



XXXVIII PhD Course in Applied Biology and Experimental Medicine

Department of Chemical, Biological, Pharmaceutical and
Environmental Sciences - SSD BIOS-11/A - Pharmacology

University of Messina

Lelio Crupi

**Developing an operational framework for high-severity
potential drug-drug interactions**

Supervisor: Prof.ssa Emanuela Esposito

Co-Supervisor: Prof. Francesco Salvo

Coordinator: Prof.ssa Emanuela Esposito

Year 2024/2025

TABLE OF CONTENTS

ABSTRACT	1
LIST OF ABBREVIATIONS.....	4
1 - GENERAL INTRODUCTION	6
1.1 DRUG-DRUG INTERACTIONS AND POTENTIAL DRUG-DRUG INTERACTIONS	7
1.1.1 Definitions and mechanisms.....	7
1.1.2 Pharmacological determinants of DDI vulnerability.....	9
1.2 PATIENT-RELATED RISK FACTORS	11
1.2.1 Number of medications.....	11
1.2.2 Multimorbidity.....	12
1.2.3 Age.....	13
1.2.4 Genetic polymorphisms	14
1.3 PREVALENCE OF DDIs AND PDDIs	15
1.4 SOURCES AND METHODS FOR IDENTIFYING PDDIs.....	17
1.4.1 Summary of Product Characteristics (SmPC).....	17
1.4.2 National or institutional reference documents	18
1.4.3 Commercial interaction checkers (ICs).....	19
1.4.4 Clinical Decision Support Systems.....	20
1.4.5 Medication review	22
1.5 STRATEGIES TO IMPROVE THE IDENTIFICATION, PRIORITIZATION, AND EVIDENCE BASE OF PDDIs.....	23
1.5.1 Standardization of evidence and information in pDDI sources	24
1.5.2 Prioritization and contextual filtering.....	25
1.5.3 AI as support for prescribing.....	26
1.5.4 Validation/Updating with Real-World Data	28
2 - THESIS OBJECTIVES	32

3 - STANDARDIZING AND COMPARING MANAGEMENT RECOMMENDATIONS FOR POTENTIAL DRUG-DRUG INTERACTIONS ACROSS DIFFERENT INTERACTION CHECKERS	35
3.1 CONTEXT AND OBJECTIVES	36
3.2 ARTICLE.....	38
<i>Abstract</i>	38
<i>Introduction</i>	39
<i>Methods</i>	40
<i>Results</i>	49
<i>Discussion</i>	55
<i>Conclusions</i>	60
3.3 LESSONS LEARNED AND METHODOLOGICAL CONSIDERATIONS	61
4 - DEVELOPMENT OF A REAL-WORLD DERIVED GOLD STANDARD FOR HIGH-SEVERITY DRUG-DRUG INTERACTIONS AND ITS APPLICATION TO THE VALIDATION OF AN AUTOMATED DETECTION TOOL	63
4.1 CONTEXT AND OBJECTIVES.....	64
4.2 ARTICLE.....	66
<i>Abstract</i>	67
<i>Introduction</i>	69
<i>Methods</i>	71
<i>Results</i>	80
<i>Discussion</i>	88
<i>Conclusions</i>	93
4.3 LESSONS LEARNED AND METHODOLOGICAL CONSIDERATIONS	94
5 - DISCORDANCE ANALYSIS TO REFINE AN IMPERFECT GOLD STANDARD FOR NLP-BASED IDENTIFICATION OF HIGH-SEVERITY POTENTIAL DRUG-DRUG INTERACTIONS	96
5.1 CONTEXT AND OBJECTIVES	97
5.2 ARTICLE.....	98
<i>Abstract</i>	98

<i>Introduction</i>	100
<i>Methods</i>	101
<i>Results</i>	106
<i>Discussion</i>	113
<i>Conclusions</i>	119
5.3 LESSONS LEARNED AND METHODOLOGICAL CONSIDERATIONS	120
6 - GENERAL DISCUSSION, PERSPECTIVES, AND CONCLUSIONS.....	122
6.1 GENERAL DISCUSSION.....	123
6.2 PERSPECTIVES.....	127
6.3 CONCLUSIONS.....	130
7 - SUPPLEMENTS	132
7.1 SUPPLEMENTS CHAPTER 3	133
7.2 SUPPLEMENTS CHAPTER 4	158
7.3 SUPPLEMENTS CHAPTER 5	163
8 - REFERENCES.....	166

ABSTRACT

Potential drug-drug interactions (pDDIs) are highly prevalent, particularly in older adults exposed to polypharmacy. They are identified using heterogeneous sources, including Summary of Product Characteristics, regulatory reference documents, and interaction checkers (ICs), often embedded in clinical decision support systems. However, the clinical relevance of many pDDIs remains uncertain, and substantial discordance exists across sources in terms of pDDI inclusion and severity grading. In addition, the large number of low-priority pDDIs presented to prescribers by these tools contributes to alert fatigue and limits their use and usefulness. The generally low level of evidence supporting many pDDIs appears to be at least partially responsible for both these issues. Improving knowledge in this field therefore requires, in parallel, strengthening the evidence base for pDDIs and improving the way they are operationalized in clinical tools.

The aim of this thesis was to develop a methodological framework to improve the identification and prioritization of high-severity pDDIs. Specifically, it aimed to: (1) quantify discordances across commonly used pDDI sources in terms of management recommendations; (2) develop a reliable gold standard for high-severity pDDIs based on the ANSM Thesaurus, the French reference document on pDDIs; and (3) evaluate and refine this gold standard through systematic comparison with a natural language processing (NLP)-based automated detection tool.

Three studies were conducted to achieve these objectives.

The first study developed a structured scale to classify pDDIs into four management-based priority categories and applied it to 218 real-world pDDIs identified in 1,923 patients. Six major pDDI sources were compared. Overall agreement between sources was moderate ($AC1 = 0.44$).

Agreement improved when analyses focused only on the classification of the highest-priority pDDIs versus all others. Nevertheless, substantial heterogeneity in management recommendations across sources was observed.

The second study developed a gold standard for high-severity pDDIs based on the ANSM Thesaurus. More than 30,000 unique drug pairs extracted from discharge prescriptions of 6,027 hospitalized patients aged ≥ 65 years were independently annotated by two experts using a predefined decision algorithm, with adjudication of discordant cases. Inter-rater agreement was very good (AC1 = 0.94). High-severity pDDIs were identified in 9.7% of patients and were strongly associated with the number of drugs prescribed. The gold standard was then used to evaluate an NLP-based detection tool (PRoSIT), which showed moderate sensitivity (63.1%) but high negative predictive value (96.0%).

The third study examined discordances between PRoSIT and the gold standard. All discrepancies were reviewed to determine whether they were due to algorithmic limitations or to human annotation errors in the gold standard. Among false negatives, 81.0% were attributable to human-related errors. After correcting these errors, sensitivity increased from 63.1% to 89.4% and the negative predictive value from 96.0% to 99.2%. Most errors were concentrated in a small number of recurring pDDI clusters.

Overall, this thesis shows that improving the identification of high-severity pDDIs requires both better references and advanced automated tools. Gold standards derived from reference documents should be considered evolving resources that require periodic revision. The concentration of high-severity pDDIs in a limited number of recurrent drug combinations also supports targeted educational strategies and focused alerting systems. Future work should combine incremental improvement of reference sources with outcome-based studies evaluating

the clinical impact of high-priority pDDIs, in order to improve knowledge about pDDIs, source quality, and operational efficiency.

LIST OF ABBREVIATIONS

ADR: adverse drug reaction

AI: artificial intelligence

ANSM: Agence Nationale de Sécurité du Médicament et des produits de santé (French national agency for medicines and health products safety)

AUC: area under the curve

BNF: British National Formulary

CDSS: clinical decision support systems

CI: confidence interval

CYP: cytochrome P450

DDI: drug-drug interactions

EHR: electronic health records

FN: false negative

FP: false positive

IC: interaction checker

NLP: natural language processing

NPV: negative predictive value

PD: pharmacodynamic

pDDI: potential drug-drug interaction

PK: pharmacokinetic

PPV: positive predictive value

PV: pharmacovigilance

Romedi: *référentiel ouvert du médicament (a French drug referential for semantic interoperability)*

SmPC: Summary of Product Characteristics

SRS: spontaneous reporting systems

TN: true negative

TP: true positive

1 - General Introduction

1.1 Drug-drug interactions and potential drug-drug interactions

1.1.1 Definitions and mechanisms

A drug-drug interaction (DDI) can be defined as a modification of the pharmacological effect or concentration of a drug caused by the presence of another substance (1). In most cases, DDIs originate from the co-administration of two drugs. However, they may also occur in the absence of direct temporal overlap when the precipitating drug has a long elimination half-life or when it induces or inhibits enzymatic systems, transporters, or receptors with effects that persist after drug discontinuation, thereby modifying the response to a subsequently administered drug (2,3). DDIs are generally classified according to their mechanism as pharmacokinetic (PK) or pharmacodynamic (PD) in nature (4).

PK DDIs occur when a *precipitant drug* modifies the absorption, distribution, metabolism, or elimination of another drug, referred to as the *object drug* (5). A classification of the main mechanisms of PK DDIs is presented in Table 1.

PD DDIs, in contrast, occur when two drugs act on the same physiological systems or molecular pathways, producing additive, synergistic, or antagonistic effects (6). A PD DDI is considered additive when the effect of a drug combination corresponds to the sum of the pharmacological effects of the individual substances. It is considered synergistic when the combined effect exceeds the sum of the individual effects. It is considered antagonistic when the combined effect is lower than expected based on the effects of the individual agents (6,7).

Table 1 – Major pharmacokinetic mechanisms of DDIs

Pharmacokinetic phase	Mechanism	Example
Absorption	Chelation with polyvalent cations	Fluoroquinolones + iron (8)
	Adsorption to polymeric resins	Patiromer + ciprofloxacin (9)
	Alteration of gastric pH	Gefitinib + omeprazole (10)
	Inhibition of intestinal uptake transporters (e.g., OATP)	Fexofenadine + grapefruit juice (11)
	Inhibition of intestinal efflux transporters (e.g., P-gp)	Dabigatran etexilate + verapamil (12,13)
	Induction of intestinal efflux transporters (e.g., P-gp)	Dabigatran etexilate + rifampicin (12)
Distribution	Inhibition of hepatic uptake transporters (e.g., OATP1B1/1B3)	Eltrombopag + rosuvastatin (14)
	Plasma protein binding displacement (generally low clinical relevance)	Dolutegravir + valproic acid (15,16)
Metabolism	Enzyme inhibition (e.g., cytochrome P450 [CYP]-mediated metabolism)	Clarithromycin + statins (17)
	Enzyme induction (e.g., CYP-mediated metabolism)	Rifampicin + crizotinib (18)
Excretion	Inhibition of renal tubular transporters (e.g., OCTs, MATE)	Dolutegravir + metformin (19)
	Interference with enterohepatic circulation	Mycophenolate mofetil + oral antibiotics (20)
	Reduction in glomerular filtration rate	Lithium + lisinopril (21)
	Alteration of urinary pH (limited clinical relevance; mainly relevant in toxicology)	Sodium bicarbonate + salicylates (22)

This list is not intended to be exhaustive; additional or secondary mechanisms of pharmacokinetic DDIs may exist.

The literature further distinguishes several subcategories of DDIs with different degrees of clinical relevance, although terminology is not fully standardized across studies.

Potential DDIs (pDDIs) represent pharmacological combinations whose concomitant or temporally close administration is associated with a theoretical risk of DDI (1). The presence of a pDDI does not necessarily imply the occurrence of an adverse drug reaction (ADR) or a clinically meaningful modification of the therapeutic effect (23). Rather, each pDDI corresponds to a condition of risk, the clinical relevance of which depends on the probability that the DDI will occur and on the magnitude of its effect. Beyond the intrinsic characteristics of the substances involved, the probability that a pDDI will occur and, more importantly, that it will translate into a clinically manifested DDI depends on multiple patient-related determinants.

Clinically relevant DDIs refer to DDIs for which a clinically measurable change in the efficacy or toxicity of one or more drugs is highly probable, taking into account not only the pharmacological combination itself but also individual patient factors such as age, sex, comorbidities, diagnosis, dose, and route of administration (23-25).

DDIs resulting in actual patient harm refer to DDIs that have produced documented clinical injury, identified on the basis of symptoms, signs, or diagnostic findings recorded in the medical record (24,25).

1.1.2 Pharmacological determinants of DDI vulnerability

Several drug-related characteristics may determine both the pharmacological vulnerability of the object drug and the interaction potential of the precipitant drug, thereby influencing the probability that a pDDI will become clinically relevant. The clinical relevance of a DDI rarely depends on a single determinant, but rather emerges from the *interaction* between multiple PK and PD properties affecting both exposure and pharmacological vulnerability (27).

Among the main determinants is the dependence of drug clearance on a single metabolic or transport pathway (27). Drugs whose clearance predominantly depends on a single isoenzyme or transporter are particularly vulnerable to DDIs because inhibition or induction of the involved pathway may produce disproportionate changes in exposure without compensatory elimination through alternative pathways (27,28). Conversely, drugs capable of inhibiting or inducing multiple metabolizing enzymes or transporters may display broad precipitant effects and high DDI potential (29).

Oral bioavailability also influences DDI magnitude. For drugs with low oral bioavailability due to extensive intestinal or hepatic first-pass metabolism, inhibition of the involved pathway may produce marked changes in systemic exposure (30). Similarly, drugs whose absorption depends on gastric pH or intestinal transport mechanisms may be highly susceptible to food- or drug-induced gastrointestinal alterations (31).

High plasma protein binding is another commonly cited determinant of PK DDIs. Competition for protein-binding sites may transiently increase the free fraction of highly bound drugs. However, protein-binding displacement alone rarely causes clinically meaningful DDIs, except when the affected drug also has a narrow therapeutic index or limited compensatory elimination capacity (16).

Drugs with a narrow therapeutic index and a steep concentration-response relationship have a limited margin between therapeutic and toxic concentrations. For these substances, even modest PK or PD perturbations, as induced by the administration of concomitant substances, may result in severe and potentially fatal complications (32,33). Typical examples include antiarrhythmics, anticoagulants, immunosuppressants, antiepileptic drugs, digoxin, lithium, and several anticancer agents (32–34).

Lastly, drugs characterized by poor receptor selectivity and broad pharmacodynamic activity profiles may be particularly prone to clinically relevant PD DDIs (35). When combined with other agents acting on the same physiological systems, these drugs may produce additive or synergistic effects.

1.2 Patient-related risk factors

Several factors, beyond the pharmacological properties of the drugs themselves, influence the likelihood that a DDI will occur. The most relevant of those include the number of prescribed medications, age, multimorbidity and organ impairment, and genetic variability affecting drug metabolism (36).

Many of these determinants are interrelated. Increasing age is often associated with a higher burden of chronic diseases and with progressive decline in organ function, particularly renal function. These conditions frequently lead to more complex therapeutic regimens and a greater use of medications (37).

1.2.1 Number of medications

The number of prescribed medications is one of the strongest predictors of DDIs. As additional drugs are added to a treatment regimen, the number of possible drug combinations increases rapidly, raising the probability that at least one pDDI may occur. Any pDDI in turn carries a potential risk of evolving into a clinically manifested DDI.

Population-based studies have consistently shown a strong association between polypharmacy, commonly defined as the use of five or more drugs, and pDDIs (38). In a large study of older adults using national prescription register data, the prevalence of clinically relevant pDDIs increased markedly with the number of dispensed medications (39). Similar findings have been

reported in hospital settings, where the number of prescribed drugs emerged as a major predictor of pDDIs among admitted patients (40). It is not yet clear whether the relationship between the number of medications and the number of pDDIs is solely due to combinatorial reasons or whether, as the number of medications increases, the ability to identify and prevent at-risk combinations decreases.

1.2.2 Multimorbidity

Multimorbidity, commonly defined as the coexistence of two or more chronic diseases in the same individual (41), increases the risk of DDIs. In addition to its obvious association with polypharmacy, the simultaneous management of several conditions often requires complex therapeutic regimens involving multiple pharmacological classes, increasing the likelihood that interacting drugs will be prescribed together (36).

Moreover, chronic diseases frequently lead to organ dysfunction, which then affects drugs PK. Renal impairment, often observed in patients with diabetes or cardiovascular disease, reduces the clearance of renally eliminated drugs and promotes drug accumulation. This may increase the clinical impact of PK DDIs and raise the risk of toxicity when interacting drugs are co-prescribed (36).

Liver dysfunction can further modify the magnitude of DDIs by altering the expression and activity of drug-metabolizing enzymes and by reducing hepatic blood flow, thereby affecting the PK of enzyme inhibition or induction (42,43).

1.2.3 Age

Advanced age is a relevant factor influencing the risk of DDIs. In many cases, age acts as a proxy reflecting the accumulation of other determinants, particularly multimorbidity and polypharmacy (37,44).

A recent systematic review of studies on older adults living in the community found that polypharmacy rates ranged from 2.6% to 86.6%. At the same time, the prevalence of multimorbidity varied from 4.8% to 91.0% (45). In studies of older adults in clinical settings, polypharmacy rates ranged from 5.0% to 85.0%, and multimorbidity rates varied from 20.0% to 93.1%. This wide range is likely due to differences in the study groups, how conditions and medications were defined, and the methods used to identify them, such as self-reports and electronic or administrative records (45).

However, the increased likelihood of DDIs in older adults cannot be attributed solely to polypharmacy or multimorbidity. Age-related reductions in the functional reserve of multiple organs can alter drug metabolism and elimination, thereby increasing susceptibility to PK DDIs (37,42).

Aging may also affect the expression of specific CYP enzymes involved in drug metabolism in an isoform-specific fashion. Evidence suggests that the activity of some enzymes, such as CYP1A2 and CYP2C19, may decline with age, whereas others, including CYP3A4, CYP2C9, and CYP2D6, appear to remain relatively stable or only slightly modified (42).

Age-related changes in body composition may also influence drug disposition. Older individuals typically exhibit an increase in body fat and a reduction in lean body mass and total body water, factors that may alter the volume of distribution of several drugs (46).

From a PD perspective, aging may also modify the response to medications. Changes in receptor sensitivity, electrolyte balance, neurotransmitter systems, blood-brain barrier permeability, metabolic regulation, and homeostatic mechanisms can alter the body's response to drugs and increase vulnerability to adverse effects (47).

1.2.4 Genetic polymorphisms

Genetic polymorphisms represent another important determinant of variability in DDIs (48). These variants, defined as stable DNA sequence differences occurring in the population with a frequency of at least 1%, can lead to quantitative or qualitative changes in the expression and activity of drug-metabolizing enzymes, transporters, or pharmacological receptors (49).

In PK, genetic variants can produce distinct metabolic phenotypes (e.g., poor, intermediate, extensive, and ultrarapid metabolizers), potentially influencing systemic drug exposure and may modify the magnitude of DDIs (50).

Differences in metabolic phenotype may amplify ADRs. For example, in individuals who are poor metabolizers, administering an enzyme inhibitor may further reduce the limited residual clearance of the affected drug (50).

In rare cases, the same drug combination may lead either to therapeutic failure or to toxicity depending on the patient's genotype, with markedly different dosing implications. A clear example is the DDI between voriconazole and atazanavir/ritonavir. Voriconazole is mainly metabolized by CYP2C19, with CYP3A4 and CYP2C9 playing a secondary role. In subjects who extensively metabolize CYP2C19, ritonavir can increase CYP2C19 activity, which can reduce voriconazole exposure to about one-third of its typical area under the curve (AUC). In contrast, in CYP2C19 poor metabolizers, the CYP3A4 pathway becomes predominant, and

strong inhibition of CYP3A4 by atazanavir/ritonavir may therefore increase voriconazole AUC leading to a higher risk of toxicity (51).

1.3 Prevalence of DDIs and pDDIs

As previously defined, the literature distinguishes between pDDIs, DDIs, clinically relevant DDIs, and DDIs that result in actual patient harm. Most epidemiological studies have focused on pDDIs only, whereas substantially fewer have assessed the remaining categories (24,25).

In hospital settings, pDDIs are consistently reported as common. In a systematic review and meta-analysis of hospitalized patients, 33.0% (95% confidence interval [CI]: 17.3-51.3%) of general inpatients experienced at least one pDDI during their hospital stay, with the estimate rising to 67.0% (95% CI: 52.7-79.1%) in intensive care units, where exposure to complex pharmacotherapy is greater (24). Individual studies included in the mentioned systematic review reported highly variable prevalence estimates, ranging from 16.3% to 71.1% in general wards and from 46.3% to 90.3% in intensive care units (24). A more recent study reported an even higher pooled prevalence of pDDIs, at 64.9% (95% CI: 61.8-73.6%) among hospitalized patients overall, again with substantial heterogeneity between studies and across clinical settings (52). Together, these findings indicate that pDDIs are very frequent in inpatient care, although their exact prevalence is highly sensitive to the population studied and to the identification method used (24,52).

Far fewer studies have evaluated clinically relevant DDIs. Among hospitalized patients, a systematic review specifically addressing clinically relevant DDIs showed that about one in ten inpatients experiences at least one DDI with documented clinical manifestation during hospitalization (pooled prevalence: 9.2%; 95% CI: 4.7-19.7) (26). However, individual

estimates ranged widely, from 1.2% to 64.0%, with the highest prevalence observed in intensive care units and lower estimates in emergency and internal medicine settings (26).

A recent multicenter study of hospitalized patients reported that while 70% of inpatients experienced at least one pDDI, fewer than 1% of DDIs resulted in actual patient harm (53).

Overall, these data suggest a marked gradient in frequency when moving from pDDIs to clinically relevant DDIs to DDIs inducing harm.

Evidence is even more limited when the outcome of interest is hospital admission or emergency attendance attributable to a pDDI or a DDI. In a systematic review and meta-analysis of actual DDIs associated with hospital use, the median prevalence of hospital admissions attributable to DDIs was 1.1%, whereas the median prevalence of hospital visits was only 0.1% (54). Among patients admitted because of ADRs, however, DDIs accounted for a large proportion, with a median prevalence of 22.2% among ADR-related hospital admissions and 8.9% among ADR-related hospital visits (54).

More recently, a cross-sectional study of emergency department admissions in patients aged 75 years or older showed that up to 6.4% of admissions involved a high-risk contraindicated or discouraged pDDI, and 0.4% to 1% of admissions were considered probably or certainly attributable to a clinically relevant DDI (55).

Outside acute-care hospital settings, pDDIs are also highly prevalent among older community-dwelling individuals and nursing home residents. A recent systematic review in community-dwelling older adults reported pooled DDI prevalence estimates ranging from 16.6% to 57.8% depending on the identification method used, with cardiovascular drugs, anticoagulants, and psychotropic agents among the most frequently implicated medications (56). Nursing home

residents appear particularly vulnerable because of extensive polypharmacy, frailty, multimorbidity, and high use of psychotropic medications. In a Spanish cross-sectional study, more than 80% of nursing home residents experienced at least one pDDI and 24.7% had at least one severe or moderate interaction (57). Similarly, an Italian study conducted in elderly populations from the Emilia-Romagna region reported exposure to at least one clinically relevant pDDI in 53.7% of nursing home residents (58).

1.4 Sources and Methods for Identifying pDDIs

Potential DDIs are typically identified through a combination of information sources and operational approaches, including reference compendia, digital interaction checkers (ICs), automated clinical decision support systems, and structured clinical processes such as medication review (59). Each of these approaches presents specific advantages and limitations in terms of coverage, usability, and clinical applicability.

1.4.1 Summary of Product Characteristics (SmPC)

The Summary of Product Characteristics (SmPC) is a standardized regulatory document for medicinal products authorized in the European Union. It forms an integral part of the marketing authorization of a medicinal product and provides healthcare professionals with the official information required for the safe and effective use of the medicine (60).

Section 4.5 (Interaction with other medicinal products and other forms of interaction) reports known or expected pDDIs based on available clinical and in vivo data at the time of authorization, as well as information derived from subsequent updates, including post-marketing data (60). Additional information may also be provided in Sections 4.3

(Contraindications) and 4.4 (Special warnings and precautions for use), which may cross-reference Section 4.5 when relevant.

Although the SmPC represents the primary regulatory source of information on drug safety, its use in routine clinical practice appears limited. Consulting individual SmPCs is often time-consuming and poorly scalable and therefore tends to play a secondary role in everyday prescribing practice (61,62). Studies suggest that SmPCs in their current format may not fully meet the needs of healthcare professionals, with difficulties reported in locating and interpreting specific information and with relatively limited use of the document as a practical source of clinical information (62,63).

Beyond these practical limitations, inconsistencies in the information provided across SmPCs have also been reported. In particular, discrepancies between originator and generic products containing the same active substances have been described, including differences in Section 4.3 - Contraindications (64).

1.4.2 National or institutional reference documents

National regulatory or institutional reference documents (e.g., the ANSM *Thésaurus des interactions médicamenteuses* in France, or the British National Formulary, BNF, in the UK) represent curated sources that synthesize information on drug therapy and pDDIs for routine clinical use (65,66). In this sense, they function as tertiary synthetic drug information resources. They generally provide a structured interaction statement for a given drug pair, often complemented by severity stratification and management recommendations (65,67). National reference documents can be used to define high-severity pDDI subsets for evidence generation and prevention strategies and as a source for clinical decision support systems (CDSS -

presented in further detail in Section 1.4.4) (68). Limitations of these resources include limited ease of consultation and lack of transparency on the underlying evidence (68,69).

1.4.3 Commercial interaction checkers (ICs)

ICs are digital tools designed to provide structured information on pDDIs using proprietary or institutional drug knowledge bases (70,71). They are typically implemented as web-based interfaces or software applications that allow users to query pDDIs between two or more drugs and retrieve summaries describing the pDDI, its severity, potential mechanism, and recommended management strategies. ICs are primarily consultation tools that provide interaction information on demand, although they may also be implemented as modules within broader CDSS integrated into electronic prescribing or health record systems (72,73).

Most ICs allow the simultaneous entry of multiple medications and generate structured reports describing the pDDIs detected among the selected drugs. The information provided commonly includes a severity classification, a brief description of the underlying mechanism, recommended management strategies, and, in some resources, an assessment of the strength or quality of the underlying evidence (69,74).

Examples of widely used ICs include commercial resources such as Micromedex and Lexidrug, as well as specialized academic databases such as the Liverpool Drug Interaction databases for HIV and hepatitis therapies (10,71,75,76)

Comparative studies consistently indicate an important degree of discordance across ICs regarding which drug pairs are flagged and how severity and management recommendations are presented (57,65-73). Consequently, the clinician's exposure to pDDI information depends on the specific tool selected and may plausibly influence prescribing decisions. This phenomenon has been described in broad comparisons of multiple ICs, in real-world

evaluations, and in domain-specific comparisons (e.g., antiretroviral pDDI resources) (69,79,81,86). This important limitation may be directly linked to the epistemic foundations of pDDIs themselves. The evidence base underlying many DDIs is heterogeneous and encompasses a wide variety of study designs, including in vitro experiments, case reports, case series, observational analyses, PK trials, and both randomized and non-randomized clinical studies, reflecting the methodological diversity of research used to establish and evaluate DDIs (87,88).

Empirical reviews illustrate the consequences of this heterogeneity. In a systematic review of psychotropic pDDIs associated with clinical adverse effects, only a minority of these pDDIs were supported by controlled studies, whereas many relied solely on uncontrolled investigations or case reports (89). Similar patterns emerge even for extensively studied drugs. In a systematic review of warfarin DDIs including 72 studies, only 11 were randomized trials, while the remaining 61 were observational designs, often retrospective and potentially subject to confounding and selection bias (90).

Together, these elements may reduce the consistency of evidence synthesis and contribute to variability in how DDI risk is interpreted across different sources.

1.4.4 Clinical Decision Support Systems

Clinical decision support systems (CDSS) are automated tools designed to assist prescribing physicians and hospital pharmacists by identifying therapies that may lead to ADRs (72). These systems can detect a variety of potential problems, including dosing errors, substances associated with allergy risk, and combinations of co-administered drugs that may be potentially harmful due to pDDIs (91).

However, the generation of an excessively high number of alerts in routine clinical practice, many of which represent false positives (FP) or low-relevance signals, has led to the phenomenon known as *alert fatigue*. Alert fatigue can be defined as the mental state resulting from continuous exposure to a large number of clinically irrelevant alerts, which may lead clinicians to inadvertently ignore important warnings together with those of minor relevance (92,93).

Empirical studies consistently demonstrate that prescribers override the majority of alerts generated by CDSS. In a recent systematic review and meta-analysis including more than half a million prescriptions, approximately 90% of DDI alerts were overridden by physicians (94,95).

Cases of serious and preventable ADRs associated with the systematic overriding of existing alerts have been reported in the literature (96). Beyond these concerns, the effectiveness of alert-based systems themselves has also been questioned. A recent multicenter controlled study by Baysari and colleagues designed to evaluate the impact of DDI alerts on hospital readmission rates in the presence of pDDIs and clinically manifested DDIs, showed a significant reduction in readmissions among patients with pDDIs, regardless of their clinical relevance (25). In contrast, no reduction was observed for readmissions associated with clinically relevant DDIs, (25).

This lack of effect may represent a practical consequence of alert fatigue itself. For an alert to influence prescribing behavior, clinicians must first recognize the relevance of the signal. In other words, when DDI alerting systems generate an excessive number of notifications, their practical utility may approach that of having no alerting system at all.

On the one hand, these findings further reinforce the importance of focusing alerting systems on high-severity pDDIs in order to improve their efficiency. On the other hand, they raise an ethical concern. If reductions are observed in the number of theoretical pDDIs without a corresponding reduction in clinically relevant DDIs, this effect may paradoxically reflect the avoidance of prescriptions that are generally safe. The possibility that excessive alerting may produce not only a neutral but also a potentially detrimental effect has been suggested (97).

1.4.5 Medication review

Medication review represents a structured clinical process, often multidisciplinary, in which a healthcare professional, typically a clinical pharmacist or clinical pharmacologist, evaluates the patient's overall pharmacotherapy with the aim of optimizing medication use (98). The objective is to prevent, identify, and resolve drug-related problems, including pDDIs (99). During this process, clinicians may decide to maintain a combination flagged as interacting when it is considered clinically justified, integrating into the evaluation factors such as comorbidities, laboratory data, and therapeutic indications, or, when appropriate, considering deprescribing strategies (100). Structured set of criteria can guide medication review in high-risk patient groups. For example, the Beers Criteria and the STOPP/START criteria are commonly used to identify potentially inappropriate medications in older adults (101,102).

Although medication review partially relies on the IC tools described above, it enables a patient-specific approach with a high degree of clinical contextualization. This allows clinicians to assess the real clinical relevance of potential interactions rather than relying solely on theoretical alerts.

The main limitation of medication review lies in its resource requirements. These approaches are time-consuming, difficult to scale to large populations, and strongly dependent on the

expertise of the professionals involved (103,104). Nevertheless, comparisons between medication review strategies and automated CDSS approaches suggest that medication review may achieve higher levels of clinical acceptance. In a large teaching hospital, Beeler and colleagues reported that a consulting pharmacist reviewing the same patients issued less than 3% of the number of recommendations that would have been generated by a CDSS while maintaining a high (>75%) rate of acceptance and implementation by treating physicians (105). However, no definitive conclusions can be drawn regarding the impact of medication review on hard clinical outcomes. Evidence from recent systematic reviews suggests that medication review and deprescribing interventions can slightly improve medication appropriateness and reduce potentially inappropriate medications in hospitalized older adults (106). These interventions appear feasible and safe and may contribute to modest but significant reductions in hospital readmissions, likely by decreasing ADRs and clinically relevant DDIs. However, their impact on harder outcomes such as mortality remains uncertain and may be limited by the availability of long-term follow-up data (106).

1.5 Strategies to Improve the Identification, Prioritization, and Evidence Base of pDDIs

Most strategies to reduce the gap between potential pDDIs and clinically relevant DDIs and to make the process of identifying and managing pDDIs less demanding converge on one of four often complementary approaches: (i) standardization of evidence and information in pDDI sources, (ii) prioritization and contextual filtering, (iii) use of artificial intelligence (AI) as support for prescribing, and (iv) validation and updating using real-world data.

1.5.1 Standardization of evidence and information in pDDI sources

Expert panels, mainly developed in the United States, have proposed structured approaches to improve the consistency and transparency of the DDI information used in clinical decision support (1,107,108). In these approaches, the knowledge bases used by ICs and CDSS should be built and maintained through standardized processes. Professional societies or academic expert groups are expected to play a central role in identifying and classifying pDDIs using structured evaluation procedures (107).

Several frameworks have therefore been proposed to support the systematic assessment of pDDI evidence. These frameworks generally address three sequential questions: whether a pDDI exists (109), whether it is clinically relevant, and how the associated risk should be managed in clinical practice.

Experts also recommend that a set of key elements should be systematically reported for each interacting drug pair. These include the classification of seriousness, which is considered more informative than the commonly used concept of severity, the potential clinical consequences of the DDI, the estimated frequency of harm or exposure, and recommended management strategies. Additional elements include factors that may modify patient susceptibility, the pharmacological mechanism of the DDI, the quality of the supporting evidence, and access to the underlying documentation (1,107).

The most critical information, especially management recommendations, should appear at the beginning of the alert. Additional details can remain available if needed, for example, through expandable sections or linked documentation (59).

Classification systems should remain simple and transparent. Limiting the number of seriousness categories can improve both interpretability and consistency across systems. The

highest category should correspond to pDDIs that trigger interruptive alerts, meaning alerts that interrupt the prescribing workflow and require clinician action (110). Lower categories may correspond to non-interruptive notifications or to pDDIs that should not generate alerts at all.

The term *contraindicated* should be used cautiously. It should be reserved for situations in which no clinical circumstance exists in which the benefits of the drug combination outweigh the risks.

In practice, however, many pDDI described as contraindicated in drug information sources represent situations that require careful management rather than absolute avoidance (110).

Extrapolation of DDIs to entire pharmacological classes should also be approached with caution. PK pDDIs are often drug-specific, and the magnitude of the effect may vary substantially between agents belonging to the same class, depending on their inhibitory or inducing potency (1).

For PK pDDIs, reporting pharmacokinetic changes alone is often insufficient. Whenever possible, the expected clinical consequences should be described, and estimates of the frequency of adverse outcomes should be provided (1).

1.5.2 Prioritization and contextual filtering

One operational approach is to limit the number of pDDIs presented to the prescriber.

This can be implemented in different ways. One option is to allow clinicians to filter alerts and display only predefined categories of pDDIs. Another approach relies on dynamic processes that incorporate physician feedback and periodic reassessment by expert panels, which review pDDIs that may have been incorrectly classified (99-102).

Interventions such as selective alert inactivation, severity reclassification, contextualization, and threshold adjustment have been associated with substantial reductions in alert volume and improved clinical acceptance rates (114). In particular, the reorganization of alerts according to

severity levels, often referred to as tiering, and the targeted suppression of low-relevance alerts have been shown to increase adherence to high-severity alerts (113).

Improving DDI alerting systems also requires increasing the specificity of alert-triggering rules. Alert systems should consider contextual factors that influence the likelihood that a pDDI will lead to clinically relevant harm (114,115). These include characteristics of drug exposure, such as dose, route of administration, and treatment duration. The clinical indication and the order in which therapies are initiated may also influence pDDI risk. Timing of administration and laboratory data may be relevant as well, although integrating these latter elements into alert systems remains challenging (92,107). These approaches, however, require stable multidisciplinary governance and regular review processes.

Another strategy considered is to abandon integrated CDSS alerts and instead make ICs freely available for voluntary consultation by prescribers. However, the advantages of this approach remain unclear. Observational experiences suggest that prescribers rarely consult such tools spontaneously when they are not integrated into the prescribing workflow, whereas their use remains high among hospital pharmacists and clinical pharmacologists (105).

1.5.3 AI as support for prescribing

NLP methods, i.e., computational methods designed to represent and analyze human language computationally (116), have been used to convert unstructured clinical narratives into structured medication data that can support clinical decision systems (117,118). NLP has also been applied to automatically extract pDDI information from regulatory and biomedical texts, although performance remains variable due to complexity and heterogeneity in drug labels (119).

More advanced approaches based on machine learning and neural networks have been proposed to improve medication alerting systems (120). In some hospital settings these methods have

shown the potential to substantially reduce alert burden, in certain cases by up to 90 percent, while improving positive predictive value compared with traditional rule-based systems (120). Some of these models use clinical, prescribing, and contextual variables to predict alert overrides and dynamically adjust alert generation (121); other approaches use data derived from electronic health records to identify patients whose prescriptions are at higher risk of medication error and to prioritize medication review (122)

Hybrid systems combining predefined clinical rules with machine learning models have been evaluated. In these systems, predefined clinical rules identify potential prescribing errors, while neural networks trained on prescription and clinical data refine the signal and filter FP based on patient-specific clinical conditions (117).

However, the current literature indicates that a majority of machine learning models have been developed using monocentric datasets and have undergone only internal validation. Real-world implementation remains limited, and systematic external validation is often lacking. These limitations restrict the generalizability of the results and suggest caution in large-scale adoption (120).

More recently, several studies have explored the use of generative large language models to extract pDDI information from unstructured sources such as product labels, with promising preliminary results (123). Nevertheless, issues such as hallucination risk, difficulties in assessing output reliability, and the environmental cost associated with large models suggest that the real advantages of these approaches remain uncertain (124).

1.5.4 Validation/Updating with Real-World Data

The use of real-world data may serve two primary purposes: reducing the structural uncertainty surrounding many pDDIs by strengthening existing evidence and identifying new pDDI signals worthy of further investigation.

