Drug Safety
https://doi.org/10.1007/s40264-026-01649-7

LEADING ARTICLE

Critical Appraisal of Artificial Intelligence for Rare-Event Recognition: Principles and Pharmacovigilance Case Studies

G. Niklas Norén[1] · Eva-Lisa Meldau[1] · Johan Ellenius[1]

Received: 30 September 2025 / Accepted: 8 January 2026
© The Author(s) 2026

Abstract

Many high-stakes artificial intelligence (AI) applications target low-prevalence events, where apparent accuracy can conceal limited real-world value. Relevant AI models range from expert-defined rules and traditional machine learning to generative large language models (LLMs) constrained for classification. As the effort and expertise required to develop modern AI decrease, there is a risk that organizations devote too little time to understanding their limitations and sources of error. We outline key dimensions for critical appraisal of AI in rare-event recognition, including problem framing and test set design, prevalence-aware statistical evaluation, robustness assessment, and integration into human workflows. In addition, we propose an approach to structured case-level examination (SCLE), to complement statistical performance evaluation, and a set of considerations to guide procurement or development of AI models for rare-event recognition. We instantiate the framework in pharmacovigilance, drawing on three studies: rule-based retrieval of pregnancy-related reports, duplicate detection combining machine learning with probabilistic record linkage, and automated redaction of person names using an LLM. We highlight pitfalls specific to the rare-event setting including optimism from unrealistic class balance and lack of difficult positive controls in test sets—and show how cost-sensitive targets align model performance with operational value. While grounded in pharmacovigilance practice, the principles generalize to domains where positives are scarce, and error costs may be asymmetric.

Key Points

Rare-event recognition poses unique challenges for AI evaluation: apparent accuracy can be misleading when prevalence is low, and test sets may fail to capture difficult positives.

We outline key dimensions for the appraisal of AI models for rare-event recognition, including prevalence-aware statistical evaluation, robustness assessment, and integration into human workflows.

Structured case-level examination (SCLE) is introduced as a novel complement to contextualize statistical performance evaluation, generating insights from human review of false positives, false negatives, and correct classifications.

A set of considerations is provided to guide procurement or development of AI models for rare-event recognition.

✉ G. Niklas Norén
niklas.noren@who-umc.org

[1] Uppsala Monitoring Centre, Uppsala, Sweden

1 Introduction

With the rapid evolution in performance and versatility of artificial intelligence systems, their use is becoming more widespread and valuable [1, 2]. The ability to appraise such systems is important for professionals and decision-makers to invest their resources wisely. A failure to effectively evaluate AI models may lead to time and effort being wasted on systems that do not deliver what they promise or even harm individuals or the public good. Whereas human operators also make mistakes and vary in their skillfulness, the impact of a single flawed human operator is typically limited, whereas a single flawed AI system can be deployed at scale with widespread impact. However, unwarranted skepticism may mean organizations and individuals miss opportunities and benefits that AI systems could bring.

In general, AI may yield improvements of the following general nature:

- Efficiency: performing tasks that humans would otherwise do—but faster, with less effort.
- Quality: performing tasks that humans would otherwise do —but more accurately and consistently.

Published online: 11 March 2026

- Capability: performing tasks that otherwise would not get done.

The distinction between quality/efficiency and capability is not always clear, and what constitutes new capability may vary between human operators. For example, an English-to-German machine translation would offer a new capability to the co-author G.N.N., who does not know German, and (possibly) efficiency or quality to the co-author E.L.M., who is a German native. Often, the value may not be achieved by the artificial intelligence system alone but by a human–AI team. Then, the aim may be intelligence *augmentation*—to support and enhance human decision-making.

Whereas AI systems can perform a variety of tasks, our focus here is on AI models for recognizing rare events[1] [3, 4] to support organizational workflows or to enrich datasets for downstream analysis. Examples include spam or phishing detection (where messages flagged for review are in the minority) [5], fraud detection (where rare suspicious transactions are escalated for audit) [6], safety monitoring in aviation or cybersecurity, and computational phenotyping as a basis for, for example, epidemiological research [7]. In contrast, applications such as clinical diagnosis, where predictions directly inform individual patient care, fall outside our scope. In pharmacovigilance, rare-event recognition plays a role both in day-to-day case processing and in data management—for example, by flagging suspected duplicates or reports that warrant expert review—and as a foundation for subsequent aggregate analyses in signal management [8, 9]. For this reason, we draw on use cases from pharmacovigilance to illustrate and ground the principles proposed in this paper.

The aims of this paper are to demonstrate that those requiring rare-event recognition must be able to critically appraise AI systems, and to equip them with practical capabilities to assess performance evaluations. Specifically, we outline key dimensions including problem framing and test set design, prevalence-aware statistical evaluation, robustness assessment, and integration into human workflows. In addition, we propose a structured approach to case-level examination to complement and contextualize statistical performance evaluation, and present a set of key considerations to support procurement or development of AI models for rare-event recognition. This should be viewed as a complement to more general guidance on the topic [10, 11].

[1] There are no generally agreed upon definitions of what constitutes a rare event. In this paper, the term is used pragmatically to refer to settings where event prevalence is sufficiently low—often well below 10%, but highly dependent on the unit of analysis and task formulation—that naïve accuracy measures can be misleading and more careful performance appraisal becomes necessary.

2 Artificial Intelligence

The use of artificial intelligence as a term to describe certain computer science applications is growing in popularity again, but interpretations and delineations vary. Whereas colloquial use of the term may imply a certain degree of autonomy and versatility, technical definitions tend to be more inclusive. For example, Aronson proposes the following definition of artificial intelligence after careful consideration of existing definitions and their limitations [12]:

> **Artificial intelligence**, *n. a branch of computer science that involves the ability of a machine, typically a computer, to emulate specific aspects of human behavior and to deal with tasks that are normally regarded as primarily proceeding from human cerebral activity.*

The OECD define an *AI system* as (adopted also in the European Union's AI Act) [13]:

> **AI system:** *An AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment.*

These definitions would include expert systems that have been explicitly programmed to perform specific tasks, as well as any application of machine learning where computational methods infer and optimize behavior on the basis of training data. The former can range from simple heuristics to highly complex and capable systems, such as IBM's Deep Blue chess-playing computer, which combined massively parallel search with expert-designed evaluation functions and human opening/endgame knowledge [14]. The latter includes both lower-dimensional data-driven linear or rule-based models and large language or image models using deep neural networks with trillions of parameters (and counting).

A wide variety of AI models may be relevant for rare-event recognition, including expert-defined rule-based methods, traditional machine learning methods [e.g., support vector machines (SVMs) or gradient boosted trees], fine-tuned LLMs with a classification layer [e.g., BERT [15]] generative LLMs [e.g., GPT4 [16]] with constrained output (e.g., through prompting or post-processing of their textual output) (Fig. 1).

