abs/2305.00050

5. Yan, C., Duan, G., Zhang, Y., Wu, FX., Pan, Y., Wang, J. (2019). *IDNDDI: An Integrated Drug Similarity Network Method for Predicting Drug-Drug Interactions.* In: Cai, Z., Skums, P., Li, M. (eds) Bioinformatics Research and Applications. ISBRA 2019. Lecture Notes in Computer Science(), vol 11490. Springer, Cham. https://doi.org/10.1007/978-3-030-20242-2_8

6. Saaty, T. L. (2008). *Decision making with the analytic hierarchy process.* International Journal of Services Sciences, 1(1), 83–98.

7. Meyboom RHB, Royer RJ. *Causality Classification in Pharmacovigilance Centres in the European Community.* Pharmacoepidemiology and Drug Safety 1992; 1:87-97.

8. T. Cover and P. Hart. *Nearest neighbor pattern classification*, IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21-27, January 1967.

9. Ren J, Zhou R, Farrow M, Peiris R, Alosi T, Guenard R, Romero-Torres S. *Application of a kNN-based similarity method to biopharmaceutical manufacturing.* Biotechnol Prog. 2020 Mar;36(2):e2945. doi:10.1002/btpr.2945. Epub 2019 Dec 16. PMID: 31811702

10. Cassisi, Carmelo, Placido Montalto, Marco Aliotta, Andrea Cannata and Alfredo Pulvirenti. *Similarity Measures and Dimensionality Reduction Techniques for Time Series Data Mining.* 2012. DOI: 10.5772/49941.

11. Piotr Indyk and Rajeev Motwani. *Approximate nearest neighbors: towards removing the curse of dimensionality*, Proceedings of the thirtieth annual ACM symposium on Theory of computing (STOC '98). ACM, New York, NY, USA, 604-613, 1998.

12. Dupuch, M., & Grabar, N. (2015). *Semantic distance-based creation of clusters of pharmacovigilance terms and their evaluation.* Journal of Biomedical Informatics, 53, 287–296. https://doi.org/10.1016/j.jbi.2014.11.012