In the context of evidence generation for pDDIs, spontaneous reporting systems (SRSs) represent a first-line tool for the timely identification of signals of potentially clinically relevant DDIs. DDIs, which are rarely fully explored during the pre-marketing phase, find in post-marketing pharmacovigilance (PV) a privileged setting for their detection through disproportionality analyses applied to large spontaneous reporting databases (125,126). Nevertheless, SRSs are affected by numerous potential biases, including notoriety bias, competition/masking effect, and channeling bias, all of which may distort both the presence and the apparent absence of signals. (115-117)

In particular, the absence of disproportionality for known DDIs should be considered an element of interest, but only a preliminary one, to be integrated with appropriately designed analytical observational studies (130). From this perspective, PV performs a triage and prioritization function: identifying previously unrecognized combinations deserving further investigation or, conversely, preliminarily signaling the need to reassess hypothesized pDDIs not corroborated by real-world data.

The integration of data derived from Electronic Health Records (EHRs) allows overcoming some intrinsic limitations of spontaneous reporting systems, particularly the lack of reliable denominators, underreporting, and limited clinical granularity (131). However, the literature shows that the application of EHR data for signal detection remains methodologically heterogeneous; studies differ in design, handling of temporality, confounding control, and

management of multiple testing, and have only sporadically been applied specifically to DDIs (131). Targeted EHR-based data mining applications have demonstrated that it is possible to identify and replicate known pDDI signals using objective clinical criteria and laboratory trajectories, as in the case of pDDIs associated with acute kidney injury (132). Such experiences suggest that real-world data may function both as a hypothesis-generating tool and as a refinement layer for pDDIs already included in knowledge bases, provided that the resulting signals are subsequently subjected to clinical review and expert contextualization (132).

A further evolution in evidence generation derives from the integration of AI techniques applied directly to clinical text. In a multicenter study based on hospital EHR data from more than 500,000 patients, Kaas-Hansen and colleagues developed an end-to-end, language-agnostic pipeline capable of associating latent information contained in clinical notes with exposures to individual drugs and drug pairs, without relying on curated ontologies or predefined dictionaries (133). This approach enabled the identification of potentially relevant signals for drug combinations, including a relevant proportion of possible pDDIs not previously described, thereby generating shortlists of signals to be submitted for expert review (133).

While various PV approaches allow for the initial identification of signals, or for the assessment of signal absence, their validation requires appropriately designed analytical observational studies. Such approaches primarily generate hypotheses regarding suspicious drug combinations, but the resulting “signal” does not constitute definitive proof; rather, it represents an alert requiring subsequent validation.

In this sense, pharmacoepidemiology provides the conceptual and operational framework for evaluating the real-world clinical outcomes of pDDIs. Nevertheless, epidemiological studies on DDIs present specific methodological challenges, including misclassification of concomitant

exposure, multiple confounders related to each drug, immortal time bias, and the distinction between the direct effect of the precipitant drug and the true DDI effect (134). These issues require careful attention and complex design strategies, including separate analyses according to the order of drug initiation, the use of negative controls, and sensitivity analyses addressing exposure definitions (134).

Interventional clinical studies represent a theoretically possible approach to DDI evaluation. The literature presents numerous examples of studies designed to assess the impact of drug co-administration (123-126). It should be emphasized that, although these studies offer the methodological advantage of operating under highly controlled conditions, they are often conducted as sub-studies within larger trials or in healthy volunteers, with limited sample sizes, selected drug pairs, predominantly PK endpoints, and controlled experimental settings.

All the strategies discussed converge on two fundamental and interconnected planes of intervention. The first is epistemic in nature and is oriented toward the consolidation, standardization, and hierarchical organization of the evidence underlying pDDIs in order to reduce heterogeneity in classification and management. The second is operational and is aimed at the concrete implementation of tools capable of identifying, filtering, and prioritizing pDDIs in real clinical settings, reducing cognitive burden and improving the clinical relevance of alerts. As detailed in the following chapters, several of these strategies were applied in an integrated manner throughout this thesis: (a) a structured approach was developed to compare and standardize pDDI management recommendations across different sources; (b) an expert-derived gold standard for high-severity pDDIs was constructed and progressively refined; (c) an NLP-

based tool for the automated identification of pDDIs in hospital prescribing data was evaluated and further informed.

Finally, this work led to the identification of a database of high-interest pDDIs in real-world prescribing contexts, which constitutes a methodologically solid foundation for future validation, evidence generation, and in-depth investigation using real-world data.

2 - Thesis objectives

This thesis aimed to address the following research questions:

1. How can management recommendations for pDDIs be systematically compared across different sources?
2. What level of consensus exists among sources regarding pDDIs management recommendations?
3. How can an expert-derived gold standard for the identification of high-severity pDDIs be constructed and refined?
4. To what extent can an NLP-based tool identify pDDIs in hospital prescribing data?

To address these questions, three studies were conducted.

1. The first study was a methodological investigation aimed at developing and testing a structured approach to standardize and compare pDDI management recommendations across different ICs commonly used in clinical practice. It also evaluated the extent of agreement among these recommendations within a set of clinically relevant pDDIs identified in a real-world hospital context.
2. The second study focused on the construction and initial validation of an expert-derived gold standard for the identification of high-priority pDDIs. This gold standard was subsequently used to perform a first evaluation of the performance metrics of an NLP-based detection tool.
3. The third study aimed to refine the limitations of the gold standard through a structured and sequential process of comparison with the aforementioned NLP tool. In addition, the information derived from this comparative analysis was used to inform subsequent improvements of the NLP system.

The overall objective of these studies was to lay the groundwork for both epistemic and operational advancement in the identification and prioritization of clinically relevant pDDIs.

3 - Standardizing and comparing management recommendations for potential drug-drug interactions across different interaction checkers

3.1 Context and objectives

Most pDDIs identified in the available sources, such as reference documents and ICs, do not translate into clinically observable events, whereas only a minority may be associated with severe outcomes. This gap between the theoretical burden of pDDIs and their measurable clinical impact generates a practical dilemma: **which pDDIs should be prioritized and brought to the clinician's attention as requiring immediate action, while simultaneously avoiding over-alerting that contributes to alert fatigue and erodes trust in decision-support tools?**

While local regulatory sources serve as essential institutional references, they may offer limited information regarding the level of evidence or provide insufficiently standardized practical guidance. Several commercial ICs are commonly used to inform clinical decisions; however, they may differ not only in the list of pDDIs identified and in severity classification, but also, and more critically, in the management recommendations they provide.

Within the existing literature, several studies have compared ICs in terms of pDDIs presence or severity grading. Nonetheless, a methodological gap persists regarding the systematic comparison of management recommendations, which represent the component most directly driving clinical action. The principal challenge lies in the absence of a shared metric capable of translating textual recommendations and heterogeneous informational structures into comparable categories applicable to real-world settings and suitable for prioritization.

This study addresses this need by developing and testing a classification scale for pDDI management recommendations, with the objectives of: (i) standardizing and enabling comparison of management recommendations across different ICs; (ii) quantifying inter-IC agreement on management recommendations, focusing on clinically relevant pDDIs identified

in a real-world hospital context; and (iii) describing, when available, variability in the reported level of evidence supporting pDDIs as a potential determinant of discordance.

The following section constitutes a faithful reproduction of the article published in the journal Drug Safety and available at the following DOI (<https://doi.org/10.1007/s40264-025-01637-3>).

The only modifications made relate to text formatting and the numbering of references, which have been adapted to ensure consistency with the formatting of the present thesis.

Note. Tables and figures follow the internal numbering system used within the study and have not been renumbered across chapters. Abbreviations are defined at their first occurrence in each article, regardless of previous definitions in the thesis.



Standardizing and Comparing Management Recommendations for Potential Drug–Drug Interactions Across Different Interaction Checkers

Lelio Crupi^{1,2} · Louis Letinier³ · Vianney Jouhet⁴ · Sébastien Cossin^{5,6} · Antoine Pariente^{1,7} · Clement Mathieu^{1,7} · Guillaume L. Martin³ · François Konschelle⁴ · Julie Rouanet^{1,7} · Massimo Carollo⁸ · Gianluca Trifirò⁸ · Emanuela Esposito² · Francesco Salvo^{1,7}

Received: 20 June 2025 / Accepted: 13 November 2025
© The Author(s), under exclusive licence to Springer Nature Switzerland AG 2025

Abstract

Background Potential drug–drug interactions (pDDIs) are frequent in clinical care, particularly among older patients. Accurate identification and management of pDDIs are essential for patient safety. Prescribers often rely on interaction checkers (ICs) to screen for pDDIs. However, these tools may provide inconsistent recommendations, potentially leading to suboptimal clinical decisions.

Objective This study aimed to develop a standardized approach for classifying and prioritizing pDDIs based on the clinical relevance of their management recommendations and to compare how these pDDIs are categorized across ICs.

Methods A scale was developed through a structured iterative process to classify pDDIs into four management categories (high priority, intermediate priority, low priority, data unavailable), based on the management recommendations extracted from six ICs. This scale was applied to 218 real-world pDDIs identified from 1923 patients, and the agreement was primarily assessed using Gwet's AC1.

Main Results Overall agreement among the ICs was moderate (Gwet's AC1 = 0.44; 95% CI 0.39–0.50), with values ranging from 0.58 (0.51, 0.65) to 0.38 (0.31, 0.44) in leave-one-out analyses. The agreement was higher in binary analyses dichotomizing the scale into high- and intermediate-priority versus low-priority pDDIs (AC1 = 0.72; 95% CI 0.65–0.79), and in the classification of high-priority versus all other pDDIs (AC1 = 0.62; 95% CI 0.54–0.69).

Conclusion This study developed and tested a structured approach to systematically compare pDDI management across ICs and prioritize clinically relevant interactions. Its application revealed a generally limited agreement between ICs, pointing to the need for harmonized approaches and further studies to support more consistent, evidence-aligned pDDI management.

✉ Francesco Salvo
francesco.salvo@u-bordeaux.fr

¹ Université de Bordeaux, INSERM, BPH, Team AHead, INSERM U1219, 33000 Bordeaux, France

² Department of Chemical, Biological, Pharmaceutical and Environmental Sciences, University of Messina, 98166 Messina, Italy

³ Synapse Médecine, 33000 Bordeaux, France

⁴ Service d'Information Médicale, Pôle de Santé Publique, CHU de Bordeaux, Team AHead, INSERM U1219, 33000 Bordeaux, France

⁵ Equipe EDS, CHU de Poitiers, 86000 Poitiers, France

⁶ Scale-Epi, Centre d'Investigation Clinique INSERM 1402, 86000 Poitiers, France

⁷ Service de Pharmacologie Médicale, Pôle de Santé Publique, CHU de Bordeaux, 33000 Bordeaux, France

⁸ Department of Diagnostics and Public Health, University of Verona, 37134 Verona, Italy

Introduction

Drug-drug Interactions (DDIs), which occur when one medication affects the activity of another, are highly prevalent in hospitalized patients (139), especially among older adults and patients on polypharmacy (39,44,56). Improving knowledge of DDIs and developing preventive strategies are key to enhancing patient safety and quality of care.

The high number of potential DDIs (pDDIs), defined as the concomitant administration of at least two potentially interacting drugs, has led to the development of several interaction checkers (ICs) (24,70), which are generally expected to screen prescribed medications, identify pDDIs, provide prescribers with actionable management advice, and help prioritize clinically relevant pDDIs. Moreover, an ideal IC should ensure transparency and regular updates. However, ICs integrated into prescription tools often generate an overwhelming number of alerts, many of which pertain to low-risk pDDIs. Although theoretically possible, many of these pDDIs are considered to have minimal impact on the efficacy or safety of the involved drugs. More importantly, they contribute to the well-documented phenomenon of “alert fatigue”, where prescribers become desensitized to warnings and may overlook critical pDDIs (92,140–142).

Despite the routine use of several ICs (70), these tools often show limited concordance in the identification and classification of pDDIs (143,144). Additionally, evidence suggests that discrepancies exist not only among ICs but also between ICs and Summaries of Product Characteristics (SmPCs), which have been shown to provide limited and sometimes incomplete information (79). Moreover, a significant proportion of pDDIs stem from theoretical or anecdotal evidence (89), potentially contributing to variability in clinical decision-making and suboptimal therapeutic outcomes.

To address these challenges, real-world studies focusing on the most clinically relevant pDDIs may be helpful. Furthermore, establishing a consensus on best practices for identifying, prioritizing, and managing pDDIs based on their clinical priority could help improve prescription quality and reduce DDI-related adverse reactions. Although several studies have compared ICs, they mostly focused on severity categories, with limited attention to information regarding their management (69,80,84,145). There remains a need for a systematic standardization of management recommendations across different ICs. The present study aims to address these gaps by developing and testing a method for prioritizing, standardizing, and comparing the management recommendations of clinically relevant pDDIs through some of the most widely used ICs. Moreover, this study aims to evaluate whether and to what extent these ICs agree on the management recommendations of pDDIs and to assess the level of evidence supporting the existence of each pDDI.

By achieving these objectives, this study seeks to contribute to more effective and consistent management of DDIs, ultimately improving prescribers' awareness and patient safety.

Methods

Development of a Scale for pDDIs Management Recommendations

Six ICs were selected for this study: the ANSM Thesaurus (*ANSM Thesaurus des interactions médicamenteuses*), the British National Formulary (BNF, No. 85 - March 2023), Drugs.com, Epocrates, UpToDate® Lexidrug™ (formerly Lexicomp™), and Medscape (146–151).

To classify and compare the management recommendations of pDDIs across these six ICs, a scale was developed based on a sample of 70 pDDIs randomly extracted from the ANSM Thesaurus (Fig. 1). This number was expected to capture the range of distinct management scenarios reported across ICs.

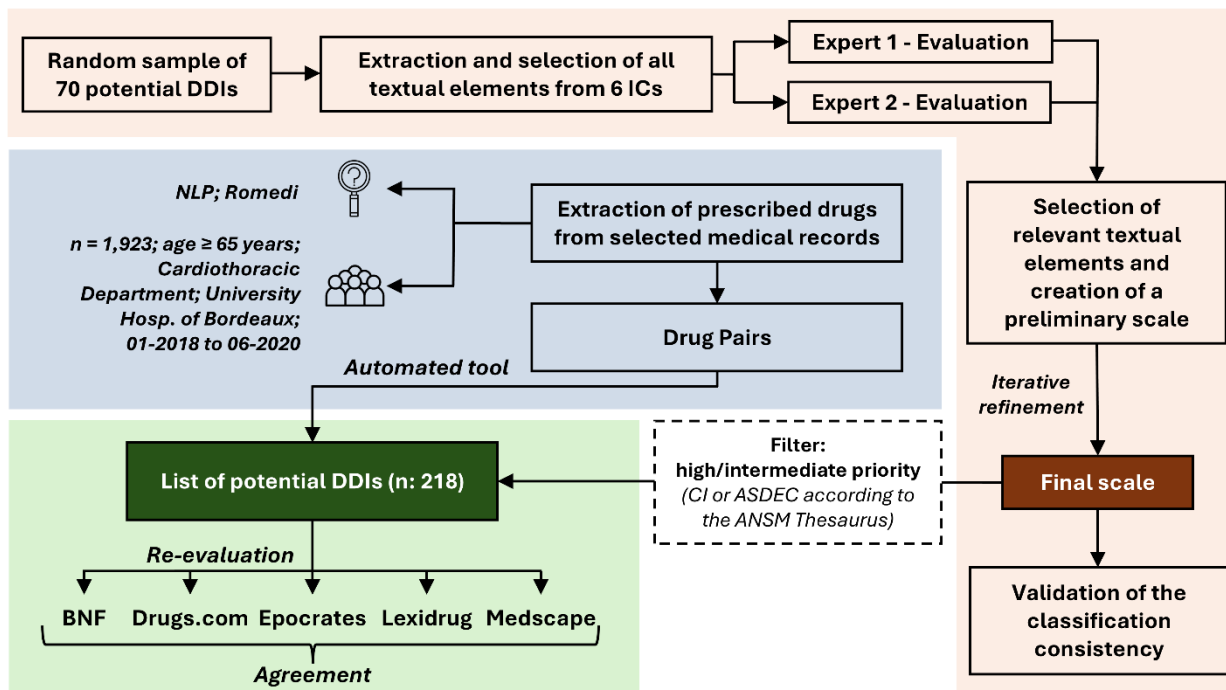


Figure 1 Workflow for the development and application of the method for classifying, prioritizing, and comparing management recommendations for potential drug-drug interactions.

ASDEC: Association not recommended; CI: Contraindicated; DDI: Drug-Drug Interaction; NLP: Natural Language Processing; Romedi: *Référentiel Ouvert du Médicament* (i.e., a French open-access drug referential for semantic interoperability).

To generate an initial list of pDDIs for the development of the scale, a single IC was selected as the reference standard. The ANSM Thesaurus was chosen because it serves as the official national standard for pDDIs in France (146), and thus reflects the clinical setting of the study's real-world application phase. The remaining five ICs were included due to their widespread use in clinical practice and the inclusion of both open-access and subscription-based tools (152). A summary of the information structure and content in each IC is provided in the Supplementary Table 1.

One author (L.C.) extracted the complete set of information on pDDIs from each IC. Two authors (L.C. and F.S.) independently reviewed the extracted data and identified the most relevant sections of each IC for classifying management recommendations (Supplementary Tables 1-4). Common patterns of management scenarios were then systematically identified by L.C. and F.S. based on recurring textual and conceptual elements. The identified scenarios included: a) prohibition of co-administration; b) context-dependent avoidance, based on clinical risk-benefit assessment; c) patient monitoring (e.g., ECG, serum potassium, INR); d) dose adjustment; e) timing or route of administration modification; f) general caution without specific intervention; g) acknowledgment of the pDDI without any recommended action; h) *absence of a documented interaction*. Combinations of these patterns occasionally occurred (e.g., monitoring AND dose adjustment). An additional set of 20 pDDIs was evaluated by one author (L.C.) to assess the completeness of these scenarios; no additional management scenarios emerged. An iterative refinement process was conducted to reduce subjectivity in the definitions of the management categories and their application conditions. The resulting classification system was consolidated into the following four categories:

- 1) **High-priority (*Avoid*):** the pDDI is explicitly described as contraindicated, or discontinuation/substitution is the only course of action mentioned or reasonably implied (e.g., “Contraindicated”, “Avoid”, “Use alternative”, “Avoid or use alternative”). It includes the above identified scenario “a”
- 2) **Intermediate-priority (*Conditional co-administration*):** co-administration may be acceptable only if specific clinical actions are implemented. These may include therapeutic monitoring, dose adjustment, or benefit-risk evaluation. Avoidance

recommendations are not absolute (e.g., “generally avoid”, “avoid or monitor”). It includes the above identified scenarios “b” to “e”

- 3) **Low-priority (*No action required*)**: the pDDI, while possibly acknowledged, does not require any specific intervention beyond general clinical awareness. This includes pDDIs explicitly stated as not clinically relevant, those listed without any associated recommendation, and associations reported as non-interacting. It includes the above identified scenarios “f” to “h”
- 4) **Data unavailable**: at least one of the two drugs was not listed in the IC’s database, preventing the association from being evaluated.

This categorization was further tested by an external evaluator, not involved in the scale development (J.R.), who reclassified the management recommendations from the same initial sample of 70 pDDIs, using the above-reported classification. The inter-rater agreement between the external evaluator and the developers of the scale was very high (Cohen’s $\kappa = 0.85$), thus supporting the clarity and internal coherence of the proposed criteria. Table 1 provides a detailed mapping of the management recommendations across the different ICs.

Identification and Classification of High-Priority and Intermediate-Priority pDDIs in Hospital Prescriptions

This phase of the study was performed using data from the Bordeaux University Hospital data warehouse. This institution, which manages approximately 250,000 hospitalized patients and 100,000 emergency visits annually, implemented a data warehouse in 2017 based on an i2b2 (Informatics for Integrating Biology and the Bedside) structure (153,154). The data warehouse, encompassing all diagnoses, laboratory results, drug prescriptions, and numerous other variables, is structured as a relational multi-table database, which allows for more efficient

querying compared to traditional compartmentalized systems (155). To perform queries, ontology alignment has been performed, and *ad hoc* natural language processing (NLP) tools have been developed, including Romedi (*Référentiel Ouvert du Médicament*), a French drug terminology tool for extracting drug information from text (27,29-31).

Potential DDIs were extracted from clinical records from hospital discharges or consultations in the Cardiothoracic Department between January 2018 and June 2020. The sample included 1,923 patients aged 65 years or older with at least one drug prescription. Medication lists were extracted from these records using NLP techniques, including Named Entity Recognition (159). A list of all possible drug-drug pairs was generated from each discharge or consultation prescription.

Detection of pDDIs from the obtained list of drug-drug pairs was performed using an automated pharmacoinformatic tool developed by *Synapse Medicine*, a spin-off from the University Hospital of Bordeaux, based on the ANSM Thesaurus. This tool has been previously used to identify pDDIs in French healthcare databases (55,160).

Table 1 Classification of management categories according to each interaction checker.

Standardized Categories	ANSM Thesaurus ^a	BNF ^a	Drugs.com	Epocrates ^a	Lexidrug ^a	Medscape
High-priority (Avoid)	Contraindicated	Action statements prohibiting the co-administration	Action statement: Contraindicated	Action statements prohibiting the co-administration	Risk rating: X - Avoid Combination	Action statements: “Contraindicated”, “Avoid or Use Alternate Drug”
Intermediate-priority (Conditional co-administration)	Avoid, unless benefits outweigh harms; Precaution of use;	Action statements discouraging the association, without prohibiting it; Action statements suggesting a therapy adjustment	Action statements: Generally Avoid; Adjust Dose/Adjust Dosing Interval; Monitor; Monitor closely;	Action statements discouraging the association, without prohibiting it; Action statements suggesting a therapy adjustment;	Risk rating: D - Consider therapy modification; C - Monitor Therapy;	Action statements: “Modify Therapy or Monitor Closely”; “Use Caution or-Monitor”;
Low-priority (No action required)	To be considered; No interaction	Action statements that do not define any clear clinical intervention; <i>No action statement; No interaction</i>	<i>No action statement; No interaction</i>	Action statements that do not define any clear clinical intervention; <i>No action statement; No interaction</i>	Risk rating: B - No Action Needed	Action statement: “Minor or Significance Unknown” <i>No action statement; No interaction</i>
Data unavailable	One or both drugs were not listed in the interaction checker’s database, preventing any evaluation of the interaction.					

^aA detailed list of the action statements/categories is reported in Supplementary Tables 1-3.

The study focused only on high- or intermediate-priority pDDIs (specifically, those classified as “Contraindicated” and “Avoid, unless benefits outweigh harms” according to the internal classification of the ANSM Thesaurus).

The detected pDDIs were manually screened and compared to the latest version of the ANSM Thesaurus, which was published on September 15, 2023.

Agreement between interaction checkers

The detected pDDIs were then searched in the five additional ICs and categorized using the developed scale. When an IC provided different management recommendations for the same pDDI based on distinct clinical subgroups or patient characteristics, the lowest-priority scenario (i.e., the one requiring less stringent requirements or less intensive monitoring) was consistently retained. For fixed drug combinations (e.g., sulfamethoxazole/trimethoprim), the management recommendation for the combination was used as a representative for the management of the individual components. Additionally, when multiple administration routes were reported, the systemic route was considered.

Level of evidence assessment

Among the six ICs included, only Lexidrug and BNF provided a structured grading scale for evidence supporting each pDDI. Drugs.com listed supporting references for pDDIs, but did not apply an explicit grading system. The remaining three ICs did not provide any information on the underlying evidence.

BNF and Lexidrug, were therefore used to summarize the evidence supporting the identified pDDIs. The classification criteria of the two ICs (Supplementary Table 5) show limited overlap. For example, BNF’s highest category (Study) includes all pDDIs based on formal studies. In contrast, Lexidrug uses four categories, three of which (i.e., Excellent, Good, and Fair) can apply to DDIs based on either observational or interventional studies. Due to these

significant differences in classification systems, defining equivalence criteria to measure agreement between the ICs was not feasible. Consequently, a descriptive analysis was conducted to examine potential overlaps in their evidence grading systems.

Statistical analyses

Descriptive analyses

A descriptive analysis was performed on pDDIs identified through the ANSM Thesaurus. Drugs were categorized using the Anatomical Therapeutic Chemical (ATC) classification, and pDDIs were described according to their potential clinical outcomes (e.g., QT prolongation, hyperkalemia).

Primary analysis

This analysis assessed agreement among ICs in classifying pDDIs according to the categories of the developed scale. The agreement was quantified using Gwet's AC1 with corresponding 95% confidence intervals (95% CIs) (161). Gwet's AC1 is a statistical measure for defining inter-rater reliability, applicable in settings with multiple raters and categorical data. It ranges from -1 to 1, with interpretative thresholds for point estimates applied as follows: <0.20 (Poor), 0.21-0.40 (Fair), 0.41-0.60 (Moderate), 0.61-0.80 (Good), and 0.81-1.00 (Very Good) (161). Agreement was assessed across all five ICs, as well as across subsets obtained by systematically excluding one IC at a time. All analyses were restricted to pDDIs classified as non-missing by at least 2 ICs.

As the BNF does not provide action statements for most pharmacodynamic (PD) pDDIs due to editorial policy, such pDDIs were excluded from inter-IC comparisons to prevent introducing systematic distortion into the agreement estimates.

Additionally, since some ICs focus on substances authorized in the UK or the US, to account for the potential geographic bias related to drug availability, an evaluation was conducted to assess whether restricting the dataset only to molecules authorized in both countries would

affect results. Given that the primary analyses already excluded pDDIs absent from any IC involved in each comparison, the resulting dataset showed substantial overlap (98.2%) with the subset of pDDIs authorized in both the UK and the US. Therefore, no additional restriction or stratification was applied.

Secondary analyses

These analyses evaluated: a) binary agreement among ICs in classifying pDDIs as requiring any clinical intervention (high- and intermediate-priority); b) binary agreement among ICs in classifying high-priority pDDIs; c) positive percent agreement (PPA) between ANSM and each IC in detecting pDDIs requiring any clinical intervention (high- and intermediate-priority); d) PPA between ANSM and each IC in detecting high-priority pDDIs.

Binary agreement and PPA analyses were based on dichotomized versions of the original scale (e.g., “high-priority” vs. all others). PPA was defined as the proportion of pDDIs classified in the same priority group as the ANSM Thesaurus (used as reference), over the number of pDDIs with an available classification in the comparator IC. Analyses excluded pDDIs labeled as “Data unavailable”.

A sensitivity analysis was conducted by restricting the dataset to PD pDDIs, as explicitly identified by the BNF, using only the four remaining ICs.

Assessment of the evidence

For each IC that provided structured evidence levels (i.e., BNF and Lexidrug), pDDIs were stratified according to their respective evidence categories. We then identified the subset of pDDIs for which both ICs reported an evidence level and compared their classifications using a heatmap, with each cell showing the absolute number and the percentage of pDDIs within each evidence-level combination.

Results

Potential DDIs extracted from discharge records

A total of 218 unique pDDIs were identified, with 40 classified as high-priority (“Contraindicated”) and 178 as intermediate-priority (specifically, “Avoid, unless benefits outweigh harms”), according to the ANSM Thesaurus. These 218 pDDIs involved 133 unique substances, as 88 drugs were related to multiple pDDIs. For example, amiodarone was involved in 19 pDDIs, apixaban in 16, valsartan in 12, and both irbesartan and escitalopram in 11. Most of the substances were cardiovascular drugs (ATC group C - n = 43; 32.3%), followed by nervous system drugs (ATC group N - n = 24; 18.0%) and anti-infectives for systemic use (ATC group J - n = 16; 12.0%). Less than one-quarter of pDDIs (n = 46; 21.1%) involved cardiovascular drugs only.

Out of 218 different pDDIs, the most common potential outcomes were cardiac rhythm and conduction disorders (n = 50; 22.9%), bradykinin-induced angioedema (n = 49; 22.5%), and bleeding (n = 29; 13.3%).

A complete list of the 218 identified pDDIs is provided in the Supplementary Table 6.

Further information on the categorization of the identified pDDIs by each additional IC, is provided in Figure 2 and in the Supplementary Figure 1.

Agreement among ICs

Overall agreement among the five ICs in classifying pDDIs across the defined categories was moderate (Gwet’s AC1: 0.44; 95% CI: 0.39-0.50) (Table 2). Agreement improved when Medscape was excluded (AC1: 0.58; 95% CI: 0.51-0.65; moderate agreement), while removing Lexidrug resulted in the lowest agreement (AC1: 0.38; 95% CI: 0.31-0.44; fair agreement).

Table 2 Agreement between the comparison ICs in classifying pDDIs.

		All the defined categories	High- or intermediate-priority vs. low-priority	High-priority vs. intermediate- or low priority
Interaction checkers	Compared pDDIs (n)	Gwet's AC1 (95% CI)	Gwet's AC1 (95% CI)	Gwet's AC1 (95% CI)
All 5 ICs	185	0.44 (0.39, 0.50)	0.72 (0.65, 0.79)	0.62 (0.54, 0.69)
Excluding BNF	183	0.44 (0.38, 0.51)	0.75 (0.68, 0.83)	0.59 (0.51, 0.67)
Excluding Drugs.com	185	0.38 (0.32, 0.45)	0.69 (0.60, 0.77)	0.54 (0.45, 0.63)
Excluding Epocrates	183	0.40 (0.33, 0.46)	0.71 (0.64, 0.79)	0.58 (0.50, 0.67)
Excluding Lexidrug	180	0.38 (0.31, 0.44)	0.71 (0.64, 0.79)	0.54 (0.45, 0.63)
Excluding Medscape	184	0.58 (0.51, 0.65)	0.72 (0.64, 0.79)	0.79 (0.73, 0.86)

Gwet's AC1 values range from -1 to 1. The agreement levels are categorized as follows: <0.20: Poor; 0.21-0.40: Fair; 0.41-0.60: Moderate; 0.61-0.80: Good; 0.81-1.00: Very Good.

In the binary classification distinguishing pDDIs requiring any form of clinical intervention (i.e., high- or intermediate-priority) from low-priority ones, the overall agreement was good (AC1: 0.72; 95% CI: 0.65-0.79). The highest agreement was observed when BNF was excluded (AC1: 0.75; 95% CI: 0.68-0.83), while the lowest was observed after excluding Drugs.com (AC1: 0.69; 95% CI: 0.60-0.77). In the binary classification focused on high-priority pDDIs, the overall agreement was slightly lower, despite remaining good (AC1: 0.62; 95% CI: 0.54-0.69). The highest estimate was observed when Medscape was excluded (AC1: 0.79; 95% CI: 0.73-0.86), whereas the lowest values was observed when Drugs.com or Lexidrug were excluded (both AC1: 0.54; 95% CI: 0.45-0.63). A sensitivity analysis was performed on PD pDDIs only, as classified by the BNF. Since action statements were generally lacking within this IC, this subset was previously excluded from agreement calculations. The resulting agreement among the remaining four ICs was moderate (AC1: 0.54; 95% CI: 0.44-0.63), slightly higher than the agreement observed in the full sample of

218 pDDIs when BNF was excluded (AC1: 0.44; 95% CI: 0.38-0.51), although confidence intervals overlapped.

Comparison with the Reference Document

Out of the 218 pDDIs classified as high/intermediate priority by the ANSM Thesaurus, 48 (22.0%) were classified as such by all five ICs, while 123 (56.4%) were classified as high/intermediate priority by at least four ICs (Fig. 2). Approximately 15% (n = 32) of the pDDIs were not considered high-priority or intermediate priority by any IC.

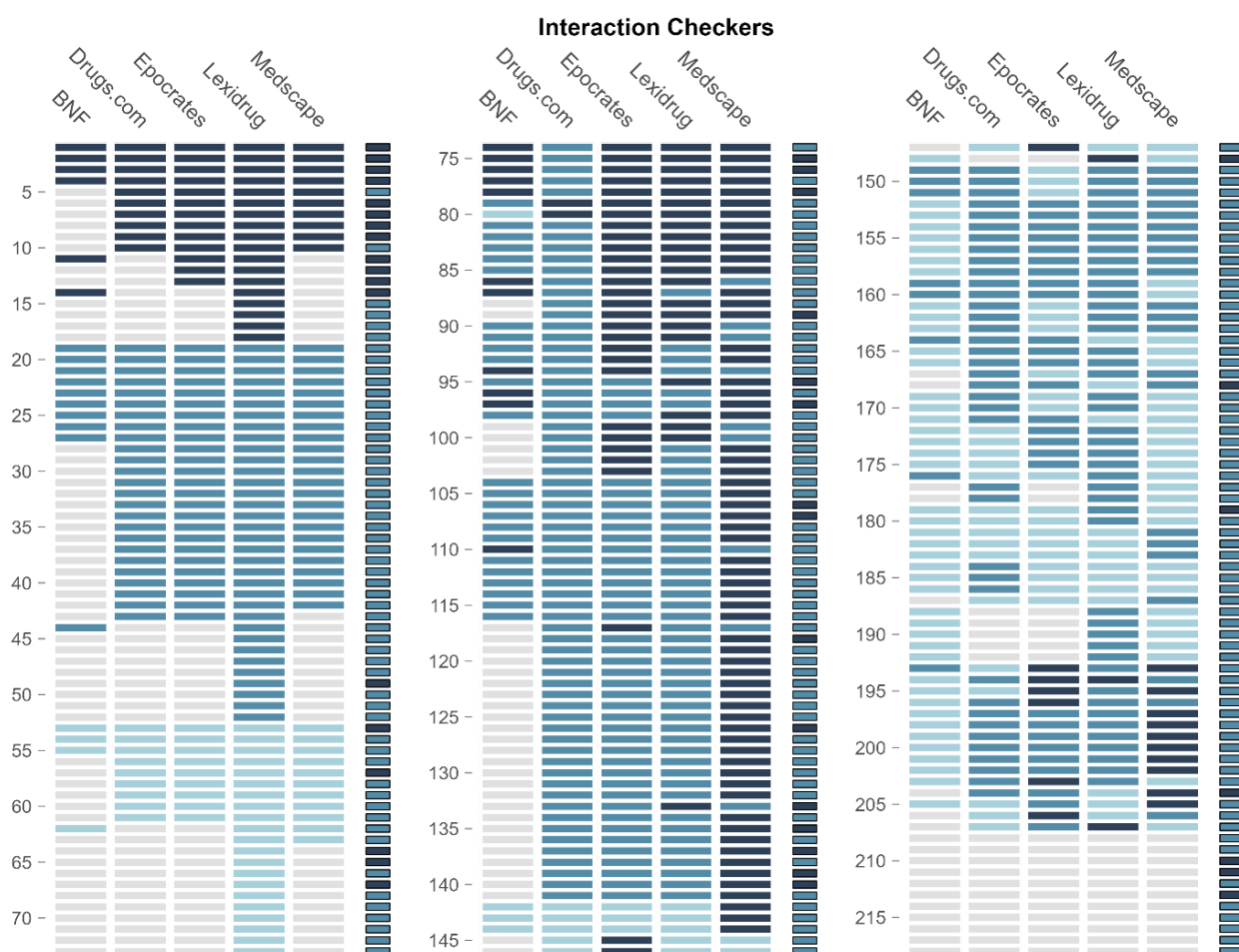


Figure 2. The table is divided into three consecutive sections for readability.

Colors represent management categories: Dark blue = high-priority, Intermediate blue = intermediate-priority, Light blue = low-priority, Grey = data unavailable. Rows are ordered by decreasing level of agreement across all five ICs (excluding “Data unavailable” classifications); in case of identical agreement, ordering follows decreasing management priority. The isolated rightmost column refers to the priority levels defined by the IC used for extraction (ANSM Thesaurus). For the BNF, 73 PD pDDIs were all classified as “Data unavailable” due to the near-systematic absence of associated management recommendations for this subset of pDDIs.

Potential DDI names are provided in Supplementary Table 6.

Only 4 of the 40 pDDIs classified as high priority by the ANSM Thesaurus were classified as such by all five ICs, while 11 (27.5%) were classified as high-priority by at least four ICs. More than a quarter of the pDDIs considered high priority by the ANSM Thesaurus were not considered as such by any other IC (11, 27.5%).

PPA in identifying high- or intermediate-priority pDDIs was highest for Epocrates (82.1%, 95% CI: 75.5-87.5), closely followed by Drugs.com (82.0%, 95% CI: 75.4-87.4). The lowest value was recorded for BNF (53.2%, 95% CI: 43.4-62.7) (Table 3).

When the analysis was restricted to high-priority pDDIs alone, the PPA was generally lower. The highest value was observed for Medscape (79.3%, 95% CI: 60.3-92.0), followed by BNF (62.5%, 95% CI: 35.4-84.8), while the lowest was observed for Drugs.com (28.6%, 95% CI: 13.2-48.7)

Table 3 Positive percent agreement between each interaction checker and the reference document in classifying pDDIs as high- or intermediate-priority, and as high-priority alone.

High- or intermediate-priority	
Interaction checkers	Positive Percent Agreement (95% CI)
ANSM Thesaurus - BNF	53.2% (43.4 - 62.7)
ANSM Thesaurus - Drugs.com	82.0% (75.4 - 87.4)
ANSM Thesaurus - Epocrates	82.1% (75.5 - 87.5)
ANSM Thesaurus - Lexidrug	80.7% (74.6 - 85.8)
ANSM Thesaurus - Medscape	76.5% (69.6 - 82.5)
High-priority	
Interaction checkers	Positive Percent Agreement (95% CI)
ANSM Thesaurus - BNF	62.5% (35.4 - 84.8)
ANSM Thesaurus - Drugs.com	28.6% (13.2 - 48.7)
ANSM Thesaurus - Epocrates	48.4% (30.2 - 66.9)
ANSM Thesaurus - Lexidrug	51.4% (34.4 - 68.1)
ANSM Thesaurus - Medscape	79.3% (60.3 - 92.0)

Positive Percent Agreement was defined as the proportion of pDDIs classified as high- and intermediate-priority, or high-priority by the ANSM Thesaurus that were also identically identified by the comparison ICs. It is expressed as a percentage with corresponding 95% confidence intervals.

