

Innovative Insights in Digital Health

AI-Powered Scoring System for Detecting Drug-Drug Interaction Insights from Scientific Literature

Charbel Alhelou^{1*}, Nour Allam², Praise Oketola², Jean François Arbona², Romain Clement¹, Julien Dufour¹, Dimitri Mikec², Pauline Molina³, Marine Poulain³, Sophie Tessier³, Anne-Sophie Lelong⁴

¹ArcaScience, Research and Development Department, Paris, France.

²ArcaScience, Technical and IT Team, Paris, France.

³Department of Medical Information Systems, VIDAL, Issy-les-Moulineaux, France.

⁴Integrated Products Team, VIDAL, Issy-les-Moulineaux, France.

Received Date: 07 Aug 2026;

Accepted Date: 14 Sep 2026;

Published Date: 15 Oct 2026.

***Correspondence:** Charbel ALHELOU, PharmD, ArcaScience, Research and Development Department, 75013 Paris, France.

Citation: Alhelou C, Allam N, Oketola P, Arbona JF, Clement R, et al. AI-Powered Scoring System for Detecting Drug-Drug Interaction Insights from Scientific Literature. *Innov Insights Digit Health*. 2026; 2(3): 1-10.

ABSTRACT

Background and Objective: Drug-drug interactions (DDIs) present major challenges in clinical care, impacting patient safety and therapeutic outcomes. Traditional detection methods, often based on curated databases, may miss newly emerging interactions or require frequent manual updates. Recent advances in natural language processing (NLP) and artificial intelligence (AI) offer novel tools for real-time, automated DDI detection directly from biomedical literature. This study introduces an AI-powered approach to identify DDIs, assign relevance scores to publications, and benchmark performance against conventional databases.

Methods: We developed an NLP pipeline incorporating sentence classification to detect safety and interaction-related content, and named entity recognition (NER) to extract drug-drug interaction types, their clinical impact, and patient features. The sentence classification module is based on a BERT architecture fine-tuned from microsoft/BiomedNLP-BiomedBERT-base-uncased-abstract-fulltext, and the NER module uses the open-source model d4data/biomedical-ner-all. A DDI relevance score was designed to quantify each article's significance based on weighted parameters derived from model outputs, with weights informed by domain experts (title relevance: 50 points, DDI impact: 40 points, publication type: 10 points). The system was validated on a manually curated dataset. We applied ANOVA and sensitivity analyses, and evaluated performance using recall, precision, and F1-score, comparing our method to established DDI resources.

Key Content and Findings: The pipeline was validated on 7,751 articles across 15 query types. ROC analysis identified a relevance score threshold of 75%. Drug-family queries achieved the highest F1-score (0.872), followed by single-drug (0.731) and drug-drug queries (0.684). ANOVA and Tukey post-hoc analysis confirmed statistically significant differences in relevance scores between annotation groups ($p < 0.05$), with full stratification across all three groups for drug-drug family queries. Compared with manually compiled expert reference lists, our system retrieved substantially more relevant articles without requiring manual updates.

Conclusion: AI-driven NLP approaches can enhance clinical decision support systems by providing comprehensive, evidence-based DDI detection and prioritization, offering scalable support for safer pharmacological decision-making.

Keywords: Drug-drug Interactions; Natural Language Processing; Artificial Intelligence; Pharmacovigilance; Clinical Decision Support.

Key Points

- Traditional drug interaction databases often miss new or rare drug combinations and require constant manual updates, leaving healthcare providers with incomplete information when prescribing.
- This study presents an AI system that automatically scans scientific literature to detect and rank drug-drug interactions, achieving high accuracy (F1-score up to 0.872) while identifying more relevant studies than existing manually curated lists.
- This approach could help doctors and pharmacists make safer prescribing decisions by providing real-time, evidence-backed references to support drug interaction alerts.

1. Introduction

1.1 Background

Drug-Drug Interactions (DDIs) represent a critical area of concern in clinical practice, as they can significantly impact patient safety and treatment outcomes [1]. DDIs occur when the pharmacological effect of one drug is altered by the presence of another, leading to potential adverse effects, reduced therapeutic efficacy, unexpected toxicity or a synergistic effect [1,2]. Drug interactions may be both pharmacokinetic (PK) and pharmacodynamic (PD) in nature [2]. The risks associated with DDIs are heightened in patients undergoing polypharmacy, as the concurrent use of multiple medications increases the complexity of managing potential interactions. According to the literature [3], DDIs are a leading cause of adverse drug reactions (ADRs) and hospital admissions worldwide, contributing to both morbidity and mortality. For example, more than one-third of older Americans regularly use five or more drugs or supplements, and 15% are at risk for serious DDIs [4]. Given that the harmful reactions from DDIs include bleeding, bone marrow suppression, arrhythmias, hypotension, rhabdomyolysis, central nervous system depression, seizures, hypoglycemia, renal failure, and so on [5], the accurate prevention and management of DDIs are crucial for avoiding adverse outcomes, optimizing treatment plans, and enhancing patient safety in clinical settings.

1.2 Rationale and knowledge gap

To detect and manage DDIs, healthcare professionals rely on a variety of tools and methods, ranging from drug interaction databases to clinical decision support systems (CDSSs). Several well-established databases, such as Vidal [6] and DrugBank [7], provide valuable resources for identifying potential interactions between drugs and their corresponding clinical implications. These tools are based on predefined interaction tables, expert-reviewed content, and available literature. However, despite their widespread use, these systems face significant limitations that compromise their clinical utility.

Most clinical DDI alert systems operate through simple table-based lookup mechanisms, matching prescribed drug combinations against predefined interaction lists [8]. When a clinician prescribes two drugs, the system checks whether that specific drug pair appears in its interaction table and generates an alert if

a match is found. Critically, this approach does not incorporate patient-specific risk modifiers such as renal or hepatic function, age, concurrent medical conditions (e.g., diabetes, heart failure, nephropathy), or genetic polymorphisms. This “one-size-fits-all” strategy means that identical alerts are generated for all patients prescribed the same drug combination, regardless of whether individual patient characteristics substantially increase or decrease the actual risk of interaction.