To determine the appropriate scope and approach for appraising an AI system the following qualities may be relevant: ambiguity of task (is the correct classification of events clear to human operators?), opacity (can humans understand the factors and inner logic based on which the AI

system arrives at its classification?), adaptiveness (can the AI system adjust its classifications to changing conditions or feedback regarding its performance?), scope (how wide a variety of events can the AI system recognize and under what conditions can it operate?), and autonomy (what level of human oversight and control is the AI system subject to?). With no ambiguity, standard computer validation should suffice, whereas with higher ambiguity, statistical performance evaluation will be required, and more care must be exercised in defining reference standards. With increasing opacity, expectations on transparency regarding the AI model's training and performance evaluation should increase [17, 18]. With higher adaptiveness, expectations on continual performance monitoring increase. However, any AI model can be vulnerable to data drift [19–21], and when AI models with broader scope are applied to new types of events or operating conditions, additional performance evaluation may be required. With higher autonomy, measures to ensure the validity, robustness, and safe use of AI must be more extensive, according to the risk-based approach.

3 Examples Used for Illustration

Throughout this article, we will refer to three recently published evaluations of artificial intelligence models in pharmacovigilance for illustration: an expert-defined rule-based method for recognition of adverse event reports related to pregnancy [22], a new version of a machine learning-based model for improved duplicate detection in large collections of adverse event reports [23], and a fine-tuned deep neural network for automated redaction of person names in case narratives [24]. The methods are different in nature but face many of the same challenges during evaluation. For an overview of intended use and test sets for our three use cases see Online Resource Table 1.

3.1 Rule-Based Retrieval of Pregnancy Reports

Sandberg et al. [22] developed a rule-based method to identify adverse event reports involving exposure to medicinal products during pregnancy, affecting either the pregnant individual or the prenatally exposed child. The method applies two sets of rules: one to exclude reports that are highly unlikely to involve pregnancy, and another to identify those that are likely relevant. This method is available as an optional search filter in VigiLyze, a tool for members of the WHO Programme for International Drug Monitoring to analyze reports in VigiBase, which is the WHO database of adverse event reports for medicines and vaccines.

3.2 Duplicate Detection with SVM and Statistical Record Linkage

Barrett et al. [23] combined statistical record linkage with traditional machine learning to improve duplicate detection in adverse event reporting systems. Duplicate reports are unlinked reports referring to the same adverse event in a given patient that can distort analyses and waste reviewer resources if not identified and removed. The new approach combines a support vector machine (SVM) classifier with principles from statistical record linkage. It compares report pairs on multiple features, some defined by experts and others derived from statistical modeling. The method rewards

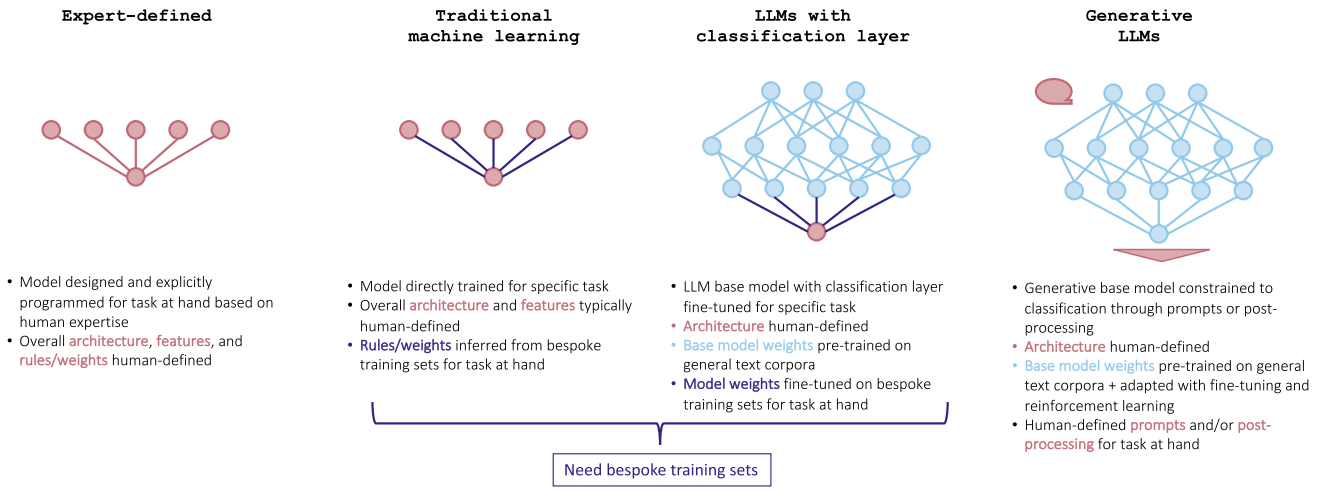


Fig. 1 Different types of AI models for rare-event recognition. Red graphical elements and text highlight model components defined on the basis of human expert input, e.g., "architecture"; dark blue elements highlight components inferred from bespoke training sets for the task at hand, e.g., "model weights"; light blue elements highlight components inferred during general purpose pre-training, e.g.,"base model weights"e

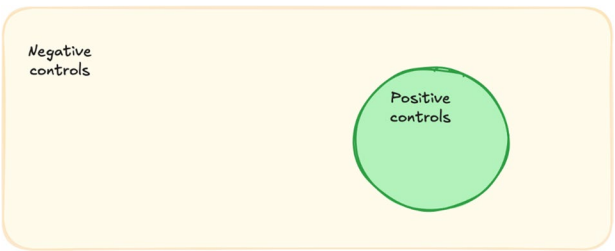
unusual similarities (e.g., two patients sharing the same rare adverse event) and penalizes unlikely mismatches. The improved model is currently being implemented to support workflows where suspected duplicates are flagged for human review or automated preprocessing, where duplicates are removed before statistical signal detection. By combining machine learning with structured statistical logic, this method aims to improve performance compared with earlier versions of vigiMatch, while remaining transparent enough to align with regulatory expectations.

3.3 Redaction of Person Names with Fine-Tuned LLM

Meldau et al. [24] fine-tuned an LLM, in this case BERT, with a classification layer to predict which token (word or other sequence of characters) in a case narrative text represents a person name. The BERT model had been pre-trained on English text including books and Wikipedia texts and was fine-tuned on combined data from a public de-identification challenge dataset called 2014 i2b2/UTHealth Corpus and case narratives from UK Yellow Card data. It is intended to either fully automate redaction or support humans redacting case narratives, i.e., to highlight parts of the text that should be masked or replaced by a placeholder. This is intended for an early step of case processing soon after reports arrive at the pharmacovigilance organization.

4 Test-Set Construction for Rare Events

In rare-event recognition, we may refer to those data points that we want to retrieve as *positive controls* and those that we do not want to retrieve as *negative controls*. We use this terminology throughout the description below.



The nature of a test set should align with the desired deployment domain. For example, in evaluating methods for pharmacovigilance signal detection, historical safety signals would typically be a better choice of positive controls than well-known, already labeled adverse drug reactions because their reporting patterns differ [25]. Similarly, if an AI model for recognizing adverse events in free text is intended for broad use across collections of adverse event reports, the test set should include reports related to a broad range of drugs and adverse events, from both patients and health professionals, etc.