Reported level of evidence supporting pDDIs

The level of evidence was reported for 159 pDDIs by Lexidrug and for 80 pDDIs by BNF (72.9% and 36.7% of the 218 analyzed pDDIs, respectively). A total of 78 (35.8%) pDDIs had an evidence level reported by both ICs (Fig. 3). The majority of pDDIs reported by BNF (n = 53, 66.3%) had the highest level of evidence (“Study”), with the remainder classified as “Anecdotal” (12, 15.0%) or “Theoretical” (15, 18.8%). Among the 80 pDDIs for which BNF provided a level of evidence, 17 were classified as high-priority (21.3%), 42 as intermediate-priority (52.5%), and 21 as low-priority (26.2%).

Among the 159 pDDIs reported by Lexidrug, 23 (14.5%) were classified with an “Excellent” level of evidence, 69 (43.4%) as “Good”, and 67 (42.1%) as “Fair”. These included 43 were pDDIs as high-priority (27.0%), 114 classified as intermediate-priority (71.7%), and 2 classified as low-priority (1.3%).

Only 6 pDDIs were classified with the highest level of evidence by both sources, two of which were categorized as high-priority, and three as intermediate-priority by both ICs. The most common combination involved pDDIs classified as having a “Good” level of evidence by Lexidrug and a “Study” level of evidence by BNF.

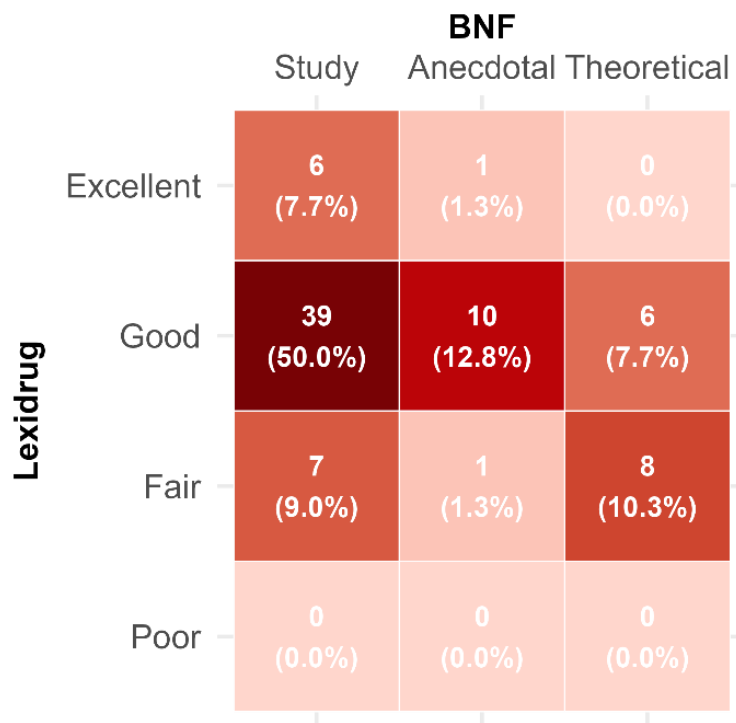


Figure 3 Comparison of DDI evidence level classifications by BNF and Lexidrug based on their respective classification scales. Each cell shows the absolute number of pDDIs, followed by the percentage relative to the total number of pDDIs classified simultaneously by both ICs ($n = 78$). No harmonization was applied between the two evidence grading systems, which differ substantially in structure and criteria. The heatmap is intended as a descriptive visualization of potential common classification patterns.

Discussion

To the best of our knowledge, this is the first study developing and testing a method for classifying, prioritizing, and comparing management recommendations across ICs. The application of this method shows, through multiple analyses, that there is moderate and variable agreement among ICs concerning the management recommendations of pDDIs. The method also identified management categories reflecting distinct and heterogeneous definitions across ICs, which can be used to prioritize pDDIs and provide information on IC consensus to support risk minimization strategies in real-world clinical practice.

The methodological approach of this study led to a restricted number of management categories to ensure applicability in real-world settings and to support better alert prioritization, helping reduce prescriber fatigue by focusing on pDDIs requiring prompt clinical action, especially in frameworks that integrate multiple ICs. (94,162). An iterative process led to the definition of four primary management categories: three reflecting distinct levels of required clinical intervention, and one additional category for drugs not listed in the IC database. To minimize the risk that differences between ICs were due to subjective interpretation, the adopted scale relied as much as possible on the presence and conditional nature of the recommended clinical actions.

The use of a reference standard enabled the assessment of the concordance of the ICs with the ANSM Thesaurus and a formal evaluation of the agreement among all other ICs.

The choice of the ANSM Thesaurus was deemed the most appropriate due to the origin of the real-world setting data used in this study (67). The focus on pDDIs that must always be avoided or that require an intervention was considered highly appropriate because it constitutes the priority for improving the quality of prescription and minimizing the risk of DDI-related adverse reactions. Additionally, the analysis was limited to pDDIs previously considered high- or intermediate priority by at least one official source (the ANSM), thus ensuring that the agreement assessment focused on a subset of pDDIs with greater clinical relevance.

Most primary and secondary analyses of this study indicate that the level of agreement between ICs in the management of clinically relevant pDDIs is limited and inconsistent, with substantial differences observed across sources. This lack of concordance persisted despite the use of a reduced set of essential management categories and the application of several mitigation strategies, including the exclusion of PD pDDIs from the BNF, and the further simplification of the classification into binary categories.

Although minor regional variations in DDI management may arise from differences in regulatory standards or clinical practice norms, such factors are unlikely to explain major discrepancies in management recommendations, which should ideally be based on common evidence-based principles (107). Missing data observed in certain ICs largely reflected their geographic scope (e.g., *BNF* for the UK, *Drugs.com* for the US), as these ICs generally did not include substances not marketed in their respective regions. Accordingly, the analyses were conducted only on complete datasets (e.g., in the main analysis, substances available in all five ICs).

These conceptual divergences remained even in specific subgroups, such as PD pDDIs. Secondary analyses on this subset showed only marginal increases in agreement, suggesting that the mechanism of interaction alone is insufficient to explain the observed inconsistencies. The differences observed between ICs are rooted in conceptual and structural divergences, not merely in differences in classification strictness. This is supported by the fact that excluding certain sources improved agreement in some binary classifications but reduced it in others (e.g., *BNF*).

While a higher agreement was observed when reducing the number of categories (e.g., high- or intermediate-priority *vs.* low-priority), this may result from reduced disagreement potential in less granular categorization schemes. It may be argued that, instead, the lower agreement observed in the main analysis reflects a misclassification of action statements. However, the classification system was independently validated by an external evaluator not involved in its development, showing very good agreement with the expert-based categorization.

The limited agreement between ICs in rating pDDI severity or level of evidence has been observed in various clinical settings over the past fifteen years (79,80,143,145,163). However, expert guidance suggests assessing interactions based on multiple elements,

including patient-specific modifying factors, quality of evidence, and management recommendations. Severity categories used in most ICs should therefore be used complementarily to management recommendations, as they may provide only generic guidance for healthcare providers. (107). For example, pDDIs defined as severe by the BNF may have several different management recommendations, such as complete avoidance of administration (e.g., fenofibrate-pravastatin), avoidance except under specific monitoring (e.g., phenobarbital-rivaroxaban), or just a dose adjustment (e.g., fenofibrate-rosuvastatin). Few studies have included information on pDDIs management proposed by different ICs (69,74,84,164), generally suggesting a lack of consistency. None of these studies has attempted to standardize and compare these recommendations using a common scale designed to prioritize pDDIs according to their management and their required level of clinical intervention. To address these issues, the present work proposed a method for prioritizing pDDIs on risk minimization measures with a direct translation into clinical practice.

This study also evaluated the clinical evidence supporting pDDIs. Among the six ICs analyzed, only BNF and Lexidrug reported this information, and it was specified by both of them in just one-third of analyzed pDDIs. Furthermore, only a very small subset of pDDIs reached the highest evidence level in both ICs (6 out of 218 pDDIs). These results align with two recent studies that compared Micromedex *vs.* Lexidrug or BNF (69,145).

The absence of transparent evidence grading in most ICs, the limited concordance between the grading systems used by BNF and Lexidrug, and the low proportion of pDDIs supported by high-level evidence all raise concerns about the transparency and reliability of the recommendations provided to clinicians. These factors may also partially explain the observed lack of agreement in management recommendations across ICs. This supports previous recommendations that DDI alerts should include explicit evidence ratings, brief

evidence summaries, and direct access to primary references whenever possible, in order to allow clinicians to directly assess the strength and clinical relevance of the information provided (107).

This study has some limitations. The random sampling approach used to construct the classification system may have resulted in underrepresentation of specific DDI mechanisms or clinical outcomes. The sample size of 218 unique pDDIs could be considered limited, but to date, this is the most extensive study on pDDI management in real-world settings. Nonetheless, a further study using other pDDIs could be useful to confirm these findings. The algorithm used to extract medications from clinical records was not formally validated, making it plausible that not all existing pDDIs were actually captured. Nevertheless, this does not impact the focus of the present work, namely the standardization and comparison of management recommendations across ICs.

The study was restricted to a single clinical field, namely cardiology, but it included all medications included in hospital prescription records. The choice of this specific setting was made because of the high prevalence of older patients with multiple comorbidities and thus highly exposed to pDDIs. Moreover, this study found that more than three-quarters of the pDDIs involved at least one non-cardiological drug, making the results at least in part generalizable. The descriptive analysis of the present study was limited to every unique pDDI, without considering their frequency of occurrence among patients. This limitation does not pertain to the methodological design of the present work, which was not designed to focus on the epidemiology of pDDIs. Finally, SmPCs were not included among the sources used for pDDI identification and agreement assessment, as they often provide broad and non-specific information on interacting substances, particularly when referring to PD effects (e.g., drugs prolonging the QT interval, drugs inducing hypokalemia), or to pharmacokinetic properties (e.g., drugs inhibiting CYP3A4 isoenzyme). Although comparing IC content with

SmPCs could be a useful approach to better clarify their added value, such an analysis would require a dedicated methodological framework and was therefore beyond the scope of this study.

The low level of agreement, coupled with limited evidence information, points to the need for standardized principles for developing ICs, grounded in evidence-based practices, to improve consistency and reliability.

Future research may extend the methodological approach developed in this study to other prescription datasets, including other high-risk areas, such as neurology, oncology, or internal medicine, to better understand pDDI management across different contexts, focusing on high- or intermediate-priority ones. It will also be useful to evaluate how feasible this approach is for prescribers and clinical pharmacologists/pharmacists working in hospital wards, as well as its impact on clinical practice, including whether it can effectively contribute to reducing alert fatigue. This approach could also potentially be integrated within automated systems that combine multiple ICs to rapidly help clinicians in identifying and managing pDDIs to be avoided or requiring a clinical action, thereby supporting real-time support for clinical decision-making.

Conclusions

This study developed and tested a method for classifying, prioritizing, and standardizing the management recommendations for pDDIs, enabling systematic comparison across multiple ICs. Its application revealed a low and inconsistent agreement among ICs, indicating the importance of comparing multiple ICs for prioritizing the pDDIs that must be avoided or specifically addressed in clinical practice.

3.3 Lessons learned and methodological considerations

The results of the first thesis study indicate that a filtering and prioritization strategy based solely on management recommendations cannot yet be implemented without caution. The initial assumption was methodologically reasonable: defining a limited number of broad and clearly delimited categories of pDDIs should reduce ambiguity and facilitate comparability across ICs, thereby enabling integration into decision-support systems.

Under conditions of sufficient agreement across these categories, a restrained clinical interface would have been feasible. In a French hospital setting, one could reasonably anchor the system to a primary reference such as the ANSM Thesaurus and alert clinicians only in cases of rare or atypical discrepancies on the management of pDDIs. This approach would have limited over-alerting while preserving transparency in genuinely controversial situations.

However, the results of this study did not support this scenario. Even after standardization, discordance between ICs remains substantial and extends beyond minor wording or regional differences. In this context, a management-based filter risks generating the opposite of its intended effect. When sources provide strongly different operational recommendations for clinically relevant pDDIs, presenting excessive discrepancies would not improve prescribing quality; conversely, it may increase cognitive load and weaken trust in the tool.

The results of this study can have several implications. First of all, the systematic mapping of discordances provided a structured inventory of instability within ICs. These discordances need to be reviewed by clinical pharmacologists, rather than being used simply as an automated filter. Frequent or clinically relevant discrepancies can be selected, analyzed in depth, and discussed to determine whether they reflect meaningful differences in evidence interpretation or limitations of the reference framework. When marked divergences or clinically complex situations are identified, targeted

educational sessions can be organized for hospital physicians. Secondly, this study found areas of persistent non-convergence that could help in defining priorities for further investigation and evidence generation. These discrepancies may reflect differences in evidentiary standards, editorial policies, and pharmacological generalizations at the class level, but they may also reflect genuine limitations in the underlying clinical evidence. More broadly, an epistemic objective remains: to orient ICs toward greater transparency and, where possible, a more explicit shared reference framework. The objectives outlined in the latter points represent valid research trajectories, but they are likely to be achieved in the long term, given the uncertainty and the variability of IC decision rules. Thirdly, and more operationally, results indicate that the most reasonable immediate step is to only use a single reference IC, having in mind a controlled, context-anchored implementation strategy, accompanied by ongoing methodological refinement. In a French hospital setting, the ANSM Thesaurus represents the most coherent choice. This consideration constituted the logical premise of the second study.

4 - Development of a real-world derived gold standard for high-severity drug-drug interactions and its application to the validation of an automated detection tool

4.1 Context and Objectives

In hospital practice, particularly at discharge, reliable identification of high-severity pDDIs is of utmost importance. Discharge is a transitional phase in which prescription oversight becomes less structured, while therapeutic complexity, especially among older, polymedicated patients, remains high. In this context, an automated system is meaningful only if grounded in an institutional standard consistent with the national regulatory framework and formalized in a robust and reproducible way.

This study arose from the need to transform an institutional source, the ANSM Thesaurus, into a structured and evaluable operational reference. In France, the ANSM Thesaurus is the official document for pDDI identification, but it does not function as a ready-to-use gold standard. Its partially free-text content must be translated into explicit decision rules, multilevel situations must be handled systematically, and an expert annotation process must be implemented.

To meet this requirement, a gold standard was constructed from real-world hospital prescriptions using double independent review with adjudication of discordant cases. This approach allowed definition of an operational reference standard for high-severity pDDI identification and quantification of inter-expert agreement.

Application of this standard to a hospitalized cohort enabled estimation of the prevalence of high-severity pDDIs in routine practice and assessment of their association with polypharmacy. The same reference standard was then used to evaluate the initial diagnostic performance of an NLP-based automated tool (PRoSIT), providing a first empirical assessment of the feasibility of a system specifically focused on high-severity pDDIs in a French setting.

The study pursued three main objectives:

- (i) to define, from real-world hospital prescriptions, an expert-derived gold standard for identifying high-severity pDDIs based on the ANSM Thesaurus, and to estimate its reliability;
- (ii) to estimate the prevalence of patients exposed to at least one high-severity pDDI and its association with polypharmacy;
- (iii) to evaluate the initial diagnostic performance of an automated pDDI identification system (PRoSIT) against the expert-derived gold standard.

Note. Tables and figures follow the internal numbering system used within the study and have not been renumbered across chapters. Abbreviations are defined at their first occurrence in each article, regardless of previous definitions in the thesis.

4.2 Article

Development of a Real-World Derived Gold Standard for High-severity Drug-Drug Interactions and Its Application to the Validation of an Automated Detection Tool

Funding: This study was part of the PROSIT project, supported by a grant from the French Ministry of Health (PREPS, 2019, 19-0170)

Ethics Approval: Data extraction was approved by the Institutional Review Board of the Bordeaux University Hospital (CER-BDX 2024-93). The study was conducted in accordance with the ethical standards defined by the 1964 Declaration of Helsinki and later amendments or comparable ethical standards.

Abstract

Background

The identification of potential drug-drug interactions (pDDIs) is a major concern. Manual detection of pDDIs is time-consuming and error-prone. Automated tools may provide essential support for rapid and accurate detection. A key limitation in the development of such tools is the evaluation of their performance compared to well-constructed and expert-derived gold standards.

Objectives

This study aimed to evaluate the reliability of a gold standard for high-severity pDDIs, based on the ANSM Thesaurus, to estimate the prevalence of high-severity pDDIs in the hospital-discharge setting, and to test it for the validation of a NLP-based detection tool.

Methods

A gold standard was developed from more than 30,000 unique drug pairs, extracted from a real-world cohort of 6,027 patients aged ≥ 65 years with a discharge prescription containing one or more individual drugs. The dataset was independently annotated by two experts, with a third resolving discordant cases, following a standardized and transparent decision algorithm. The agreement between evaluators was calculated using the Gwet's AC1 statistic. Moreover, the gold standard was used to calculate the prevalence of high-severity pDDIs and its related factors, as well as the diagnostic performance of a NLP-based detection tool, particularly focused on Sensitivity (Se) and Negative Predictive Value (NPV), thus validating a tool based on a rule-out approach.

Results

Agreement between evaluators was very good (Gwet's AC1 = 0.94; 95% CI: 0.94-0.96), even when prescriptions with just one drug were excluded (AC1 = 0.93; 95% CI: 0.92-0.94). High-

severity pDDIs were identified in 9.7% of patients, and their occurrence was significantly associated with the number of drugs reported in the discharge prescription.

An experimental NLP-based pharmacoinformatic tool designed to detect high-severity pDDIs showed suboptimal performance (Sensitivity 63.1%; Negative Predictive Value 96.0%). The performance increased when pDDIs initially related to a discordance among experts were excluded (Sensitivity 75.7%; Negative Predictive Value: 98.4%).

Conclusions

This study provided a reliable gold standard based on the ANSM Thesaurus for the validation of tools targeting high-severity pDDIs in discharge hospital context. Artificial intelligence tools could be used in routine practice, but need to be better tailored for complex pDDIs.

Introduction

Drug-Drug interactions (DDIs) refer to situations where the effect of a drug is modified by the administration of another substance (53).

Potential DDIs (pDDIs) occur when two co-administered substances are known to have interaction risks, potentially leading to adverse drug reactions (ADRs). DDIs and pDDIs are particularly relevant in clinical practice, with reported incidences of up to 17.7% and 64.9% among hospitalized patients, respectively, although substantial variability across studies has been reported (139). A French study analyzing the incidence of ADR-related hospitalizations found that up to 150,000 admissions per year are linked to ADRs, with 30% of these cases attributed to DDIs (165). Elderly patients face an increased risk of both DDIs and pDDIs (166). One of the main contributing factors is polypharmacy, commonly defined as the use of at least five medications (39,167). As the number of prescribed medications grows, the potential combinations of drugs increase exponentially, raising the likelihood of inadvertently co-administering interacting substances. At the same time, aging leads to a gradual decline in liver and kidney function, reducing drug clearance and altering distribution, metabolism, and excretion. These changes may increase not only the risk but also the severity of DDIs, by raising drug exposure and toxicity (46). It is estimated that DDIs account for 2% to 5% of hospital admissions among older patients (168,169).

To address the clinical need for more effective detection and prevention of pDDIs in elderly patients, various interaction checkers (ICs) integrated within clinical decision support systems (CDSS) have been developed, assisting clinicians by analyzing clinical data and identifying pDDIs (70). However, the effectiveness of these tools is often limited by the frequent generation of low-severity alerts, which can overwhelm healthcare providers. This “alert fatigue” may result in low-severity warnings being ignored, reducing the effectiveness of safety alerts (162,170). As a result, the real potential of ICs within CDSS remains limited,

potentially exposing patients to an increased risk of DDIs. In this context, hospital discharge represents one of the most critical phases in patient care. While medications prescribed during hospitalization are expected to carry a lower risk of pDDIs, since they undergo a dual review by both the prescribing physician and the hospital pharmacist, the hospital discharge prescription, defined as the list of medications prescribed at discharge or following a consultation (whether new or ongoing treatments), does not receive the same level of oversight, thereby carrying a higher risk of pDDIs.

A 2015 survey supervised by the French Health Authority identified two main barriers to the broader adoption of “medication reconciliation”: a shortage of human resources and the lack of suitable tools and software (171). Moreover, existing CDSS platforms are not designed to support discharge processes and may not provide an adequate automated identification of pDDIs during this phase of patient care. These limitations emphasize the need for automated pharmacoinformatic tools that prioritize high-severity pDDIs to minimize the risk of interaction-related ADRs while also mitigating the issue of alert fatigue.

To address these challenges, well-defined gold standards are essential, as they provide a reliable reference for the repeatable evaluation of automated tools (172). In other words, the development of pharmacoinformatic solutions for the automated detection of pDDIs requires a validated reference against which the information produced by such tools can be compared. In France, a reference document for the identification of pDDIs is available: the ANSM Thesaurus. This thesaurus, published by the French National Agency for Medicines and Health Products Safety (ANSM), lists all pDDIs to be considered in clinical practice and represents a widely used and reputable source for pDDI screening in the French healthcare system (66,67). Its primary purpose is to facilitate and secure drug prescribing. For a pDDI to be included, it must have a clinically significant impact, either by causing or exacerbating adverse effects, or by reducing therapeutic efficacy through impaired drug activity. The

ANSM Thesaurus further specifies whether the pDDI applies to a specific drug or an entire therapeutic class, the mechanism of interaction, and the recommended preventive measures. Given its characteristics, the ANSM Thesaurus provides a basis for the development of automated tools for pDDI research. It was for example used to identify DDIs leading to emergency department visits and hospitalizations in population-based studies (55,160).

However, no gold standards derived directly from the ANSM Thesaurus are currently available to evaluate the performance of such tools in pDDI identification.

On this basis, the main objective of this study was to evaluate the reliability a real-world derived gold standard, primarily intended for the validation of any kind of tool that aims at identifying high-severity pDDIs using hospital discharge prescriptions. The gold standard was built following a standardized and transparent procedure, based on the information provided by the ANSM Thesaurus.

As secondary objectives, the study aimed to describe the population used for gold standard development, quantify the degree of polypharmacy, determine the prevalence of high-severity pDDIs, and identify the drugs and therapeutic classes most frequently prescribed and/or involved in pDDIs. Finally, the gold standard was applied to test the diagnostic performance of an experimental tool based on Natural Language Processing (NLP) techniques, specifically designed for the identification of high-severity pDDIs.

Methods

Source of information for the definition of pDDIs: ANSM Thesaurus

The ANSM Thesaurus classifies pDDIs into four levels of severity (146), defined according to the potential severity of the pDDI and the appropriate response measures:

- Contraindicated (*Contre-indication*, CI): the association must always be avoided.

- Avoid unless benefits outweigh harms (*Association déconseillée*, ASDEC): the association should be avoided unless the benefit-risk ratio has been thoroughly evaluated and the patient is strictly monitored.
- Needing cautious use (*Précaution d'emploi*, PE): the two drugs may be co-administered; however, specific recommendations should be followed to prevent adverse reactions (e.g., dosage adaptation, enhanced monitoring).
- To be taken into account (*À prendre en compte*, APEC): a risk of adverse reactions exists, but the prescriber should assess the appropriateness of the association.

In some cases, the ANSM Thesaurus provides more than one severity level for the same pDDI (hereafter defined as multilevel pDDIs), when a single severity level cannot be assigned due to substance-related circumstances (e.g., dose, route of administration) or patient-related factors (e.g., age, renal function).

Within the framework of this project, pDDIs belonging to the first two categories (CI and ASDEC) were defined as high-severity pDDIs, as they correspond either to combinations that must be strictly avoided or to associations requiring close monitoring, which cannot be systematically ensured after hospital discharge. By contrast, pDDIs classified as APEC or PE were defined as low-severity. The latest version of the ANSM Thesaurus available at the beginning of the gold standard development (January 2025) was used for the purposes of this study.

Source of clinical data

Clinical records were extracted from the CHU de Bordeaux data warehouse. This institution manages approximately 250,000 hospitalized patients and 100,000 emergency admissions annually. In 2017, a data warehouse based on the i2b2 structure was implemented (153,155). The warehouse contains diagnoses, laboratory results, drug prescriptions, and various other

variables, and is organized as a multi-table relational database, allowing more efficient queries compared with traditional compartmentalized systems (155). To perform queries, an ontological alignment was performed, and NLP tools such as Romedi, a French terminology system for extracting drug-related information from text, were developed (157).

Study population

Prescriptions were defined as the list of drugs prescribed either at hospital discharge or during a consultation, including both new and ongoing treatments, as recorded in the hospital information system.

A prescription was included if all of the following conditions were met:

- Patient characteristics: age ≥ 65 years
- Hospitalization or consultation context: departments of cardiology, internal medicine, or neurology
- Period: January 1, 2020 - December 31, 2023
- Prescription requirements: at least one prescribed drug.

Since each prescription corresponded to one patient the latter term was preferentially used throughout the text.

Manual data extraction and development of the gold standard

Patients were randomly selected from the hospital data warehouse across three departments, Cardiothoracic, Internal medicine, Neuroscience (Figure 1 for a workflow of the extraction process). The random selection was designed to preserve the age and sex distribution of the source population.

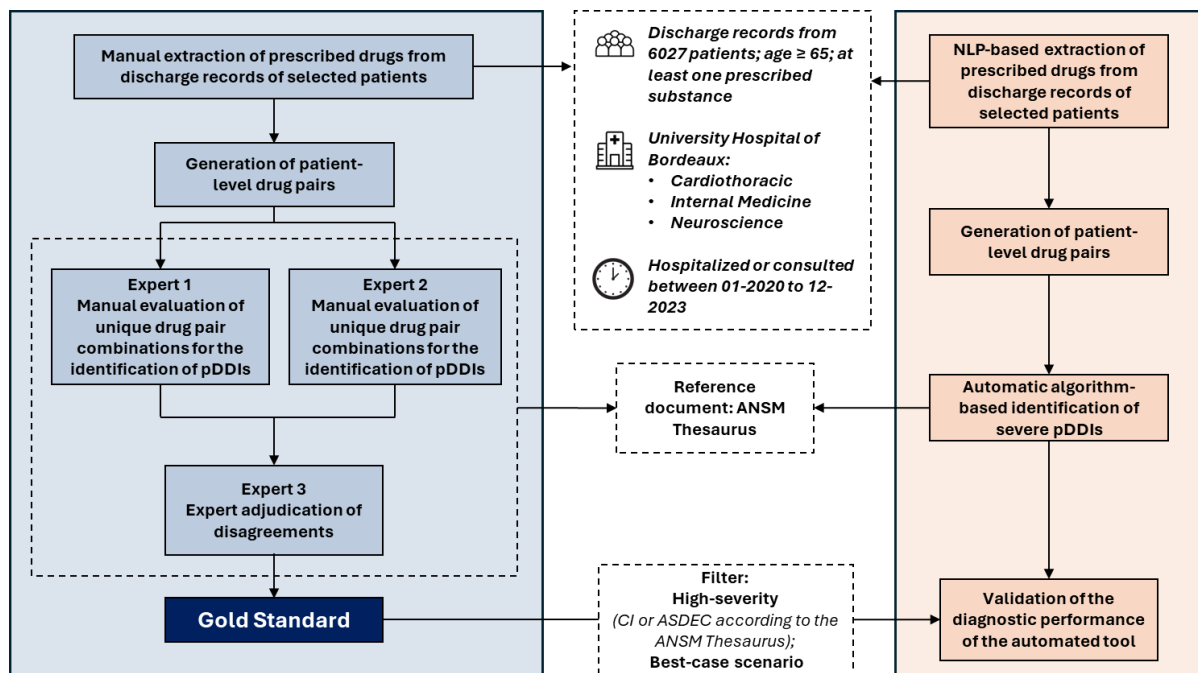


Figure 1. Workflow of the gold standard construction (blue) and of the automated tool which was validated against the gold standard (orange).

The following steps were undertaken to prepare the gold standard:

- Two clinical research assistants retrospectively retrieved and anonymized the prescriptions from the discharge records of selected patients. Extracted information included the hospital service, unit, and department where the patient was hospitalized or consulted, as well as the list of prescribed drugs.
- After data extraction, two members of the Methodological Support Unit for Clinical and Epidemiological Research (USMR) at the University Hospital of Bordeaux compiled a list of all unique drug pairs at the International Nonproprietary Name level. This list was then further reviewed to eliminate duplicates, yielding a final database of 30,960 unique drug pairs.

Prior to the establishment of the gold standard, a drug pair evaluation module was developed by the USMR, then tested and validated by two evaluators. In addition, a supplementary web-based tool was created to facilitate navigation of the ANSM Thesaurus.

For each unique drug pair, several types of information were independently extracted by both evaluators and compiled in the corresponding drug pair evaluation module (in detail, Supplementary File 1). The gold standard was targeted to the identification of patients with ≥ 1 high-severity pDDI. A patient was considered to have a high-severity pDDI if at least one of their drug pairs was classified as high-severity. In cases where a pDDI could assume different severity levels depending on therapy or clinical context, a best-case scenario approach was applied. This approach assumed that the physician operated under optimal clinical conditions (patient characteristics, dosage, route of administration) (Figure 2).

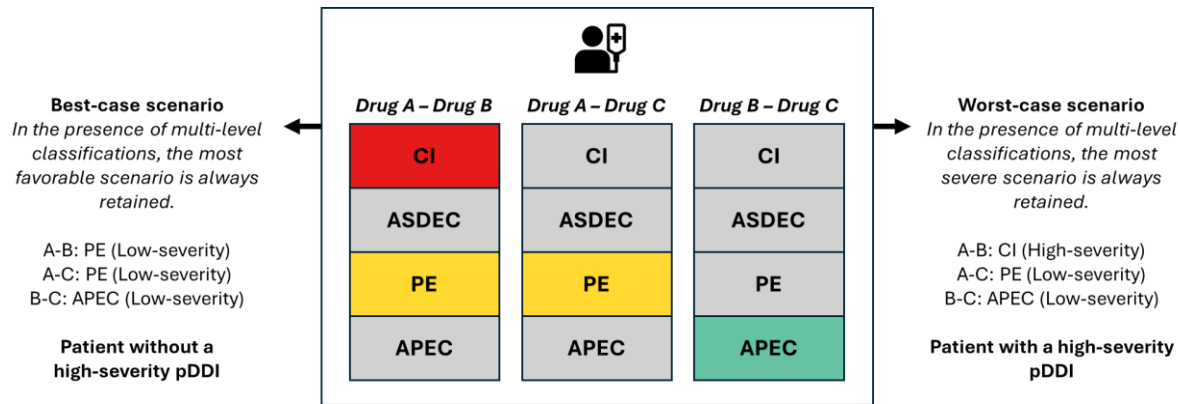


Figure 2 - Assignment of severity levels according to best-case and worst-case scenarios.
The worst-case scenario, in which any potential high-severity classification is considered, is presented for clarity but was not applied in the analyses.

In parallel, additional information was collected (e.g., classification of low-severity pDDIs), although this was not the primary objective. These ancillary data were included to enable further comparative and sensitivity analyses in subsequent phases. The construction of the gold standard was carried out in parallel by two evaluators, each blinded to the other's decisions. To ensure consistency across evaluators, a predefined decisional algorithm was applied. Three possible scenarios were defined:

- Positive concordance: both evaluators agreed on the severity of the pDDI (i.e., both CI or both ASDEC).
- Negative concordance: both evaluators agreed that the drug pair did not represent a pDDI.
- Discordance: any situation where the evaluators provided different classifications, including disagreement on the presence of a pDDI or on its specific severity level.

In all cases of discordance, irrespective of whether the disagreement involved high-severity or low-severity categories, a third evaluator independently reassessed the drug pair, acting as a tiebreaker (see Supplementary File 1 for details). Once all disagreements had been resolved, the resulting dataset was assumed as the ground truth, and the gold standard was considered established.

REDCap Version 14.7.1 was used for data extraction and manual identification of drugs and pDDIs.

Automated tool for the extraction and identification of pDDIs

The tested automated tool proceeded to automatically extract drugs using an algorithm based on NLP techniques, including Named Entity Recognition (159). Thereafter, for each discharge prescription a list of all possible drug-drug pairs was generated. Detection of pDDIs from these lists was then carried out using an automated pharmacoinformatic tool, based on the ANSM Thesaurus (55,160).

Statistical analysis

Description of the population

The study population was analyzed in its entirety (n = 6027), and stratified according to the corresponding hospital department. The distribution of individual prescribed drugs was

described both globally and by department, with identification of the most frequently prescribed drugs.

The degree of polypharmacy was assessed by counting the number of distinct drugs per patient. Polypharmacy was defined as the use of ≥ 5 drugs, while hyperpolypharmacy was defined as the use of ≥ 10 drugs (38). The prevalence of both conditions was estimated overall and by department, and 95% CI were calculated.

The number of drugs per patient was summarized as median (IQR), and compared across departments using the Kruskal-Wallis test. Pairwise comparisons were performed using Dunn's test with Holm correction for multiple testing.

The most frequent high-severity pDDIs were identified both globally and stratified by hospital department.

The prevalence of patients presenting with at least one high-severity pDDI was calculated for the entire population and stratified by department, and 95% CI were calculated.

To characterize the drugs most frequently involved, the following were calculated: a) the absolute number of high-severity pDDIs per drug; b) the proportion of patients with ≥ 1 high-severity pDDI among the total patients exposed to each drug, considering two minimum thresholds of 5 and 20 exposed patients.

Consistency and reliability of the gold standard

To evaluate the consistency and quality of the gold standard, inter-expert agreement was primarily evaluated using Gwet's AC1 statistic (161). Gwet's AC1 values range from -1 to 1. The agreement levels were categorized as follows: <0.20 : Poor; 0.21-0.40: Fair; 0.41-0.60: Moderate; 0.61-0.80: Good; 0.81-1.00: Very Good (161). In addition, positive and negative specific agreement were reported, as they provide complementary information on concordance within the minority (high-severity) and majority

(not high-severity) classes, which is especially relevant in imbalanced datasets (173,174). All metrics were calculated at patient level.

A secondary analysis was conducted using the same metrics (Gwet's AC1, positive and negative specific agreement), excluding patients with only one prescribed drug, as their inclusion could have artificially inflated the results.

Application of the gold standard to an NLP-based tool (PRoSIT)

The performance of the tested tool in identifying pDDIs was assessed against the gold standard, using the following parameters:

- True Positive (TP): number of patients with high-severity pDDI identified both by the gold standard and the automated tool.
- True Negative (TN): number of patients without high-severity pDDI identified both by the gold standard and the automated tool.
- False Positive (FP): number of patients without high-severity pDDI according to the gold standard but identified as having pDDIs by the automated tool.
- False Negative (FN): number of patients with high-severity pDDI according to the gold standard but not identified by the automated tool.

These four parameters formed the contingency table used for performance evaluation.

The following performance metrics of the tested tool were therefore calculated:

- Sensitivity (Se): proportion of patients with at least one high-severity pDDI correctly identified by the automated tool out of all patients with a high-severity pDDI according to the gold standard (175).
- Negative Predictive Value (NPV): proportion of patients classified without a high-severity pDDIs by the automated tool that were confirmed as not having a high-severity pDDIs by the gold standard.

- Specificity (Sp): proportion of patients without a high-severity pDDI correctly identified by the automated tool out of all patients without a high-severity pDDI according to the gold standard.
- Positive Predictive Value (PPV): proportion of patients classified with a high-severity pDDIs by the automated tool that were confirmed as having a high-severity pDDIs by the gold standard.
- Accuracy: overall proportion of patients correctly classified, combining both positive and negative cases (176).
- F1-score: harmonic mean of sensitivity and PPV, used to summarize performance when correct detection of positives is of primary interest and the data are unbalanced.
- Youden's J index: a summary statistic of diagnostic effectiveness, independent of pDDI prevalence, combining sensitivity and specificity into a single measure.

Among the performance metrics, Se and NPV were considered of primary importance given the rule-out objective of the tool, while the other indices were used to provide a comprehensive assessment of its overall performance.

The sample size estimation was based on the precision of the sensitivity (Se) confidence interval, calculated using an exact binomial method. Assuming an expected sensitivity of 98% with the lower bound of the confidence interval above 95%, the study required 173 patients (95% confidence interval for sensitivity: [0.950; 0.996]). Using literature estimates of the frequency of patients with high-severity pDDI at approximately 3% (177), the minimum number of patients to include was 5,769, corresponding to 1,923 patients per department. To account for variability, a small excess of patients was considered, resulting in 6,027 patients finally included.

For all metrics, the 95% two-sided confidence interval was calculated. As an exploratory post-hoc analysis, the performance metrics of the tool were recalculated separately considering only patients for whom the two experts independently reached the same conclusion regarding the presence of at least one high-severity pDDI at the patient level, without the need for adjudication by the third evaluator. This definition of concordance should not be confused with the “positive concordance” described in Section 2.5, which referred to drug-pair-level agreement on the classification of pDDIs, across all severity categories (i.e., both high- and low-severity pDDIs)

The evaluation adhered, for the items applicable, to the STARD (Standards for Reporting Diagnostic Accuracy Studies) guidelines (178).

Results

Description of population

The overall population included 6,027 patients. The distribution by department showed 2,066 subjects (34.3%) in the Cardiothoracic, 2,073 (34.4%) in Internal Medicine, and 1,812 (30.0%) in Neurosciences. A total of 76 patients (1.3%) were not assigned to any department and were excluded from all the subsequent stratified analyses.

Prescription patterns

A total of 775 unique drugs were identified. The most common drugs were reported in Table 1.