Consequently, clinicians are frequently confronted with alerts that lack clinical actionability for their specific patient, contributing to alert fatigue and potentially dangerous override behavior, a phenomenon where providers routinely ignore warnings due to excessive false positives. A systematic review by Van der Sijs et al. [9] found that drug safety alerts are overridden by clinicians in 49% to 96% of cases across different clinical settings, indicating widespread dismissal of system-generated warnings. The magnitude of inappropriate alerting is further illustrated by Phansalkar et al. [10], who identified that 36% of DDI alerts displayed in one institution were low-priority interactions that did not warrant interruptive alerts, contributing unnecessarily to alert burden. This over-alerting problem represents a fundamental barrier to the stated goal of using DDI detection systems to support clinical decision-making, as excessive false positives erode clinician trust and reduce compliance with legitimate warnings.

Beyond this personalization gap, current DDI detection methods face structural limitations. The reliance on static, predefined drug pairs or families restricts the ability to detect novel or rare interactions that fall outside established combinations. This challenge is amplified by the exponential growth in pharmaceutical agents: The latest release of DrugBank [11] Online (version 5.1.13, released January 2, 2025) contains 19,775 drug entries, including 3,016 approved small molecule drugs, 1,768 approved biologics (proteins, peptides, vaccines, and allergenics), 135 nutraceuticals, and over 8,924 experimental (discovery-phase) drugs. Maintaining comprehensive, manually curated interaction tables for all possible combinations is increasingly infeasible, leaving clinicians with incomplete information precisely when the therapeutic landscape demands greater vigilance.

One additional problem related to these databases is that they often don’t reference the interactions they provide: the clinician or pharmacist may face an alert about an interaction without being able to easily validate it, or understand and quantify the extent to which it is clinically relevant.

Beyond traditional databases, researchers have explored various AI-driven computational approaches to predict DDIs, including traditional machine learning models (e.g., random forests, support vector machines) [12,13], deep learning models utilizing NLP techniques [14,15], and score function-based models [16]. However, these prediction-focused methods face significant limitations: they lack transparent literature referencing to support their predictions; traditional models require extensive training data and may not generalize to new drugs; deep learning approaches struggle

with negative sample scarcity; and score function-based models depend heavily on known DDI databases and probability assumptions. This gap motivates the need for systems that can systematically extract, evaluate, reference and prioritize DDI information directly from published research in real-time.

1.3 Objective

In light of these challenges, the present study aims to test whether a novel DDI detection and referencing approach from scientific literature can overcome key limitations of traditional methods. Specifically, we seek to evaluate whether this system can: generate dynamic, evidence-based DDI reports that are not constrained by static, predefined drug pair lists; enhance clinical decision support system (CDSS) alerts by providing accurate, literature-backed references that substantiate interaction warnings; and improve alert prioritization by quantifying the strength and quality of evidence supporting each potential interaction.

2. Materials and Methods

2.1 Data Collection

The data used in this study were collected from three primary sources: PubMed, MEDLINE, and ClinicalTrials.gov, without any restrictions on publication date. A flexible query strategy was implemented to retrieve articles relevant to drug-drug interactions (DDIs). Queries were designed to capture various types of relationships, including “Drug 1 AND Drug 2,” “Drug 1 AND Drug Family 2,” “Drug Family 1 AND Drug 2,” “Drug Family 1 AND Drug Family 2,” as well as single-drug queries. All searches were conducted without timeframe limitations and were applied across all available sections of each article (title, abstract, full text when available, etc.). Any article matching a query (regardless of where in the text the match occurred) was automatically included in the Datamart. The specificity of queries was further refined based on user input. Users could request a focus on specific drugs or broader drug families. In addition, when a query returned too few results, the system broadened its scope using drug families defined by MeSH (Medical Subject Headings), which organizes drugs hierarchically by pharmacological class or therapeutic use. All queries were constructed using a proprietary drug ontology developed in-house, which integrates information from authoritative sources such as MeSH and the National Cancer Institute Thesaurus (NCIT), enabling consistent and accurate drug identification. This data retrieval process, termed Datamart Creation, served as the foundation for all subsequent analyses. The resulting Datamart includes metadata for each article (e.g., title, abstract, publication date, source, authors), the full text when available (around 35% of the articles), and all associated drug entities identified through the ontology, ensuring comprehensive and systematic coverage of DDI-related literature.

2.2 Data Points Extraction and Processing

2.2.1 Publication Type Classification

The publication type classification is built upon a BERT-based architecture, fine-tuned from the pretrained encoder microsoft/BiomedNLP-BiomedBERT-base-uncased-abstract-fulltext. This encoder is specifically trained with PubMed full-text articles. To

Table 1: Inclusion Criteria for Article Selection in the Datamart.

Criterion	Details
Data Sources	PubMed, MEDLINE, ClinicalTrials.gov
Language	English or French
Timeframe	No restrictions on publication date
Text Scope	Query terms searched across all sections (title, abstract, full text when available)
Query Structure	Includes combinations such as: • Drug 1 AND Drug 2 • Drug 1 AND Drug Family 2 • Drug Family 1 AND Drug 2 • Drug Family 1 AND Drug Family 2 • Single-drug queries
Inclusion Rule	Any article matching a query is included

enhance classification accuracy, different classification head variants were tested and compared. The best-performing architecture comprised a bi-directional Long Short-Term Memory (bi-LSTM) layer processing the concatenated last 2 hidden states (input dimension: 1,536), with the bi-LSTM output fed to a linear layer (output dimension: 10 classes). Since the context window is limited to 512 tokens, the model only processes the most relevant content of an article, ensuring only the most pertinent sections are fed to the model for analysis. The model is trained on a large dataset from PubMed and PubMed Central (PMC), where articles are classified into 10 distinct categories. The classification follows the global evidence pyramid, prioritizing Meta-Analyses, Systematic Reviews, and Randomized Controlled Trials above lower-evidence sources like observational studies and author opinions. There is no predefined threshold used in the model; rather, it leverages this evidence hierarchy to determine the importance of a publication in relation to drug-drug interactions.

2.2.2 DDI Sentence Classification

To identify sentences that mention drug-drug interactions (DDIs), a specialized sentence classification model based on the BERT architecture is employed. This model is fine-tuned from the pretrained encoder microsoft/BiomedNLP-BiomedBERT-base-uncased-abstract-fulltext, with a simple linear layer serving as the classification head (input dimension: 768, output dimension: 3 classes; dropout rate: 0.1 applied to the pooled [CLS] token representation).