Sometimes, boundaries between positive and negative controls are not clear-cut, which yields additional sources of ambiguity. For example, different organizations may have different requirements on how strong the conviction should be that two individual case reports refer to the same event for them to be classified as suspected duplicates. Similar considerations apply in defining negative controls: one may use all data points which do not meet the criteria to qualify as positive controls or establish a margin-of-error in which data points that are difficult to classify are set aside and do not count as either positive or negative controls. For example, because of ambiguity in the presentation and encoding of adverse events, one may choose to exclude from the negative controls any terms adjacent to a positive control (e.g., MedDRA Preferred Terms in the same High-Level Term group).

For transparency, the nature of test sets should be clearly specified and design choices made during their development documented and communicated [17]. It may be relevant to present descriptive statistics for positive and negative controls. This is important especially when a method may be used for variations of the intended use case in the future, and to assess the diversity and representativeness of the test set. In our experience, the development of annotation guidelines outlining the principles based on which data points are classified as positive or negative controls in creating test sets sometimes improves future quality and consistency of human processing by harmonizing and making explicit decision processes that were previously implicit and variable within an organization.

For rare events, straight random samples of data points will contain a low proportion of positive controls, and it may be impossible or too costly in time and effort to annotate a sufficient random sample to obtain enough positive controls. Conversely, if enrichment strategies are used to increase the proportion of positive controls in the test set, this may yield misleading performance estimates, if not appropriately accounted for in the analysis, as discussed below.

We focus here on test sets, because our interest is primarily in performance evaluation. Reference sets are typically divided into training, validation, and test splits, where the former two are used for model training, fine-tuning, and hyper-parameter tuning. Many considerations in this section are relevant also during model training and validation. However, more liberty can be allowed in those phases, and some approaches that would not be acceptable in test-set construction can be applied with care. For example, in fine-tuning the BERT model for the Uppsala Monitoring Centre (UMC) redaction of narratives method, we took advantage of the tendency of person names to cluster within narratives,

by directing our annotators toward narratives that a simpler classifier had suggested contained at least one person name. Similarly, the training and validation data for the SVM model in vigiMatch duplicate detection was actively enriched with positive controls (some of which were identified by earlier versions of the method under development). Each of these practices could have resulted in suboptimal performance or generalization of the trained models, and it is critical to scrutinize fit models and their interim performance evaluation results on validation data during development to recognize and course correct when this happens. As an example, during the development of vigiMatch, inspection of an interim SVM model showed that it would *penalize* two reports if they came from the *same* country, which is counterintuitive. The reason, it turned out, was an over-representation of reports from one specific country, which had had an excessive rate of false positives for an earlier version of the model. This had led many non-duplicate report pairs from this country to be annotated and included in our training set, which could then be addressed.

5 Prevalence-Aware Statistical Performance Evaluation

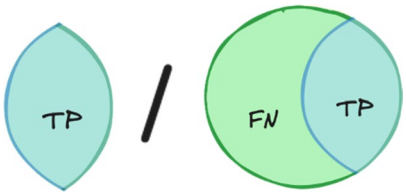
For rare-event recognition, where we cannot require or expect perfect performance, and where the task itself may be ambiguous (i.e., humans may not be able to classify all data points with certainty), evaluation of AI models will typically be statistical.

The basis for statistical performance evaluation for a binary classification task is a cross-classification of data points comparing the predictions of an AI model with the annotations in the test set. A positive control predicted positive by the AI model is referred to as a true positive (TP). A positive control predicted negative by the AI model is referred to as a false negative (FN). A negative control predicted positive by the AI model is referred to as a false positive (FP) and a negative control predicted negative is referred to as a true negative (TN) (Fig. 2).

The above cross-classification presumes that, for AI models with continuous output, a decision threshold has been set. The selection of the threshold should account for the relative costs of different types of errors (false positives versus false negatives) and may be guided by domain experts' input on acceptable performance levels. Detailed performance analyses may assess a range of decision thresholds, as discussed below. The relative costs of errors may vary with the intended use of a given method, and so may the relevant decision thresholds. For example, the cost of false positives in duplicate detection may be time and effort (and possibly alert fatigue) in a use case where suspected duplicates are forwarded to human operators for review, and may be missed/delayed signals in a use case where suspected duplicates are automatically removed prior to statistical signal detection.

In rare-event recognition with a focus on organizational impact and downstream analyses, recall, precision—and in certain circumstances, specificity—are core performance metrics. Together, they account for false negatives (recall/sensitivity) and false positives (precision/positive predictive value or specificity). Negative predictive value (the probability that a predicted negative is correct), as with recall/sensitivity, focuses on false negatives but asks how much an individual negative prediction can be trusted. It is important in medical diagnostics, for example, but less relevant for the applications discussed here. However, it does provide useful information on the purity of the set of predicted negatives for downstream analyses, which may inform to what extent misclassified positive controls may bias subsequent analyses. As discussed below, precision provides similar information regarding misclassification of negative controls. For a complementary discussion of pitfalls in AI performance evaluation, with a slightly different scope, see Hicks et al. [26].

5.1 Recall

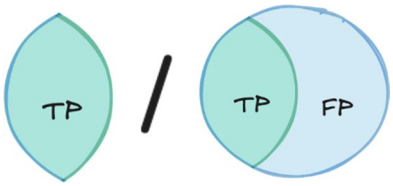


Recall is the proportion of positive controls that are predicted positive by the AI model: $\frac{TP}{TP+FN}$. It measures how many of the data points that we are interested in are correctly classified (recalled). Sensitivity and true positive rate are synonyms

Recall is relevant in applications where it is important to identify events of interest as completely as possible, and false negatives are costly, for example, if unrecognized events mean appropriate actions are not taken or are delayed. For downstream analyses, low recall means that not all true events are recognized, which may reduce statistical power.

To compute recall requires a representative set of positive controls. In rare event settings, annotation of random samples can be costly, and if enrichment heuristics are used to increase the proportion of positive controls in the test set, recall may be over-estimated if positive controls that are harder to recognize for the AI model are also less likely be found with the heuristic. If positive controls are not easily recognized/classified by human operators upon review, misclassification in the annotation can lead to optimistic recall estimates (if AI struggles with the same cases as humans). Table 1 describes how recall was evaluated in the three examples.

5.2 Precision



Precision is the proportion of predicted positives that are positive controls: $\frac{TP}{TP+FP}$. It measures how many of the data points are classified as "of interest" by the AI model, which are correct according to the reference standard. Positive predictive value (PPV) is a synonym

Computing precision requires a representative set of data points that are predicted positive by the AI model of interest. Precision is relevant in applications where the proportion of false positives among predicted positives matters—for example, when there are humans in the loop and each false positive requires human effort to process. Low precision then reduces efficiency because of time spent processing false positives and may lead to alert fatigue. For downstream analyses, precision reflects the purity of the set of predicted events, and the risk of misclassification bias from incorrectly flagged events.

Importantly, the estimated precision is highly dependent on the prevalence of positive controls in the test set, and if test sets have been enriched with positive controls, naive test set precision estimates will be optimistic and not reflect real-world performance. Note that the expected estimated precision of random guessing equals the prevalence of positive controls (so for a test set balanced on positive and negative controls, 50% precision is the baseline).