Table 1. List of the 10 most frequent prescribed drugs in the overall population and by department.

Rank	Overall (n: 6,027; %)	Cardiothoracic (n: 2,066; %)	Internal Medicine (n: 2,073; %)	Neurosciences (n: 1,812; %)
1	Acetylsalicylic acid (1470; 24.4%)	Bisoprolol (704; 34.1%)	Paracetamol (652; 31.5%)	L-dopa - benserazide (417; 23.0%)
2	Paracetamol (1383; 22.9%)	Acetylsalicylic acid (665; 32.2%)	Cholecalciferol (584; 28.2%)	Paracetamol (360; 19.9%)
3	Bisoprolol (1126; 18.7%)	Furosemide (617; 29.9%)	Acetylsalicylic acid (450; 21.7%)	Acetylsalicylic acid (349; 19.3%)
4	Lansoprazole (1079; 17.9%)	Lansoprazole (511; 24.7%)	Prednisone (422; 20.4%)	Atorvastatin (334; 18.4%)
5	Atorvastatin (961; 15.9%)	Apixaban (444; 21.5%)	Lansoprazole (383; 18.5%)	Lansoprazole (182; 10.0%)
6	Furosemide (961; 15.9%)	Atorvastatin (424; 20.5%)	Valaciclovir (352; 17.0%)	Bisoprolol (151; 8.3%)
7	Cholecalciferol (846; 14.0%)	Paracetamol (368; 17.8%)	Trimethoprim - sulfamethoxazole (271; 13.1%)	Apixaban (150; 8.3%)
8	Apixaban (840; 13.9%)	Spirolactone (295; 14.3%)	Bisoprolol (269; 13.0%)	Macrogol (148; 8.2%)
9	Levothyroxine (593; 9.8%)	Amiodarone (288; 13.9%)	Folic acid (266; 12.8%)	Levothyroxine (125; 6.9%)
10	Prednisone (555; 9.2%)	Potassium chloride (279; 13.5%)	Furosemide (250; 12.1%)	Perindopril (124; 6.8%)

Values are reported as absolute numbers with percentages relative to each group. 76 patients could not be classified into any of the three departments.

The median number of drugs per patient was 5 (IQR = 2 - 9; range 1-24). Stratified values were: 7 (IQR = 4 - 9) in Cardiothoracic, 6 (IQR = 2 - 10) in Internal Medicine, and 3 (IQR = 2 - 6) in Neurosciences. The distribution of the number of drugs per patient differed significantly across departments (Kruskal-Wallis $\chi^2 = 346.9$; $p < 0.0001$). Pairwise tests showed lower numbers of drugs in the Neurosciences compared to both other departments (adj-p < 0.0001). Additional information on the degree of polypharmacy in the population is

reported in Table 2. The distribution of patients by number of drugs received, both in the overall population and stratified by department is reported in Figure 3.

Table 2. Prevalence of polypharmacy and hyperpolypharmacy in the overall population and by department

	Polypharmacy, ≥ 5 drugs Prevalence % (95% CI)	Hyperpolypharmacy, ≥ 10 drugs Prevalence % (95% CI)
Overall	55.4% (54.2-56.7)	19.6% (18.6-20.6)
Cardiothoracic	67.8% (65.8-69.8)	22.9% (21.2-24.8)
Internal Medicine	58.0% (55.8-60.1)	25.0% (23.2-26.9)
Neurosciences	40.0% (37.7-42.2)	10.2% (8.8-11.6)

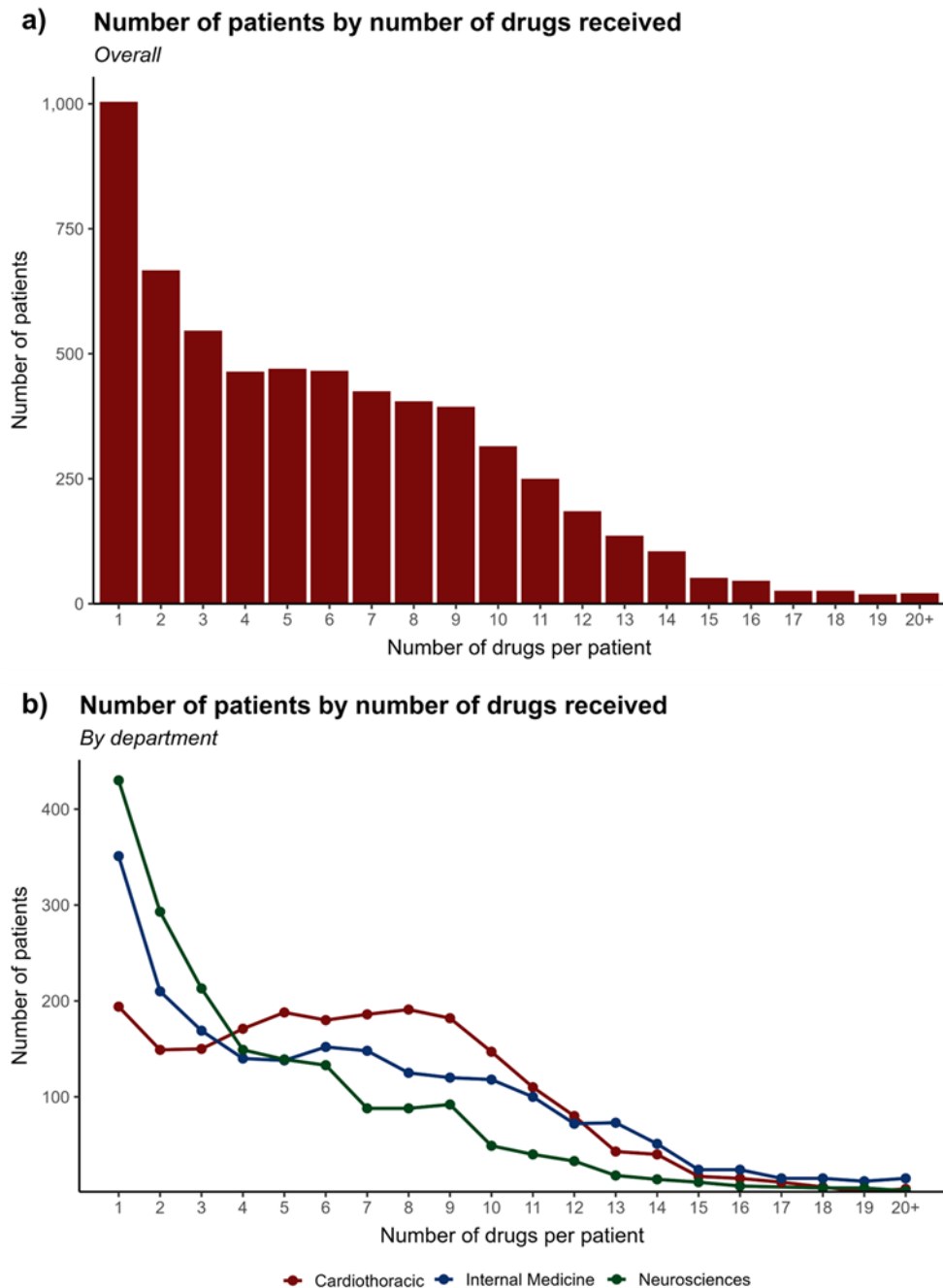


Figure 3 - Distribution of patients by number of drugs received.

In detail: panel a) Overall distribution of patients according to the number of drugs per patient; panel b) Distribution stratified by department

High-severity pDDIs

A total of 586 out of 6027 patients presented at least one high-severity pDDI, corresponding to a prevalence of 9.7% (95% CI 8.99-10.50). 18.3% of these patients presented more than

one high-severity pDDI (n=107). Logistic regression showed an association between the number of drugs and the risk of pDDIs: compared to 2-4 drugs, patients with 5-9 drugs had OR = 5.18 (95% CI 3.62-7.42; $p < 0.0001$), while those with ≥ 10 drugs had OR = 17.58 (12.3-25.05; $p < 0.0001$). The proportion of the population exposed to at least one high-severity pDDI, according to the minimum number of drugs taken, is shown in Figure 4.

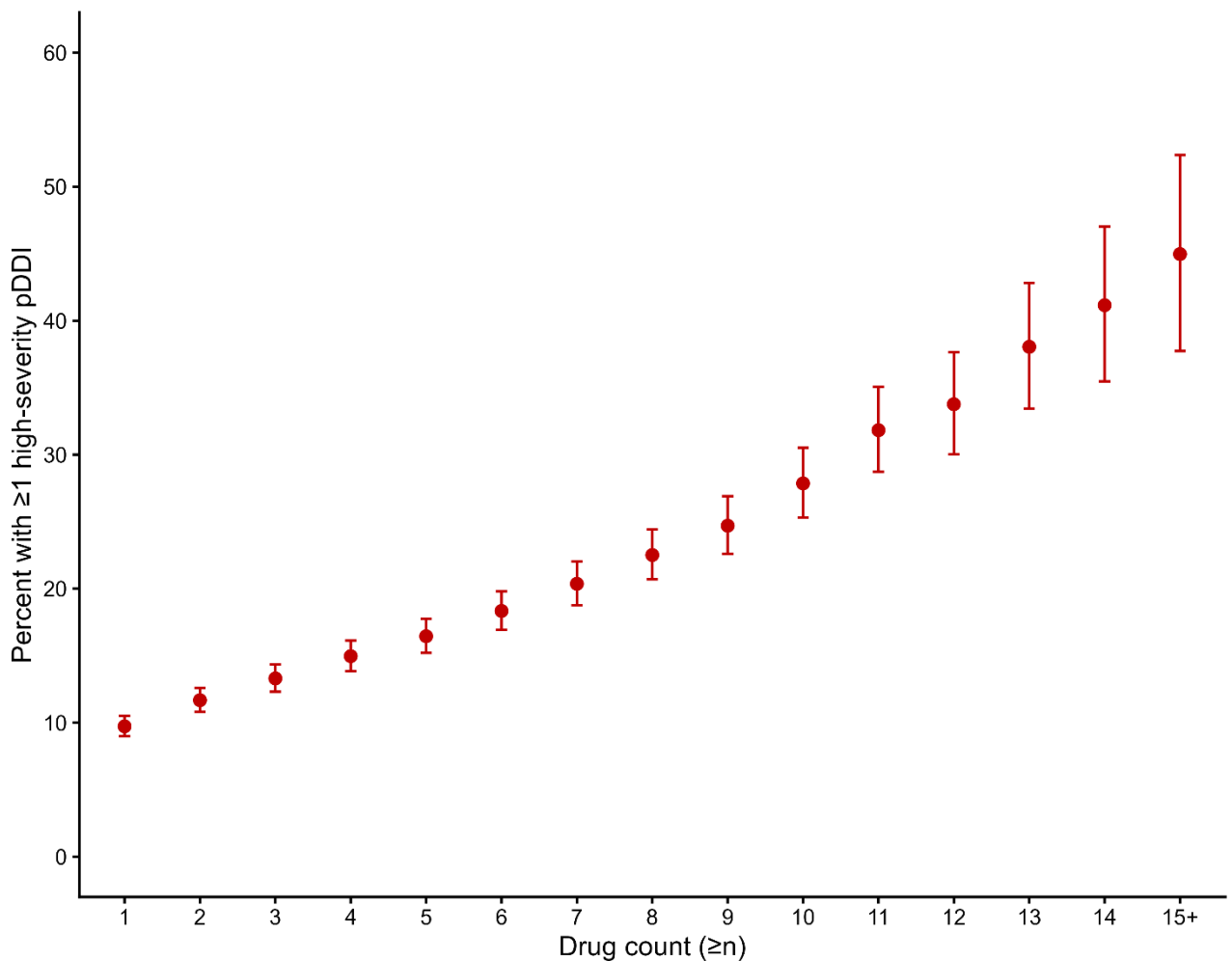


Figure 4 - Prevalence (with 95% CI) of high-severity pDDIs by cumulative drug thresholds.

The prevalence of patients with at least one high-severity pDDI varied across departments: 230/2,066 (11.1%; 95% CI 9.8-12.6) in Cardiothoracic, 225/2073 (10.9%; 95% CI 9.6-12.3)

in Internal Medicine, and 126/1812 (6.9%; 95% CI 5.8-8.2) in Neurosciences. The most frequent high-severity pDDIs identified are reported in Table 3.

Table 3. Top five most frequent high-severity pDDIs (overall and stratified by department)

Rank	Overall	Cardiothoracic	Internal Medicine	Neurosciences
1	Potassium chloride + Ramipril (40)	Potassium chloride + Valsartan and Sacubitril (28)	Amiodarone + Trimethoprim-sulfamethoxazole (15)	Acetylsalicylic acid + Acetylsalicylic acid - Clopidogrel (16)
2	Potassium chloride + Valsartan and Sacubitril (29)	Potassium chloride + Ramipril (25)	Potassium chloride + Ramipril (13)	Levodopa-benserazide + Quetiapine (11)
3	Amiodarone + Trimethoprim-sulfamethoxazole (19)	Ciclosporin + Ezetimibe (11)	Repaglinide + Trimethoprim-sulfamethoxazole (12)	Perindopril + Macrogol combination containing potassium chloride (7)
4	Macrogol combination containing potassium chloride + Ramipril (19)	Ciclosporin + Potassium chloride (9)	Acetylsalicylic acid + Anagrelide (11)	Duloxetine + Tramadol (6)
5	Amiodarone + Escitalopram (18)	Amiodarone + Escitalopram (9)	Methotrexate + Trimethoprim-sulfamethoxazole (11)	Irbesartan + Macrogol combination containing potassium chloride (6)

The drugs most frequently associated with high-severity pDDIs, in terms of absolute number of events, were potassium chloride (135 patients with ≥ 1 high-severity pDDI), trimethoprim-sulfamethoxazole (81), ramipril (72), macrogol combination containing potassium chloride (59), amiodarone (48), valsartan-sacubitril (43).

After accounting for exposure prevalence, and restricting the analysis to drugs present in at least 5 patients, the highest rates of high-severity pDDIs were observed for nicorandil (9/10 exposed patients, 90.0%), phenobarbital (5/7, 71.4%), ketoconazole (4/6, 66.7%), anagrelide (12/21, 57.1%), cobimetinib (4/7, 57.1%), and lithium (4/7, 57.1%). When further restricting to drugs with ≥ 20 exposed patients, the following emerged: anagrelide (12/21, 57.1%), quetiapine (14/27, 51.9%), metformin/sitagliptin (15/30, 50.0%), ciclosporin (28/57, 49.1%), racecadotril (18/48, 37.5%), and ketoprofene (9/25, 36.0%).

Reliability of the gold standard

The agreement between the two experts, assessed using Gwet's AC1, was 0.94 (95% CI 0.94-0.96). Positive specific agreement was 0.67 (95% CI 0.63-0.70), while negative specific agreement was 0.97 (95% CI 0.97-0.98), reflecting excellent concordance for non-high-severity cases and good concordance for the less prevalent high-severity cases.

When prescriptions with just one drug were excluded, the agreement remained very high (AC1 = 0.93; 95% CI: 0.92-0.94). Positive specific agreement was unchanged at 0.67 (95% CI, 0.63-0.70), whereas negative specific agreement decreased slightly to 0.96 (95% CI, 0.96-0.97).

Diagnostic performance of the tested automated tool (PRoSIT)

Among the 586 patients who presented at least one high-severity pDDI, the automated tool correctly identified 370 (63.1%, TP) and missed 216 (36.9%, FN). Among the 5,441 patients without a high-severity pDDI, the tool correctly classified 5,165 (94.9%, TN) and misclassified 276 (5.1%, FP). The automated tool thus showed a Se of 63.1% and a NPV reached 96.0%. Also, Sp was of 94.9% and PPV was 57.3%. Additional information on the tool's performance is reported in Table 4.

Table 4 - Performance of the NLP-based automated tool compared with the gold standard.

Metric	Estimate	95% CI
Sensitivity	63.1%	59.2-66.9
NPV	96.0%	95.4-96.5
Specificity	94.9%	94.3-95.5
PPV	57.3%	53.4-61.0
Accuracy	91.8%	91.1-92.5
F1-score	0.60	0.57-0.63
Youden's J	0.58	0.54-0.62

The post-hoc analysis showed that among the 370 TP patients, concordance between the two experts was 70.8% (262 patients). Conversely, among the 216 FN, concordance was 38.9% (84 patients), with a high proportion of discordance (61.1%, 132 patients). Among the 276 FP, concordance was 92.0% (254 patients), while among the 5,165 TN, concordance reached 99.4% (5,133 patients).

Overall, expert discordance was observed in 294 patients (4.9% of the study population). After excluding these discordant cases and restricting the analysis to patients for whom both evaluators agreed (n=5,733), the automated tool was able to correctly identify 262 (75.7%, TP) and missed 84 (24.3%, FN) of the 246 patients with at least one high-severity pDDI; Among the 5,387 patients without a high-severity pDDI, the tool correctly classified 5,133 (95.3%, TN) and misclassified 254 (4.7%, FP). Sensitivity increased to 75.7% (95% CI 70.9-80.1), and the negative predictive value (NPV) was 98.4% (95% CI 98.0-98.7).

Additional details on the performance of the tool against the gold standard, restricted to expert-concordant patients, are provided in Table 5.

Table 5- Performance of the NLP-based automated tool compared with the gold standard restricted to patients for whom both initial evaluators were in agreement.

Metric	Estimate	95% CI
Sensitivity	75.7%	70.9-80.1
NPV	98.4%	98.0-98.7
Specificity	95.3%	94.7-95.8
PPV	50.8%	46.4-55.2
Accuracy	94.1%	93.5-94.7
F1-score	0.61	0.57-0.65
Youden's J	0.71	0.66-0.76

Discussion

This study developed a gold standard based on the ANSM Thesaurus for validating automated tools aimed at identifying high-severity pDDIs.

The gold standard was built following established practices, with standardized guidance, double independent blinded annotation, and adjudication by a third evaluator to resolve discrepancies. This ensured both transparency and reproducibility of the process. This framework is consistent with evaluation standards used in previous studies conducted with different reference sources (155,172,179,180).

The definition of the existence and the severity of pDDIs was based on the ANSM Thesaurus (67). Previous reference sets have been constructed using information from this source

(67,181). However, to the best of our knowledge, this is the first expert-annotated gold standard derived directly and solely from the ANSM Thesaurus for tool validation. The focus on high-severity pDDIs reflects recommendations for including pDDIs in CDSS, which endorse the development of methods to safely reduce less relevant alerts in order to mitigate the risk of alert fatigue for clinicians (107).

For the evaluation of inter-expert consistency, Gwet's AC1 coefficient was adopted alongside positive and negative specific agreement, as suggested by methodological literature for applications with rare positives (182). Gwet's AC1 was preferred over Cohen's kappa because it is more robust in the context of unbalanced category distributions, as in this dataset, where most cases were classified as not high-severity by both experts. Under such conditions, the traditional kappa tends to underestimate the true level of agreement, whereas Gwet's AC1 provides a more stable and interpretable estimate, closer to the actual level of observed concordance.

Analyses conducted on the evaluations of the two independent experts demonstrated a very high level of reliability, even reducing the dataset to patients with at least two prescribed drugs. The population used for the construction of the gold standard was selected due to its high risk of presenting pDDIs (65-67), therefore representing a valid population for testing automated tools. The high prevalence of polypharmacy, exceeding 55% in the overall cohort and peaking in the cardiothoracic department, reflects the clinical complexity and directly increases their risk of pDDIs (38,183,185). The distribution of prescribed drugs confirms that each department has its own "fingerprint" in terms of prescribing profiles, consistent with the types of patients managed, and with the existing evidence (186).

A key finding was the prevalence of high-severity pDDIs, equal to 9.7%, and higher than the prevalence reported in previous studies (3%) using the same primary source (177). This discrepancy likely reflects greater fragility of the study population. Unlike prior estimates

derived from general practice databases, the current population, although in the same age range, included individuals consulted or hospitalized in hospital departments, and therefore at higher risk.

The gold standard was applied to an experimental NLP-based automated tool. This system showed suboptimal sensitivity performance, although it achieved good specificity and excellent negative predictive value. In other words, the system reliably excluded not high-severity cases but failed to detect almost one-third of high-severity pDDIs. The high overall accuracy therefore primarily reflects high specificity in a context of relatively low prevalence, rather than a genuine discriminative ability in clinically relevant cases. Operationally, the NLP tool cannot yet be considered a reliable tool for direct clinical use.

In our view, algorithms targeting only high-severity pDDIs, which represent a minority of patients but carry the greatest clinical risk, should primarily function as rule-out (187), high-sensitivity tools: systems capable of automatically detecting a manageable set of suspected cases (i.e., maintaining specificity reasonably high), leaving final clinical assessment to the clinician. Even with an increase in FP, such an approach would allow the clinician to re-examine only a fraction of the overall population, thus reducing the review workload compared to the manual re-evaluation of all patient records by a hospital pharmacist or clinical pharmacologist.

The post-hoc analysis revealed that most FN occurred in patients with pDDIs where the severity classification had been discordant between the experts. This feature, which likely contributed to the difficulties of the tool in aligning with the gold standard evaluations, suggests potential ambiguities of the ANSM Thesaurus in providing easily interpretable information. This represents however an area of interest for the development of tools capable of transparently and unequivocally managing such ambiguity.

From a methodological standpoint, the use of the best-case scenario presents both advantages and limitations. On the one hand, it allows restriction of the analysis to a subset of pDDIs unambiguously high-severity in all clinical contexts, making their non-identification clinically unacceptable. On the other hand, it risks underestimating the real number of potentially relevant high-severity pDDIs and inflating the share of FP (i.e., when the NLP tool flags pDDIs as high-severity only under specific conditions). A useful strategy could consist of maintaining the best-case scenario as the primary validation criterion, but limiting ambiguous cases and implementing tools capable of accounting for common clinical circumstances (e.g., high-dose vs low-dose ASA). In this sense, a best-case approach would be particularly useful when tools are integrated with scenario-aware rules that update severity according to dose, route of administration, or organ function, when such data are available (188).

Conversely, the use of the worst-case scenario does not appear a valid alternative. This choice would lead to an artificial inflation in the number of pDDIs classified as high-severity, including circumstances that in real clinical practice would rarely justify a high-severity alert. For example, the ANSM Thesaurus considers the co-prescription of acetylsalicylic acid and oral anticoagulants as high-severity only at analgesic doses; such a scenario is less likely to occur in hospital discharge prescriptions, especially in the cardiothoracic department, where acetylsalicylic acid is likely used mainly as an antiplatelet agent. Including circumstances like this within high-severity pDDIs would therefore result in an overload of clinically marginal alerts, increasing the risk of alert fatigue without improving actual patient safety, undermining one of the key goals of creating automated tools (107).

The gold standard nevertheless presents some limitations. Its construction was limited to severity classes, without including information on reported mechanisms or specific management approaches. This choice was dictated, on the one hand, by the inconsistent

reporting of such information in the ANSM Thesaurus, and on the other by the risk of expert fatigue from overload during gold standard compilation (174). Although Gwet's AC1 agreement was very good, a non-negligible proportion of discordance was observed regarding the positive specific agreement, with the two evaluators disagreeing in approximately one quarter of the cases in which at least one evaluator classified a pDDI as high-severity. The presence of a third expert as tie-breaker provided additional support to the robustness of the gold standard. Another theoretical limitation is represented by the remote possibility that some drug pairs may have been concordantly but incorrectly classified by both raters, and therefore not intercepted by the tiebreaker. The high observed agreement between evaluators suggests that such an occurrence, although possible, will have a negligible impact on the present results. Finally, the inclusion of patients ≥ 65 years and from only three hospital departments, albeit chosen for their risk profile, may limit the generalizability of the results to other clinical contexts, such as oncology, which is characterized by different prescribing patterns. However, the inclusion of all drugs taken by patients and the presence of a high number of unique drugs (764 drugs distributed across more than 30,000 evaluated pairs) substantially mitigates this limitation.

Future analyses should focus on a systematic analysis of FN to identify the main causes of algorithmic failure and guide the development of more sensitive detection strategies. The clinical acceptability of such improvements should be assessed through their impact on error reduction and workflow integration. In parallel, semi-automated extraction of structured information from the ANSM Thesaurus could support the creation of a fully annotated, surrogate gold standard, extending beyond severity to encompass the full range of ANSM Thesaurus data for large-scale validation.

Conclusions

This study developed a gold standard based on the ANSM Thesaurus for the validation of automated tools to identify high-severity pDDIs. The defined gold standard may represent a valuable resource for testing automated tools in French clinical contexts. Artificial intelligence tools based on the ANSM Thesaurus have the potential to be used in routine clinical practice, but need to be better tailored for complex pDDIs.

4.3 Lessons learned and methodological considerations

The first consideration concerns the nature of the reference document. The ANSM Thesaurus is a regulatory text designed to guide clinical decision-making. It was not created as a computable ontology or a rule-based system. Transforming it into a gold standard therefore required formalization: narrative statements had to be translated into operational rules, ambiguities clarified, contextual conditions specified, and aggregation criteria defined.

Within this context, adopting a best-case scenario logic was methodologically justified in order to reduce the risk of alert fatigue. Restricting the standard to clinically plausible high-severity configurations avoided inflating alert burden with rare or extreme situations. The trade-off is straightforward: one may miss a rare pDDI occurring under worst-case conditions or systematically generate implausible severe pDDIs. The former represented the more proportionate compromise and was supported by the study findings. Post-hoc analyses (data not shown) showed that, among the ten most frequent drug pairs associated with multiple severity levels, eight involved high-dose acetylsalicylic acid. In the study population, approximately 93% of prescriptions corresponded to the antiplatelet formulation, supporting the reliability of the best-case operationalization.

Moreover, this study highlighted the need to evaluate gold standard reliability. Double independent review with adjudication allowed estimation of inter-expert agreement, but it does not guarantee ontological validity in pDDIs evaluation. Although a high Gwet's AC1 indicates strong internal consistency of the annotations, it does not exclude the possibility of shared systematic error. A gold standard can therefore be reproducible while remaining imperfect. Although the dual-review consensus process meets accepted methodological standards, residual uncertainty persists when interpretation is complex.

The patient-level definition (≥ 1 pDDI) also introduces a structural transformation. The unit of analysis shifts from drug pair to patient. In this framework, a single pair-level

misclassification can reclassify the entire patient. In low-prevalence settings, this effect is non-linear and can influence sensitivity and specificity disproportionately.

As a consequence, the apparent underperformance of the NLP tool (Se at 63%) should be considered together with the limitations of the gold standard construction process. When the gold standard derives from a complex regulatory text, the distinction between model error and formalization error becomes blurred. Therefore, this study produced an operationally robust standard while raising a broader methodological question: how can computational limitations be separated from epistemic limitations embedded in the gold standard itself? This question is addressed in the third study of the thesis.

5 - Discordance analysis to refine an imperfect gold standard for NLP-based identification of high-severity potential drug-drug interactions

5.1 Context and objectives

The study presented in the previous chapter established a structured gold standard and quantified tool performance, revealing discrepancies between automated detection and expert classification. However, even well-constructed gold standards are rarely completely free of error (189). When the gold standard itself is derived from regulatory non-algorithmic texts and subject to ambiguity, misclassification within the gold standard may contribute substantially to observed performance estimates (190). In contexts where the gold standard is not fully independent or inherently error-free, performance metrics may reflect gold standard bias rather than intrinsic model performance (190).

The third study of the thesis aimed to address the following question: to what extent are observed FN and FP attributable to algorithmic limitations *vs.* imperfections in the gold standard?

To investigate this issue, the third study was conducted to implement a structured, causal audit of discordances between P_{Ro}SIT and the initial gold standard (GS-v1). Through iterative refinement (GS-v2, GS-v3, GS-v4), human-related misclassifications were corrected, allowing quantification of their impact on performance estimates.

Therefore, the primary objective of the third study was to estimate the proportion of patient-level discordances attributable to human-related errors in gold standard construction. The secondary objectives were: (i) to characterize recurrent patterns of human and algorithmic error; and (ii) to quantify how iterative refinement of the gold standard impacts P_{Ro}SIT performance.

Note. Tables and figures follow the internal numbering system used within the study and have not been renumbered across chapters. Abbreviations are defined at their first occurrence in each article, regardless of previous definitions in the thesis.

5.2 Article

-Discordance Analysis to Refine an Imperfect Gold Standard for NLP-Based Identification of High-Severity Potential Drug-Drug Interactions

Lelio Crupi

Abstract

Introduction

Automated systems for detecting potential drug-drug interactions (pDDIs) depend on expert-derived gold standards for validation, but imperfections in these reference standards can distort performance estimates. In a recently developed gold standard for high-severity pDDIs from the ANSM Thesaurus, we found that part of the underperformance of an AI tool (PRoSIT) might stem from imperfections in the human-annotated reference standard. This study aimed to quantify how many discrepancies were due to human annotation errors and to evaluate how correcting these errors impacts PRoSIT performance.

Methods

We conducted a methodological study to refine a double-blind, expert-derived gold standard based on the ANSM Thesaurus. PRoSIT performance was initially evaluated in a cohort of 6,027 French hospitalized patients. All discordances were reviewed to determine whether they reflected human-related or algorithmic errors. The gold standard was iteratively refined by correcting only human-related errors. The primary outcome was the proportion of human-related errors among false negatives (i.e., cases where PRoSIT failed to identify an expert-annotated pDDI), while the secondary outcome was the corresponding proportion among false positives (i.e., cases where PRoSIT identified a pDDI not confirmed by experts). Sensitivity and negative predictive value (NPV) were recalculated after refinement.

Results

Among 216 false-negative patients, 175 (81.0%) were attributable to human-related errors, while among 315 false-positive patients, 17 (5.4%) were human-related. Refinement substantially improved performance: sensitivity increased from 63.1% to 89.4%, while NPV increased from 96.0% to 99.2%. A small number of errors accounted for the vast majority of imperfections, mostly related to interactions resulting in hyperkalemia and bradykinin-mediated angioedema.

Conclusion

A large proportion of PProSIT underperformance was attributable to imperfections in the gold standard. Structured, auditable refinement significantly modified diagnostic performance estimates, providing empirical evidence of reference standard bias. Validation of automated pDDI detection systems should account for reference standard imperfection and incorporate systematic discordance audits.

Introduction

Expert-derived gold standards constitute an essential reference for the evaluation of automated tools for potential drug-drug interactions (pDDIs) identification (172). However, when the gold standard is not fully valid or not directly verifiable, it effectively functions as an imperfect reference standard, with the potential to introduce bias into estimates of diagnostic accuracy (190). Under such conditions, performance metrics of the index test may be systematically distorted, making it difficult to discriminate algorithmic limitations from misclassification within the gold standard itself (189).

This issue is particularly relevant in contexts characterized by high semantic complexity of regulatory sources or by low-prevalence events. In the case of high-severity pDDIs, the operational definition of a pDDI directly depends on interpretative choices regarding molecular granularity, pharmacological class, clinical context, and applicability conditions. In such a setting, discordance between an NLP tool and expert annotation cannot automatically be interpreted as algorithmic error; rather, it may reflect intrinsic limitations of the initial gold standard.

From the tool perspective, potential sources of discordance include difficulties in interpreting linguistic context, such as negations, hypothetical statements, or temporal conditions, both within the regulatory reference document and within the clinical record, as well as limitations in mapping between individual molecules and broader pharmacological classes (191,192). From the human perspective, discordance may arise from under-specification of decision rules, ambiguities within the regulatory source, or inter-expert variability in interpretation. Within this framework, systematic analysis of discordances aimed at understanding how and why disagreements arise represents a well-established methodological strategy for iterative refinement of the gold standard (193). Previous studies have shown that error analyses

against a gold standard can help identify not only algorithmic error sources but also errors in the human-curated gold standard itself (192).

An expert-derived gold standard for the identification of high-severity pDDIs, based on the ANSM Thesaurus, had previously been developed and used to validate an NLP tool (PRoSIT). Its initial application yielded modest performance (194).

The primary objective of the present study was to determine whether the observed discordances were mainly attributable to algorithmic limitations or to imperfections in the gold standard. In parallel, the study aimed to characterize error patterns and to quantify the impact of gold standard refinement on the performance metrics of the PRoSIT.

Methods

Study design and setting

This retrospective methodological study was based on a structured audit of discordances between PRoSIT and an initial expert-derived gold standard (GS-v1) for the identification of high-severity pDDIs. Starting from GS-v1, the gold standard was iteratively updated (GS-v2, GS-v3, GS-v4) by correcting exclusively errors attributed to the human annotation process. No modifications were made to the NLP tool throughout the refinement process.

The study used the same cohort (6,027 hospitalized patients), the same version of the ANSM Thesaurus, and the same automated NLP tool used in a previous study (194). Details regarding cohort characteristics, inclusion criteria, and the operational definition of high-severity pDDIs are reported in extenso in that publication (194).

Briefly, each unique drug pair was evaluated independently and in parallel by two human reviewers, each blinded to the other's assessment. For each unique drug pair, predefined information elements were independently extracted by two evaluators and recorded in a structured drug-pair evaluation module. In all cases of discordance, a third evaluator

independently reassessed the drug pair while blinded to the decisions of the first two reviewers and acted as a consensus adjudicator.

Each patient was considered positive if at least one drug pair was classified as high-severity.

Identification of discordances and gold standard refinement strategy

Phase 1: False negatives (GS-v1 -> GS-v2)

False negatives (FN) were defined as patients classified as positive in GS-v1 and negative in PRoSIT. All FN patients were identified. For each FN patient, all drug pairs labelled as high-severity in GS-v1 were extracted and subjected to structured audit.

The audit included:

1. Reassessment of the appropriateness of high-severity classification according to the ANSM Thesaurus and the operational rules adopted in the gold standard.
2. Review of the prior adjudication process and re-examination of the original human annotation.
3. Cross-check of PRoSIT output to determine whether a comparable pair had been detected in: i) the same patient, with different severity; ii) or in other patients in the dataset, as severe.
4. Final causal attribution of each discordance (human-related vs algorithm-related).

If the discordance was attributed to human-related error, the gold standard was corrected. Algorithm-related discordances did not lead to modification of the algorithm. Corrections generated GS-v2, and performance metrics were recalculated.

Phase 2: False positives (GS-v2 -> GS-v3)

False positives (FP) were defined as patients classified as positive in PRoSIT and negative in GS-v2. All FP patients were identified. For each FP patient, all pairs flagged by PRoSIT as high-severity were reviewed.

For pairs identified by PRoSIT but absent from the gold standard, the audit included:

1. Evaluation of the pDDI according to the ANSM Thesaurus and the framework adopted in the gold standard.
2. Assessment of entity extraction and normalization (entity recognition, disambiguation, and substance-to-class mapping).
3. Determination of whether the pair was present in the gold standard with a different severity level or represented a false detection; manual drug extraction for gold standard construction was assumed to have negligible error.
4. Final causal attribution.

Human-related errors led to correction of the gold standard and generation of GS-v3. Algorithm-related discordances did not modify the gold standard. Diagnostic performance was recalculated.

Phase 3: Structural analysis of true positives (GS-v3 -> GS-v4)

True positives (TP) were defined as patients classified as positive in both PRoSIT and GS-v3. For TP patients, the set of high-severity pairs identified by PRoSIT was compared with the set recorded in GS-v3. Five mutually exclusive agreement categories were defined based on pair-level correspondence (detailed in Section 2.3.3).

All cases without exact correspondence underwent targeted audit of discordant pairs using the same causal attribution procedure described above. Human-related corrections generated the final version of the gold standard (GS-v4), after which performance metrics were recalculated.

The overall workflow of the study is presented in **Figure 1**.

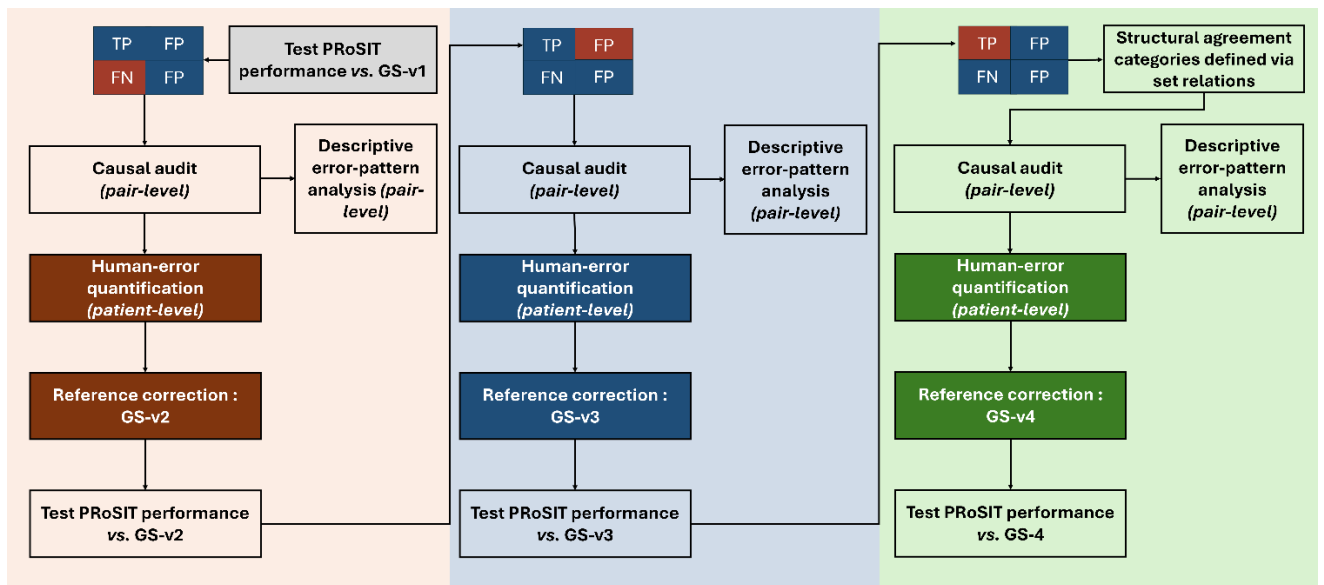


Figure 1. Study workflow. The refinement process followed a sequential and unidirectional logic applied across three iterative cycles. Each cycle included: (i) identification of patient-level discordances between PProSIT and the current version of the gold standard; (ii) structured pair-level audit with causal attribution; (iii) estimation of the proportion of discordances attributable to human-related errors; (iv) correction of the gold standard exclusively for human-related errors; and (v) recalculation of diagnostic performance against PProSIT, generating a new version of the gold standard. The NLP tool remained unchanged throughout the process.