The DDI sentence classifier was trained on a dataset of approximately 3,500 annotated sentences compiled from two sources: sentences manually extracted from biomedical publications indexed in PubMed and PMC, and sentences generated through language-model-based augmentation to improve coverage of underrepresented classes. The agreement between annotators was assessed using Cohen’s Kappa ($\kappa = 0.82$), indicating high inter-annotator reliability. The dataset was partitioned using a stratified split strategy to preserve the class distribution (DDI-positive, DDI-possible, DDI-negative) across subsets: 70% for training, 15% for validation, and 15% for testing. The test set was composed exclusively of sentences extracted from real publications, with no augmented data, to ensure that held-out evaluation reflects true in-distribution performance.

The model was trained using the weighted cross-entropy loss function, standard for multi-class classification. For an input sentence

with true class label y and predicted class probabilities p over $C = 3$ classes, the base loss is defined as $L = -\sum_{c=1 \text{ to } C} y_c \cdot \log(p_c)$, where y_c is the one-hot encoded ground-truth indicator for class c and p_c is the softmax probability output by the final classification head. To mitigate class imbalance (DDI-negative sentences being more prevalent than DDI-positive or DDI-possible), class weights inversely proportional to class frequency were applied, yielding the weighted variant $L_w = -\sum_{c=1 \text{ to } C} w_c \cdot y_c \cdot \log(p_c)$, where $w_c = N / (C \cdot N_c)$ and N_c is the number of samples in class c .

The model was fine-tuned using the AdamW optimizer with an initial learning rate of 2×10^{-5} , $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1 \times 10^{-8}$, and weight decay of 0.01. Training was conducted over 4 epochs with a batch size of 16. A linear learning rate warm-up was applied over the first 10% of total training steps, followed by linear decay to zero. The maximum input sequence length was set to 512 tokens, consistent with the BERT architecture’s positional embedding limit. Early stopping was employed with a patience of 2 epochs, monitoring validation loss to prevent overfitting.

Model training was conducted on a Linux server equipped with an NVIDIA A10G GPU (24 GB VRAM) and 64 GB of system RAM. The software environment used Python 3.11 with PyTorch 2.1, Hugging Face Transformers 4.35, and scikit-learn 1.3. CUDA 12.1 was used for GPU acceleration. Reproducibility was ensured by setting fixed random seeds across Python, NumPy, and PyTorch. The base pre-trained model was loaded from the Hugging Face model hub.

Sentences are classified into three categories: relevant, which contain explicit mentions of DDIs and their associated pharmacokinetic implications (e.g., “Coadministration of ketoconazole, a CYP3A4 inhibitor, significantly increased plasma levels of simvastatin, suggesting a pharmacokinetic interaction”); irrelevant, which do not mention DDIs at all (e.g., “Simvastatin is commonly prescribed for hyperlipidemia and is well tolerated by most patients”); and possibly relevant, which discuss the concomitant use of drugs without directly confirming or rejecting a DDI (e.g., “The patient was treated with simvastatin and clarithromycin during hospitalization”). The model is evaluated using precision, recall, and F1 scores, all of which exceed 0.85, and the overall accuracy surpasses 0.90, ensuring reliable DDI detection.

2.2.3 Safety Model

The safety model identifies sentences reporting safety concerns, specifically adverse events, by combining sentence classification with Named Entity Recognition (NER). Each sentence is classified as relevant, possibly relevant, or non-relevant. For sentences labeled relevant or possibly relevant, the NER component extracts the specific adverse event mentioned, using the MedDRA ontology. The training dataset consists of articles selected primarily by clinicians and pharmacists, who annotated sentences for safety relevance and adverse event types.

2.2.4 Patient Information

This model was applied to relevant articles identified in order to extract patient information related to the populations affected by the detected drug-drug interactions, including only the age, sex, and medical conditions. The model used is an open-source tool, d4data/biomedical-ner-all, which has been fine-tuned on a domain-specific corpus to recognize over 100 biomedical entities, enabling accurate extraction of patient-related data from the text.

2.2.5 Pipeline Integration

These models are integrated into a seamless, multi-stage pipeline that operates in sequence. The process begins with the classification of publication types, followed by sentence extraction to identify mentions of drug-drug interactions (DDIs) and safety information. Each model generates its output, which is then aggregated into a comprehensive scoring system. This ensures that the final output reflects both the relevance of the publication and its implications for safety or DDIs. Once relevant articles are identified through the scoring system, the patient information Named Entity Recognition (NER) model described in Section 2.2.4, which is distinct from the classifiers used for scoring, is applied to extract essential biomedical entities, such as patient data, from those articles.

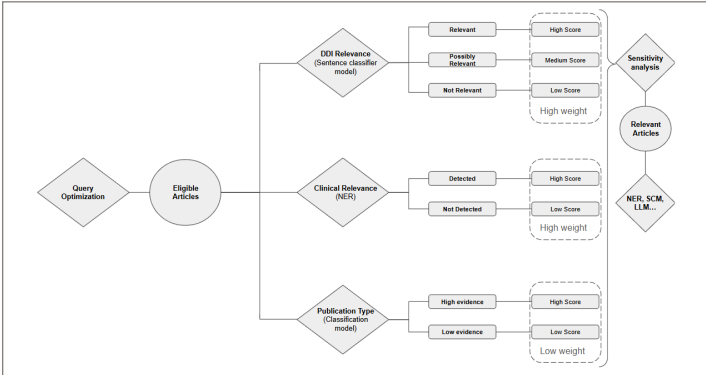


Figure 1: Diagram illustrating the pipeline overview. NER: named entity recognition; SCM: sentence classification model; LLM: large language model.

2.3 Scoring System Development

The scoring system was developed through extensive discussions with Vidal’s team and independent experts to identify the most critical components and assign appropriate weights to each. These experts included six pharmacists and two toxicologists. This process involved multiple iterations and fine-tuning of the coefficients attributed to each element to optimize the system’s accuracy and relevance. During this iterative process, the scoring components were defined first through expert consensus, followed by iterative adjustment of the weight coefficients against the gold standard dataset. The final framework integrated several key components, including title relevance (weight of 50 points), determined by the DDI sentence classifier (Section 2.2.2); mention of DDI impact (weight of 40 points), primarily derived from the safety model (Section 2.2.3); and publication type (weight of 10 points), calculated as the publication type score multiplied by a coefficient of 0.10. The rationale for assigning higher weights to title relevance and DDI impact reflects their stronger discriminative power in identifying clinically relevant articles. The resulting relevance

score ranges from 0 to 100, where higher scores indicate greater relevance of the article to the queried drug-drug interaction.