Precision is, in principle, straightforward and relatively cheap (in time and effort) to estimate for a given AI model by applying it to a random sample of data points and annotating all predicted positives. However, such *model-specific precision test sets* will need to be replaced or updated if the AI model is changed or another model is considered. As such, they are not suitable to reuse for benchmarking. As model-specific precision test sets depend on the interplay between AI models and source data sets, descriptive statistics for the set of predicted positives can be valuable to identify possible bias or over-representation.

"Precision@k" measures precision for the highest decision threshold which results in k predicted positives. It can

Table 1 Evaluation of recall for each of the three examples used for illustration

vigiBase pregnancy algorithm
Published study annotated random sample of 7874 reports, after restriction on the basis of patient age and sex (i.e., no active enrichment with positive controls, but downsampling of negative controls). Demonstrated that the impact of the restriction by age and sex on estimates of recall (by potentially excluding hard positive controls) was negligible
vigiMatch duplicate detection
Used sets of known duplicates already identified by four different national regulators to study recall (i.e., no new annotations and therefore no enrichment)
Acknowledged that these test sets may fail to include true duplicates not recognized as such by human operators (either for lack of sufficient information on individual reports or because there exist too many pairs for humans to assess all possible pairs), which may lead to optimistic recall estimates
UMC redaction method
Annotated random sample of 5042 case narratives identifying 179 NAME tokens in 71 narratives (i.e., no enrichment)
Conservative classification of edge cases as NAMEs when in doubt, which should lead to conservative (possibly pessimistic) recall estimates

[a]ICH E2B is an international standard for the electronic transmission of Individual Case Safety Reports used in pharmacovigilance to report adverse events associated with medicines, defined by the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use

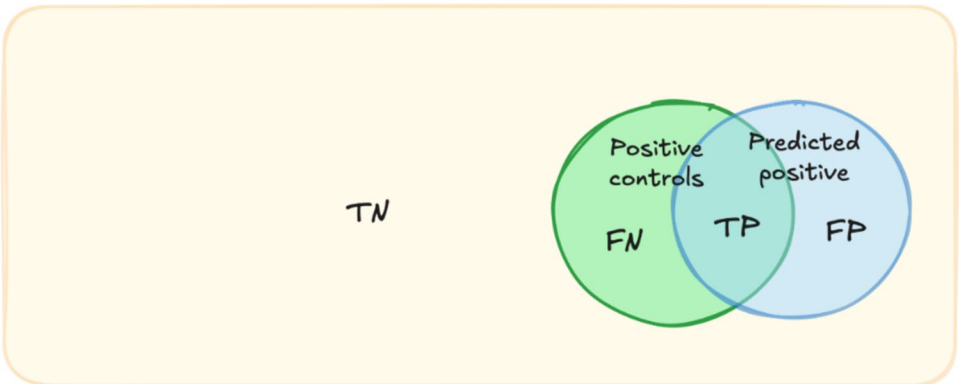
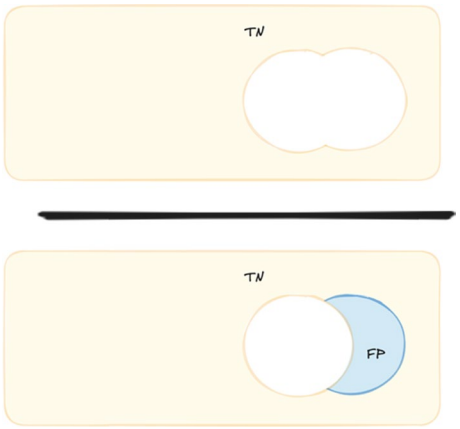


Fig. 2 Basis for statistical performance evaluation in a binary classification task. Any data point outside the blue ellipse is predicted negative, and any point outside the green ellipse is a negative control. TP, true positive; FP, false positive; FN, false negative; TN, true negative

be relevant when existing operating points are of that nature, for example, when teams operate under a fixed review budget. It can also enable comparative analyses between different models or across time, when decision thresholds are difficult to calibrate or unstable. Table 2 describes how precision was evaluated in the three examples.

5.3 Specificity



Specificity is the proportion of negative controls that are predicted negative by the AI model: $\frac{TN}{TN+FP}$. It measures how many of the data points we are not interested in are correctly classified as negative. True negative rate is a synonym

When some action is taken automatically for every predicted positive, specificity may be more relevant than precision to assess the impact of false positives—or at least serve as a useful complement. One advantage is its independence of the prevalence of positive controls in the test set. On the basis of estimates of sensitivity and specificity, precision can be computed for different levels of (assumed) prevalence of positive controls, using Bayes theorem. However, to estimate specificity requires a representative (and large enough) set of negative controls, and since negative controls far outnumber positive controls in rare-event recognition, specificity needs to come very close to 1 to achieve acceptable precision in many applications. This in turn requires a very large number of negative controls, which can be impossible, or at least very costly to obtain by annotating a random sample. This limits the usefulness of specificity as a performance metric in many rare-event recognition applications. Table 3 describes considerations regarding the evaluation of specificity for two of the examples.

5.4 Composite Performance Metrics

$F1$-score (sometimes just F-score) aggregates the estimated recall and precision for a specific threshold into a single measure of predictive performance. It is computed as the harmonic mean (a special type of average) between precision (p) and recall (r):

$$F1 = \frac{2 \times p \times r}{p + r}.$$

It inherits the possible limitations of precision and recall described above. Moreover, it assumes equal costs of false positives and false negatives, which is appropriate only in special cases. Variations of the F-score, such as the $F2$-score, make different assumptions of error costs but are not straightforward to interpret.

Precision-recall curves are a tool to display threshold-independent performance by plotting precision and recall for different values of the decision threshold. They can be valuable to describe performance at different operating points, when precision and recall are the relevant metrics. However, they suffer from the same limitations and risks of misleading results as their component metrics, when test sets are enriched with positive controls.

Receiver-operating characteristic (ROC) curves are another tool to display threshold-independent performance

Table 2 Evaluation of precision for each of the three examples used for illustration

vigiBase pregnancy algorithm
A random sample of 30,000 reports were retrieved from the full dataset, and all 448 predicted positives by the algorithm or its benchmark within this sample were annotated. This sample size was chosen on the basis of simulations showing it would provide sufficient statistical power (80% at a 5% significance level) to detect a true relative difference in precision of at least 10% between the algorithm and its benchmark
For transparency, precision was also calculated and presented on the basis of the downsampled dataset used to study recall.
vigiMatch duplicate detection
Model-specific precision tests with ~ 100 predicted positives for the method of interest and its benchmark were annotated. Careful reuse of random report pair sequences ensured maximal overlap between the two model-specific precision tests, reducing the number of required annotations
A too low prevalence of negative controls in training and validation sets during early model development gave models with unacceptably low precision in real-world settings—revealed by model-specific precision tests during model training and validation (unpublished results).
UMC redaction of narratives method
Precision was computed on the completely annotated randomly sampled test set described above, which contained 263,272 NON-NAME negative control tokens along with the 179 positive control NAME tokens. Precision is most relevant for a use case where human operators review all flagged tokens (reflecting what proportion of redacted tokens correspond to NAMEs)

by plotting recall (sensitivity) against (1 − specificity) for different values of the decision threshold. Performance better than chance is above the diagonal, and the closer the curve comes to the top left corner, the better the performance. Overall performance is often measured as the area under the curve of the ROC curve (AUC or AUROC), which corresponds to the probability that a randomly selected positive control is ranked above a randomly selected negative control by the assessed method. AUC values greater than 0.5 are better than chance.