All classifications were performed at the patient level (≥ 1 high-severity pDDI), while causal attribution was conducted at the pair level.

Statistical analysis

Primary analysis: human-related error among FN

The primary endpoint was the proportion of FN patients, identified with respect to GS-v1, whose discordance was attributable to human-related error in the gold standard.

For this analysis:

- The numerator was the number of FN patients in whom at least one discordant drug pair led to correction of the gold standard due to a human-related cause.
- The denominator was the total number of FN patients relative to GS-v1.

Secondary analysis: human-related error among false positives

A secondary endpoint was the proportion of FP patients, identified with respect to GS-v2, whose discordance was attributable to human-related error.

For this phase:

- The numerator was the number of FP patients in whom at least one discordant pair led to correction of the gold standard due to a human-related cause.
- The denominator was the total number of FP patients relative to GS-v2.

Structural concordance analysis among true positives

Among patients classified as TP with respect to GS-v3, a structural concordance analysis was conducted at the pair level to assess agreement between the set of high-severity drug pairs identified by P_{Ro}SIT and those recorded in the gold standard.

Patients were categorized into five mutually exclusive agreement categories (Figure 2):

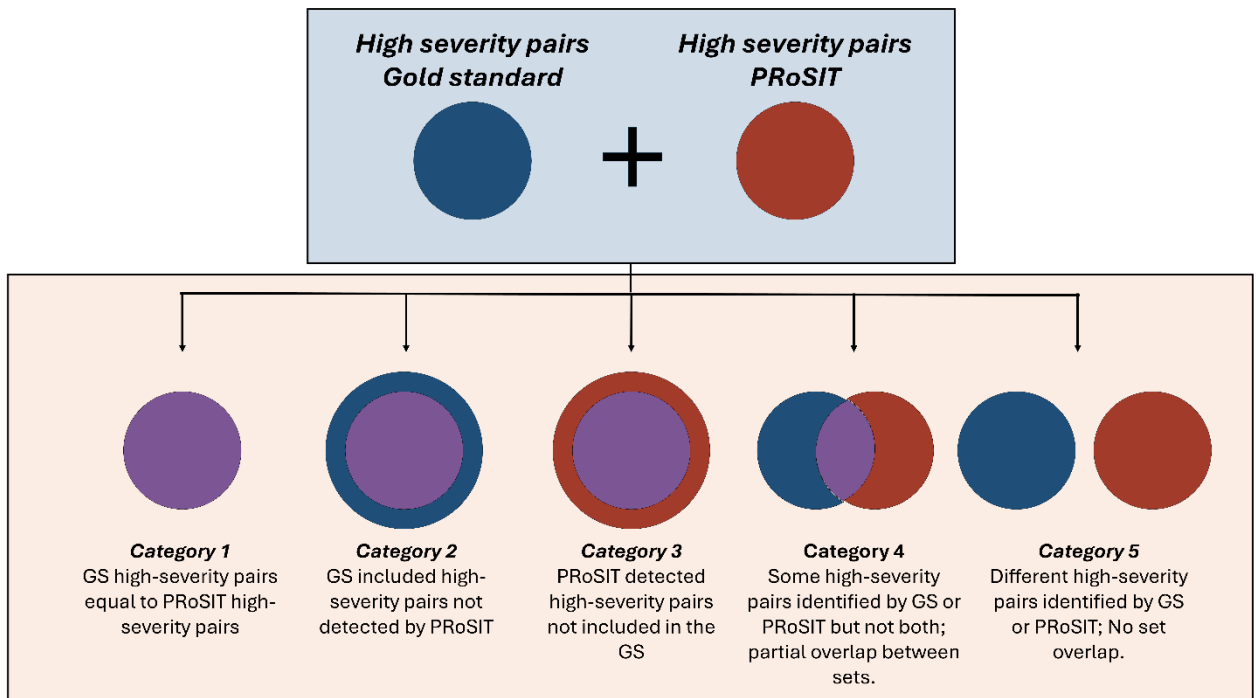


Figure 2. Definition of the structural agreement categories among TP patients. *Blue indicates pairs identified by the GS-v3, red indicates pairs identified by P_{Ro}SIT, and purple indicates overlapping pairs.*

Patients without complete pair-level agreement (i.e., categories 2-5) underwent targeted audit to determine the nature of the discordances and to guide final refinement (GS-v4). Findings from this component were interpreted as residual structural disagreement at the pair level rather than as patient-level misclassification.

Descriptive analysis

A qualitative descriptive analysis was conducted at the pair level across all audit phases to characterize dominant patterns of human-related and algorithm-related errors, including recurrent substances and systematic rule misinterpretations.

Impact of refinement on diagnostic performance

For each version of the gold standard (GS-v1 through GS-v4), the patient-level contingency table was reconstructed.

Sensitivity and negative predictive value (NPV) were defined a priori as primary performance metrics, consistent with the intended rule-out application of PRoSIT. Specificity, positive predictive value (PPV), accuracy, F1-score, and Youden's J were considered secondary metrics. Absolute changes (Δ) between successive gold standard versions were reported.

All metrics were reported with 95% confidence intervals (95% CI).

Confidence intervals for binomial proportions (sensitivity, specificity, PPV, NPV, and accuracy) were calculated using the Wilson score method. Confidence intervals for F1-score and Youden's J were estimated using non-parametric bootstrap.

Results

Patient-level causal audit of discordances

FN with respect to GS-v1

A total of 216 patients were classified as FN with respect to GS-v1. Following structured audit, 175/216 (81.0%) were attributed to human-related error, whereas 41/216 (19.0%) were attributed to algorithm-related error.

The transition from GS-v1 to GS-v2 resulted in reclassification of 175 patients initially labelled as FN (Table 1). As a consequence of the gold standard update 39 patients previously classified as TP were reclassified as FP.

Table 1. Contingency tables comparing PRoSIT classifications with successive versions of the gold standard (GS-v1 to GS-v4).

GS-v1			GS-v2		
	GS+ (n=586)	GS- (n=5,441)		GS+ (n=372)	GS- (n=5,655)
PRoSIT+ (n=646)	TP 370 (63.1%)	FP 276 (5.1%)	PRoSIT+ (n=646)	TP 331 (88.9%)	FP 315 (5.6%)
PRoSIT- (n=5,381)	FN 216 (36.9%)	TN 5,165 (94.9%)	PRoSIT- (n=5381)	FN 41 (11.1%)	TN 5,340 (94.4%)
GS-v3			GS-v4		
	GS+ (n=389)	GS- (n=5,638)		GS+ (n=386)	GS- (n=5,641)
PRoSIT+ (n=646)	TP 348 (89.5%)	FP 298 (5.3%)	PRoSIT+ (n=646)	TP 345 (89.4%)	FP 301 (5.3%)
PRoSIT- (n=5,381)	FN 41 (10.5%)	TN 5,340 (94.7%)	PRoSIT- (n=5,381)	FN 41 (10.6%)	TN 5,340 (94.7%)

Percentages are calculated relative to the column denominator (GS+ or GS-). Row totals remain constant across versions because the tool was unchanged and modifications were applied exclusively to the gold standard.

False positives with respect to GS-v2

With respect to GS-v2, 315 patients were classified as FP. After audit, 17/315 (5.4%) were attributed to human-related error, whereas 298/315 (94.6%) were attributed to algorithm-related error.

The transition from GS-v2 to GS-v3 resulted in reclassification of 17 FP patients (Table 1). The update also modified the underlying pair composition in 12 patients classified as TP, without changing their patient-level status.

Structural concordance analysis among true positives (GS-v3)

A total of 348 patients were classified as TP in GS-v3. A structural concordance analysis comparing the list of drug pairs identified by P_{Ro}SIT with those recorded in the gold standard was performed. The distribution across the five pre-specified agreement categories is shown in Figure 3.

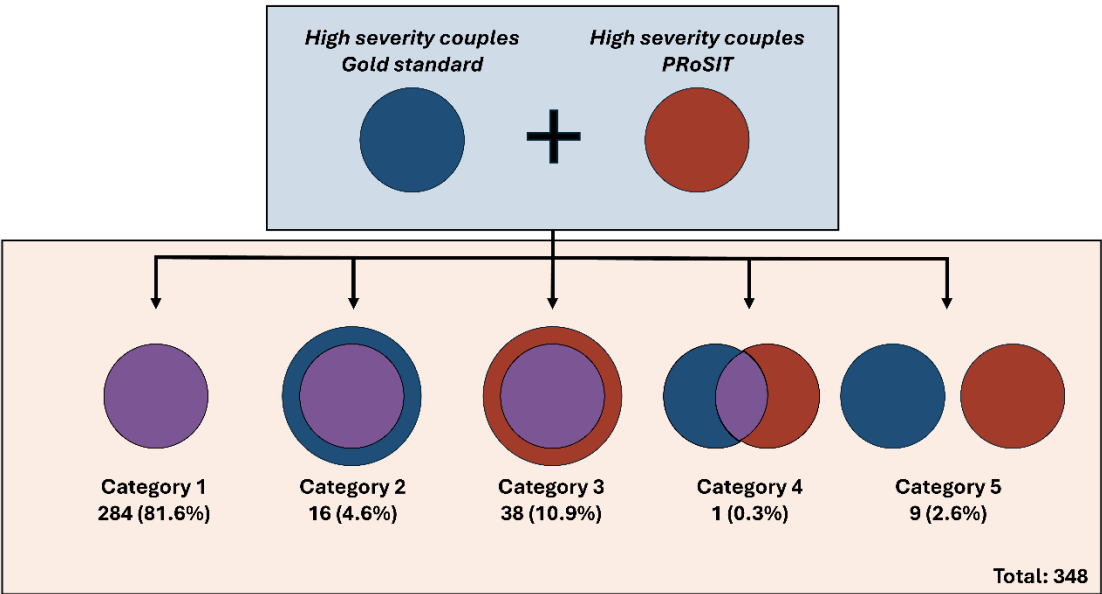


Figure 3. *Distribution of true-positive patients across the five pre-specified structural agreement categories. For each patient, concordance was defined based on the relationship between the set of high-severity pairs identified by the gold standard and by P_{Ro}SIT. Blue indicates pairs identified by the GS-v3, red indicates pairs identified by P_{Ro}SIT, and purple indicates overlapping pairs.*

Patients without complete pair-level agreement (categories 2-5; n = 64) underwent targeted audit.

At the patient level:

- Category 2 (n = 16): five patients had at least one human-related error; remaining discrepancies were algorithm-related.
- Category 3 (n = 38): ten patients had at least one human-related error.
- Category 4 (n = 1): the case involved both human-related and algorithm-related errors.
- Category 5 (n = 9): in three patients, neither pDDI was confirmed as valid; in the remaining cases, the pDDI recorded in the GS-v3 was confirmed, whereas the algorithm-detected pDDI was unsupported.

Final refinement (GS-v3 -> GS-v4) resulted in modification of information in 19 patients and final reclassification of three previously TP patients as FP (Table 1).

Pair-level qualitative audit of discordances

Qualitative analysis was conducted at the pair level, including all discordant pairs identified across audit phases.

False negatives with respect to GS-v1

Among the 216 FN patients, 27 (12.5%) had more than one discordant pair (range 2-6), corresponding to 248 discordant drug pairs. Of the latter, 202 (81.5%) were attributed to human-related error, and 46 (18.5%) to algorithm-related error.

- Among the 202 human-related pairs, 8 (4.0%) were due to joint misclassification by the initial reviewers, whereas 194 (96.0%) emerged during the consensus phase.
- Among the 46 algorithm-related pairs, one pDDI had been identified but classified as non-severe. In 15 cases, the same pharmacological combination had been correctly identified and classified in other patients within the dataset. The remaining pairs were attributable to rule misinterpretation or normalization issues.

Human-related patterns. The majority of human-related pairs (165/202; 81.7%) were linked to ANSM Thesaurus lexical ambiguity in potassium-containing combinations. In these cases, ANSM Thesaurus specified a single severity level accompanied by a note excluding applicability in the presence of hypokalemia, leading to inconsistent operational interpretation.

Fifteen pairs involved acetylsalicylic acid combined with clopidogrel or diclofenac, with discrepancies related to distinctions between validated indications, off-label use, and dose-dependent conditions.

Thirteen pairs concerned a broadly defined pharmacological class (“medicinal products administered orally”), associated with a general recommendation to reduce absorption and separate administration (>2 hours). The remaining nine pairs showed no recurring pattern.

Algorithm-related patterns. Among the 46 algorithm-related pairs, 14 (30.4%) involved substances associated with bradykinin-mediated angioedema risk. Thirteen pairs (28.3%) included levodopa. Four pairs (8.7%) involved apixaban combined with non-steroidal anti-inflammatory drugs or azole antifungals. The remaining pairs were heterogeneous.

False positives with respect to GS-v2

Among the 315 FP patients, 17 (5.4%) had more than one discordant pair (range 2-3), corresponding to 333 discordant pairs.

Of these, 17/333 (5.1%) were attributed to human-related error and 316/333 (94.9%) to algorithm-related error.

Human-related patterns. All 17 human-related pairs had undergone consensus review. Ten (58.8%) were due to incorrect severity attribution for drugs associated with bradykinin-mediated angioedema. The remaining pairs did not form a recurring cluster.

Algorithm-related patterns. Among the 316 algorithm-related pairs:

- 197 (62.3%) were attributable to incorrect interpretation of pDDIs between potassium and either angiotensin II receptor blockers, ACE-inhibitors, or diuretics, with systematic classification as positive irrespective of contextual conditions.
- 64 (20.3%) resulted from incorrect substances identification in the clinical record, including 44 cases where one member of the pair was absent and 20 where both were absent.
- 38 (12.0%) resulted from incorrect interpretation of pDDIs involving drugs associated with bradykinin-mediated angioedema, where co-presence of two listed substances was treated as an automatic pDDI.
- 15 (4.7%) were due to incorrect substance-to-category mapping (ex. cholecalciferol) within the ANSM Thesaurus.

Discordant pairs among non-exact match true positives (GS-v3)

Across the 64 patients in categories 2-5, a total of 80 discordant pairs were analyzed.

In category 2 (n = 16), 19 pairs were present in the gold standard but not detected by PROSIT. Five were human-related, all established through consensus and related to ambiguities involving antihypertensive-potassium combinations. The remaining 14 were algorithm-related; in four cases, a comparable combination had been identified in other patients. Seven of the 14 undetected pairs included levodopa.

In category 3 (n = 38), 42 pairs were identified by PROSIT but absent from the gold standard. Eleven were human-related (ten from consensus and one from initial joint misclassification). The remaining 31 were algorithm-related: 18 due to incorrect severity attribution, nine due to identification of non-documented substances, and four due to substance-to-class linkage errors within the ANSM Thesaurus. Eplerenone and/or spironolactone were involved in 20 of these 31 pairs.

The single category 4 patient had two discordant pairs, one on each side, involving both human-related and algorithm-related components.

In category 5 ($n = 9$), nine pairs were present in the gold standard but not detected by PRoSIT, and nine were detected by PRoSIT but absent from the gold standard. Among gold standard pairs, three were found to be incorrect.

Impact of gold standard refinement on diagnostic performance

Performance after FN correction (GS-v2)

In GS-v2, 372 patients were classified as having at least one high-severity pDDI. Of these, 331 (88.9%) were correctly identified and 41 (11.1%) were missed. Sensitivity increased to 88.9% (95% CI: 85.4-91.8), a 25.8% increase compared to GS-v1; NPV reached 99.2% (95% CI: 99.0-99.4), a 3.2% increase compared to the previous GS version. Detailed results are shown in Table 2.

Performance after FN and FP correction (GS-v3)

In GS-v3, 389 patients were classified as positive. PRoSIT correctly identified 348 (89.5%) and missed 41 (10.5%). Sensitivity was 89.5% (95% CI: 86.0-92.1), NPV remained 99.2% (99.0-99.4), specificity slightly increased to 94.7% (94.1-95.3), and PPV was 53.9% (50.0-57.7) (Table 2).

3.3.3 Performance after final refinement (GS-v4)

Following final refinement, 386 patients were classified as positive. PRoSIT correctly identified 345 (89.4%) and missed 41 (10.6%). Sensitivity was 89.4% (95% CI: 85.9-92.1), NPV remained unchanged at 99.2% (95% CI: 99.0-99.4).

Table 2. Performance of PProSIT across gold standard versions

Metric	GS-v1	GS-v2	Δ	GS-v3	Δ	GS-v4	Δ
Sensitivity (%)	63.1 (59.2-66.9)	88.9 (85.4-91.8)	+25.8	89.5 (86.0-92.1)	+26.4	89.4 (85.9-92.1)	+26.3
NPV (%)	96.0 (95.4-96.5)	99.2 (99.0-99.4)	+3.2	99.2 (99.0-99.4)	+3.2	99.2 (99.0-99.4)	+3.2
Specificity (%)	94.9 (94.3-95.5)	94.4 (93.8-95.0)	-0.5	94.7 (94.1-95.3)	-0.2	94.7 (94.0-95.2)	-0.2
PPV (%)	57.3 (53.4-61.0)	51.2 (47.4-55.1)	-6.1	53.9 (50.0-57.7)	-3.4	53.4 (49.6-57.2)	-3.9
Accuracy (%)	91.8 (91.1-92.5)	94.1 (93.5-94.7)	+2.3	94.4 (93.8-94.9)	+2.6	94.3 (93.7-94.9)	+2.5
F1-score	0.60 (0.57-0.63)	0.65 (0.61-0.68)	+0.05	0.67 (0.64-0.70)	+0.07	0.67 (0.63-0.70)	+0.07
Youden's J	0.58 (0.54-0.62)	0.83 (0.80-0.86)	+0.25	0.84 (0.81-0.87)	+0.26	0.84 (0.81-0.87)	+0.26

Estimates with 95% CI in parentheses; Δ expressed in percentage points relative to GS-v1.

Discussion

This study applied a structured and iterative refinement process to a previously developed gold standard (GS-v1) used to validate the NLP tool PProSIT for identifying high-severity pDDIs. The main finding is that a substantial proportion of errors initially attributed to PProSIT were in fact due to errors within the gold standard itself.

The most relevant variation across GS versions resulted from the reclassification of 175 patients from FN to TN in the evaluation of PProSIT between GS-v1 and GS-v2. As PProSIT itself was not modified, the shift in performance metrics reflects a direct effect of refining and correcting the gold standard. During the same transition, 39 patients were also reclassified from TP to FP.

This latter result is less intuitive when considered only at the patient level. Two elements should be considered:

First, pDDI identification occurs at the drug-pair level, whereas performance metrics were calculated at the patient level using a ≥ 1 -event logic. Under this framework, a patient's classification may depend on different combinations of drug pairs. Without a dedicated analysis of the underlying pairs, patient-level evaluation alone may therefore represent an imprecise or spurious proxy of true algorithmic performance.

Second, whenever a pair was modified for a FN or FP patient, all other instances of that pair were modified, regardless of the patient's status.

As a consequence, the final analysis of discordances among TP does not reflect only the last adjudication step (GS-v3 $\rightarrow$ GS-v4) but also cumulative modifications introduced during earlier stages of reference standard refinement. For this reason, the proportion of human-related errors observed among TP was not a directly measured study outcome, as it is partly influenced by upstream structural updates of the reference standard.

A relevant result of the study was that error causes were not symmetrically distributed. Most FN were related to human errors, often emerging during the consensus phase, due to lexical and conditional ambiguities in the regulatory source. In contrast, the analysis of FP revealed a predominantly structural, tool-related error. PRoSIT tended to interpret certain regulatory entries as unconditional positives, whereas the ANSM Thesaurus defined them as context-dependent, for example, combinations allowed under specific monitoring conditions.

The audit of discordances allowed a distinction between random and structural errors. The predominance of human-related errors during the consensus stage, when the two initial reviewers disagreed, indicates an intrinsic interpretative difficulty in applying certain ANSM Thesaurus entries rather than isolated individual mistakes.

In addition, the distribution of discordant pairs followed a clear Pareto-like pattern. A small number of drug combinations accounted for most errors. Among FN, more than 80% of human-related pairs were concentrated in potassium-containing combinations affected by lexical ambiguity in the ANSM Thesaurus. Among FP, over 60% of algorithm-related pairs resulted from systematic misinterpretation of potassium-related pDDIs. Other clusters, such as bradykinin-mediated angioedema risk, levodopa, and substance identification errors, contributed smaller but recurrent proportions. This pattern suggests that both human reviewers and the algorithm are sensitive, for different reasons, to lexical ambiguities that affect highly prevalent drug pairs or groups of pairs.

These pair-level findings raise a major concern. If ambiguous entries systematically generate errors in frequent co-prescriptions, similar misinterpretations may occur in routine clinical practice. At the same time, the fact that it was very rare for both reviewers to commit the same mistake supports the overall robustness of the gold standard.

A secondary but methodologically important finding concerns the impact of gold standard refinement on system performance. Following correction of human-related cases, sensitivity increased substantially, reaching approximately 89%, with an improvement of more than 25 percentage points, while the negative predictive value approached 99%. Importantly, this improvement was not the result of any modification to the tool but of stabilization of the gold standard. The combination of high sensitivity, preserved specificity, and near-perfect NPV suggests that, in its final configuration, the system may be compatible with a rule-out logic, although external validation in independent settings remains necessary.

Concerning the secondary metric of performance evaluation, PPV showed a modest decrease after FN correction, reflecting the reduction in gold standard-defined prevalence. Conversely, the accuracy increase needs to be evaluated with caution, as in low-prevalence settings this metric is strongly influenced by the large proportion of TN. The F1-score improved only

moderately, as the gain in sensitivity was partially offset by a decrease in PPV. Consistently, Youden's J increased markedly, mainly driven by the substantial gain in sensitivity after gold standard refinement.

Furthermore, a specific clarification on the rationale of the overall framework adopted must be made. The diagnostic accuracy literature has long recognized that the so-called "gold standard" is often imperfect and that misclassification of the reference can substantially distort estimates of index test performance. Reporting frameworks such as STARD and risk-of-bias tools emphasize the importance of transparently describing how the gold standard is constructed and of explicitly acknowledging its potential sources of bias (195,196). When the gold standard is not perfectly sensitive and/or specific, conventional contingency tables yield naïve estimates that reflect agreement with an imperfect label rather than with the underlying clinical truth. Methodological work on imperfect gold standards proposes several approaches, including imputation or adjustment for missing verification, algebraic correction for misclassification when the operating characteristics of the reference are known, construction of composite or panel-based references, and latent class modeling when no acceptable gold standard exists (190). Algebraic correction for reference misclassification requires known operating characteristics of the reference standard, which cannot be defined for a regulatory document translated into operational rules; composite reference standards rely on multiple independent sources measuring the same condition, whereas the ANSM Thesaurus represents the single French normative source defining high-severity pDDIs; latent class modeling would have required strong assumptions of conditional independence and multiple statistically independent measurements, conditions not met by the dataset (197). A deterministic and fully auditable refinement of the gold standard, using a form based on discrepancies, was thus chosen, with an explicit tracking of all modifications (193).

Several limitations must be acknowledged. First, the audit of discordances was not fully blinded to algorithmic or human classifications, which introduces the possibility of review bias. Predefined audit rules and rigorous tracking of all modifications were implemented. Reassessment according to the ANSM Thesaurus was initially performed with knowledge limited to the binary classification (i.e., positive, or not positive) of the pair according to P_{RoSIT} and the gold standard. Entities and pharmacological classes were independently re-evaluated before considering additional information from the gold standard or P_{RoSIT}. Residual bias, however, cannot be excluded.

An additional assumption concerns the correctness of pharmacological entity extraction by human annotators. Identified entities were assumed to be correct, and discrepancies in substances were attributed to NLP-related errors, including possible false detections due to misinterpretation of negated statements such as drug discontinuation. This assumption was not formally verified. However, entity extraction was performed in a structured manner by trained personnel, and the task appears inherently more error-prone for an NLP system than for adequately trained human reviewers.

Approaches similar to the one conducted in this study, based on selective re-examination of discordant cases have historically been criticized on methodological grounds (198). A clear distinction is therefore required. In classical discrepant analysis, the index test or a closely related assay is used to redefine disease status in discordant cases, thereby introducing incorporation bias and potentially producing systematically inflated estimates of sensitivity and specificity (199). The main concern is that the new test contributes, directly or indirectly, to the construction of the reference standard, thus violating independence between index and reference.

Although the present study shares a structural element with discrepant approaches, namely the identification and targeted audit of discordant cases, the framework differs in key aspects.

First, the NLP tool or any related algorithmic output was never used to redefine the presence of a high-severity pDDI. All corrections were anchored exclusively to the institutional source, the ANSM Thesaurus, and to predefined operational rules. Second, corrections were limited to errors clearly attributable to human annotation and were documented through a structured proof log to ensure traceability. Third, refinement was not restricted to FN or FP. TP cases also underwent structural concordance analysis, which allowed identification of residual pair-level discrepancies independent of patient-level agreement. The present strategy, therefore, is more accurately described as a discordance-based refinement rather than classical discrepant analysis. Similar targeted reviews of discordant cases have been reported in other diagnostic AI evaluations, where independent expert re-examination of cases initially classified as discordant revealed labeling errors in the reference standard and led to post hoc reanalysis of system performance after correction of the reference labels (200,201).

This strategy renders the refinement of the gold standard transparent and rational; however, the complete absence of residual errors cannot be excluded. Given the systematic audit of FN, FP, and TP cases, potential residual misclassification would most plausibly concern concordant TP or TN cases, i.e., instances where both PRoSIT and the gold standard indicate absence or presence of high-severity pDDIs. Such errors would require a simultaneous failure of both human application of the operational rules and automated detection, which is expected to be uncommon. Consistently, during the refinement process, it was rare for both reviewers to commit the same error, suggesting that most human misclassifications were not systematically shared. Nonetheless, residual errors cannot be fully excluded. Targeted re-evaluation of a random sample of concordant cases could be performed in future work to further assess this possibility.

Applying the same iterative framework to tools based on different architectures may contribute to a final test/refinement of the gold standard, as comparing models that show different structural errors may help to better identify residual human errors.

Conclusions

This study confirms that evaluation of AI tools for identifying high-severity pDDIs cannot be separated from an explicit analysis of gold standard imperfection and data structure. Performance metrics must be interpreted as a joint function of the tested tool, the gold standard, and the aggregation logic adopted. A systematic discordance analysis, grounded in a uniform causal taxonomy and a clear distinction between patient-level and pair-level evaluation, represents an essential methodological instrument to make the gold standard more coherent, auditable, and progressively closer to clinical ground truth.

5.3 Lessons learned and methodological considerations

This study showed that accuracy metrics of automated tools cannot be interpreted independently from the structure of the gold standard. As the gold standard is often imperfect when dealing with free text regulatory or institutional sources, performance estimates reflect both tool behavior and gold standard instability. In this study, FN and sensitivity were affected by human-related errors in the gold standard, while most FP were linked to structural characteristics of the tool, such as the unconditional application of context-dependent rules. Performance metrics should therefore be read as the result of an interaction between algorithm, reference, and aggregation logic.

The audit also showed that errors were not randomly distributed. Discordances followed a Pareto-like pattern, with a limited number of semantic clusters accounting for the majority of discrepancies. Potassium-related pDDIs and bradykinin-mediated angioedema represented recurrent instability points. This concentration indicates that imperfections arise from specific ambiguity nodes in the translation of regulatory text into operational rules, rather than from diffuse annotation noise or reference ambiguity. Both human reviewers and the algorithm were sensitive, for different reasons, to these high-prevalence and lexically ambiguous pDDI groups.

The study further clarified the limits of patient-level aggregation. Agreement at the patient level may occur even when the identified drug pairs do not overlap. Without systematic pair-level auditing, performance can appear artificially improved or degraded for structural reasons. A robust gold standard therefore requires dual evaluation: patient-level metrics to preserve clinical interpretability, and pair-level analysis to maintain structural consistency. When aggregation is performed at the patient level, the possibility of spurious agreement should be considered. The review of TP requires however particular caution. As corrections at the pair level propagate across all occurrences of the same combination, some TP are

indirectly modified before any dedicated TP analysis. A final review of TP is therefore time-consuming and may provide only marginal additional gains in structural coherence. Its necessity should be evaluated pragmatically, balancing methodological benefit against operational burden.

As the refinement process modified the underlying gold standard used in the second study, the results presented in that study should be interpreted in light of the refined gold standard. Updated results derived from the refined gold standard are provided in the Supplementary Material.

In conclusion, the study shows that in complex clinical systems the gold standard may not be a fixed truth but a progressively stabilized construct. Iterative discordance analysis represents a practical strategy to improve its internal coherence. In this perspective, evaluation of NLP systems is not strictly unidirectional. While the tool is assessed against the gold standard, the analysis of the gold standard against the tool can also reveal weaknesses in the gold standard itself and contribute to its refinement.

6 - General discussion, perspectives, and conclusions

6.1 General discussion

This thesis addressed the problem of high-severity pDDIs through a structured progression aligned with its primary objectives: to improve comparability of sources and clarify the definition of ground truth; to construct and refine a gold standard based on real-world data; and to evaluate an automated detection system for the identification of high-severity pDDIs. As discussed in previous chapters, when dealing with pDDIs, the epistemic dimension (i.e., what we consider valid knowledge) and the operational dimension (i.e., how we address the issue based on current knowledge) must realistically advance in parallel.

The first study therefore focused on management recommendations of pDDIs across different sources, after standardization into comparable categories. The underlying premise was that the “strength of intervention” associated with recommendations is central for clinical prioritization. Risk information becomes clinically relevant only when it is translated into action. The study showed that even when pDDIs are recognized across multiple databases, the recommended actions may diverge substantially. In a set of pDDIs identified in discharge records, overall agreement across five ICs was moderate and varied by source. When granularity was deliberately reduced to a binary distinction, any intervention *vs.* no intervention, agreement improved but remained suboptimal, given that simplification mechanically reduces opportunities for disagreement. Divergences also persisted when restricting the analysis to the most critical pDDIs.

These discrepancies could already have concrete implications for clinical practice. In French hospital settings, several ICs may be routinely consulted, while SmPCs and the ANSM Thesaurus should provide a regulatory background. However, in clinical practice, systematic consultation of SmPCs or free-text sources such as the ANSM Thesaurus is time-consuming and not always feasible. As a result, identical prescriptions may trigger, or fail to trigger, interventions in a non-uniform way depending on the source consulted.

Beyond these considerations, the observed discrepancies also raise a more fundamental methodological question concerning the reliability of the underlying knowledge base. Some degree of variability is expected. Public health decisions derive from national regulatory frameworks, and different sources target different audiences (e.g., international ICs vs. national regulatory references). However, the magnitude of disagreement observed cannot reasonably be attributed to contextual factors alone. Most pDDIs depend on fundamental PD/PK properties of co-administered substances. Large divergences therefore suggest either limited robustness or limited transparency in the underlying evidence base.

This observation has direct implications for methodological choices in operational research on pDDIs. Although the results are informative for prioritization, it would be inappropriate to treat cross-database consensus as an immediate proxy for clinical ground truth. Integrating information from multiple secondary sources in contexts of high disagreement, without a clear hierarchy, risks increasing confusion among prescribers rather than supporting decision-making.

A further difficulty is the extremely limited transparency regarding the supporting evidence. Without access to the rationale underlying individual recommendations, it becomes difficult to determine why discordances arise; it is not reasonable to expect clinicians to resolve strong discrepancies when the sources and justifications required for that judgement are not provided.

From this perspective, basing operational studies on a national reference, despite their limits, appears to be the most coherent strategy for implementation within a defined clinical context. Systematic integration of divergences across sources remains a promising but complex challenge, requiring a careful balance between informational overload, adherence to regulatory guidance, and the ethical responsibility to highlight potential omissions or ambiguities when discrepancies are substantial.

Nonetheless, elements of common convergence, although rare, may represent areas where the literature is likely to be consistently in agreement. These areas of convergence could allow the creation of short lists of *never events* that can be further investigated and used as a starting point for targeted prevention efforts.

The second study thus moved to an operational construction for the identification of high-severity pDDIs based on a single, established source, the ANSM Thesaurus. This was achieved through construction of an expert-derived gold standard and its application to a real-world hospital cohort.

A double independent review and adjudication process enabled the creation of a transparent, and auditable gold standard, constructed using more than 6,000 hospitalized patients and over 30,000 unique pharmacological pairs. Initial results on its reliability showed high inter-rater agreement between the two experts, indicating that only a small proportion of cases required the intervention of the consensus. Secondly, the findings of the study show the added value of analyzing real prescribing patterns. Approximately 6-10% of patients were exposed to at least one high-severity pDDI (figures varying based on gold standard version). These pDDIs were strongly associated with polypharmacy, and concentrated within a limited number of drug classes and recurrent prescribing clusters. This result supports pharmacist-to-clinician information strategies focused on a restricted core of clinically critical pDDIs. Similarly, the heterogeneity observed across different care settings suggests that information strategies should be context-specific and aligned with local prescribing patterns.

The initial evaluation of P_{Ro}SIT against the original gold standard showed insufficient performance for reliable identification of at-risk patients. The tool missed a clinically relevant proportion of cases, rendering direct deployment as a rule-out automated interception system unacceptable. At the same time, the distribution of discordances suggested that FN and FP could not be interpreted solely as algorithmic limitations. A

relevant proportion of FN regarded patients without initial expert agreement, indicating a plausible component of error linked to the ANSM Thesaurus and therefore suggesting the need to further discriminate between tool error and uncertainty or ambiguity within the reference.

Thus, the third study was designed to systematically investigate the nature and determinants of the observed discordances. Results showed that measured performance was a joint function of both the tested tool and the gold standard, and that error patterns were asymmetrically distributed between FN and FP, reflecting distinct underlying mechanisms. Analysis of recurrent clusters suggested intrinsic limitations of the regulatory source, particularly ambiguous formulations and conditional notes that proved difficult to operationalize consistently by a NLP tool. Regulatory documents are primarily intended for communicative purposes within clinical practice; they are not designed as formal computational rule sets (e.g., the structure used by guidelines). Their transformation into executable decision logic may introduce variability for both clinicians and algorithms.

A substantial proportion of patients classified as having a high-severity pDDI exclusively by the gold standard (i.e., FN) were initially interpreted as algorithmic failures. However, structured audit demonstrated that these cases largely reflected inconsistencies in the construction of the gold standard itself. Specifically, a small number of complex or ambiguous reference document formulations accounted for most of the variability across reviewers, thus generating misclassification at the gold standard level. For example, potassium-related pDDIs included conditional notes excluding applicability in specific conditions, which were interpreted inconsistently during consensus review. Similarly, combinations involving acetylsalicylic acid with clopidogrel or diclofenac required distinction between validated indications, off-label use, and dose-dependent criteria.

In contrast, the audit of FP identified a different mechanism. Here, the gold standard classification was mostly confirmed as appropriate, whereas PRoSIT had systematically applied pDDI rules without incorporating contextual qualifiers of the regulatory source (i.e., applicability restricted to specific patient categories or clinical conditions). These errors were structural and implementation-related, arising from rigid execution of formalized rules that required a more nuanced contextual interpretation.

This iterative process conducted in the third study recontextualized validation as a refinement process. Discordances were treated as informative signals about the system as a whole, including the regulatory source, the gold standard, the algorithm, and the unit of analysis. The results illustrate the presence of reference standard bias in pDDI research and show that performance metrics are a joint function of tool behavior and source document limits.

Taken together, the results of these studies indicate that identification of high-severity pDDIs must rest on explicit, traceable, and contextually grounded criteria. In hospital settings characterized by polypharmacy and clinical complexity, the absence of structured, automated tools risks heterogeneous decision-making and inconsistent application of interventions.

6.2 Perspectives

This work identified several practical and research perspectives, regarding gold standard development, educational intervention, evidence generation and improvement of the reference source.

A first perspective concerns the gold standard as a research infrastructure. The ANSM Thesaurus, in its current form, presents structural limitations, including conditional formulations and partial overlap between severity and management categories. Nevertheless, it remains the only reference explicitly anchored to the French regulatory and clinical context. International databases do not systematically include drugs approved exclusively at

the national level, nor do they consistently reflect recommendations issued by French authorities. For tools intended for use in French hospitals, reliance on this reference document is therefore still essential and the developed gold standard represents a means for their validation.

At the same time, the ANSM Thesaurus undergoes gradual revisions, and validation in other healthcare systems may require adaptation to different regulatory frameworks. For these reasons, the gold standard should not be treated as a fixed entity but as an ongoing research infrastructure.

Within this framework, incremental extension represents a coherent strategy to balance external validity with difficulty of manual annotation. The existing version can be used as a stable base, while previously evaluated drug pairs are retained as validated. Expert review can then focus on newly emerging combinations, for example across different hospital units or time periods. This controlled accumulation allows the reference to evolve without re-opening the entire structure at each iteration.

Sequential testing of additional tools against the same reference may also further help identify residual structural weaknesses that persist over time.