2.4 Validation Tests

To validate the scoring system’s ability to detect DDIs from scientific literature, the evaluation process was based on a representative gold standard collaboratively developed by two pharmacists from ArcaScience. This gold standard was created using rigorous guidelines to ensure objectivity and minimize bias. Specifically, the annotators were provided with the full list of 7,751 article titles retrieved for 15 sample queries: 5 Drug-Drug queries, 5 Drug-Drug Family queries, and 5 Single Drug queries. These examples were selected to represent a diverse range of DDI scenarios, enabling a comprehensive assessment of the system’s performance. The annotators were required to categorize each article using one of three labels: relevant, possibly relevant, or non-relevant. Relevance was determined based on the article title alone. A title was considered relevant if it explicitly indicated that the article dealt with the DDI in question; possibly relevant if the title suggested a potential relation, but did not directly mention the DDI topic; and non-relevant if the title was unrelated to the DDI subject. It is important to note that these guidelines slightly differed depending on the query type and were iteratively refined during the process based on ambiguous or borderline cases encountered, ensuring progressive alignment between annotators. The differences concerned specifically the definition of ‘possibly relevant’, which was adapted for each query type to account for family-level mentions and interaction context specificity. Inter-annotator agreement was assessed using Cohen’s kappa coefficient, which yielded a value of $\kappa = 0.83$, indicating almost perfect agreement between the two annotators. This process led to four possible outcomes: true positive, false positive, true negative, and false negative. These results were then used to calculate the sensitivity and specificity of the models for each possible threshold. A ROC/AUC evaluation was conducted via a Receiver Operating Characteristic (ROC) curve to identify the optimal relevancy score threshold, ensuring the accurate identification of relevant articles. Sensitivity and specificity were calculated for every 5-point difference in the threshold.

Our evaluation strategy aimed to balance high precision with adequate recall, as both metrics address distinct clinical needs in DDI detection. Precision is critical for reducing over-alerting, a major limitation of existing DDI systems that leads to alert fatigue and reduced clinician compliance. Simultaneously, recall remains important to ensure that clinically significant interactions are not missed. However, we recognized that truly important DDIs are typically reported across multiple publications in the literature; therefore, extremely high recall was deprioritized in favor of precision, as critical interactions are unlikely to be represented by only a single article. This approach optimizes the practical utility of the system by providing clinicians with a focused, high-quality set of relevant articles rather than an overwhelming volume of results.

The efficiency of the scoring system across components was further evaluated using one-way ANOVA, with Tukey’s post-hoc test applied for pairwise comparisons (significance threshold: $p < 0.05$ for both tests). Standard performance metrics, including precision,

which measures the proportion of correctly predicted positive instances among all predicted positives; recall, which quantifies the ability to identify all actual positives; F1-score, the harmonic mean of precision and recall that balances both metrics; and accuracy, which represents the proportion of correctly classified instances among all predictions, were calculated based on the gold standard to provide detailed insights into procedural effectiveness.

Finally, the results of the scoring system (for the optimal threshold) were benchmarked against Vidal’s experts lists (a list of references computed manually by the team for internal purposes), allowing for direct comparison of performance and coverage. This benchmarking process highlighted the model’s competitive edge and areas for potential improvement.

3. Results

3.1 Performance Metrics

3.1.1 Threshold Detection

The results of the individual sensitivity analyses (for each query type independently) and the ROC/AUC evaluation for the global dataset both exhibited a typical sensitivity curve, indicating an optimal threshold of 75%.

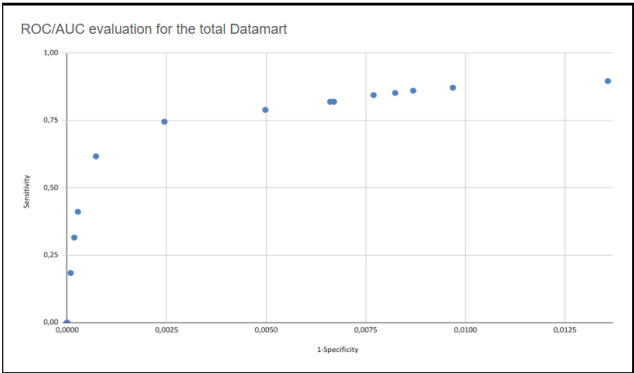


Figure 2: ROC/AUC evaluation (ROC Curve) for the global data mart derived from the three types of queries, illustrating that the optimal threshold is 75%. The analysis was conducted using threshold values ranging from 100% to 0%, with intervals of 5% units between consecutive points.

3.1.2 Drug-Drug Family Query

The performance metrics for the drug-drug family query demonstrated strong precision (0.941), while maintaining robust recall (0.813) and achieving an F1 score of 0.872 (Table 2). The differentiation between relevant, possibly relevant, and non-relevant articles was validated through ANOVA and Tukey post-hoc analysis, confirming the system’s ability to stratify articles accurately (Figure 3).

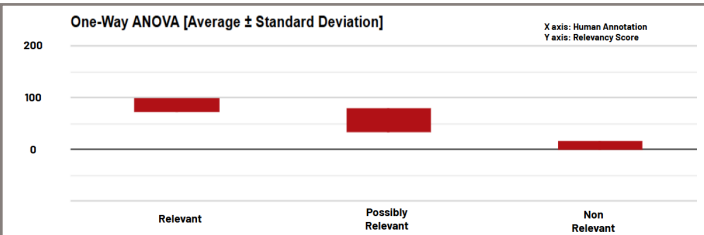


Figure 3: Bar graph of ANOVA results for the drug-drug family query, showing a significant difference between the three groups: Relevant, Possibly Relevant, and Non-Relevant ($p < 0.05$).

3.1.3 Drug-Drug Query

The performance metrics for drug-drug queries varied, with F1 scores differing based on the level of available documentation (average: 0.684) (Table 2). Differentiation between relevant, possibly relevant, and non-relevant articles was validated using ANOVA and Tukey post-hoc analysis (Figure 4).