ROC curves inherit the limitations of recall/sensitivity and specificity highlighted above. Whereas specificity can sometimes be relevant in rare event settings, only very limited portions of the ROC curve are of interest for rare events, which limits their usefulness. Because overall measures, such as AUC, are based on the entire range of specificity values from 0 to 1, they will be dominated by portions of the decision curve of no consequence in a rare-event setting. Table 4 describes use of composite performance metrics in the three examples.

5.5 Robustness Analyses

Most classifiers for rare-event recognition are deterministic. Once trained and fine-tuned (if machine-learning-based), they will generate the same output for any given input. However, classifiers based on generative LLMs include stochastic components (sometimes controlled with hyper-parameters such as "temperature") and may classify the same data point differently in repeated executions.

Stability is a general performance metric reflecting the tendency of a nondeterministic AI model to generate similar or the same output under limited (or no) perturbations of the underlying data. There is not a single stability metric. Rather, stability must be defined with a specific application in mind. For example, stability in rare-event recognition may be assessed as the proportion of data points assigned the same labels in repeated execution of a nondeterministic AI model

More broadly, the robustness of training or fine-tuning a machine learning model can be assessed by considering the variability in fitted models and their parameters (or their individual predictions or overall performance) when trained on different subsets of training data—e.g., folds of a cross-validation or bootstrap samples.

A further aspect of robustness concerns performance on specific subsets of the data. Global performance metrics may mask important heterogeneity: a classifier may perform well overall, while failing systematically for subgroups of cases defined by, for example, their features, source, demographics, or time period. Subset-specific analyses are relevant to ensure both fairness and equity as well as validity and robustness. From a fairness perspective, it is important not to under-serve or explicitly bias against certain subgroups. For downstream analyses, differential performance in recognizing an event of interest across subsets may introduce bias. Robustness analysis should thus consider stratified performance estimates and, where feasible, assess whether the model's errors are randomly distributed within subsets or cluster around some data characteristics. Such analyses do not necessarily require new metrics, but rather a systematic breakdown of standard measures (e.g., precision, recall, specificity) across relevant subsets.

Transparency of robustness analyses helps support trustworthy use of AI by clarifying the boundaries of reliable performance [17]. If limitations are identified, these can guide mitigations—such as targeted threshold adjustments, abstain/triage strategies, or prioritization of further training data collection. Table 5 describes robustness analyses for the three examples.

5.6 Benchmarks

Comparison to relevant benchmark methods provides an important point of reference when evaluating novel AI models, especially for bespoke test sets, whose nature, scope, strengths, and limitations may not be easy to assess for outsiders. For more complex benchmark methods, appropriate

Table 3 Consideration of specificity for performance evaluation in two of the examples used for illustration

vigiMatch duplicate detection
Specificity *not* reported as part of the study. Specificity at report *pair* level by necessity extremely close to 1 and practically impossible to estimate. However, specificity at the report-level (how many nonduplicates are incorrectly flagged by the method) could be valuable in assessing the collateral impact of removing suspected duplicates prior to subsequent analyses
UMC redaction of narratives method
False positive rate (= 1 − specificity) estimated to 0.05% used as a measure of the impact on the utility of the redacted narrative, assuming no human in the loop. Impact would be high if many NON-NAME tokens were removed, making narratives harder to read and potentially removing clinically relevant information. Note that in this case, the specificity of 99.95% resulted in a precision of only 55% because of the low prevalence of positive controls. Using Bayes formula, we see that if specificity had instead been 98% with prevalence and recall unchanged, precision would drop to 3%. In other words, almost all redacted tokens would be NON-NAMEs, even though the nominal specificity may seem high

Table 4 Use of composite performance metrics in performance evaluation for each of the three examples used for illustration

vigiBase pregnancy algorithm
No composite performance metrics presented as part of the study. The pregnancy algorithm does not rely on a tunable decision threshold, as it consists of a fixed set of rules. Potential improvements in recall and precision could be achieved by modifying the rule set to be more or less inclusive
vigiMatch duplicate detection
No composite performance metrics presented as part of the study, since relative costs of false positives and negatives considered to vary with the use case
Threshold-independent analyses such as precision-recall graphs may have helped assess the impact of decision threshold but were not obtained as part of the study
Specificity at report pair-level extremely close to 1 and ROC curves considered not relevant or informative
UMC redaction of narratives method
*F*1-scores computed to allow comparison to other studies and to summarize precision/recall into a single score. However, recall was seen as the most important (and method optimized for recall during development)
Precision-recall graphs used during model development to choose threshold in the BERT classification layer
Specificity of 0.995 for selected threshold so ROC curves considered not relevant or informative

hyper-parameter-tuning should be ensured to obtain relevant results.

Where available, public benchmark test sets offer additional value by enabling standardized comparisons across studies. In specialized applications such resources may be less common, and benchmarks from adjacent domains could be considered, though such comparisons require careful interpretation given differences in context and data characteristics. Unfortunately, benchmark methods and test sets remain scarce in specialized domains, such as pharmacovigilance, limiting opportunities for standardized comparison. Table 6 describes benchmark methods and reference sets for the three examples.

6 Structured Case-Level Examination (SCLE)

We propose a structured case-level examination (SCLE) to complement statistical performance evaluation in critical appraisal of AI models with case-level review of errors and correct classifications. SCLE extends and systematizes a practice we have applied informally in the past, including in the three example studies presented in this paper.

Summary performance metrics go only so far in enabling us to assess and understand the performance of an AI model. Equally important is to inspect representative, concrete examples of an AI model's classifications. Such examples should be analyzed during AI model development, evaluation, and performance monitoring/retraining. They should ideally be communicated to end users and in scientific publications when performance evaluation results are described. Importantly, SCLE is meant as a complement to, not a replacement for, rigorous statistical performance evaluation.

An SCLE should include both correct and incorrect classifications. Examining false positives and false negatives can each give useful insights regarding the strengths and limitations of the AI model and its evaluation, and they contextualize statistical performance evaluation. For example, if a false negative in redacting case narratives corresponds to a full name preceded by "Mr," this may undermine end users' trust in the AI model, even if the aggregate recall estimate is excellent. However, if the false negative is "AF" and it is

Table 5 Robustness analyses for each of the three examples used for illustration

vigiBase pregnancy algorithm
The method's recall increased from 75% overall to 91% when applied specifically to reports adhering to the ICH E2B format[a]. This subset-specific analysis reveals how structural differences in input data can significantly affect performance, underscoring the importance of stratified evaluation
vigiMatch duplicate detection
To assess performance for individual countries, small-scale model-specific precision tests were performed for two African countries and one European country. This was done because the benchmark method had been found to perform less well in some countries with adverse event and drug distributions that differed substantially from the global pattern
UMC redaction of narratives method
Sensitivity analyses focused on the one false negative corresponding to a full person name, which was of Indian origin (to assess possible issues with fairness and equity). Data manipulation experiments showed that the method was able to recall other names (of various origins) when inserted in the narrative of interest (except, strangely, the first name in "John Smith") and that the name of interest was appropriately redacted when inserted in one of the other narratives in the study. This indicated that the failure was due to some interaction between the name and the narrative

hard to know for a human specialist from the surrounding text if these are initials or an abbreviation for *atrial fibrillation*, then the overall precision metric may be viewed as potentially conservative in view of the ambiguity. However, SCLE should not be used to rationalize poor overall performance. Review of correctly classified data points may in turn give insights regarding the difficulty of the tasks successfully performed by an AI model. This may be especially important when there is no benchmark method that can provide baseline comparator, and we may not understand from overall performance metrics the difficulty of the task at hand.