Although the methodological framework developed is potentially transferable to other healthcare systems, several components remain intrinsically dependent on the French clinical and regulatory context. The gold standard was derived from the ANSM Thesaurus, a national regulatory reference whose classifications and formulations may differ substantially from those used in other countries. Similarly, PROSIT relies on French-language ontology and prescription normalization procedures, limiting immediate applicability to other linguistic environments without substantial adaptation. In addition, prescribing patterns, drug availability, reimbursement policies, and integration of clinical pharmacy services vary

across healthcare systems and may modify both the prevalence and operational relevance of specific pDDIs.

A second perspective concerns intervention in patient care. As shown in the empirical analyses, high-severity pDDIs were not evenly distributed but concentrated in a limited number of recurrent combinations. After gold standard refinement, more than one quarter of corrected cases involved only ten pDDIs, several of which shared common PD/PK mechanisms. For example, QT prolongation emerged as a recurrent pattern, notably through combinations involving amiodarone, escitalopram, hydroxyzine, and trimethoprim-sulfamethoxazole. Psychotropic-related combinations involving quetiapine, tramadol, duloxetine, or paroxetine also represented another important subgroup, more markedly in neurosciences settings.

These clusters offer a practical basis for action. A restricted core of recurrent, high-impact pDDIs can support focused alerting strategies, targeted educational initiatives, and structured pharmacist-to-clinician feedback in departments characterized by high polypharmacy. Prioritizing a limited number of highly prevalent pDDIs is likely to produce greater clinical benefit than dispersing attention across a large number of lower frequency, high-severity pDDIs. It also reduces the volume of signals presented to prescribers, making decision support more manageable and more likely to be used consistently.

A third perspective concerns evidence generation. For many pDDIs, the link between theoretical risk and clinical outcome remains insufficient. Classifications may rely on pharmacological plausibility more than on robust outcome data. Concentrating research efforts on the identified high-priority high-prevalence core offers a pragmatic pathway. PV signal assessment can be conducted using large spontaneous reporting databases (130) and signals can then be evaluated through observational studies focusing on clinically relevant endpoints in routine care (134).

A final perspective concerns the structured comparison of pDDI sources as a tool for evidence development. Divergences between national and international ICs should be treated as indicators of areas requiring further investigation. When substantial differences in management classification are observed, these pDDIs can be prioritized for documentation review, targeted research, and expert reassessment. Evidence generated through this process can then potentially inform updates of the ANSM Thesaurus and progressively improve its quality. Validated discordances may be incorporated into NLP systems through controlled integration procedures. In this way, cross-source comparison may become a mechanism for identifying research priorities and refining the regulatory reference.

6.3 Conclusions

This thesis developed a methodological framework to improve the identification and prioritization of pDDIs, from the comparison of existing sources to the construction of an operational reference and the evaluation of an automated detection tool.

The results show that management recommendations for pDDIs may diverge substantially across sources with limited reported evidence, limiting the use of cross-source information as an immediate prioritization strategy or as a reliable basis for clinical decision-making. They also show that reliable identification of high-severity pDDIs requires explicit and traceable operational criteria. When evaluating automated tools, performance metrics cannot be interpreted independently from the quality and stability of the reference used for validation.

The work presented here thus suggests that system implementation and refinement should proceed in parallel with evidence generation. Improving the identification and management of high-severity pDDIs requires systems that are not only technically accurate, but also transparent, auditable, and continuously revisable. Only through the continuous refinement

of both operational criteria and the epistemic foundations on which they rest can pDDI management become more coherent, sustainable, and clinically meaningful.

7 - SUPPLEMENTS

7.1 Supplements Chapter 3

Chapter 3 - Supplementary Table 1: Classification of pDDIs according to each comparison IC.

ANSM Thesaurus.

This document, published by the *Agence Nationale de sécurité des Médicaments et des produits de santé* (ANSM), is regularly updated to reflect the latest information on drug interactions in France.

The ANSM Thesaurus classifies DDIs into four categories, each associated with different management recommendations, defined as follows:

- **Contraindicated:** the drug combination must be avoided in all cases.
- **Avoid, unless benefits outweigh harms:** the drug combination requires careful consideration of the benefit-risk ratio and strict patient monitoring.
- **Needing precaution of use:** co-administration is generally allowed but requires specific precautions to minimize adverse reactions (e.g., dosage adjustments, enhanced monitoring).
- **To be considered:** a risk of adverse reactions theoretically exists, and it is up to the prescriber to evaluate the appropriateness of the combination.

It also includes, for a majority of pDDIs, a description of the mechanism or clinical effect of the interaction; for some interactions, it also provides a detailed description of the recommended management. The categories (associated with management recommendations) were used for this study (Table 1)

British National Formulary.

It provides extensive advice on therapeutics and pharmacology and lists numerous DDIs, primarily focusing on drugs available under the UK National Health Service (NHS). Most pDDIs are listed under specific entries for each substance/class. It provides a grading for both the severity of the DDI and the level of evidence. It provides short details regarding the mechanism of the DDI and an action statement for its management. Additionally, twenty tables are dedicated to listing pharmacodynamic pDDIs. Nonetheless, they rarely provide any information on their management (i.e., only interactions reported under Table 9). The action statements were used in this study (Table 1 and, in detail, Table S2).

Drugs.com

The DDIs are categorized by a classification that defines their severity (Major, Moderate, Minor, Unknown). It also includes an action statement defining the recommended behavior (e.g., Contraindicated, Generally Avoid, Monitor Closely, Adjust Dose) and a specific discussion on mechanisms, risk factors, and management. The general action statements were used for this study (Table 1).

Epocrates Interaction Check (*Epocrates*)

This IC includes only U.S. FDA-approved products and is meant to reflect U.S.-specific clinical practices. It allows for searching combinations of drugs, herbs, and supplements, listing for each pDDI a short description of the potential adverse effects, the action required for clinical management, and the mechanism of the pDDI. It classifies pDDIs through general categories, broadly indicating the recommended actions. (i.e., Contraindicated, Avoid/Use Alternative, Monitor/Modify treatment, Caution advised, None, Therapeutic advantage). It also includes specific action statements, which were used for this study (Table 1 and, in detail, in Table S3).

UpToDate® Lexidrug™ (*Lexidrug*)

It classifies drugs both through a severity scale and by a “Risk Rating,” defining the recommended actions. Detailed information regarding the effect, the mechanism, the risk factors, and the management of the pDDI are also provided. For this study, the risk rating system was adopted (Table 1):

X - Avoid Combination: The interaction is clinically significant and the risks clearly outweigh the benefits. D - Consider Therapy Modification: The interaction is significant. Co-administration may be justified depending on patient-specific factors, but therapeutic modifications (e.g., monitoring, dose changes, alternative drugs) are required to reduce risk. C - Monitor Therapy: The interaction exists but the benefits of co-administration usually outweigh the risks. Clinical monitoring is recommended; dosage adjustment may be needed in some patients. B - No Action Needed: An interaction may exist, but there is little to no clinical concern. No specific management is required.

A - No Known Interaction: No pharmacokinetic or pharmacodynamic interaction has been identified between the agents.

Medscape Drug Interaction Checker (*Medscape*). It classifies DDIs based on a category regarding the recommended action to follow by the clinician (e.g., Contraindicated, Serious-Use alternative), together with a more specific action statement and general details on the pDDI. The action statements were used for this study (Table 1).

Chapter 3 - Supplementary Table 2: Classification of Common action statements from the BNF.

High-priority (Always avoid)

- Avoid;
- Avoid D1/D2 for (n amount of time) before or (n amount of time) after D1/D2;
- Avoid D1/D2 for at least (n amount of time) before or (n amount of time) after D1/D2;
- Avoid for (n amount of time) after stopping D1/D2;
- Avoid for at least (n amount of time) after stopping treatment;
- Avoid for at least (n amount of time) before and up to (n amount of time) after stopping treatment;

Intermediate-priority (Conditional co-administration)

- Adjust D1 and D2 dose;
- Adjust D1 dose with D2;
- Adjust D1/D2 and monitor adverse effects;
- Adjust D1/D2 dose and separate administration by (n amount of time);
- Adjust D1/D2 dose;
- Adjust D1/D2 starting dose and monitor;
- Adjust D1/D2 starting dose;
- Adjust dose and monitor (physiological/biochemical parameter);
- Adjust dose with D1/D2;
- Administer D1/D2 with an acidic beverage;
- Avoid (specific route of administration);
- Avoid another dose of D1/D2 within (n amount of time) of concurrent use;

- Avoid D1/D2 or adjust D1/D2 dose and monitor for (symptoms);
- Avoid D1/D2 or adjust D1/D2 dose;
- Avoid depending on (physiological/biochemical parameter) - consult product literature;
- Avoid depending on other drugs taken - consult product literature;
- Avoid except in an emergency situation - consult product literature;
- Avoid or adjust D1/D2 dose - consult product literature;
- Avoid or adjust D1/D2 dose and monitor (physiological/biochemical parameter);
- Avoid or adjust D1/D2 dose and monitor concentration;
- Avoid or adjust dose with D1/D2 - consult product literature;
- Avoid or monitor (symptoms/biochemical parameter);
- Avoid or monitor adverse effects and consider (other drug) as an alternative;
- Avoid or monitor adverse effects;
- Avoid or monitor D1/D2 concentration and adjust dose accordingly;
- Avoid or monitor efficacy;
- Avoid or monitor for toxicity;
- Avoid or monitor;
- Avoid or take D1/D2 first and separate administration by at least (n amount of time);
- Avoid unless patient can be monitored for signs of (signs/symptoms);
- D1/D2 should be taken (n amount of time) after D1/D2;
- D1/D2 should be taken (n amount of time) before D1/D2;
- D1/D2 should be taken (n amount of time) before or (n amount of time) after D1/D2;
- Manufacturer advises caution or adjust D1/D2 dose with D1/D2;
- Manufacturer advises caution or adjust D1/D2 dose;
- Manufacturer advises take both drugs at the same time;
- Manufacturer advises take D1/D2 (n amount of time) before or (n amount of time) after D1/D2;

- Manufacturer advises take D1/D2 with food;
- MHRA advises avoid unless (physiological/biochemical parameter) can be monitored closely;
- Monitor (physiological/biochemical parameter) and adjust dose;
- Monitor adverse effects and adjust dose;
- Monitor and adjust D1/D2 dose;
- Monitor and adjust dose;
- Monitor concentration and adjust dose;
- Monitor for toxicity and adjust dose;
- Separate administration by at least (n amount of time);
- UKHSA advises separating administration by (n amount of time);
- Use with caution or avoid.

Low-priority (No action required)

- Manufacturer advises caution depending on genotype;
- Manufacturer advises caution depending on other drugs taken;
- *No action statements;*
- *No interactions;*
- Refer to patient information leaflet for missed pill advice
- Refer to specialist literature.

This table presents the most common action statements from the BNF and their classification according to the scale used in the study. While this list covers all the action statements of the pDDIs included in this study, it should not be considered exhaustive. D1/D2 refers to the first and/or the second substance in an interacting pair.

Chapter 3 - Supplementary Table 3: Classification of Common action statements from Epocrates.

High-priority (Always avoid)

- avoid combo;
- avoid combo during and (n amount of time) after D1/D2; avoid combo, especially with (dose of D1/D2); contraindicated;
- contraindicated during and (n amount of time) after D1/D2;
- contraindicated, hold D1/D2 if urgent use, monitor (physiological/biochemical parameter);
- contraindicated w/in (n amount of time) before or (n amount of time) after D1/D2 use;
- contraindicated w/in (n amount of time) of D1/D2 use;
- contraindicated w/in (n amount of time) of D1/D2 use, hold D1/D2 if urgent use, monitor (physiological/biochemical parameter);
- hold D1/D2 while on D1/D2;
- use alternative (substance/class)
- use alternative or hold D1/D2 if urgent use;

Intermediate-priority (Conditional co-administration)

- avoid combo w/ (specific formulation/dose of D1/D2), otherwise, adjust D1/D2 to (dose);
- co-administer simultaneously;
- consider alternative or monitor (physiological/biochemical parameter);
- consider alternative;
- consider decr. D1/D2 dose;
- consider decr. dose of one or both drugs;
- consider holding D1/D2 during D1/D2 tx, otherwise monitor (physiological/biochemical parameter);
- contraindicated in (medical condition), otherwise use lowest effective D1/D2 dose and monitor for (physiological/biochemical parameter);
- decr. D1/D2 dose (specific %) when D1/D2 initiated;
- decr. D1/D2 dose by (amount), consider holding D1/D2 while on D1/D2;
- decr. D1/D2 dose by (amount), otherwise monitor (physiological/biochemical parameter);
- decr. D1/D2 dose by (specific dose) if not tolerated;
- doses not estab. for combo of D1/D2 w/ D1/D2;

- give D1/D2 (specific dose) w/ D1/D2 in tx-naive pts;
- generally avoid concurrent use, withhold D1/D2 for at least (n amount of time) before (therapy modification);
- give (specific dose of D1/D2);
- if D1/D2 not admin. w/ (other drug), monitor (physiological/biochemical parameter) w/ D1/D2;
- increase D1/D2 dose to (amount) during and (n amount of time) after D1/D2;
- max D1/D2 dose (max dose);
- max D1/D2 dose = (specific dose);
- monitor (physiological/biochemical parameter), consider decr. D1/D2 dose, max D1/D2 dose = (specific dose);
- monitor (physiological/biochemical parameter), consider decr. D1/D2 dose; monitor (physiological/biochemical parameter), consider decr. D1/D2 start dose;
- monitor (physiological/biochemical parameter), consider monitoring (physiological/biochemical parameter);
- monitor (physiological/biochemical parameter), especially during initiation/titration;
- monitor (physiological/biochemical parameter), especially if D1/D2 dose > (specific dose), consider D1/D2 dose adjustment;
- monitor (physiological/biochemical parameter), if also combined with (other drug), consider decr. D1/D2 dose;
- monitor (physiological/biochemical parameter), if also combined with (other drug), decr. D1/D2 dose or avoid combo if already on (D1/D2 dose);
- monitor (physiological/biochemical parameter), max D1/D2 dose = (specific dose);
- monitor (physiological/biochemical parameter);
- monitor (physiological/biochemical parameters), decr. D1/D2 dose (specific %) when D1/D2 initiated;
- monitor (physiological/biochemical parameters), may need to incr. D1/D2 dose;
- use alternative or monitor (physiological/biochemical parameter), consider decr. D1/D2 dose;
- monitor (physiological/biochemical parameters), use lowest effective D1/D2 dose; monitor for (physiological/biochemical parameter);
- monitor for (signs/symptoms); use alternative or increase D1/D2 dose to (specific dose);
- use alternative or monitor (physiological/biochemical parameter), consider decr. D1/D2 dose during and (n amount of time) after D1/D2 tx;
- use alternative or monitor (physiological/biochemical parameter), consider lower D1/D2 doses;

<ul style="list-style-type: none"> - use alternative or monitor (physiological/biochemical parameter), especially if D1/D2 dose > (specific dose); - use alternative or monitor (physiological/biochemical parameter); - use alternative or monitor (physiological/biochemical parameter, symptoms), especially during initiation/titration; - use alternative or monitor (physiological/biochemical parameter, symptoms), use max D1/D2 dose; - use alternative or monitor (physiological/biochemical parameter, symptoms); - use alternative or monitor (physiological/biochemical parameters), consider holding D1/D2 during D1/D2 tx; - use alternative or monitor (physiological/biochemical parameters), especially during initiation/titration; - use alternative or monitor (physiological/biochemical parameters, including (specific parameter)), risk may be lower with (alternative formulation); - use lowest effective D1/D2 dose, monitor (physiological/biochemical parameter).
<p style="text-align: center;">-</p> <p>Low-priority (No action required)</p> <ul style="list-style-type: none"> - caution advised; - no significant interaction(s) known or found for selected drugs.

This table presents the most common action statements from the Epocrates and their classification according to the scale used in the study. While this list covers all the action statements of the pDDIs included in this study, it should not be considered exhaustive. D1/D2 refers to the first and/or the second substance in an interacting pair

Chapter 3 - Supplementary Table 4 - List of 70 sample pDDIs extracted to construct the scale, with relevant sections regarding management.

pDDI	ANSM Thesaurus	BNF	Lexidrug	Drugs.com	Epocrates	Medscape
ciclosporin- everolimus	Needing precaution of use	Avoid or adjust dose	D - Consider therapy modification	Generally Avoid	admin. PCP prophylaxis; monitor renal fxn;	Modify Therapy/Monitor Closely
rifampicin - voxelotor	Avoid, unless benefits outweigh harms	Avoid	D - Consider therapy modification	Generally Avoid	use alternative or incr. voxelotor dose to 2500 mg qd	No Interactions found
dabigatran- tacrolimus	Avoid, unless benefits outweigh harms	Avoid	No interactions found	Monitor	consider separating admin. at least 2h apart:	Use Caution/Monitor
atovaquone - rifabutin	To be considered	Avoid	X - Avoid	Generally Avoid	avoid combo	Use Caution/Monitor
rifampicin - rivaroxaban	Avoid, unless benefits outweigh harms	Avoid unless patient can be monitored for signs of thrombosis	X - Avoid	Generally Avoid	avoid combo	Use Caution/Monitor
codeine - paroxetine	Avoid, unless benefits outweigh harms	No action statements	C - Monitor	Monitor	use alternative or monitor resp. rate, serotonin syndrome sx, especially during initiation/titration, withdrawal sx; decr. dose of one or both drugs, use lowest effective doses and shortest duration of conc. tx	Use Caution/Monitor

pDDI	ANSM Thesaurus	BNF	Lexidrug	Drugs.com	Epocrates	Medscape
midazolam - diltiazem	Needing precaution of use	Monitor adverse effects and adjust dose	D - Consider therapy modification	Monitor	monitor BP, resp. rate, consider decr. midazolam dose	Use Caution/Monitor
atazanavir - efavirenz	Avoid, unless benefits outweigh harms	Refer to specialist literature	D - Consider therapy modification	Adjust Dose	avoid combo in tx-experienced pts; give atazanavir 400 mg qd w/ ritonavir or cobicistat in tx- naive pts	Avoid or Use Alternate Drug
rifampicin- nintedanib	Needing precaution of use	No interactions found	X - Avoid	Generally Avoid	avoid combo: combo may decr. nintedanib levels, efficacy hepatic metab. induced; P-gp- mediated transport induced	Avoid or Use Alternate Drug
rifabutin - posaconazole	Avoid, unless benefits outweigh harms	Avoid	D - Consider therapy modification	Generally Avoid	use alternative or monitor breakthrough fungal infection, CBC, uveitis s/sx	Avoid or Use Alternate Drug
rifampicin - morphine	Précaution d'emploi	No action statements	C - Monitor	Monitor	caution advised	Minor/Significance Unknown
oxcarbazepine- topiramate	Needing precaution of use	No interactions found	No interactions found	Monitor	caution advised	No Interactions found
fluoxetine- tamoxifen	Avoid, unless benefits outweigh harms	Avoid	D - Consider therapy modification	Generally Avoid	Caution advised	Use Caution/Monitor
rolapitant - tamoxifen	Avoid, unless benefits outweigh harms	One or more substances missing in the IC	D - Consider therapy modification	Generally Avoid	caution advised	Use Caution/Monitor

pDDI	ANSM Thesaurus	BNF	Lexidrug	Drugs.com	Epocrates	Medscape
metoprolol - darifenacin	Needing precaution of use	No interactions found	C - Monitor	Monitor	caution advised	Use Caution/Monitor
carbamazepine - simvastatin	Association Déconseillée	Monitor and adjust dose	C - Monitor	Generally Avoid	consider alternative	Avoid or Use Alternate Drug
rifampicin - simvastatin	Association Déconseillée	No action statements	C - Monitor	Generally Avoid	consider alternative	Avoid or Use Alternate Drug
bleomycin - brentuximab	Contraindicated	Avoid	X - Avoid	Contraindicated	contraindicated	Contraindicated
voriconazole - rifampicin	Contraindicated	Avoid	X - Avoid	Contraindicated	contraindicated	Contraindicated
voriconazole- quinidine	Contraindicated	One or more substances missing in the IC	X - Avoid	Contraindicated	contraindicated	Contraindicated
colchicine - fluconazole	Avoid, unless benefits outweigh harms	Adjust Colchicine dose with moderate CYP3A4inhibitors	D - Consider therapy modification	Adjust Dose	give colchicine 0.3-0.6 mg/day divided qd-bid	Avoid or Use Alternate Drug
amiodarone- afatinib	Needing precaution of use	No action statements	D - Consider therapy modification	Adjust Dose	decr. afatinib dose by 10 mg/day if not tolerated	Avoid or Use Alternate Drug

pDDI	ANSM Thesaurus	BNF	Lexidrug	Drugs.com	Epocrates	Medscape
afatinib - verapamil	Needing precaution of use	No action statements	D - Consider therapy modification	Adjust Dose	decr. afatinib dose by 10 mg/day if not tolerated:	Avoid or Use Alternate Drug
selexipag - clopidogrel	Needing precaution of use	Adjust selexipag dose	D - Consider therapy modification	Adjust Dose	decr. selexipag dose to once daily	Use Caution/Monitor
fosamprenavir - efavirenz	Needing precaution of use	Refer to specialist literature	D - Consider therapy modification	Adjust Dose	doses not estab. for combo of fosamprenavir w/ efavirenz; if using qd fosamprenavir plus ritonavir, incr. ritonavir to 300 mg qd	Use Caution/Monitor
verapamil - tenofovir alafenamide	Needing precaution of use	No interactions found	No interactions found	No drug-drug interactions were found between the drugs.	if tenofovir alafenamide not admin. w/ cobicistat, monitor renal fxn w/ oral verapamil	No Interactions found
afatinib - rifampicin	Needing precaution of use	No action statements	D - Consider therapy modification	Adjust Dose	incr. afatinib dose by 10 mg/day	Avoid or Use Alternate Drug
propranolol - ergotamine	Needing precaution of use	No action statements	C - Monitor	Monitor	monitor BP	No Interactions found
rifampicin - metoprolol	To be considered	No action statements	C - Monitor	Monitor	monitor BP	Use Caution/Monitor
nebivolol - paroxetine	Needing precaution of use	No action statements	C - Monitor	Monitor	monitor BP, HR	Avoid or Use Alternate Drug

pDDI	ANSM Thesaurus	BNF	Lexidrug	Drugs.com	Epocrates	Medscape
metoprolol - bupropion	Needing precaution of use	No action statements	C - Monitor	Adjust Dose	monitor BP, HR; consider lower metoprolol dose	Use Caution/Monitor
metoprolol - duloxetine	Needing precaution of use	No action statements	C - Monitor	Monitor	monitor BP, HR; consider lower metoprolol dose	Use Caution/Monitor
fluvoxamine - olanzapine	Needing precaution of use	Adjust dose	C - Monitor	Monitor Closely	monitor BP; consider decr. olanzapine dose:	Avoid or Use Alternate Drug
carbamazepine- topiramate	Needing precaution of use	No action statements	C - Monitor	Adjust Dose	monitor carbamazepine levels, especially w/ topiramate doses>200 mg/day; consider topiramate dose adjustment	Use Caution/Monitor
itraconazole - busulfan	Avoid, unless benefits outweigh harms	Monitor and adjust dose	C - Monitor	Monitor	monitor CBC	Use Caution/Monitor
acetylsalicylic acid - deferiprone	To be considered	No action statements	No interactions found	Monitor Closely	monitor CBC, renal fxn, bleeding s/sx, ototoxicity	Use Caution/Monitor
rolapitant - rosuvastatin	Needing precaution of use	One or more substances missing in the IC	C - Monitor	Generally Avoid	monitor CK, myopathy sx; use lowest effective rosuvastatin dose	Use Caution/Monitor
verapamil - digoxin	Needing precaution of use	Monitor and adjust dose	C - Monitor	Adjust Dose	monitor digoxin levels, HR	Avoid or Use Alternate Drug

pDDI	ANSM Thesaurus	BNF	Lexidrug	Drugs.com	Epocrates	Medscape
digoxin - sulfasalazine	Needing precaution of use	No action statements	C - Monitor	Adjust Dosing Interval	monitor digoxin levels, may need to incr. digoxin dose 20- 40%	Use Caution/Monitor
ciprofloxacin- clozapine	Needing precaution of use	No interactions found	D - Consider therapy modification	Monitor Closely	monitor ECG, BP, HR; decr. clozapine to 1/3 usual dose	Avoid or Use Alternate Drug
propafenone- esmolol	Needing precaution of use	Use with caution or avoid	C - Monitor	No drug-drug interactions were found between the drugs.	monitor HR, BP	No Interactions found
ozanimod - clopidogrel	To be considered	No action statements	C - Monitor	Generally Avoid	monitor HR, ECG	No Interactions found
quinidine- cobicistat	Contraindicated	One or more substances missing in the IC	C - Monitor	Monitor Closely	monitor quinidine levels, ECG	Modify Therapy/Monitor
carbamazepine- acetazolamide	Needing precaution of use	No interactions found	No interactions found	Monitor	monitor sodium	Minor/Significance Unknown
docetaxel - apalutamide	Avoid, unless benefits outweigh harms	No action statements	C - Monitor	Generally Avoid	No significant interactions known or found for selected drugs	Avoid or Use Alternate Drug
dextromethorphan- cinacalcet	Contraindicated	One or more substances missing in the IC	C - Monitor	Generally Avoid	No significant interactions known or found for selected drugs	No Interactions found

pDDI	ANSM Thesaurus	BNF	Lexidrug	Drugs.com	Epocrates	Medscape
dihydroergotamine- triclabendazole	Contraindicated	One or more substances missing in the IC	No interactions found	No drug-drug interactions were found between the drugs.	No significant interactions known or found for selected drugs	No Interactions found
bictegravir - st. john's wort	Contraindicated	Avoid	X - Avoid	Generally Avoid	One or more substances missing in the IC	Avoid or Use Alternate Drug
telithromycin - rifampicin	Avoid, unless benefits outweigh harms	One or more substances missing in the IC	One or more substances missing in the IC	Generally Avoid	One or more substances missing in the IC	No Interactions found
norfloxacin- caffeine	To be considered	One or more substances missing in the IC	C - Monitor	Monitor	One or more substances missing in the IC	No Interactions found
dalfopristin - dihydroergotamine	Contraindicated	One or more substances missing in the IC	One or more substances missing in the IC	Monitor	One or more substances missing in the IC	No Interactions found
diprophylline- probenecid	Needing precaution of use	One or more substances missing in the IC	C - Monitor	No action statements provided.	One or more substances missing in the IC	No Interactions found
ciclosporin - lercanidipine	Needing precaution of use	Use with caution or avoid	X - Avoid	One of the drugs is not available inside the U.S.; no data available	One or more substances missing in the IC	No Interactions found
midazolam - roxithromycin	To be considered	One or more substances missing in the IC	C - Monitor	One or more substances missing in the IC	One or more substances missing in the IC	No Interactions found

pDDI	ANSM Thesaurus	BNF	Lexidrug	Drugs.com	Epocrates	Medscape
dasabuvir - ethinylestradiol	Contraindicated	One or more substances missing in the IC	X - Avoid	Contraindicated	One or more substances missing in the IC	One or more substances missing in the IC
agomelatine- ciprofloxacin	Avoid, unless benefits outweigh harms	No action statements	X - Avoid	One of the drugs is not available inside the U.S; no data available	One or more substances missing in the IC	One or more substances missing in the IC
mequitazine- paroxetine	Avoid, unless benefits outweigh harms	One or more substances missing in the IC	X - Avoid	One or more substances missing in the IC	One or more substances missing in the IC	One or more substances missing in the IC
josamycin- pimozide	Contraindicated	One or more substances missing in the IC	One or more substances missing in the IC	One or more substances missing in the IC	One or more substances missing in the IC	One or more substances missing in the IC
roxadustat - calcium	Needing precaution of use	Roxadustat should be taken at least 1 hour after calcium	D - Consider therapy modification	One or more substances missing in the IC	One or more substances missing in the IC	One or more substances missing in the IC
itraconazole - st. john's wort	Contraindicated	No interactions found	C - Monitor	Generally Avoid	One or more substances missing in the IC	Use Caution/Monitor
risperidone - carbamazepine	Needing precaution of use	Adjust dose	D - Consider therapy modification	Adjust Dose	monitor BP; consider incr. risperidone dose up to double; IM RISPERIDONE: monitor BP; consider incr. dose or additional oral risperidone; SC RISPERIDONE: monitor BP; consider incr. risperidone to next highest dose; consider additional oral risperidone if on highest SC dose	Use Caution/Monitor
zidovudine- ganciclovir	Needing precaution of use	No interactions found	C - Monitor	Generally Avoid	use alternative or monitor CBC	Contraindicated

pDDI	ANSM Thesaurus	BNF	Lexidrug	Drugs.com	Epocrates	Medscape
foscarnet - pentamidine	Needing precaution of use	No action statements	D - Consider therapy modification	Monitor Closely	use alternative or monitor CBC, renal fxn, electrolytes incl. calcium, ECG; risk may be lowerw/ nebulized pentamidine	Avoid or Use Alternate Drug
atorvastatin - erythromycin	Needing precaution of use	Monitor and adjust dose	C - Monitor	Adjust Dose	use alternative or monitor ECG, CK, myopathy sx	Avoid or Use Alternate Drug
fluconazole- amiodarone	Needing precaution of use	No action statements	D - Consider therapy modification	Generally Avoid	use alternative or monitor ECG, esp. if fluconazole dose >800 mg	Avoid or Use Alternate Drug
ondansetron- tramadol	To be considered	No action statements	C - Monitor	Monitor Closely	use alternative or monitor ECG, serotonin syndrome sx, especially during initiation/titration	No Interactions found
lithium - mannitol	Needing precaution of use	No interactions found	No interactions found	No drug-drug interactions were found between the drugs.	use alternative or monitor lithium levels, renal fxn; consider holding lithium during mannitol tx:	No Interactions found
verapamil - quinidine	Avoid, unless benefits outweigh harms	One or more substances missing in the IC	C - Monitor	Monitor	use alternative or monitor quinidine levels, ECG, BP, HR	Use Caution/Monitor
tramadol - bupropion	Avoid, unless benefits outweigh harms	No action statements	C - Monitor	Monitor Closely	use alternative or monitor resp. rate, ECG, withdrawal sx; consider lower bupropion doses	Use Caution/Monitor
methadone- voriconazole	Needing precaution of use	Adjust dose	D - Consider therapy modification	Monitor Closely	use alternative or monitor resp. rate, ECG; consider decr. methadone dose	Modify Therapy/Monitor Closely

Chapter 3 - Supplementary Table 5 - Classification of levels of evidence according to the BNF and Lexidrug.

<p style="text-align: center;">BNF</p> <p>Study: DDIs based on formal studies (both on the substance or other drugs with the same mechanism) are available.</p> <p>Anecdotal: DDIs based on one or a few case reports.</p> <p>Theoretical: DDIs based on theoretical considerations, which may derive from preclinical studies or information on other class members.</p>
<p style="text-align: center;">Lexidrug</p> <p>Excellent: DDIs based on multiple well-controlled studies (e.g., randomized controlled trials). The evidence is coherent and supports the existence of the DDI.</p> <p>Good: DDIs based on at least one well- controlled study (e.g., randomized controlled trial) or significant supporting evidence from non-RCTs. Strong evidence supports the existence of the DDI.</p> <p>Fair: DDIs based on one or more of the following: case studies/series, retrospective reviews, pilot investigations with low sample sizes, safety monitoring data, drug labeling, etc. This class also includes DDIs described only in drug labeling and DDIs based on known PK/PD properties and/or preclinical data. Poor: DDIs based on a single case report with an uncertain mechanistic basis. This class also includes theoretical DDIs without strong mechanistic or clinical evidence or DDIs with highly contradictory evidence supporting their existence.</p>

Chapter 3 - Supplementary Table 6 - List of identified pDDIs. This table follows the same order as Fig. 2 in the main document.

n	pDDI	n	pDDI	n	pDDI
1	azathioprine - febuxostat	74	dabigatran - rifampicin	147	lisinopril - ramipril
2	ciclosporin - simvastatin	75	diltiazem - ivabradine	148	ciprofibrate - fenofibrate
3	ketoconazole - simvastatin	76	erythromycin - ivabradine	149	itraconazole - prednisolone
4	itraconazole - ticagrelor	77	atorvastatin - gemfibrozil	150	ketoconazole - prednisolone
5	linezolid - tramadol	78	metoclopramide - ropinirole	151	prednisolone - voriconazole
6	riociguat - tadalafil	79	alfuzosin - clarithromycin	152	daptomycin - rosuvastatin
7	lisinopril - sacubitril	80	amitriptyline - linezolid	153	atorvastatin - daptomycin
8	perindopril - sacubitril	81	clarithromycin - fluticasone	154	daptomycin - pravastatin
9	ramipril - sacubitril	82	apixaban - rifampicin	155	ciclosporin - repaglinide
10	linezolid - venlafaxine	83	clarithromycin - tamsulosine	156	sulfamethoxazole - warfarin
11	crizotinib - domperidone	84	apixaban - carbamazepine	157	posaconazole - prednisolone
12	amiodarone - domperidone	85	apixaban - phenytoin	158	fluconazole - rivaroxaban
13	domperidone - sotalol	86	phenytoin - ticagrelor	159	furosemide - lithium
14	fusidic acid - simvastatin	87	fenofibrate - pravastatin	160	fluconazole - oxycodone
15	alfuzosin - urapidil	88	ciclosporin - spironolactone	161	everolimus - perindopril
16	prazosin - urapidil	89	citalopram - escitalopram	162	everolimus - ramipril

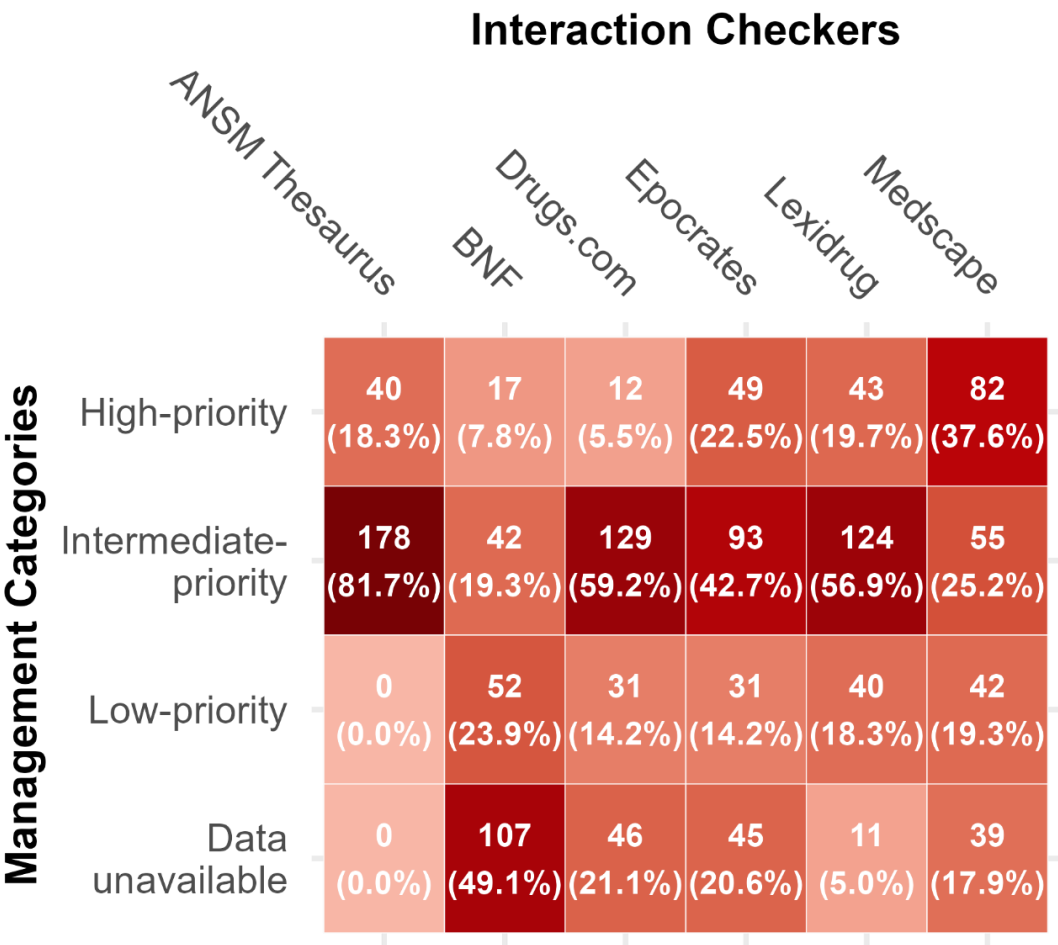
17	silodosin - urapidil	90	carbamazepine - rivaroxaban	163	everolimus - trandolapril
18	tamsulosine - urapidil	91	rifampicin - rivaroxaban	164	amiodarone - ibrutinib
19	raltegravir - rifampicin	92	atorvastatin - fenofibrate	165	apixaban - erythromycin
20	amiodarone - flecainide	93	phenobarbital - rivaroxaban	166	everolimus - sitagliptin
21	atorvastatin - darunavir	94	apixaban - itraconazole	167	itraconazole - prednisone
22	repaglinide - trimethoprim	95	disopyramide - sotalol	168	escitalopram - hydroxyzine
23	lithium - perindopril	96	methotrexate - trimethoprim	169	ezetimibe - fenofibrate
24	lithium - ramipril	97	ciclosporin - rosuvastatin	170	rifampicin - sertraline
25	amiodarone - ciclosporin	98	everolimus - posaconazole	171	flecainide - lidocaine
26	oxycodone - rifampicin	99	alfuzosin - doxazosin	172	irbesartan - sitagliptin
27	itraconazole - oxycodone	100	celecoxib - diclofenac	173	losartan - sitagliptin
28	spironolactone - tacrolimus	101	enalapril - irbesartan	174	sitagliptin - telmisartan
29	amiloride - valsartan	102	irbesartan - perindopril	175	sitagliptin - valsartan
30	amiloride - losartan	103	irbesartan - ramipril	176	nalmefene - tramadol
31	celecoxib - rivaroxaban	104	atorvastatin - rifampicin	177	diclofenac - tinzaparin
32	dabigatran - ketoprofen	105	carbamazepine - simvastatin	178	ketoprofen - tinzaparin
33	ibuprofen - rivaroxaban	106	allopurinol - azathioprine	179	escitalopram - sulfamethoxazole
34	ketoprofen - rivaroxaban	107	amiodarone - sotalol	180	saxagliptin - valsartan

35	apixaban - imatinib	108	ciclosporin - colchicine	181	chlorpromazine - sulfamethoxazole
36	apixaban - ibrutinib	109	colchicine - verapamil	182	chlorpromazine - trimethoprim
37	prednisone - voriconazole	110	gefitinib - phenytoin	183	haloperidol - trimethoprim
38	amiloride - irbesartan	111	propranolol - verapamil	184	everolimus - irbesartan
39	amiloride - telmisartan	112	sotalol - verapamil	185	everolimus - losartan
40	telmisartan - triamterene	113	fenofibrate - rosuvastatin	186	everolimus - valsartan
41	amiloride - perindopril	114	nebivolol - verapamil	187	finasteride - rifampicin
42	enoxaparin - ketoprofen	115	bisoprolol - verapamil	188	acetylsalicylic acid - nicorandil
43	carteolol - diltiazem	116	atorvastatin - ritonavir	189	irbesartan - vildagliptin
44	acenocoumarol - capecitabine	117	diltiazem - verapamil	190	mianserin - rifampicin
45	acenocoumarol - diclofenac	118	chlorpromazine - escitalopram	191	telmisartan - vildagliptin
46	amitriptyline - rilmenidine	119	apixaban - celecoxib	192	valsartan - vildagliptin
47	amiodarone - levomepromazine	120	apixaban - diclofenac	193	apixaban - phenobarbital
48	amiodarone - zuclopenthixol	121	apixaban - flurbiprofen	194	daptomycin - simvastatin
49	escitalopram - levomepromazine	122	apixaban - ibuprofen	195	apixaban - primidone
50	amiodarone - tiapride	123	apixaban - ketoprofen	196	acetazolamide - acetylsalicylic acid
51	perindopril - racecadotril	124	apixaban - indometacin	197	rifampicin - simvastatin
52	racecadotril - ramipril	125	amiodarone - amisulpride	198	apixaban - idelalisib

53	hydroxyzine - sulfamethoxazole	126	amiodarone - escitalopram	199	atenolol - diltiazem
54	sacubitril - saxagliptin	127	amisulpride - haloperidol	200	bisoprolol - diltiazem
55	sacubitril - sitagliptin	128	irbesartan - lisinopril	201	posaconazole - tacrolimus
56	finasteride - phenobarbital	129	lisinopril - valsartan	202	ciclosporin - ezetimibe
57	escitalopram - trimethoprim	130	perindopril - valsartan	203	imatinib - warfarin
58	enalapril - perindopril	131	ramipril - valsartan	204	hydroxychloroquine - hydroxyzine
59	enalapril - ramipril	132	trandolapril - valsartan	205	amiodarone - trimethoprim
60	perindopril - ramipril	133	amiodarone - citalopram	206	irbesartan - telmisartan
61	perindopril - trandolapril	134	amiodarone - haloperidol	207	sacubitril - valsartan
62	sacubitril - vildagliptin	135	escitalopram - sotalol	208	amiodarone - cyamemazine
63	sitagliptin - vildagliptin	136	amiodarone - hydroxychloroquine	209	amisulpride - pipotiazine
64	amiodarone - spiramycin	137	amiodarone - hydroxyzine	210	colchicine - pristnamycin
65	escitalopram - spiramycin	138	hydroxychloroquine - sotalol	211	cyamemazine - escitalopram
66	everolimus - racecadotril	139	hydroxyzine - sotalol	212	cyamemazine - hydroxychloroquine
67	hydroxyzine - spiramycin	140	escitalopram - hydroxychloroquine	213	cyamemazine - hydroxyzine
68	irbesartan - racecadotril	141	amiodarone - pentamidine	214	cyamemazine - sotalol
69	racecadotril - sirolimus	142	amiodarone - sulfamethoxazole	215	fluindione - flurbiprofen
70	racecadotril - sitagliptin	143	sotalol - sulfamethoxazole	216	fluindione - imatinib

71	racecadotril - telmisartan	144	sotalol - trimethoprim	217	fluindione - ketoprofen
72	racecadotril - valsartan	145	irbesartan - valsartan	218	fluindione - sulfamethoxazole
73	ramipril - zofenopril	146	lisinopril - perindopril		

Chapter 3 - Supplementary Figure 1 - Distribution of Management Categories Across Interaction Checkers



7.2 Supplements Chapter 4

Chapter 4 - Supplementary File 1 - Operational Guidelines for Gold Standard

Development

To ensure consistency among evaluators, additional standard procedures were defined during a series of expert meetings.