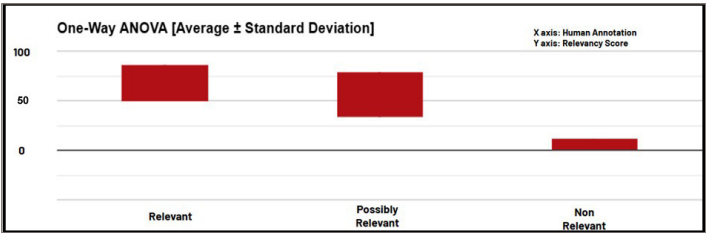


Figure 4: Bar graph of ANOVA results for the drug-drug query, showing a significant difference between two groups: Relevant with Possibly Relevant and Non-Relevant ($p<0.05$).

3.1.4 Single Drug Query

For single drug queries, the F1 score was 0.731 (Table 2; formulas provided in Table 3). Differentiation between relevant, possibly relevant, and non-relevant articles was validated through ANOVA and Tukey post-hoc analysis, further confirming the robustness of the system’s classification (Figure 5).

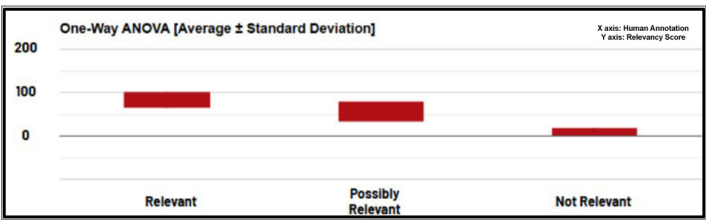


Figure 5: Bar graph of ANOVA results for the single drug query, showing a significant difference between two groups: Relevant with Possibly Relevant and Non-Relevant ($p<0.05$).

Table 2: Performance Metrics of the Scoring System for Each Query Type.

Metric	Drug-Drug Family	Drug-Drug	Single Drug
Recall	0.813 (0.764 - 0.856)	0.583 (0.489 - 0.673)	0.682 (0.462 - 0.850)
Precision	0.941 (0.901 - 0.969)	0.829 (0.726 - 0.904)	0.789 (0.544 - 0.929)
F1	0.872 (0.835 - 0.905)	0.684 (0.602 - 0.758)	0.731 (0.546 - 0.870)
Accuracy	0.983 (0.976 - 0.986)	0.993 (0.991 - 0.995)	0.992 (0.987 - 0.996)
Specificity	0.996 (0.993 - 0.998)	0.998 (0.997 - 0.999)	0.997 (0.993 - 0.999)
Negative Predictive Value	0.984 (0.979 - 0.989)	0.994 (0.992 - 0.996)	0.995 (0.991 - 0.998)

Table 3: Calculation formulas for the validation metrics.

Metric	Formula
Recall (Sensitivity)	$TP / (TP + FN)$
Precision	$TP / (TP + FP)$
F1-score	$2 \times (Precision \times Recall) / (Precision + Recall)$
Accuracy	$(TP + TN) / (TP + TN + FP + FN)$
Specificity	$TN / (TN + FP)$
Negative Predictive Value (NPV)	$TN / (TN + FN)$

3.1.5 Enhanced Flexibility and Comprehensive Coverage in DDI Analysis

The comparison with the lists provided by Vidal’s scientific teams was carried out using two different types of queries. In the first example, focusing on proton pump inhibitors and clopidogrel, ArcaScience identified 155 relevant articles, whereas Vidal’s list contained 43 relevant articles. Among these, 26 articles were common to both sources. The 17 missing articles from ArcaScience were due to unsupported languages or databases, as we currently only process French and English content and use PMC and Medline as our data sources, as well as removed articles and limitations in the algorithm’s recall. Another example, examining the interaction between risperidone and selective serotonin reuptake inhibitors, showed that ArcaScience provided approximately twice as many relevant articles as Vidal’s manual list. These examples demonstrate that ArcaScience’s pipeline provides substantially broader coverage than manually curated databases, while maintaining high precision. This advantage can be attributed to the dynamic nature of article selection, the absence of a need for manual annotation, and the flexibility the algorithm provides.

3.2 Case Studies

3.2.1 Specific drug-drug queries

A well-known example of drug-family interactions is the interaction between macrolides and statins. Searching for their interaction can generate hundreds of articles, ranging from general discussions to studies focused on specific drug pairs. However, not all drugs within the same family share the same interaction profile: for instance, azithromycin has a lower impact on cytochromes and thus interacts less compared to clarithromycin and erythromycin. To address this variability, users may need to find evidence specific to a particular drug interaction. Our algorithm optimizes article ranking by adjusting their order based on drug modifications, ensuring that users receive the most relevant and accurate information (Table 4).

Table 4: The first three articles found for each drug pair, demonstrating that directly related articles are prioritized at the top as key references.

Drug Pair	Relevant Articles
Simvastatin + Clarithromycin	- Assessment of potential interaction between simvastatin and clarithromycin in healthy adult male subjects [17]. - Rhabdomyolysis in association with simvastatin and dosage increment in clarithromycin [18]. - Rhabdomyolysis secondary to a drug interaction between simvastatin and clarithromycin [19].
Erythromycin + Atorvastatin	- Statin toxicity from macrolide antibiotic coprescription: a population-based cohort study [20]. - Risk management of simvastatin or atorvastatin interactions with CYP3A4 inhibitors [21]. - Drug safety of macrolide and quinolone antibiotics in a tertiary care hospital: administration of interacting co-medication and QT prolongation [22].
Lovastatin + Azithromycin	- Lovastatin-induced rhabdomyolysis possibly associated with clarithromycin and azithromycin [23]. - Statin-induced rhabdomyolysis from azithromycin interaction in a patient with heterozygous SLCO1B1 polymorphism [24]. - Antibiotics and Lipid-Modifying Agents: Potential Drug-Drug Interactions and Their Clinical Implications [25].