For optimal use of human resources, we propose, for rare-event recognition, to examine a stratified random sample of individual (i) false positives, (ii) false negatives, and (iii) true positives. True negatives are the majority case and typically of less interest in in rare-event recognition. If there are subgroups of special interest in the performance evaluation, one may sub-stratify to ensure relevant coverage of these. It may also be relevant to sub-stratify by distance to the decision boundary—for example, "confident" misclassifications can be reviewed as part of sanity checking. Generally, a risk-based approach should determine the relative importance of (i), (ii), and (iii); the corresponding sample sizes; and the stringency of the human review and labeling.

Human review may consider a set of diagnostic tags relevant to the case at hand, in addition to free text notes. For example:

- "Never event" (misclassifications that would not be acceptable or could severely undermine trust in AI).
- Unexpected error (may point to opportunities to improve model and/or issues with training set).
- Input data issue (surprising quality issues, unexpected nature/scope of data points, human classification on the basis of information not available to AI).
- Test set issue (incorrect or ambiguous labels, too high or low granularity compared with intended use).
- Triviality (consider a simple metric such as non-trivial/trivial/unclear, for use in aggregate).

When there is a benchmark method, one may focus the SCLE on data points that are differentially classified by the method of interest and the benchmark, to better understand the nature of differences in performance between the two. In such circumstances, consider oversampling the disagreement cells (Model+, Benchmark− and Model−, Benchmark+). It may also be relevant to include diagnostic tags related to unexpected performance compared with benchmark.

Possible remedial action in response to findings in an SCLE include retraining (changes to training set and/or model architecture), modifying decision thresholds, data quality improvements, updates to annotation guidelines, and refined communication of performance evaluation. An over-representation of certain features or types of data among the misclassified may motivate systematic performance evaluations in specific subgroups. During model development SCLE can be instrumental in identifying areas of possible improvement, especially for more opaque methods that do not allow inspection/interpretation. A high rate of triviality of correct classifications may call for closer scrutiny of the test set.

SCLE is relevant not just in pre-deployment but throughout an AI model's lifecycle, for example, following retraining or in response to identified data drift [10]. Table 7 describes SCLE for the three examples.

7 Performance Evaluation of Human–AI Teams

Many AI systems are designed for intelligence augmentation. This means they are intended to support and enhance human decision-making rather than replace it or fully automate a task. In this context, evaluation could also consider the performance and dynamics of the human–AI team.

One important aspect is the assessment of team-level outcomes rather than evaluating the AI model in isolation. Metrics such as recall and precision can still be informative. In addition, decision efficiency, including the time and resources required, may be factored in. However,

Table 6 Benchmark methods and reference sets for each of the three examples used for illustration

vigiBase pregnancy algorithm
Benchmark method: Standardized MedDRA Query (SMQ) for pregnancy-related terms
Benchmark reference set: not available
vigiMatch duplicate detection
Benchmark method: earlier version of vigiMatch, representing current state-of-the-art (for adverse event reports related to drugs; for vaccine reports, no benchmark method available)
Benchmark reference set: not available
UMC redaction of narratives method
Benchmark method: not available for the specific intended use case
Benchmark reference set: test set from i2b2 challenge (acknowledging that this is not fully aligned with the intended use case and not what the method has been optimized for) and compared with state-of-the-art methods

experimental designs must consider the challenge that the same dataset cannot typically be used to simultaneously evaluate both human-only and human plus AI performance. Once a human has reviewed a case, their subsequent decisions may be influenced by prior exposure, which would bias the comparison.

Another important aspect for evaluation is how well the AI system integrates into the human workflow. Measures such as "decision concordance" and "override rate" by a human-in-the-loop can reveal the degree of alignment between AI recommendations and human judgment. The format of the AI output, whether visual or textual, can significantly influence usability and trust [27]. Evaluation methods that focus on the user experience, such as usability testing, are useful for understanding how users adapt to and benefit from the system over time.

Trust in AI systems is affected by their opacity. Users are more likely to trust systems whose reasoning they can understand [18]. Evaluating human–AI teams requires frameworks that go beyond individual technology acceptance, focusing instead on how AI systems integrate into collaborative workflows. Human factors engineering offers an approach for this purpose. As explained by Sujan et al. (2022), human factors engineering principles—such as automation bias, explanation, trust, and human–AI teaming—are essential for designing and assessing AI systems that support effective human–AI collaboration [28].

8 Critical Appraisal Considerations

The critical appraisal considerations presented in Table 8 are intended to support systematic evaluation of AI models for rare-event recognition. They are intentionally phrased as descriptive prompts for aspects to be examined, rather than as binary criteria to be met. This reflects the context-dependent nature of AI model evaluation and is intended to promote transparent reporting and critical appraisal rather than implicit pass/fail judgments or checklist-style compliance.

9 Discussion

The cost in time and effort to develop artificial intelligence systems is rapidly decreasing, with the advent of pretrained generative language models that offer unprecedented versatility and can be deployed without fine-tuning [1]. In contrast, rigorous performance evaluation remains resource intensive. This imbalance creates a risk that organizations devote less time to understanding limitations and sources of error, while being tempted to cut corners in appraisal. The danger is not only wasted resources but also negative impact on stakeholders and eroded trust if systems underperform in practice. A risk-based approach is essential: bold experimentation may be appropriate where costs of errors are contained and possible to reverse, whereas careful evaluation remains indispensable in settings where costs are substantial or irreversible. Our framework is intended as a backbone that can be adapted to context, rather than a rigid one-size-fits-all prescription.