The decisions made concerned the following aspects: standardization and transparency of the process, preservation of the information contained in the ANSM Thesaurus in order to facilitate the diagnosis of issues within a validated tool, and the possibility of developing a gold standard flexible enough to be applied to tools with variable requirements.

1. The annotation format is based on a standardized form developed through RedCap software, allowing for standardized entries.
2. The information to be collected mainly concerns severity levels. Additional information is gathered only when necessary to disambiguate pDDI with multiple severity levels.
3. Experts constructing the gold standard must classify pDDI across all four severity levels (CI, ASDEC, PE, APEC). Although the primary objective remains the assessment of high-severity pDDI (CI or ASDEC according to the ANSM Thesaurus), the gold standard must nevertheless allow subsequent analyses at different levels.
4. For each unique substance pair, the following information must be collected and added to the respective online module (also, Supplementary Figure 1):
 - a) Presence/absence of pDDI in the ANSM Thesaurus
 - *Procedure:* If the pair is reported in the Thesaurus, the evaluator indicates the presence of the pDDI in the corresponding form and proceeds with

classification. If it is not reported, “No” is selected and the evaluator moves on to the next pair.

- *Rationale:* This step provides a binary distinction on whether or not the pair falls within the scope of the Thesaurus. Only pairs identified as pDDI are analyzed in detail.

b) Single severity level

- *Procedure:* If the Thesaurus assigns only one severity level, the evaluator selects one of the four severity levels.

c) Multiple severity levels (multilevel)

- *Procedure:* If the pair is associated with more than one level, the evaluator checks the Multilevel box, records all severity levels indicated in the Thesaurus, and marks the relevant options with justification.
- *Identified causes of multilevel classification:*
 1. Renal impairment
 2. Drug dosage
 3. Route of administration
 4. Therapeutic indication
 5. Multiple classifications for molecules within the same pharmacological class (see point d)
 6. Other unforeseen conditions (see point e)
- *Rationale:* Certain pairs cannot be reduced to a single severity level because the presence or absence of the pDDI depends on clinical or contextual parameters. At this stage, best- or worst-case scenarios are not selected, in order to maintain adaptability for tools with different needs. Moreover, identifying the typical causes allows: (a) more accurate diagnosis of the

circumstances leading to multilevel classifications and potential tool errors, and (b) limitation of free-text fields to exceptional cases only.

d) Multilevel within a specific molecule

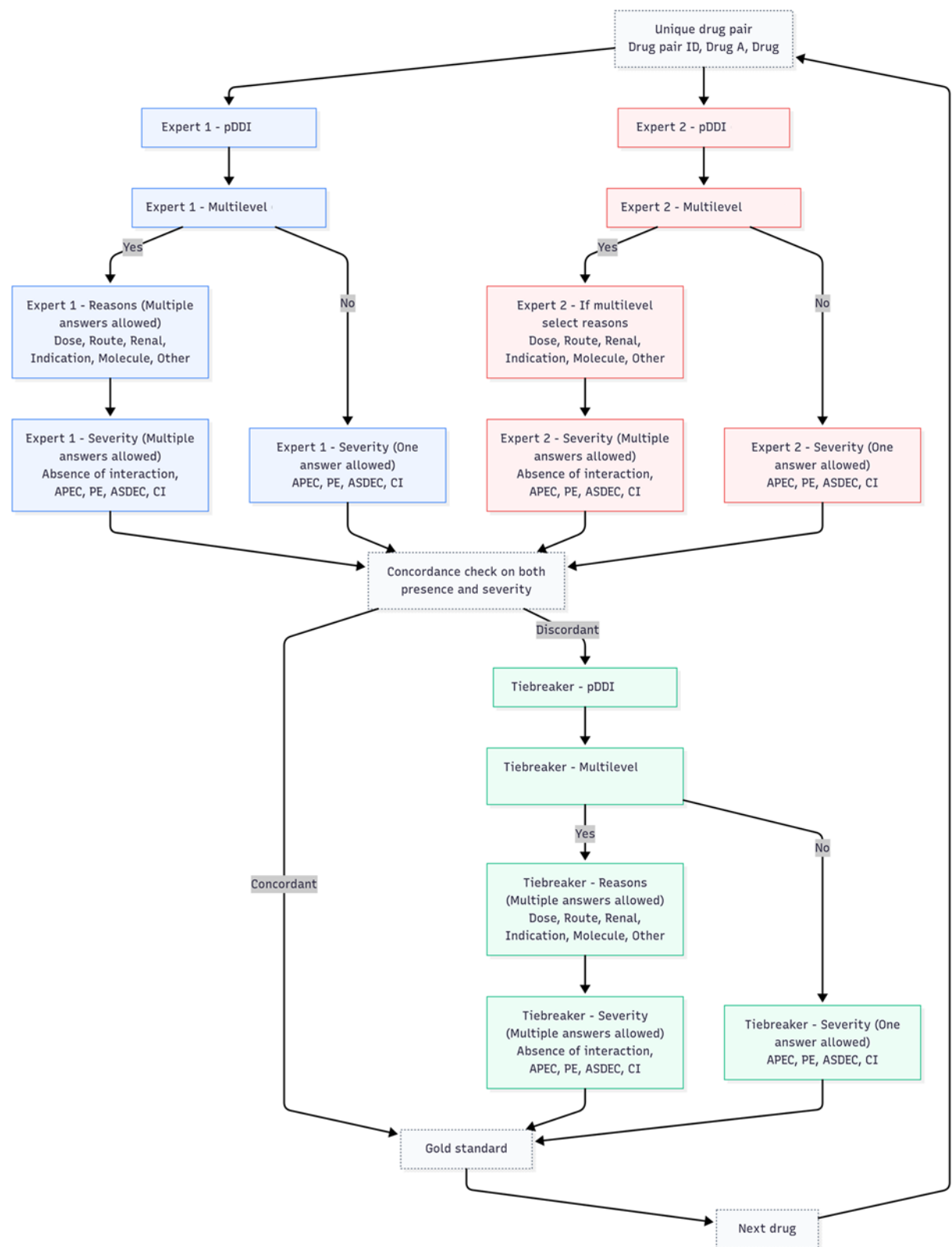
- *Procedure:* If multiple levels concern only one molecule within a pharmacological class, the evaluator still reports the presence of multilevel classification. However, for the specific molecule, only one severity level is selected.
- *Rationale:* This rule preserves information on the presence of multiple severity levels in the molecule's pharmacological class, while still allowing for a unique evaluation where it is genuinely possible.

e) Other unforeseen circumstances

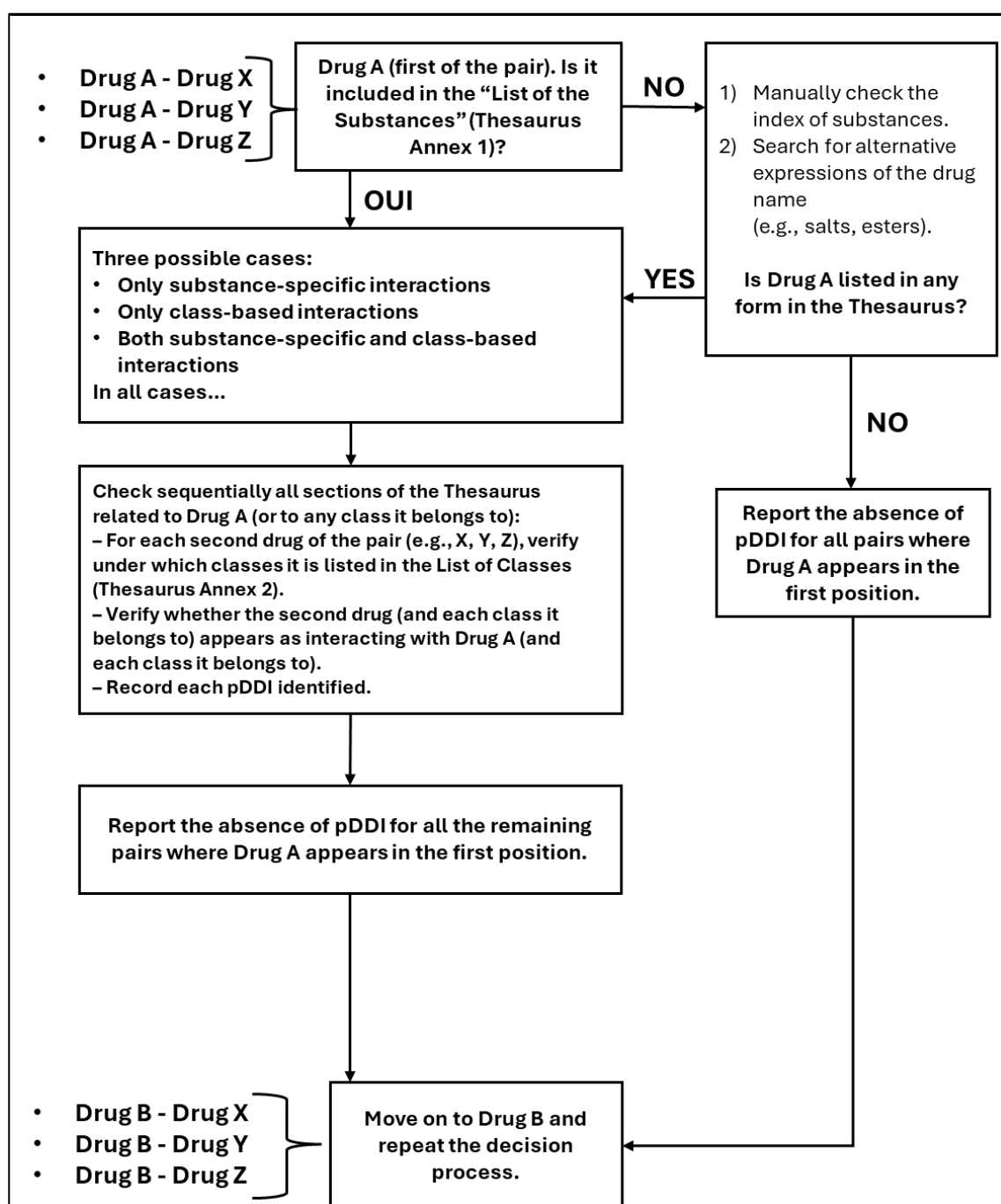
- *Procedure:* If multilevel classification depends on factors not covered by the categories above, the evaluator selects "Other" and provides a detailed explanation in a free-text field.

5. The evaluation of pairs by experts proceeds in alphabetical order and follows a defined decision tree, shown in the Supplementary Figure 2.

Chapter 4 - Supplementary Figure 1 - Annotation workflow for drug pairs and adjudication of discordant cases



Chapter 4 - Supplementary Figure 2. Decision Flowchart - Gold Standard



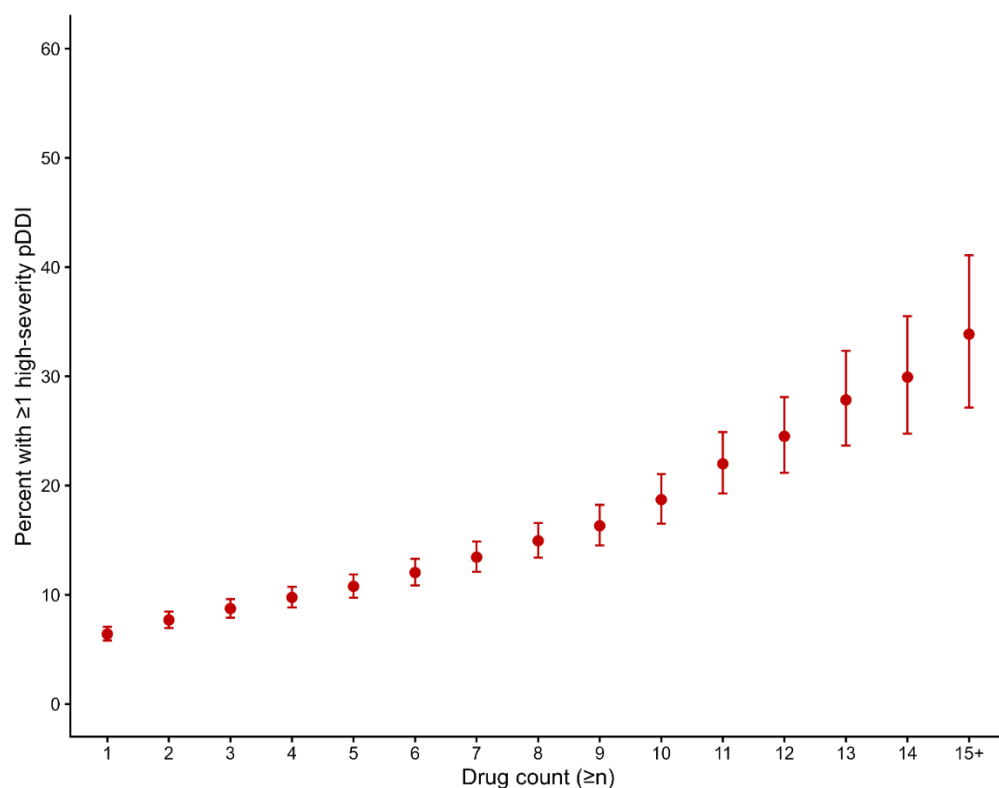
7.3 Supplements Chapter 5

The results from the study presented in Chapter IV were recalculated based on the refined gold standard and are reported below.

High-severity pDDIs

A total of 386 out of 6027 patients presented at least one high-severity pDDI, corresponding to a prevalence of 6.40% (95% CI 5.80-7.05). 16.58% of these patients presented more than one high-severity pDDI (n=64).. Logistic regression showed an association between the number of drugs and the risk of pDDIs: compared to 2-4 drugs, patients with 5-9 drugs had OR = 4.36 (95% CI 2.86-6.66; $p < 0.0001$), while those with ≥ 10 drugs had OR = 14.60 (9.65-22.09; $p < 0.0001$). The proportion of the population exposed to at least one high-severity pDDI, according to the minimum number of drugs taken, is shown in Supplementary

Figure 1



Supplementary Figure 1- Prevalence (with 95% CI) of high-severity pDDIs by cumulative drug thresholds.

The prevalence of patients with at least one high-severity pDDI varied across departments: 168/2073 (8.1%; 95% CI 7.0-9.4) in Internal Medicine, 147/2,066 (7.1%; 95% CI 6.0-8.3) in Cardiothoracic, and 66/1812 (3.7%; 95% CI 2.8-4.6) in Neurosciences. The most frequent high-severity pDDIs identified are reported in Supplementary Table 1.

Supplementary Table 1. Top five most frequent high-severity pDDIs (overall and stratified by department)

Rank	Overall	Cardiothoracic	Internal Medicine	Neurosciences
1	Amiodarone + Trimethoprim- sulfamethoxazole (19)	Ciclosporin + Ezetimibe (11)	Amiodarone + Trimethoprim- sulfamethoxazole (15)	Levodopa- benserazide + Quetiapine (9)
2	Amiodarone + Escitalopram (18)	Amiodarone + Escitalopram (9)	Repaglinide + Trimethoprim- sulfamethoxazole (12)	Duloxetine + Tramadol (6)
3	Levodopa- benserazide + Quetiapine (12)	Bisoprolol + Rilmenidine (5)	Acetylsalicylic acid + Anagrelide (11)	Amiodarone + Escitalopram (3)
4	Methotrexate + Trimethoprim- sulfamethoxazole (12)	Racecadotril + Valsartan/Sacubitril (5)	Methotrexate + Trimethoprim- sulfamethoxazole (11)	Hydroxyzine + Escitalopram (3)
5	Ciclosporin + Ezetimibe (12)	Everolimus + Ramipril (5)	Escitalopram + Trimethoprim- sulfamethoxazole (11)	Paroxetine + Tramadol (3)

Among the 386 patients with a high-severity pDDI, 83 (18.9%) presented one of the five most frequent pDDIs, and 121 (31.3%) one of the ten most frequent.

The drugs most frequently associated with high-severity pDDIs, in terms of absolute number of events, were trimethoprim-sulfamethoxazole (83 patients with ≥ 1 high-severity pDDI), amiodarone (48), escitalopram (38), ciclosporin (24), apixaban (21).

After accounting for exposure prevalence, and restricting the analysis to drugs present in at least 5 patients, the highest rates of high-severity pDDIs were observed for nicorandil (9/10 exposed patients, 90.0%), everolimus (15/20, 75.0%), phenobarbital (5/7, 71.4%), ketoconazole (4/6, 66.7%), anagrelide (12/21, 57.1%), and lithium (4/7, 57.1%). When further restricting to drugs with ≥ 20 exposed patients, the following emerged: everolimus (15/20 exposed patients, 75.0%), anagrelide (12/21, 57.1%), quetiapine (14/27, 51.9%), metformin/sitagliptin (15/30, 50.0%), ciclosporin (24/57, 42.1%), racecadotril (18/48, 37.5%).

8 - REFERENCES

1. Scheife RT, Hines LE, Boyce RD, Chung SP, Momper JD, Sommer CD, et al. Consensus Recommendations for Systematic Evaluation of Drug–Drug Interaction Evidence for Clinical Decision Support. *Drug Saf.* 2015 Feb;38(2):197–206. doi:10.1007/s40264-014-0262-8
2. Inui N, Akamatsu T, Uchida S, Tanaka S, Namiki N, Karayama M, et al. Chronological Effects of Rifampicin Discontinuation on Cytochrome P450 Activity in Healthy Japanese Volunteers, Using the Cocktail Method. *Clin Pharmacol Ther.* 2013 Dec;94(6):702–8. doi:10.1038/clpt.2013.167
3. Skov K, Falskov B, Jensen EA, Dorff MH. Supratherapeutic rivaroxaban levels: A persistent drug-drug interaction after discontinuation of amiodarone. *Basic Clin Pharma Tox.* 2020 Oct;127(4):351–3. doi:10.1111/bcpt.13419
4. Corrie K, Hardman JG. Mechanisms of drug interactions: pharmacodynamics and pharmacokinetics. *Anaesthesia & Intensive Care Medicine.* 2014 Jul;15(7):305–8. doi:10.1016/j.mpaic.2014.04.005
5. Peng Y, Cheng Z, Xie F. Evaluation of Pharmacokinetic Drug-Drug Interactions: A Review of the Mechanisms, In Vitro and In Silico Approaches. *Metabolites.* 2021 Jan 27;11(2):75. doi:10.3390/metabo11020075 PubMed PMID: 33513941; PubMed Central PMCID: PMC7912632.
6. Niu J, Straubinger RM, Mager DE. Pharmacodynamic Drug–Drug Interactions. *Clin Pharma and Therapeutics.* 2019 Jun;105(6):1395–406. doi:10.1002/cpt.1434
7. Sun L, Mi K, Hou Y, Hui T, Zhang L, Tao Y, et al. Pharmacokinetic and Pharmacodynamic Drug-Drug Interactions: Research Methods and Applications. *Metabolites.* 2023 Jul 29;13(8):897. doi:10.3390/metabo13080897 PubMed PMID: 37623842; PubMed Central PMCID: PMC10456269.
8. Hughes S, Heard K, Mughal N, Moore LSP. Burden of enteral supplement interactions with common antimicrobial agents: a single-centre observational analysis. *Eur J Hosp Pharm.* 2022 Sep;29(5):280–3. doi:10.1136/ejhpharm-2020-002445 PubMed PMID: 33414257; PubMed Central PMCID: PMC9660598.
9. Lesko LJ, Offman E, Brew CT, Garza D, Benton W, Mayo MR, et al. Evaluation of the Potential for Drug Interactions With Patiromer in Healthy Volunteers. *J Cardiovasc Pharmacol Ther.* 2017 Sep;22(5):434–46. doi:10.1177/1074248417691135
10. Xu ZY, Li JL. Comparative review of drug–drug interactions with epidermal growth factor receptor tyrosine kinase inhibitors for the treatment of non-small-cell lung cancer. *OTT.* 2019 Jul;Volume 12:5467–84. doi:10.2147/OTT.S194870
11. Akamine Y, Miura M. An update on the clinical pharmacokinetics of fexofenadine enantiomers. *Expert Opinion on Drug Metabolism & Toxicology.* 2018 Apr 3;14(4):429–34. doi:10.1080/17425255.2018.1459565
12. Mar PL, Gopinathannair R, Gengler BE, Chung MK, Perez A, Dukes J, et al. Drug Interactions Affecting Oral Anticoagulant Use. *Circ: Arrhythmia and Electrophysiology.* 2022 Jun;15(6). doi:10.1161/CIRCEP.121.007956

13. Härtter S, Sennewald R, Nehmiz G, Reilly P. Oral bioavailability of dabigatran etexilate (P radaxa[®]) after co-medication with verapamil in healthy subjects. *Brit J Clinical Pharma*. 2013 Apr;75(4):1053–62. doi:10.1111/j.1365-2125.2012.04453.x
14. Allred AJ, Bowen CJ, Park JW, Peng B, Williams DD, Wire MB, et al. Eltrombopag increases plasma rosuvastatin exposure in healthy volunteers. *Br J Clin Pharmacol*. 2011 Aug;72(2):321–9. doi:10.1111/j.1365-2125.2011.03972.x PubMed PMID: 21434975; PubMed Central PMCID: PMC3162661.
15. Bollen PDJ, Prins HAB, Colbers A, Velthoven-Graafland K, Rijnders BJA, De Vries-Sluijs TEMS, et al. The dolutegravir/valproic acid drug–drug interaction is primarily based on protein displacement. *Journal of Antimicrobial Chemotherapy*. 2021 Apr 13;76(5):1273–6. doi:10.1093/jac/dkab021
16. Benet LZ, Hoener B. Changes in plasma protein binding have little clinical relevance. *Clin Pharma and Therapeutics*. 2002 Mar;71(3):115–21. doi:10.1067/mcp.2002.121829
17. Hirota T, Fujita Y, Ieiri I. An updated review of pharmacokinetic drug interactions and pharmacogenetics of statins. *Expert Opinion on Drug Metabolism & Toxicology*. 2020 Sep 1;16(9):809–22. doi:10.1080/17425255.2020.1801634
18. Gay C, Toulet D, Le Corre P. Pharmacokinetic drug-drug interactions of tyrosine kinase inhibitors: A focus on cytochrome P450, transporters, and acid suppression therapy: Pharmacokinetic interactions of TKIs. *Hematological Oncology*. 2017 Sep;35(3):259–80. doi:10.1002/hon.2335
19. Song IH, Zong J, Borland J, Jerva F, Wynne B, Zamek-Gliszczyński MJ, et al. The Effect of Dolutegravir on the Pharmacokinetics of Metformin in Healthy Subjects. *J Acquir Immune Defic Syndr*. 2016 Aug 1;72(4):400–7. doi:10.1097/QAI.0000000000000983 PubMed PMID: 26974526; PubMed Central PMCID: PMC4935531.
20. Simoons M, Naipal KAT, De Jong H, Den Hoed CM, De Winter BCM, Mulder MB. Oral antibiotics lower mycophenolate mofetil drug exposure, possibly by interfering with the enterohepatic recirculation: A case series. *Pharmacology Res & Perspec*. 2023 Jun;11(3):e01103. doi:10.1002/prp2.1103
21. Schneir A, Masom CP. Lithium Toxicity from the Addition of an ACE Inhibitor with an Unexpected Type I Brugada Pattern ECG: Case Files of the Medical Toxicology Fellowship at the University of California, San Diego. *J Med Toxicol*. 2020 Jul;16(3):321–8. doi:10.1007/s13181-020-00773-9 PubMed PMID: 32297151; PubMed Central PMCID: PMC7320113.
22. American College of Medical Toxicology. Guidance Document: Management Priorities in Salicylate Toxicity. *J Med Toxicol*. 2015 Mar;11(1):149–52. doi:10.1007/s13181-013-0362-3
23. Fitzmaurice MG, Wong A, Akerberg H, Avramovska S, Smithburger PL, Buckley MS, et al. Evaluation of Potential Drug–Drug Interactions in Adults in the Intensive Care Unit: A Systematic Review and Meta-Analysis. *Drug Saf*. 2019 Sep 5;42(9):1035–44. doi:10.1007/s40264-019-00829-y

24. Zheng WY, Richardson LC, Li L, Day RO, Westbrook JI, Baysari MT. Drug-drug interactions and their harmful effects in hospitalised patients: a systematic review and meta-analysis. *Eur J Clin Pharmacol*. 2018 Jan;74(1):15–27. doi:10.1007/s00228-017-2357-5
25. Baysari MT, Hilmer SN, Day RO, Van Dort BA, Zheng WY, Quirk R, et al. Effectiveness of computerised alerts to reduce drug–drug interactions (DDIs) and DDI-related harm in hospitalised patients: a quasi-experimental controlled pre–post study. *BMJ Qual Saf*. 2025 Dec;34(12):788–97. doi:10.1136/bmjqs-2024-018243
26. Gonzaga De Andrade Santos TN, Mendonça Da Cruz Macieira G, Cardoso Sodré Alves BM, Onozato T, Cunha Cardoso G, Ferreira Nascimento MT, et al. Prevalence of clinically manifested drug interactions in hospitalized patients: A systematic review and meta-analysis. *Lampe JN, editor. PLoS ONE*. 2020 Jul 1;15(7):e0235353. doi:10.1371/journal.pone.0235353
27. Bechtold B, Clarke J. Multi-factorial pharmacokinetic interactions: unraveling complexities in precision drug therapy. *Expert Opin Drug Metab Toxicol*. 2021 Apr;17(4):397–412. doi:10.1080/17425255.2021.1867105 PubMed PMID: 33339463; PubMed Central PMCID: PMC8110027.
28. Lee J, Beers JL, Geffert RM, Jackson KD. A Review of CYP-Mediated Drug Interactions: Mechanisms and In Vitro Drug-Drug Interaction Assessment. *Biomolecules*. 2024 Jan 12;14(1):99. doi:10.3390/biom14010099 PubMed PMID: 38254699; PubMed Central PMCID: PMC10813492.
29. Hakkola J, Hukkanen J, Turpeinen M, Pelkonen O. Inhibition and induction of CYP enzymes in humans: an update. *Arch Toxicol*. 2020 Nov;94(11):3671–722. doi:10.1007/s00204-020-02936-7 PubMed PMID: 33111191; PubMed Central PMCID: PMC7603454.
30. Galetin A, Hinton L, Burt H, Obach R, Houston J. Maximal Inhibition of Intestinal First-Pass Metabolism as a Pragmatic Indicator of Intestinal Contribution to the Drug-Drug Interactions for CYP3A4 Cleared Drugs. *CDM*. 2007 Oct 1;8(7):685–93. doi:10.2174/138920007782109805
31. Shih CY, Chen CY, Lin HT, Liao YJ, Liang YJ. Oral Bioavailability and Pharmacokinetics of Sildenafil Orally Disintegrating Tablets under Various Gastric pH Levels Following Administration of Omeprazole in Rats. *Life (Basel)*. 2023 Oct 27;13(11):2126. doi:10.3390/life13112126 PubMed PMID: 38004266; PubMed Central PMCID: PMC10671963.
32. Tamargo J, Le Heuzey JY, Mabo P. Narrow therapeutic index drugs: a clinical pharmacological consideration to flecainide. *Eur J Clin Pharmacol*. 2015 May;71(5):549–67. doi:10.1007/s00228-015-1832-0
33. Habet S. Narrow Therapeutic Index drugs: clinical pharmacology perspective. *Journal of Pharmacy and Pharmacology*. 2021 Oct 1;73(10):1285–91. doi:10.1093/jpp/rgab102

34. Gozzo L, Caraci F, Drago F. Bioequivalence, Drugs with Narrow Therapeutic Index and the Phenomenon of Biocreep: A Critical Analysis of the System for Generic Substitution. *Healthcare*. 2022 Jul 26;10(8):1392. doi:10.3390/healthcare10081392
35. Huggins DJ, Sherman W, Tidor B. Rational approaches to improving selectivity in drug design. *J Med Chem*. 2012 Feb 23;55(4):1424–44. doi:10.1021/jm2010332 PubMed PMID: 22239221; PubMed Central PMCID: PMC3285144.
36. Harsh S, Tripathi S, Venuturumilli R, K SP, Reddy GHV, Mukherjee B, et al. Adverse Drug Reactions and Drug Interactions in Multimorbid Patients: A Review of Current Evidence. *Cureus*. 2025 Nov;17(11):e97640. doi:10.7759/cureus.97640 PubMed PMID: 41450405; PubMed Central PMCID: PMC12729518.
37. Maher RL, Hanlon J, Hajjar ER. Clinical consequences of polypharmacy in elderly. *Expert Opinion on Drug Safety*. 2014 Jan;13(1):57–65. doi:10.1517/14740338.2013.827660
38. Guillot J, Maumus-Robert S, Bezin J. Polypharmacy: A general review of definitions, descriptions and determinants. *Therapies*. 2020 Sep;75(5):407–16. doi:10.1016/j.therap.2019.10.001
39. Johnell K, Klarin I. The Relationship between Number of Drugs and Potential Drug-Drug Interactions in the Elderly: A Study of Over 600 000 Elderly Patients from the Swedish Prescribed Drug Register. *Drug Safety*. 2007;30(10):911–8. doi:10.2165/00002018-200730100-00009
40. Bethi Y, Shewade DG, Dutta TK, Gitanjali B. Prevalence and predictors of potential drug–drug interactions in patients of internal medicine wards of a tertiary care hospital in India. *Eur J Hosp Pharm*. 2018 Nov;25(6):317–21. doi:10.1136/ejhpharm-2017-001272
41. Skou ST, Mair FS, Fortin M, Guthrie B, Nunes BP, Miranda JJ, et al. Multimorbidity. *Nat Rev Dis Primers*. 2022 Jul 14;8(1):48. doi:10.1038/s41572-022-00376-4
42. Konstandi M, Johnson EO. Age-related modifications in CYP-dependent drug metabolism: role of stress. *Front Endocrinol (Lausanne)*. 2023;14:1143835. doi:10.3389/fendo.2023.1143835 PubMed PMID: 37293497; PubMed Central PMCID: PMC10244505.
43. Palatini P. Pharmacokinetic drug interactions in liver disease: An update. *WJG*. 2016;22(3):1260. doi:10.3748/wjg.v22.i3.1260
44. Błeszyńska E, Wierucki Ł, Zdrojewski T, Renke M. Pharmacological Interactions in the Elderly. *Medicina*. 2020 Jun 28;56(7):320. doi:10.3390/medicina56070320
45. Nicholson K, Liu W, Fitzpatrick D, Hardacre KA, Roberts S, Salerno J, et al. Prevalence of multimorbidity and polypharmacy among adults and older adults: a systematic review. *The Lancet Healthy Longevity*. 2024 Apr;5(4):e287–96. doi:10.1016/S2666-7568(24)00007-2

46. Drenth-van Maanen AC, Wilting I, Jansen PAF. Prescribing medicines to older people—How to consider the impact of ageing on human organ and body functions. *Brit J Clinical Pharma*. 2020 Oct;86(10):1921–30. doi:10.1111/bcp.14094
47. Ngcobo NN. Influence of Ageing on the Pharmacodynamics and Pharmacokinetics of Chronically Administered Medicines in Geriatric Patients: A Review. *Clin Pharmacokinet*. 2025 Jan 11. doi:10.1007/s40262-024-01466-0
48. Predicting Drug-Drug and Drug-Gene Interactions in a Community Pharmacy Population. *Am J Manag Care*. 2022 Nov 1;28(11):566–71. doi:10.37765/ajmc.2022.89259
49. Gummadi AC, Guddati AK. Genetic Polymorphisms in Pharmaceuticals and Chemotherapy. *World J Oncol*. 2021 Oct;12(5):149–54. doi:10.14740/wjon1405 PubMed PMID: 34804277; PubMed Central PMCID: PMC8577603.
50. Tannenbaum C, Sheehan NL. Understanding and preventing drug–drug and drug–gene interactions. *Expert Review of Clinical Pharmacology*. 2014 Jul;7(4):533–44. doi:10.1586/17512433.2014.910111
51. Zhu L, Brüggemann RJ, Uy J, Colbers A, Hruska MW, Chung E, et al. *CYP2C19* Genotype-Dependent Pharmacokinetic Drug Interaction Between Voriconazole and Ritonavir-Boosted Atazanavir in Healthy Subjects: The Journal of Clinical Pharmacology / Vol XX No X 2016. *The Journal of Clinical Pharmacology*. 2017 Feb;57(2):235–46. doi:10.1002/jcph.798
52. Aksoy N, Ozturk N. A meta-analysis assessing the prevalence of drug–drug interactions among hospitalized patients. *Pharmacoepidemiology and Drug*. 2023 Dec;32(12):1319–30. doi:10.1002/pds.5691
53. Li L, Baker J, Quirk R, Deidun D, Moran M, Salem AA, et al. Drug-Drug Interactions and Actual Harm to Hospitalized Patients: A Multicentre Study Examining the Prevalence Pre- and Post-Electronic Medication System Implementation. *Drug Saf*. 2024 Jun;47(6):557–69. doi:10.1007/s40264-024-01412-w PubMed PMID: 38478349; PubMed Central PMCID: PMC11116265.
54. Dechanont S, Maphanta S, Butthum B, Kongkaew C. Hospital admissions/visits associated with drug–drug interactions: a systematic review and meta-analysis. *Pharmacoepidemiology and Drug*. 2014 May;23(5):489–97. doi:10.1002/pds.3592
55. Letinier L, Pujade I, Duthoit P, Evrard G, Salvo F, Gil-Jardine C, et al. Emergency department admissions induced by drug–drug interactions in the elderly: A cross-sectional study. *Clinical Translational Sci*. 2022 Jun;15(6):1472–81. doi:10.1111/cts.13262
56. Hughes JE, Waldron C, Bennett KE, Cahir C. Prevalence of Drug–Drug Interactions in Older Community-Dwelling Individuals: A Systematic Review and Meta-analysis. *Drugs Aging*. 2023 Feb;40(2):117–34. doi:10.1007/s40266-022-01001-5