3.2.2 Drug-Drug Family Interactions

Our algorithm demonstrates its capacity to detect not only individual drug-drug interactions but also broader drug-drug family interactions, such as those occurring between risperidone and SSRIs. It successfully identifies both pharmacokinetic interactions (such as CYP2D6 inhibition leading to altered risperidone plasma levels) and pharmacodynamic interactions, including additive serotonergic or cardiac effects. Importantly, it also captures the clinical implications of these interactions, which range from metabolic and endocrine disturbances to neurotoxic and cardiovascular complications. These complications include, for example, severe weight gain, documented in “Severe weight gain induced by combination treatment with risperidone and paroxetine” [26], and gynecomastia, as described in “Gynecomastia with risperidone-fluoxetine combination” [27]. Other PD-related consequences include serotonin syndrome, reported in “Serotonin syndrome during treatment with paroxetine and risperidone” [28], and torsade de pointes, noted in “Torsade de pointes during combined treatment with risperidone and citalopram” [29]. These findings highlight the algorithm’s robustness in surfacing clinically relevant interactions that span mechanistic categories and therapeutic risks within a drug-family interaction framework.

3.2.3 Drug interaction profiling

An additional need for interaction profiling of a single drug could arise when prescribing new medications with limited research. In such cases, it is essential to back up prescriptions with solid evidence. Having a list of significant articles related to the drug’s interactions becomes crucial. For instance, Siponimod, a drug with fewer than 500 articles mentioning it, has only around 10 relevant studies about its potential DDIs. Examples of these studies include: “Siponimod pharmacokinetics, safety, and tolerability in combination with the potent CYP3A4 inhibitor itraconazole in healthy subjects with different CYP2C9 genotypes” [30], “Effect of Fluconazole Coadministration and CYP2C9 Genetic Polymorphism on Siponimod Pharmacokinetics in Healthy Subjects” [31], among others. The algorithm effectively identified these articles and summarized the interaction profile of this recent drug.

4. Discussion

4.1 Key findings

This study shows the potential of the scoring system to improve how drug-drug interactions are detected, referenced, and managed by addressing major limitations of current methods, namely slow update cycles and limited coverage. It works alongside existing static references, adding the ability to generate on-demand, evidence-based reports when data is available, while still allowing for expert input when it’s not. By integrating diverse metrics, such as publication type, pharmacokinetic and pharmacodynamic effects, and patient-specific factors, the system enhances the clinical value of the insights it delivers.

4.2 Strengths and limitations

The system’s enhanced flexibility and comprehensive coverage further distinguish it from traditional tools. This system dynamically analyzes the latest literature, ensuring up-to-date and com-

prehensive coverage of potential DDIs. This feature is particularly valuable in the context of rare or emerging drug interactions that may not yet be documented in conventional databases.

With strong performance metrics, an F1 score of 0.872 for drug-drug family queries and a precision of 0.941, the system effectively prioritizes relevant interactions and the supporting evidence. These capabilities have significant implications for pharmacovigilance, as they enable healthcare professionals to identify potential safety concerns more efficiently and tailor treatment regimens to individual patient needs.

Despite its promising performance, our scoring system has some limitations. In particular, its recall could be improved by increasing the size of the training datasets and refining the scoring methods. Currently, the recall for drug-drug family queries is 0.813, which is noticeably higher than the recall for drug-drug queries (0.583) and single drug queries (0.682). Enhancing the model’s performance for drug-drug and single drug queries to match that of drug-drug family queries remains an important goal.

In addition, the system relies solely on evidence from published literature, which inherently includes gaps and missing data and lags real-world clinical phenomena due to publication timelines. Incorporating new data sources, such as real-world evidence, and additional scientific databases, may improve its comprehensiveness and performance. However, electronic health records carry substantial limitations, including data quality issues, coding errors, and confounding factors that require careful methodological consideration. Also, this system does not predict DDIs for new molecules, as it depends on existing evidence, introducing a time lag between drug approval and literature availability. It’s also important to note that pharmacodynamic (PD) interactions are often underrepresented in the literature, precisely because many of them are mechanistically predictable and do not always warrant publication as clinical alerts.

4.3 Comparison with similar research

When compared to other systems in the DDI domain, such as the SFINX database (INXBASE from Medbase) [32], ArcaScience scoring system demonstrates several advantages. SFINX is recognized for its robust and evidence-based approach to DDI detection, relying on expert-reviewed content and predefined drug pairs (8000 pairs). However, it shares limitations with other traditional systems, including a static database structure and limited flexibility to incorporate newly emerging drug interactions or adapt to novel queries.

The literature mentions more than 10 databases and web servers that address the subject of drug-drug interactions, each with its own approach to handling and presenting the data [33]. These platforms, such as DDInter, SuperDRUG2, DIDB, DrugComb, Drugs.com, Drugbank, and others, offer various methods for recording and displaying DDIs, along with additional drug-related information like pharmacokinetics, side effects, and drug targets. For example, DDInter, available only in China, not only records 236 834

DDIs involving 1833 drugs but also documents the detailed information about each DDI, such as mechanisms and risk levels [34].

A study by Kin Wah Fung et al. compared three commercial knowledge bases (First DataBank, Micromedex, and Multum) for DDI detection and found that, while varying in size and alert generation, all provided 99.8–99.9% coverage of the Office of the National Coordinator for Health Information Technology (ONC) list, highlighting the need for standardization to reduce variations among knowledge sources [35]. Similarly, another study from Elpidia Kotsioti et al. explored the level of agreement on DDI information across three major online drug information resources (DIRs), the British National Formulary (BNF) [36], ANSM (Agence nationale de sécurité du médicament et des produits de santé) Thesaurus [37], and Micromedex [38], and found considerable variation in the drug pairs involved, severity ratings, evidence ratings, and clinical management recommendations. This variation emphasizes the need for harmonization and standardization in DDI information across different knowledge sources [39]. A concrete example is a comparison between EPIC and the Seegnal platform, which further highlights these limitations: EPIC generated over 90% more alerts with markedly lower specificity, contributing to alert fatigue, whereas Seegnal produced fewer, more clinically relevant alerts and could have prevented several adverse events [40].