The principles described here are largely method-agnostic and apply to rare-event recognition in general. However, the extent of performance evaluation should account for the ambiguity of the task, and the opacity, adaptiveness, scope, and autonomy of the AI system. For example, identifying pregnancy-related reports using a generative LLM rather than an expert-defined rule-based algorithm may necessitate additional evaluation and governance measures, such

Table 7 SCLE for each of the three examples used for illustration

vigiBase pregnancy algorithm
Case-level examination of false negatives identified that preprocessing errors were the most frequent source of these errors. The most critical issue was incomplete mapping to MedDRA during data preprocessing of pregnancy-related terms provided as free text
Case-level examination also confirmed known algorithm-level limitations including the inability to process pregnancy information confined to free-text fields, meaning relevant data was present but not accessible to the rule-based logic
vigiMatch duplicate detection
Benchmark-aware case-level examination of true positives, false positives, and false negatives presented for vigiMatch model for drugs
Case-level examination of false positives and false negatives presented for vigiMatch model for vaccines (where no benchmark was available)
Scope limited to handful of cases per classification category; human review subjective in nature without prespecified diagnostic tags
Scientific publication included examples of true positives, false negatives, and false positives illustrating the abilities and mistakes of the method
UMC redaction of narratives method
Systematic case-level examination of *all* false negatives classifying into diagnostic tags for directly, indirectly, and nonidentifiable narratives to determine the risk of re-identification. This identified that the only leaked full name was of Indian origin, leading to a follow-up analysis to determine if there was a problem with under-serving individuals with names of certain origins
Systematic case-level examination of all false positives, with diagnostic tags identifying masked tokens containing clinically relevant information
Scientific publication included examples of true positives, false negatives, and false positives with scrambled personal identifiers illustrating the abilities and mistakes of the method

Table 8 Key considerations in the procurement or development of AI models for rare-event recognition to ensure validity, robustness, and fairness

Consideration	Aspects
Test sets	Describe the nature of the information and the scope, size, and composition of test sets used for evaluation. Explain how these characteristics align with the intended use case. Describe the selection, number, and characteristics of positive and negative controls, and how they represent events in scope and out of scope for the intended use
Bias	Describe any known or suspected sources of bias that may influence the AI model, including how each of these were or were not assessed. Describe how the performance of the AI model was assessed under varying conditions and across data subsets relevant to the intended use (to identify possibly under-served groups)
Annotation process	Describe the criteria used to define positive and negative controls and the annotation process applied. Describe measures taken to ensure and assess annotation quality and consistency, and how edge cases or ambiguous instances are identified and handled
Choice of metrics	Describe the performance metrics used and their relevance to the intended use case. Explain which aspects of performance they are intended to capture (e.g., false positives, false negatives, stability) and which aspects may not be fully captured
Decision thresholds	Describe the decision thresholds considered in performance evaluation and how they relate to the intended use case, including assumptions about the relative costs of different types of errors (e.g., asymmetrical misclassification costs)
Evaluation of recall	If recall is evaluated, describe how the test sets reflect the spectrum of positive controls relevant to the intended use (types, difficulty, etc.). Describe any enrichment with positive controls and how this is accounted for in the interpretation of results
Evaluation of precision	If precision is evaluated, describe the prevalence of positive controls in the test sets and how this relates to the expected prevalence in the intended use case. Describe any enrichment with positive controls and how this is accounted for in the interpretation of results
Evaluation of specificity	If specificity is evaluated, describe how the test sets reflect the spectrum of negative controls relevant to the intended use (types, difficulty, etc.)
Benchmarks	Describe any comparisons made to relevant benchmark methods, where applicable, including how benchmark methods were implemented and optimized. Describe the availability (or not) of relevant benchmark test sets for the intended use case, and the performance of the proposed AI model on these
Performance drift	Describe any measures in place to identify, monitor, and respond to data, model, or performance drift. For AI models incorporating third-party components, describe whether and how sensitivity to updates or version changes of these components has been considered
Non-triviality	Describe the nature of true positive outputs generated by the AI model, including whether they reflect detection of non-trivial events relevant to the intended use case (e.g., events unlikely to be detected by simpler methods)
Types of errors	Describe the nature and patterns of false positives and false negatives observed. Explain how these error types relate to the intended use case, including whether they raise concerns regarding the validity, fairness, or downstream consequences
Human–AI interaction	Describe the intended human–AI interaction, including how outputs are presented, processed, and acted upon by users. Explain how this interaction has been accounted for in the performance evaluation

as systematic assessment of output stability, monitoring of model drift (including that induced by LLM versioning), and evaluation of risks related to leakage of personal or sensitive data via LLM prompts. That said, prevalence-aware statistical evaluation, transparency in test set design and annotation, and the use of our proposed structured case-level examination to complement summary metrics are broadly relevant. So is performance assessment across data subsets and other approaches to robustness analysis. It is worth noting that AI models producing ranked outputs effectively become binary classifiers once a threshold is applied. Similarly, while methods for mapping free text verbatims to standard terminologies and other multiclass classification problems are not rare-event recognition per se, prevalence-aware statistical evaluation can still be relevant. For example, overall evaluation of a method for mapping free text to adverse event terms may be dominated by more common terms, with strong general performance hiding a failure to effectively recognize rare adverse events that can be of crucial importance in pharmacovigilance. It can therefore be helpful to perform complementary performance analyses considering a set of binary event-recognition classification tasks, using the principles outlined here.

While the principles underpinning SCLE have been derived from real-world experience, the more systematic framework proposed here needs testing and refinement under real-world conditions. It is meant as a complement that should help contextualize statistical performance evaluation but not replace it. It should also be noted that by design, case-level examination is vulnerable to sampling variability, and observed patterns may require confirmation in systematic follow-up analyses. These are not reasons to forgo this aspect of appraisal but rather to motivate careful design, transparent reporting, and ongoing methodological

refinement. Qualitative research traditions demonstrate that such analyses can be conducted with rigor, transparency, and reproducibility.

Looking ahead, two developments deserve special attention. First, there is growing interest in using AI itself to support performance evaluation, for example, by employing generative LLMs to annotate test sets or review output from (simpler) AI models [29]. Such use of LLMs as a judge may help reduce the resource burden of evaluation and allow performance evaluation at entirely new scales [30], but their validity must be ensured and demonstrated via human calibration. They may be particularly valuable in continual performance monitoring after deployment, since human specialists may be more difficult to engage at this stage. Second, generative applications are becoming more widespread where AI models produce free text rather than numerical or categorical output [31]. It is possible that some of the systems that perform rare-event recognition will be replaced by other modalities. For example, redaction of person names could be addressed as a text editing task instead of as binary token classification. This would require a different approach to statistical evaluation, but SCLE would still be relevant—with different categories for the stratified random sampling. SCLE may also be valuable if relying on an LLM-as-a-judge for performance evaluation. Many workflows involving human operators will likely continue to rely on dichotomous decisions (spam/no spam, human review or not). The same is true of use cases where event recognition is the basis for subsequent analyses. However, processes that evolve to a more fluid back-and-forth exchange between human operators and AI [32] will require a different evaluation paradigm, assessing real-world decision-making by human–AI teams, accounting for user experience, cognitive ergonomics, and decision efficiency.

As AI technologies and practices evolve, so too must the standards for their evaluation. What remains constant is the need for appraisal that balances efficiency with rigor, enabling organizations to harness the benefits of AI for rare-event recognition while safeguarding validity, robustness, and fairness.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s40264-026-01649-7.

Acknowledgements The authors would like to thank Lovisa Sandberg and Jim Barrett for helpful comments on the manuscript preprint (https://doi.org/10.48550/arXiv.2510.04341), especially content related to the case studies. Some of the ideas and visualizations in the paper were presented to the CIOMS XIV working group on Artificial Intelligence in Pharmacovigilance during one of its meetings, and Niklas Norén would like to acknowledge the members of the working group for their critical review and constructive feedback. The views expressed are those of the authors, and not necessarily those of the Uppsala Monitoring Centre or any other organization.