In the academic area, recent studies have leveraged transformer-based models and deep learning architectures to achieve high performance in both drug named entity recognition and DDI classification from literature. For instance, transformer frameworks such as SciBERT and BioBERT have been employed to sequentially extract drug entities and classify their interactions within scientific articles, reporting strong F1-scores for both tasks [41,42]. Similarly, long short-term memory (LSTM) networks augmented by sentiment analysis [43] have not only detected DDI events but also evaluated interaction severity, which is a functionality closely aligned with our system's safety and adverse event extraction components. Comparative evaluations across rule-based, deep learning, and language model-based approaches [44] further underscore the potential of modern NLP models to capture subtle, context-dependent drug-event relationships from large, heterogeneous text corpora.

Simultaneously, a wave of graph-based and multi-view fusion models has emerged, integrating molecular, network, and pharmacological data to predict or classify DDIs [45–48]. These approaches often surpass traditional similarity-based or profile-driven algorithms by capturing more complex inter-drug relationships and contextual features, although, unlike our literature-centric pipeline, many focus primarily on data-driven event prediction rather than dynamic literature extraction and interpretability.

Taken together, these recent advancements highlight the rapid progress in AI-powered DDI detection and the growing recognition of literature mining as a vital complement to static knowledge bases. Our approach contributes to this landscape by combining robust NLP-driven extraction with a clinically oriented scoring and ranking system, aiming to bridge the gap between high-throughput

computational methods and the nuanced, context-rich evidence required for real-world pharmacotherapy decision support, while also offering unique capabilities such as multilingual text processing, real-time literature updates, and patient-centered risk scoring to enhance clinical relevance and adaptability.

4.4 Explanations of findings

These findings demonstrate that, thanks to artificial intelligence, it is now possible to generate an unlimited list of publications related to drug-drug interactions, as long as relevant literature exists. Unlike the current state of the art, which relies on automatic queries but still depends heavily on manual contributions, this AI-powered approach significantly reduces the need for manual input.

4.5 Implications and actions needed

The use of this technology in clinical decision support systems is a crucial step to enable clinicians to benefit from the comprehensive results, thereby improving the safety of their prescriptions. Additionally, integrating this approach into benefit-risk assessments would be particularly important, especially for polymedicated patients. Finally, feedback loops from end-users can refine the methodology and ensure clinical alignment.

5. Conclusions

This study demonstrates that an AI-powered scoring system can successfully overcome key limitations of traditional DDI detection methods. The system generates dynamic, evidence-based DDI reports unconstrained by static drug pair lists, provides literature-backed references to substantiate clinical alerts, and enables evidence-based alert prioritization through quantitative relevance scoring. With high precision across all query types and a validated relevance threshold, the system delivers accurate, clinically actionable insights. By leveraging diverse metrics, including publication type, clinical impacts, and patient-specific contexts, this approach addresses critical gaps in traditional tools and offers a scalable solution for enhancing clinical decision support systems.

Conflict of Interest and Funding

The authors declare no conflict of interest. This study was a collaborative effort between ArcaScience and VIDAL, with funding provided by VIDAL. The decision to submit the manuscript was made jointly by both parties.

Acknowledgements

The authors thank the pharmacists and toxicologists who contributed their expertise to the development of the scoring system's weighting parameters.

References

1. Magro L, Moretti U, Leone R. Epidemiology and characteristics of adverse drug reactions caused by drug-drug interactions. *Expert Opin Drug Saf.* 2012; 11: 83–94.
2. Niu J, Straubinger RM, Mager DE. Pharmacodynamic drug-drug interactions. *Clin Pharmacol Ther.* 2019; 105: 1395–1406.

3. Dhabali AAH, Awang R, Zyoud SH. Clinically important drug-drug interactions in primary care. *J Clin Pharm Ther.* 2012; 37: 426-430.
4. Qato DM, Wilder J, Schumm LP, Gillet V, Alexander GC. Changes in prescription and over-the-counter medication and dietary supplement use among older adults in the United States, 2005 vs 2011. *JAMA Intern Med.* 2016; 176: 473-482.
5. Day RO, Snowden L, McLachlan AJ. Life-threatening drug interactions: what the physician needs to know. *Intern Med J.* 2017; 47: 501-512.
6. VIDAL. Available at: <https://www.vidal.fr/>
7. DrugBank. Available at: <https://go.drugbank.com/>
8. Smithburger PL, Buckley MS, Bejian S, Burenheide K, Kane-Gill SL. A critical evaluation of clinical decision support for the detection of drug-drug interactions. *Expert Opin Drug Saf.* 2011; 10: 871-882.
9. van der Sijs H, Aarts J, Vulto A, Berg M. Overriding of drug safety alerts in computerized physician order entry. *J Am Med Inform Assoc.* 2006; 13: 138-147.
10. Phansalkar S, van der Sijs H, Tucker AD, Desai AA, Bell DS, et al. Drug-drug interactions that should be non-interruptive in order to reduce alert fatigue in electronic health records. *J Am Med Inform Assoc.* 2013; 20: 489-493.
11. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* 2018; 46: D1074-D1082.
12. Gill J, Moullet M, Martinsson A, Miljković F, Williamson B, et al. Evaluating the performance of machine-learning regression models for pharmacokinetic drug-drug interactions. *CPT Pharmacometrics Syst Pharmacol.* 2023; 12: 122-134.
13. Ryu JY, Kim HU, Lee SY. Deep learning improves prediction of drug-drug and drug-food interactions. *Proc Natl Acad Sci U S A.* 2018; 115: E4304-E4311.
14. Bouraoui A, Jamoussi S, Ben Hamadou A. A comprehensive review of deep learning for natural language processing. *Int J Data Min Model Manag.* 2022; 14: 149-182.
15. Otter DW, Medina JR, Kalita JK. A survey of the usages of deep learning for natural language processing. *IEEE Trans Neural Netw Learn Syst.* 2021; 32: 604-624.
16. Vilar S, Harpaz R, Uriarte E, Santana L, Rabadan R, et al. Drug-drug interaction through molecular structure similarity analysis. *J Am Med Inform Assoc.* 2012; 19: 1066-1074.
17. Kaleem Z, Khan JA, Mushtaq Z, Altaf S, Javed I. Assessment of potential interaction between simvastatin and clarithromycin in healthy adult male subjects. *Pak J Pharm Sci.* 2018; 31: 801-806.
18. Page SR, Yee KC. Rhabdomyolysis in association with simvastatin and dosage increment in clarithromycin. *Intern Med J.* 2014; 44: 690-693.
19. Lee AJ, Maddix DS. Rhabdomyolysis secondary to a drug interaction between simvastatin and clarithromycin. *Ann Pharmacother.* 2001; 35: 26-31.
20. Patel AM, Shariff S, Bailey DG, Juurlink DN, Gandhi S, et al. Statin toxicity from macrolide antibiotic coprescription: a population-based cohort study. *Ann Intern Med.* 2013; 158: 869-876.
21. Molden E, Skovlund E, Braathen P. Risk management of simvastatin or atorvastatin interactions with CYP3A4 inhibitors. *Drug Saf.* 2008; 31: 587-596.
22. Niedrig D, Maechler S, Hoppe L, Corti N, Kovari H, et al. Drug safety of macrolide and quinolone antibiotics in a tertiary care hospital: administration of interacting co-medication and QT prolongation. *Eur J Clin Pharmacol.* 2016; 72: 859-867.
23. Grunden JW, Fisher KA. Lovastatin-induced rhabdomyolysis possibly associated with clarithromycin and azithromycin. *Ann Pharmacother.* 1997; 31: 859-863.
24. Burns H, Russell L, Cox ZL. Statin-induced rhabdomyolysis from azithromycin interaction in a patient with heterozygous SLCO1B1 polymorphism. *J Clin Pharm Ther.* 2021; 46: 853-855.
25. Spanakis M, Alon-Ellenbogen D, Ioannou P, Spornovasilis N. Antibiotics and lipid-modifying agents: potential drug-drug interactions and their clinical implications. *Pharmacy (Basel).* 2023; 11: 130.
26. Fukui H, Murai T. Severe weight gain induced by combination treatment with risperidone and paroxetine. *Clin Neuropharmacol.* 2002; 25: 269-271.
27. Benazzi F. Gynecomastia with risperidone-fluoxetine combination. *Pharmacopsychiatry.* 1999; 32: 41.
28. Hamilton S, Malone K. Serotonin syndrome during treatment with paroxetine and risperidone. *J Clin Psychopharmacol.* 2000; 20: 103-105.
29. Blaschke D, Parwani AS, Huemer M, Rolf S, Boldt LH, et al. Torsade de pointes during combined treatment with risperidone and citalopram. *Pharmacopsychiatry.* 2007; 40: 294-295.
30. Gardin A, Shakeri-Nejad K, Feller A, Huth F, Neelakantham S, et al. Siponimod pharmacokinetics, safety, and tolerability in combination with the potent CYP3A4 inhibitor itraconazole in healthy subjects with different CYP2C9 genotypes. *Eur J Clin Pharmacol.* 2019; 75: 1565-1574.
31. Gardin A, Ufer M, Legangneux E, Rossato G, Jin Y, et al. Effect of fluconazole coadministration and CYP2C9 genetic polymorphism on siponimod pharmacokinetics in healthy subjects. *Clin Pharmacokinet.* 2019; 58: 349-361.
32. Böttiger Y, Laine K, Andersson ML, Korhonen T, Molin B, et al. SFINX: a drug-drug interaction database designed for clinical decision support systems. *Eur J Clin Pharmacol.* 2009; 65: 627-633.

33. Zhao Y, Yin J, Zhang L, Zhang Y, Chen X. Drug-drug interaction prediction: databases, web servers and computational models. *Brief Bioinform.* 2023; 25: bbad445.
34. Xiong G, Yang Z, Yi J, Wang N, Wang L, et al. DDInter: an online drug-drug interaction database towards improving clinical decision-making and patient safety. *Nucleic Acids Res.* 2022; 50: D1200-D1207.
35. Fung KW, Kapusnik-Uner J, Cunningham J, Higby-Baker S, Bodenreider O. Comparison of three commercial knowledge bases for detection of drug-drug interactions in clinical decision support. *J Am Med Inform Assoc.* 2017; 24: 806-812.
36. National Institute for Health and Care Excellence. BNF: British National Formulary. Available at: <https://bnf.nice.org.uk/>
37. Agence nationale de sécurité du médicament et des produits de santé. Thésaurus des interactions médicamenteuses. Available at: <https://ansm.sante.fr/documents/reference/thesaurus-des-interactions-medicamenteuses-1>
38. IBM Watson Health. Micromedex. Available at: <https://www.micromedexsolutions.com/>
39. Kontsioti E, Maskell S, Bensalem A, Dutta B, Pirmohamed M. Similarity and consistency assessment of three major online drug-drug interaction resources. *Br J Clin Pharmacol.* 2022; 88: 4067-4079.
40. Shah SN, Seger DL, Fiskio JM, Horn JR, Bates DW. Comparison of medication alerts from two commercial applications in the USA. *Drug Saf.* 2021; 44: 661-668.
41. Zaikis D, Kokkas S, Vlahavas IP. Transforming drug-drug interaction extraction from biomedical literature. In: *Proceedings of the 12th Hellenic Conference on Artificial Intelligence.* New York, NY: Association for Computing Machinery; 2022: 1-8.
42. Zaikis D, Vlahavas I. TP-DDI: Transformer-based pipeline for the extraction of drug-drug interactions. *Artif Intell Med.* 2021; 119: 102153.
43. Salman M, Munawar HS, Latif K, Akram MW, Khan SI, et al. Big data management in drug-drug interaction: a modern deep learning approach for smart healthcare. *Big Data Cogn Comput.* 2022; 6: 30.
44. Mahendran D, McInnes BT. Extracting adverse drug events from clinical notes. *AMIA Jt Summits Transl Sci Proc.* 2021; 2021: 420-429.
45. Lin K, Kang L, Yang F, Lu P, Lu J. MFDA: multiview fusion based on dual-level attention for drug interaction prediction. *Front Pharmacol.* 2022; 13: 1021329.
46. Yan X, Gu C, Feng Y, Han J. Predicting drug-drug interaction with graph mutual interaction attention mechanism. *Methods.* 2024; 223: 16-25.
47. Pan D, Lu P, Wu Y, Kang L, Huang F, et al. Prediction of multiple types of drug interactions based on multi-scale fusion and dual-view fusion. *Front Pharmacol.* 2024; 15: 1354540.
48. Kang LP, Lin KB, Lu P, Yang F, Chen JP. Multitype drug interaction prediction based on the deep fusion of drug features and topological relationships. *PLoS One.* 2022; 17: e0273764.