Funding Not applicable.

Declarations

Conflicts of interest G. Niklas Norén is an Editorial Board member of *Drug Safety*. G. Niklas Norén was not involved in the selection of peer reviewers for the manuscript nor in any of the subsequent editorial decisions. Eva-Lisa Meldau and Johan Ellenius declare no conflicts of interest relevant to this work.

Ethics approval Not applicable; the paper does not involve human subjects or analysis of personal data.

Consent to participate Not applicable.

Consent for publication Not applicable.

Availability of data and material Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

Code availability Not applicable.

Author contributions GNN conceptualized the paper and wrote its first draft. ELM and JE critically reviewed all content and wrote specific sections. All authors read and approved the final version.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License, which permits any non-commercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc/4.0/.

References

1. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. Adv Neural Inf Process Syst. 2020;33:1877–901.
2. Green A. The great acceleration: CIO perspectives on generative AI. MIT Technology Review Insights; 2023. Available from: https://www.databricks.com/sites/default/files/2023-07/ebook_mit-cio-generative-ai-report.pdf. Accessed 15 Sept 2025.
3. Weiss GM. Mining with rarity: a unifying framework. ACM SIGKDD Explor Newsl. 2004;6(1):7–19.
4. Shyalika C, Wickramarachchi R, Sheth AP. A comprehensive survey on rare event prediction. ACM Comput Surv. 2025;57(3):1–39.
5. Basit A, Zafar M, Liu X, Javed AR, Jalil Z, Kifayat K. A comprehensive survey of AI-enabled phishing attacks detection techniques. Telecommun Syst. 2021;76(1):139–54.
6. Bello OA, Ogundipe A, Mohammed D, Folorunso A, Alonge OA. AI-driven approaches for real-time fraud detection in us financial transactions: challenges and opportunities. Eur J Comput Sci Inf Technol. 2023;11(6):84–102.
7. Richesson RL, Sun J, Pathak J, Kho AN, Denny JC. Clinical phenotyping in selected national networks: demonstrating the need

for high-throughput, portable, and computational methods. Artif Intell Med. 2016;71:57–61.
8. Kjoersvik O, Bate A. Black swan events and intelligent automation for routine safety surveillance. Drug Saf. 2022;45(5):419–27.
9. Botsis T, Ball R, Norén GN. Editorial: Computational methods and systems to support decision making in pharmacovigilance. Front Drug Saf Regul. 2023;21(3):1188715.
10. Yeung K. Recommendation of the council on artificial intelligence (OECD). Int Leg Mater. 2020;59(1):27–34.
11. Council for International Organizations of Medical Sciences (CIOMS). Artificial intelligence in pharmacovigilance - Report of the CIOMS Working Group XIV. Geneva: CIOMS; 2025. Available from: https://cioms.ch/artificial-intelligence-inpv/. Accessed 15 Jan 2026.
12. Aronson JK. Artificial intelligence in pharmacovigilance: an introduction to terms, concepts, applications, and limitations. Drug Saf. 2022;45(5):407–18.
13. Organisation for Economic Co-operation and Development (OECD). Explanatory memorandum on the updated OECD definition of an AI system. Paris: OECD Publishing; 2024 Mar. (OECD Artificial Intelligence Papers; vol. 8). Available from: https://www.oecd.org/en/publications/explanatory-memorandum-on-the-updated-oecd-definition-of-an-ai-system_623da898-en.html. Accessed 15 Sept 2025.
14. Campbell M, Hoane AJ, Hsu F. Deep blue. Artif Intell. 2002;134(1–2):57–83.
15. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805; 2019.
16. OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, et al. GPT-4 technical report. arXiv preprint arXiv:2303.08774; 2023.
17. Haibe-Kains B, Adam GA, Hosny A, Khodakarami F, Massive Analysis Quality Control (MAQC) Society Board of Directors, Shraddha T, et al. Transparency and reproducibility in artificial intelligence. Nature. 2020;586(7829):E14–6.
18. Von Eschenbach WJ. Transparency and the black box problem: why we do not trust AI. Philos Technol. 2021;34(4):1607–22.
19. Ditzler G, Roveri M, Alippi C, Polikar R. Learning in non-stationary environments: a survey. IEEE Comput Intell Mag. 2015;10(4):12–25.
20. Bayram F, Ahmed BS, Kassler A. From concept drift to model degradation: an overview on performance-aware drift detectors. Knowl Based Syst. 2022;245:108632.
21. Sahiner B, Chen W, Samala RK, Petrick N. Data drift in medical machine learning: implications and potential remedies. Br J Radiol. 2023;96(1150):20220878.
22. Sandberg L, Vidlin SH, K-Pápai L, Savage R, Raemaekers BC, Taavola-Gustafsson H, et al. Uncovering pregnancy exposures in pharmacovigilance case report databases: a comprehensive evaluation of the VigiBase pregnancy algorithm. Drug Saf. 2025;48(10):1103–18.
23. Barrett JW, Erlanson N, China JF, Norén GN. A scalable predictive modelling approach to identifying duplicate adverse event reports for drugs and vaccines. arXiv preprint arXiv:2504.03729; 2025.
24. Meldau EL, Bista S, Melgarejo-González C, Norén GN. Automated redaction of names in adverse event reports using transformer-based neural networks. BMC Med Inform Decis Mak. 2024;24(1):401.
25. Noren GN, Caster O, Juhlin K, Lindquist M. Zoo or savannah? Choice of training ground for evidence-based pharmacovigilance. Drug Saf. 2014;37(9):655–9.
26. Hicks SA, Strümke I, Thambawita V, Hammou M, Riegler MA, Halvorsen P, et al. On evaluation metrics for medical applications of artificial intelligence. Sci Rep. 2022;12(1):5979.
27. Spiker J, Kreimeyer K, Dang O, Boxwell D, Chan V, Cheng C, et al. Information visualization platform for postmarket surveillance decision support. Drug Saf. 2020;43(9):905–15.
28. Sujan M, Pool R, Salmon P. Eight human factors and ergonomics principles for healthcare artificial intelligence. BMJ Health Care Inform. 2022;29(1):e100516.
29. Zheng L, Chiang WL, Sheng Y, Zhuang S, Wu Z, Zhuang Y, et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In: Oh A, Naumann T, Globerson A, Saenko K, Hardt M, Levine S, editors. Advances in neural information processing systems. Curran Associates; 2023. p. 46595–623.
30. Schuemie MJ, Ostropolets A, Zhuk A, Korsik U, Seo SI, Suchard MA, et al. Standardized patient profile review using large language models for case adjudication in observational research. Npj Digit Med. 2025;8(1):18.
31. Van Veen D, Van Uden C, Blankemeier L, Delbrouck JB, Aali A, Bluethgen C, et al. Adapted large language models can outperform medical experts in clinical text summarization. Nat Med. 2024;30(4):1134–42.
32. Petersen M, Alaa A, Kıcıman E, Holmes C, Van Der Laan M. Artificial intelligence-based copilots to generate causal evidence. NEJM AI. 2024;1(12).