57. Díez R, Cadenas R, Susperregui J, Sahagún AM, Fernández N, García JJ, et al. Drug-Related Problems and Polypharmacy in Nursing Home Residents: A Cross-Sectional Study. *IJERPH*. 2022 Apr 4;19(7):4313. doi:10.3390/ijerph19074313
58. Burato S, Leonardi L, Antonazzo IC, Raschi E, Ajolfi C, Baraghini M, et al. Comparing the Prevalence of Polypharmacy and Potential Drug-Drug Interactions in Nursing Homes and in the Community Dwelling Elderly of Emilia Romagna Region. *Front Pharmacol*. 2021 Feb 11;11:624888. doi:10.3389/fphar.2020.624888
59. Russ-Jara AL, Elkhadragey N, Arthur KJ, DiIulio JB, Militello LG, Ifeakor AP, et al. Cognitive task analysis of clinicians' drug–drug interaction management during patient care and implications for alert design. *BMJ Open*. 2023 Dec;13(12):e075512. doi:10.1136/bmjopen-2023-075512
60. Summary of product characteristics (SmPC) Guidelines| European Medicines Agency (EMA) [Internet]. 2013 [cited 2026 Jan 9]. Available from: <https://www.ema.europa.eu/en/human-regulatory-overview/marketing-authorisation/product-information-requirements/how-prepare-review-summary-product-characteristics>
61. Raynor DK, De Veene P, Bryant D. The Effectiveness of the Summary of Product Characteristics (SmPC) and Recommendations for Improvement. *Ther Innov Regul Sci*. 2014 Mar;48(2):255–65. doi:10.1177/2168479013501311
62. Drelich E, Religioni U, Chung K, Kaźmierczak J, Blicharska E, Neumann-Podczaska A, et al. The Quality and Reliability of Information in the Summaries of Product Characteristics. *Int J Environ Res Public Health*. 2022 Feb 15;19(4):2185. doi:10.3390/ijerph19042185 PubMed PMID: 35206371; PubMed Central PMCID: PMC8872284.
63. Vromans L, Doyle G, Petak-Opel S, Rödiger A, Röttgermann M, Schlüssel E, et al. Shaping medicinal product information: a before and after study exploring physicians' perspectives on the summary of product characteristics. *BMJ Open*. 2013 Aug 21;3(8):e003033. doi:10.1136/bmjopen-2013-003033 PubMed PMID: 23965932; PubMed Central PMCID: PMC3753505.
64. Thoenes A, Cariolato L, Spierings J, Pinçon A. Discrepancies Between the Labels of Originator and Generic Pharmaceutical Products: Implications for Patient Safety. *Drugs - Real World Outcomes*. 2020 Jun;7(2):131–9. doi:10.1007/s40801-020-00187-7
65. Bellerby A, Needham DS. Using the British National Formulary effectively. *Nursing Standard*. 2016 Sep 21;31(4):56–62. doi:10.7748/ns.2016.e10472
66. ANSM [Internet]. [cited 2025 Aug 31]. Thésaurus des interactions médicamenteuses. Available from: <https://ansm.sante.fr/documents/reference/thesaurus-des-interactions-medicamenteuses-1>
67. Moreau F, Simon N, Walther J, Dambrine M, Kosmalski G, Genay S, et al. Does DDI-Predictor Help Pharmacists to Detect Drug-Drug Interactions and Resolve Medication Issues More Effectively? *Metabolites*. 2021 Mar 17;11(3):173. doi:10.3390/metabo11030173

68. Létinier L, Cossin S, Mansiaux Y, Arnaud M, Salvo F, Bezin J, et al. Risk of Drug-Drug Interactions in Out-Hospital Drug Dispensings in France: Results From the DRUG-Drug Interaction Prevalence Study. *Front Pharmacol.* 2019 Mar 22;10:265. doi:10.3389/fphar.2019.00265
69. Kontsioti E, Maskell S, Bensalem A, Dutta B, Pirmohamed M. Similarity and consistency assessment of three major online drug–drug interaction resources. *Brit J Clinical Pharma.* 2022 Sep;88(9):4067–79. doi:10.1111/bcp.15341
70. Roblek T, Vaupotic T, Mrhar A, Lainscak M. Drug-drug interaction software in clinical practice: a systematic review. *Eur J Clin Pharmacol.* 2015 Feb;71(2):131–42. doi:10.1007/s00228-014-1786-7
71. Ng JY, Munford V, Thakar H. Web-based online resources about adverse interactions or side effects associated with complementary and alternative medicine: a systematic review, summarization and quality assessment. *BMC Med Inform Decis Mak.* 2020 Dec;20(1):290. doi:10.1186/s12911-020-01298-5
72. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ Digit Med.* 2020;3:17. doi:10.1038/s41746-020-0221-y PubMed PMID: 32047862; PubMed Central PMCID: PMC7005290.
73. Holbrook AM, Silva JM, Faruque JAY, Deng J, Schneider T, Jaffer A. Effect of electronic drug-drug interaction alerts on patient and clinician outcomes: a systematic review. *Journal of the American Medical Informatics Association.* 2025 Oct 1;32(10):1617–28. doi:10.1093/jamia/ocaf139
74. Tukukino C, Eriksson AL, Hamdan W, Wallin Hybelius F, Wallerstedt SM. Interaction alerts: A comparison of classifications and recommendations for clinical management between *J anusmed* and three other knowledge resources. *Basic Clin Pharma Tox.* 2023 May;132(5):403–15. doi:10.1111/bcpt.13843
75. Spanakis M, Alexakis K, Ioannou P. Antiretroviral therapy and associated drug interactions with cardiovascular drugs: a critical review. *Front Pharmacol.* 2025 Nov 21;16:1685710. doi:10.3389/fphar.2025.1685710
76. Marcath LA, Xi J, Hoylman EK, Kidwell KM, Kraft SL, Hertz DL. Comparison of Nine Tools for Screening Drug-Drug Interactions of Oral Oncolytics. *JOP.* 2018 Jun;14(6):e368–74. doi:10.1200/JOP.18.00086
77. Patel RI, Beckett RD. Evaluation of resources for analyzing drug interactions. *jmla.* 2017 Jan 5;104(4). doi:10.5195/jmla.2016.142
78. Beckett RD, Stump CD, Dyer MA. Evaluation of drug information resources for drug-ethanol and drug-tobacco interactions. *jmla.* 2019 Jan 4;107(1). doi:10.5195/jmla.2019.549
79. Carollo M, Crisafulli S, Ciccimarra F, Andò G, Diemberger I, Trifirò G. Exploring the level of agreement among different drug-drug interaction checkers: a comparative study

- on direct oral anticoagulants. *Expert Opinion on Drug Metabolism & Toxicology*. 2024 Mar 3;20(3):157–64. doi:10.1080/17425255.2024.2322134
80. Carollo M, Crisafulli S, Selleri M, Piccoli L, L'Abbate L, Trifirò G. Agreement of Different Drug-Drug Interaction Checkers for Proton Pump Inhibitors. *JAMA Netw Open*. 2024 Jul 9;7(7):e2419851. doi:10.1001/jamanetworkopen.2024.19851
 81. Drwiega EN, Badowski ME, Michienzi S. Antiretroviral drug-drug interactions: A comparison of online drug interaction databases. *J Clin Pharm Ther*. 2022 Oct;47(10):1720–4. doi:10.1111/jcpt.13750 PubMed PMID: 36059105; PubMed Central PMCID: PMC9826109.
 82. Bektay MY, Buker Cakir A, Gursu M, Kazancioglu R, Izzettin FV. An Assessment of Different Decision Support Software from the Perspective of Potential Drug-Drug Interactions in Patients with Chronic Kidney Diseases. *Pharmaceuticals (Basel)*. 2024 Apr 28;17(5):562. doi:10.3390/ph17050562 PubMed PMID: 38794132; PubMed Central PMCID: PMC11124126.
 83. 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 Jul 1;24(4):806–12. doi:10.1093/jamia/ocx010 PubMed PMID: 28339701; PubMed Central PMCID: PMC6080681.
 84. Shariff A, Belagodu Sridhar S, Abdullah Basha NF, Bin Taleth Alshemeil SSH, Ahmed Aljallaf Alzaabi NA. Assessing Consistency of Drug-Drug Interaction-Related Information Across Various Drug Information Resources. *Cureus*. 2021 Mar 8. doi:10.7759/cureus.13766
 85. Liu Y, Wang J, Gong H, Li C, Wu J, Xia T, et al. Prevalence and associated factors of drug-drug interactions in elderly outpatients in a tertiary care hospital: a cross-sectional study based on three databases. *Ann Transl Med*. 2023 Jan 15;11(1):17. doi:10.21037/atm-22-5463 PubMed PMID: 36760261; PubMed Central PMCID: PMC9906203.
 86. Yaowaluk T, Tangpanithandee S, Techapichetvanich P, Khemawoot P. Discordance in Drug-Drug Interaction Alerts for Antidotes: Comparative Analysis of Electronic Databases and Interpretive Insights from AI Tools. *Drug Des Devel Ther*. 2025;19:7427–43. doi:10.2147/DDDT.S543827 PubMed PMID: 40896200; PubMed Central PMCID: PMC12399100.
 87. Utecht J, Brochhausen M, Judkins J, Schneider J, Boyce RD. Formalizing Evidence Type Definitions for Drug-Drug Interaction Studies to Improve Evidence Base Curation. *Stud Health Technol Inform*. 2017;245:960–4. PubMed PMID: 29295242; PubMed Central PMCID: PMC5765984.
 88. Hoang L, Boyce RD, Bosch N, Stottlemeyer B, Brochhausen M, Schneider J. Automatically classifying the evidence type of drug-drug interaction research papers as a step toward computer supported evidence curation. *AMIA Annu Symp Proc*. 2020;2020:554–63. PubMed PMID: 33936429; PubMed Central PMCID: PMC8075461.

89. Nguyen T, Liu X, Abuhashem W, Bussing R, Winterstein AG. Quality of Evidence Supporting Major Psychotropic Drug-Drug Interaction Warnings: A Systematic Literature Review. *Pharmacotherapy*. 2020 May;40(5):455–68. doi:10.1002/phar.2382
90. Wang M, Zeraatkar D, Obeda M, Lee M, Garcia C, Nguyen L, et al. Drug–drug interactions with warfarin: A systematic review and meta-analysis. *Brit J Clinical Pharma*. 2021 Nov;87(11):4051–100. doi:10.1111/bcp.14833
91. Chen Z, Liang N, Zhang H, Li H, Yang Y, Zong X, et al. Harnessing the power of clinical decision support systems: challenges and opportunities. *Open Heart*. 2023 Nov;10(2):e002432. doi:10.1136/openhrt-2023-002432
92. Wasylewicz ATM, Van De Burgt BWM, Manten T, Kerskes M, Compagner WN, Korsten EHM, et al. Contextualized Drug–Drug Interaction Management Improves Clinical Utility Compared With Basic Drug–Drug Interaction Management in Hospitalized Patients. *Clin Pharma and Therapeutics*. 2022 Aug;112(2):382–90. doi:10.1002/cpt.2624
93. Wilken M, Hüske-Kraus D, Klausen A, Koch C, Schlauch W, Röhrig R. Alarm Fatigue: Causes and Effects. *Stud Health Technol Inform*. 2017;243:107–11. PubMed PMID: 28883181.
94. Alanazi A, Alalawi W, Aldosari B. An Evaluation of Drug-Drug Interaction Alerts Produced by Clinical Decision Support Systems in a Tertiary Hospital. *Cureus*. 2023 Aug 8. doi:10.7759/cureus.43141
95. Felisberto M, Lima GDS, Celuppi IC, Fantonelli MDS, Zanotto WL, Dias De Oliveira JM, et al. Override rate of drug-drug interaction alerts in clinical decision support systems: A brief systematic review and meta-analysis. *Health Informatics J*. 2024 Apr;30(2):14604582241263242. doi:10.1177/14604582241263242
96. Carspecken CW, Sharek PJ, Longhurst C, Pageler NM. A Clinical Case of Electronic Health Record Drug Alert Fatigue: Consequences for Patient Outcome. *Pediatrics*. 2013 Jun 1;131(6):e1970–3. doi:10.1542/peds.2012-3252
97. Strom BL, Schinnar R, Abera F, Bilker W, Hennessy S, Leonard CE, et al. Unintended Effects of a Computerized Physician Order Entry Nearly Hard-Stop Alert to Prevent a Drug Interaction: A Randomized Controlled Trial. *Arch Intern Med*. 2010 Sep 27;170(17). doi:10.1001/archinternmed.2010.324
98. Penm J, Vaillancourt R, Pouliot A. Defining and identifying concepts of medication reconciliation: An international pharmacy perspective. *Research in Social and Administrative Pharmacy*. 2019 Jun;15(6):632–40. doi:10.1016/j.sapharm.2018.07.020
99. Lias N, Lindholm T, Holmström AR, Uusitalo M, Kvarnström K, Toivo T, et al. Harmonizing the definition of medication reviews for their collaborative implementation and documentation in electronic patient records: A Delphi consensus study. *Research in Social and Administrative Pharmacy*. 2024 Jun;20(6):52–64. doi:10.1016/j.sapharm.2024.01.016

100. Reeve E, Gnjjidic D, Long J, Hilmer S. A systematic review of the emerging definition of “deprescribing” with network analysis: implications for future research and clinical practice. *Br J Clin Pharmacol*. 2015 Dec;80(6):1254–68. doi:10.1111/bcp.12732 PubMed PMID: 27006985; PubMed Central PMCID: PMC4693477.
101. By the 2023 American Geriatrics Society Beers Criteria® Update Expert Panel. American Geriatrics Society 2023 updated AGS Beers Criteria® for potentially inappropriate medication use in older adults. *J American Geriatrics Society*. 2023 Jul;71(7):2052–81. doi:10.1111/jgs.18372
102. O’Mahony D, Cherubini A, Guiteras AR, Denkinger M, Beuscart JB, Onder G, et al. STOPP/START criteria for potentially inappropriate prescribing in older people: version 3. *Eur Geriatr Med*. 2023 May 31;14(4):625–32. doi:10.1007/s41999-023-00777-y
103. Duncan P, Cabral C, McCahon D, Guthrie B, Ridd MJ. Efficiency versus thoroughness in medication review: a qualitative interview study in UK primary care. *Br J Gen Pract*. 2019 Mar;69(680):e190–8. doi:10.3399/bjgp19X701321 PubMed PMID: 30745357; PubMed Central PMCID: PMC6400610.
104. Abuzour AS, Wilson SA, Woodall AA, Mair FS, Clegg A, Shantsila E, et al. A qualitative exploration of barriers to efficient and effective structured medication reviews in primary care: Findings from the DynAIRx study. Bennett K, editor. *PLoS ONE*. 2024 Aug 30;19(8):e0299770. doi:10.1371/journal.pone.0299770
105. Beeler PE, Eschmann E, Rosen C, Blaser J. Use of an On-demand Drug–Drug Interaction Checker by Prescribers and Consultants: A Retrospective Analysis in a Swiss Teaching Hospital. *Drug Saf*. 2013 Jun;36(6):427–34. doi:10.1007/s40264-013-0022-1
106. Carollo M, Crisafulli S, Vitturi G, Besco M, Hinek D, Sartorio A, et al. Clinical impact of medication review and deprescribing in older inpatients: A systematic review and meta-analysis. *J American Geriatrics Society*. 2024 Oct;72(10):3219–38. doi:10.1111/jgs.19035
107. Tilson H, Hines LE, McEvoy G, Weinstein DM, Hansten PD, Matuszewski K, et al. Recommendations for selecting drug–drug interactions for clinical decision support. *American Journal of Health-System Pharmacy*. 2016 Apr 15;73(8):576–85. doi:10.2146/ajhp150565
108. Payne TH, Hines LE, Chan RC, Hartman S, Kapusnik-Uner J, Russ AL, et al. Recommendations to improve the usability of drug-drug interaction clinical decision support alerts. *Journal of the American Medical Informatics Association*. 2015 Nov 1;22(6):1243–50. doi:10.1093/jamia/ocv011
109. Brozek JL, Canelo-Aybar C, Akl EA, Bowen JM, Bucher J, Chiu WA, et al. GRADE Guidelines 30: the GRADE approach to assessing the certainty of modeled evidence—An overview in the context of health decision-making. *Journal of Clinical Epidemiology*. 2021 Jan;129:138–50. doi:10.1016/j.jclinepi.2020.09.018
110. Hatton RC, Rosenberg AF, Morris CT, McKelvey RP, Lewis JR. Evaluation of Contraindicated Drug-Drug Interaction Alerts in a Hospital Setting. *Ann Pharmacother*. 2011 Mar;45(3):297–308. doi:10.1345/aph.1P533

111. Bhakta SB, Colavecchia AC, Haines L, Varkey D, Garey KW. A systematic approach to optimize electronic health record medication alerts in a health system. *American Journal of Health-System Pharmacy*. 2019 Apr 8;76(8):530–6. doi:10.1093/ajhp/zxz012
112. Horn JR, Hansten PD, Osborn JD, Wareham P, Somani S. Customizing clinical decision support to prevent excessive drug–drug interaction alerts. *American Journal of Health-System Pharmacy*. 2011 Apr 15;68(8):662–4. doi:10.2146/ajhp100465
113. Parke C, Santiago E, Zussy B, Klipa D. Reduction of clinical support warnings through recategorization of severity levels. *American Journal of Health-System Pharmacy*. 2015 Jan 15;72(2):144–8. doi:10.2146/ajhp140095
114. Ledger TS, Brooke-Cowden K, Coiera E. Post-implementation optimization of medication alerts in hospital computerized provider order entry systems: a scoping review. *Journal of the American Medical Informatics Association*. 2023 Nov 17;30(12):2064–71. doi:10.1093/jamia/ocad193
115. Eppenga WL, Derijks HJ, Conemans JMH, Hermens WAJJ, Wensing M, De Smet PAGM. Comparison of a basic and an advanced pharmacotherapy-related clinical decision support system in a hospital care setting in the Netherlands. *J Am Med Inform Assoc*. 2012 Jan;19(1):66–71. doi:10.1136/amiajnl-2011-000360
116. Khurana D, Koli A, Khatter K, Singh S. Natural language processing: state of the art, current trends and challenges. *Multimed Tools Appl*. 2023;82(3):3713–44. doi:10.1007/s11042-022-13428-4 PubMed PMID: 35855771; PubMed Central PMCID: PMC9281254.
117. Jouffroy J, Feldman SF, Lerner I, Rance B, Burgun A, Neuraz A. Hybrid Deep Learning for Medication-Related Information Extraction From Clinical Texts in French: MedExt Algorithm Development Study. *JMIR Med Inform*. 2021 Mar 16;9(3):e17934. doi:10.2196/17934
118. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Wren J, editor. Bioinformatics*. 2020 Feb 15;36(4):1234–40. doi:10.1093/bioinformatics/btz682
119. Demner-Fushman D, Fung KW, Do P, Boyce RD, Goodwin TR. Overview of the TAC 2018 Drug-Drug Interaction Extraction from Drug Labels Track.
120. Graafsma J, Murphy RM, Van De Garde EMW, Karapinar-Çarkit F, Derijks HJ, Hoge RHL, et al. The use of artificial intelligence to optimize medication alerts generated by clinical decision support systems: a scoping review. *Journal of the American Medical Informatics Association*. 2024 May 20;31(6):1411–22. doi:10.1093/jamia/ocae076
121. Poly TN, Islam MdM, Muhtar MS, Yang HC, Nguyen PA (Alex), Li YC (Jack). Machine Learning Approach to Reduce Alert Fatigue Using a Disease Medication–Related Clinical Decision Support System: Model Development and Validation. *JMIR Med Inform*. 2020 Nov 19;8(11):e19489. doi:10.2196/19489

122. Corny J, Rajkumar A, Martin O, Dode X, Lajonchère JP, Billuart O, et al. A machine learning–based clinical decision support system to identify prescriptions with a high risk of medication error. *Journal of the American Medical Informatics Association*. 2020 Nov 1;27(11):1688–94. doi:10.1093/jamia/ocaa154
123. Gisladdottir U, Zietz M, Kivelson S, Tanaka Y, Sirdeshmukh G, Brown KL, et al. Leveraging Large Language Models in Extracting Drug Safety Information from Prescription Drug Labels. *Drug Saf*. 2026 Feb;49(2):177–93. doi:10.1007/s40264-025-01594-x
124. Salvo F, Crupi L, Cholle C. Language models in pharmacovigilance: Applications, promises and limits. *Therapies*. 2026 Jan;S0040595726000168. doi:10.1016/j.therap.2025.12.003
125. Deljehier T, Pariente A, Miremont-Salamé G, Haramburu F, Nguyen L, Rubin S, et al. Rhabdomyolysis after co-administration of a statin and fusidic acid: an analysis of the literature and of the WHO database of adverse drug reactions. *Brit J Clinical Pharma*. 2018 May;84(5):1057–63. doi:10.1111/bcp.13515
126. Noguchi Y, Tachi T, Teramachi H. Review of Statistical Methodologies for Detecting Drug–Drug Interactions Using Spontaneous Reporting Systems. *Front Pharmacol*. 2019 Nov 8;10:1319. doi:10.3389/fphar.2019.01319
127. Gravel CA, Krewski D, Mattison DR, Momoli F, Douros A. Concomitant use of statins and sodium-glucose co-transporter 2 inhibitors and the risk of myotoxicity reporting: A disproportionality analysis. *Brit J Clinical Pharma*. 2023 Aug;89(8):2430–45. doi:10.1111/bcp.15711
128. Alkabbani W, Pelletier R, Beazely MA, Labib Y, Quan B, Gamble JM. Drug–Drug Interaction of the Sodium Glucose Co-Transporter 2 Inhibitors with Statins and Myopathy: A Disproportionality Analysis Using Adverse Events Reporting Data. *Drug Saf*. 2022 Mar;45(3):287–95. doi:10.1007/s40264-022-01166-3
129. Raschi E, Poluzzi E, Salvo F, Pariente A, De Ponti F, Marchesini G, et al. Pharmacovigilance of sodium-glucose co-transporter-2 inhibitors: What a clinician should know on disproportionality analysis of spontaneous reporting systems. *Nutrition, Metabolism and Cardiovascular Diseases*. 2018 Jun;28(6):533–42. doi:10.1016/j.numecd.2018.02.014
130. Raschi E, Poluzzi E, De Ponti F. Spotlight commentary: The value of spontaneous reporting systems to detect (the lack of) clinically relevant drug–drug interactions in clinical practice. *Brit J Clinical Pharma*. 2023 Aug;89(8):2365–8. doi:10.1111/bcp.15780
131. Davis SE, Zabotka L, Desai RJ, Wang SV, Maro JC, Coughlin K, et al. Use of Electronic Health Record Data for Drug Safety Signal Identification: A Scoping Review. *Drug Saf*. 2023 Aug;46(8):725–42. doi:10.1007/s40264-023-01325-0
132. Girardeau Y, Trivin C, Durieux P, Le Beller C, Louet Agnes LL, Neuraz A, et al. Detection of Drug–Drug Interactions Inducing Acute Kidney Injury by Electronic

- Health Records Mining. *Drug Saf.* 2015 Sep;38(9):799–809. doi:10.1007/s40264-015-0311-y
133. Kaas-Hansen BS, Placido D, Rodríguez CL, Thorsen-Meyer H, Gentile S, Nielsen AP, et al. Language-agnostic pharmacovigilant text mining to elicit side effects from clinical notes and hospital medication records. *Basic Clin Pharma Tox.* 2022 Oct;131(4):282–93. doi:10.1111/bcpt.13773
 134. Chen C, Pham Nguyen TP, Hughes JE, Hennessy S, Leonard CE, Miano TA, et al. Evaluation of Drug–Drug Interactions in Pharmacoepidemiologic Research. *Pharmacoepidemiology and Drug.* 2025 Jan;34(1):e70088. doi:10.1002/pds.70088
 135. Derks M, Abt M, Phelan M. Lack of clinically relevant drug–drug interactions when dalcetrapib is co-administered with ezetimibe. *Brit J Clinical Pharma.* 2010 Dec;70(6):825–33. doi:10.1111/j.1365-2125.2010.03763.x
 136. Egerer G, Eisenlohr K, Gronkowski M, Burhenne J, Riedel K, Mikus G. The NK₁ receptor antagonist aprepitant does not alter the pharmacokinetics of high-dose melphalan chemotherapy in patients with multiple myeloma. *Brit J Clinical Pharma.* 2010 Dec;70(6):903–7. doi:10.1111/j.1365-2125.2010.03792.x
 137. Schoedel KA, Pope LE, Sellers EM. Randomized Open-Label Drug-Drug Interaction Trial of Dextromethorphan/Quinidine and Paroxetine in Healthy Volunteers: Clinical Drug Investigation. 2012 Mar;32(3):157–69. doi:10.2165/11599870-000000000-00000
 138. Merali S, Sychterz C, Perera V, Gaohua L, Florea V, Murthy B. Drug–Drug Interaction Potential of Mavacamten with Midazolam: Combined Results from Clinical and Model-Based Studies. *The Journal of Clinical Pharma.* 2025 May;65(5):598–606. doi:10.1002/jcph.6175
 139. Aksoy N, Ozturk N. A meta-analysis assessing the prevalence of drug–drug interactions among hospitalized patients. *Pharmacoepidemiology and Drug.* 2023 Dec;32(12):1319–30. doi:10.1002/pds.5691
 140. Pirnejad H, Amiri P, Niazkhani Z, Shiva A, Makhdoomi K, Abkhiz S, et al. Preventing potential drug-drug interactions through alerting decision support systems: A clinical context based methodology. *International Journal of Medical Informatics.* 2019 Jul;127:18–26. doi:10.1016/j.ijmedinf.2019.04.006
 141. Bagri H, Dahri K, Legal M. Hospital Pharmacists’ Perceptions and Decision-Making Related to Drug-Drug Interactions. *CJHP.* 2019 Aug 21;72(4). doi:10.4212/cjhp.v72i4.2915
 142. Bischof T, Al Jalali V, Zeitlinger M, Jorda A, Hana M, Singeorzan K, et al. Chat GPT vs. Clinical Decision Support Systems in the Analysis of Drug–Drug Interactions. *Clin Pharma and Therapeutics.* 2025 Apr;117(4):1142–7. doi:10.1002/cpt.3585
 143. Monteith S, Glenn T. Comparison of potential psychiatric drug interactions in six drug interaction database programs: A replication study after 2 years of updates. *Human Psychopharmacology.* 2021 Nov;36(6):e2802. doi:10.1002/hup.2802

144. Acton EK, Willis AW, Gelfand MA, Kasner SE. Poor concordance among drug compendia for proposed interactions between enzyme-inducing antiepileptic drugs and direct oral anticoagulants. *Pharmacoepidemiology and Drug*. 2019 Nov;28(11):1534–8. doi:10.1002/pds.4896
145. Günay A, Demirpolat E, Ünal A, Aycan MB. A comparison of four drug–drug interaction databases for patients undergoing haematopoietic stem cell transplantation. *Clinical Pharmacy Therapeu*. 2022 Oct;47(10):1711–9. doi:10.1111/jcpt.13728
146. Agence nationale de sécurité du médicament et des produits de santé (ANSM). Thesaurus des interactions médicamenteuses. <https://ansm.sante.fr/documents/reference/thesaurus-des-interactions-medicamenteuses-1>. Accessed 18 January 2025.
147. Medscape. About Medscape. <https://www.medscape.com/public/about>. Accessed 25 January 2025.
148. Joint Formulary Committee. BNF 85 (British National Formulary) March 2023. London: BMJ Publishing and the Royal Pharmaceutical Society; 2023.
149. Epocrates. <https://www.epocrates.com/about>. Accessed 20 January 2025.
150. Wolters Kluwer. Facts and Comparisons drug interactions module. <https://www.wolterskluwer.com/en/solutions/uptodate/resources/facts-comparisons/drug-interactions>. Accessed 31 January 2025.
151. Drugs.com. Drug interaction checker. https://www.drugs.com/drug_interactions.html. Accessed 25 January 2025.
152. Rambaran KA, Huynh HA, Zhang Z, Robles J. The Gap in Electronic Drug Information Resources: A Systematic Review. *Cureus*. 2018 Jun 22. doi:10.7759/cureus.2860
153. Murphy SN, Mendis ME, Berkowitz DA, Kohane I, Chueh HC. Integration of clinical and genetic data in the i2b2 architecture. *AMIA Annu Symp Proc*. 2006;2006:1040. PubMed PMID: 17238659; PubMed Central PMCID: PMC1839291.
154. Ferté T, Jouhet V, Griffier R, Hejblum BP, Thiébaud R, Bordeaux University Hospital Covid-19 Crisis Task Force, et al. The benefit of augmenting open data with clinical data-warehouse EHR for forecasting SARS-CoV-2 hospitalizations in Bordeaux area, France. *JAMIA Open*. 2022 Oct 4;5(4):ooac086. doi:10.1093/jamiaopen/ooac086
155. Post AR, Ai M, Kalsanka Pai A, Overcash M, Stephens DS. Architecting the Data Loading Process for an i2b2 Research Data Warehouse: Full Reload versus Incremental Updating. *AMIA Annu Symp Proc*. 2017;2017:1411–20. PubMed PMID: 29854210; PubMed Central PMCID: PMC5977612.
156. Ferté T, Cossin S, Schaefferbeke T, Barnette T, Jouhet V, Hejblum BP. Automatic phenotyping of electronic health record: PheVis algorithm. *Journal of Biomedical Informatics*. 2021 May;117:103746. doi:10.1016/j.jbi.2021.103746

157. Cossin Sébastien, Lebrun Luc, Lobre Grégory, Loustau Romain, Jouhet Vianney, Griffier Romain, et al. Romedi: An Open Data Source About French Drugs on the Semantic Web. In: *Studies in Health Technology and Informatics* [Internet]. IOS Press; 2019 [cited 2025 Aug 31]. Available from: <https://www.medra.org/servlet/aliasResolver?alias=iospressISBN&isbn=978-1-64368-002-6&spage=79&doi=10.3233/SHTI190187> doi:10.3233/SHTI190187
158. Nikiema JN, Griffier R, Jouhet V, Mougin F. Aligning an interface terminology to the Logical Observation Identifiers Names and Codes (LOINC®). *JAMIA Open*. 2021 Apr 19;4(2):ooab035. doi:10.1093/jamiaopen/ooab035
159. Wu Y, Jiang M, Xu J, Zhi D, Xu H. Clinical Named Entity Recognition Using Deep Learning Models. *AMIA Annu Symp Proc*. 2017;2017:1812–9. PubMed PMID: 29854252; PubMed Central PMCID: PMC5977567.
160. Létinier L, Bezin J, Jarne A, Pariente A. Drug–Drug Interactions and the Risk of Emergency Hospitalizations: A Nationwide Population-Based Study. *Drug Saf*. 2023 May;46(5):449–56. doi:10.1007/s40264-023-01283-7
161. Gwet K. *Handbook of inter-rater reliability: the definitive guide to measuring the extent of agreement among raters*. 4th ed. Gaithersburg (MD): Advanced Analytics, LLC; 2014.
162. Bryant AD, Fletcher GS, Payne TH. Drug interaction alert override rates in the Meaningful Use era: No evidence of progress. *Appl Clin Inform*. 2014;05(03):802–13. doi:10.4338/ACI-2013-12-RA-0103
163. Vitry AI. Comparative assessment of four drug interaction compendia. *Brit J Clinical Pharma*. 2007 Jun;63(6):709–14. doi:10.1111/j.1365-2125.2006.02809.x
164. Moreira PR, De Farias LT, Feitosa AR, Silva LT, Ferreira TXAM, Provin MP, et al. Concordance analysis of two databases to search for potential drug interactions in onco-hematologic patients. *J Oncol Pharm Pract*. 2025 Feb;31(1):90–7. doi:10.1177/10781552231225187
165. Bénard-Larivière A, Miremont-Salamé G, Pérault-Pochat M, Noize P, Haramburu F, the EMIR Study Group on behalf of the French network of pharmacovigilance centres. Incidence of hospital admissions due to adverse drug reactions in France: the EMIR study. *Fundamental Clinical Pharma*. 2015 Feb;29(1):106–11. doi:10.1111/fcp.12088
166. Rasool MF, Rehman AU, Khan I, Latif M, Ahmad I, Shakeel S, et al. Assessment of risk factors associated with potential drug-drug interactions among patients suffering from chronic disorders. De Vincentis A, editor. *PLoS ONE*. 2023 Jan 24;18(1):e0276277. doi:10.1371/journal.pone.0276277
167. Delara M, Murray L, Jafari B, Bahji A, Goodarzi Z, Kirkham J, et al. Prevalence and factors associated with polypharmacy: a systematic review and meta-analysis. *BMC Geriatr*. 2022 Jul 19;22(1):601. doi:10.1186/s12877-022-03279-x
168. Becker ML, Kallewaard M, Caspers PW, Visser LE, Leufkens HG, Stricker BHc. Hospitalisations and emergency department visits due to drug–drug interactions: a

- literature review. *Pharmacoepidemiology and Drug*. 2007 Jun;16(6):641–51. doi:10.1002/pds.1351
169. Olivier P, Bertrand L, Tubery M, Lauque D, Montastruc JL, Lapeyre-Mestre M. Hospitalizations because of Adverse Drug Reactions in Elderly Patients Admitted through the Emergency Department: A Prospective Survey. *Drugs & Aging*. 2009 Jun;26(6):475–82. doi:10.2165/00002512-200926060-00004
 170. Van Der Sijs H, Aarts J, Vulto A, Berg M. Overriding of Drug Safety Alerts in Computerized Physician Order Entry. *Journal of the American Medical Informatics Association*. 2006 Mar 1;13(2):138–47. doi:10.1197/jamia.M1809
 171. DGOS. Ministère du Travail, de la Santé, des Solidarités et des Familles [Internet]. [cited 2025 Feb 6]. La conciliation médicamenteuse : enquête sur son déploiement national. Available from: <https://sante.gouv.fr/soins-et-maladies/qualite-securite-et-pertinence-des-soins/qualite-des-soins/article/la-conciliation-medicamenteuse-enquete-sur-son-deploiement-national>
 172. Herrero-Zazo M, Segura-Bedmar I, Martínez P, Declerck T. The DDI corpus: An annotated corpus with pharmacological substances and drug–drug interactions. *Journal of Biomedical Informatics*. 2013 Oct;46(5):914–20. doi:10.1016/j.jbi.2013.07.011
 173. Hripcsak G. Agreement, the F-Measure, and Reliability in Information Retrieval. *Journal of the American Medical Informatics Association*. 2005 Jan 31;12(3):296–8. doi:10.1197/jamia.M1733
 174. Liang JJ, Tsou CH, Devarakonda MV. Ground Truth Creation for Complex Clinical NLP Tasks - an Iterative Vetting Approach and Lessons Learned. *AMIA Jt Summits Transl Sci Proc*. 2017;2017:203–12. PubMed PMID: 28815130; PubMed Central PMCID: PMC5543376.
 175. Monaghan TF, Rahman SN, Agudelo CW, Wein AJ, Lazar JM, Everaert K, et al. Foundational Statistical Principles in Medical Research: Sensitivity, Specificity, Positive Predictive Value, and Negative Predictive Value. *Medicina (Kaunas)*. 2021 May 16;57(5):503. doi:10.3390/medicina57050503 PubMed PMID: 34065637; PubMed Central PMCID: PMC8156826.
 176. Rainio O, Teuho J, Klén R. Evaluation metrics and statistical tests for machine learning. *Sci Rep*. 2024 Mar 13;14(1):6086. doi:10.1038/s41598-024-56706-x
 177. Létinier L, Cossin S, Mansiaux Y, Arnaud M, Bezin J, Pariente A. Prévalence et description des situations à risque d’interactions médicamenteuses en France à partir des données de l’Assurance maladie. *Revue d’Épidémiologie et de Santé Publique*. 2018 Mar;66:S22. doi:10.1016/j.respe.2018.01.044
 178. Cohen JF, Korevaar DA, Altman DG, Bruns DE, Gatsonis CA, Hooft L, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ Open*. 2016 Nov;6(11):e012799. doi:10.1136/bmjopen-2016-012799
 179. Murphy RM, Dongelmans DA, De Keizer NF, Jongeneel RJ, Koster CH, Jager KJ, et al. Creation of a gold standard Dutch corpus of clinical notes for adverse drug event

- detection: the Dutch ADE corpus. *Lang Resources & Evaluation*. 2025 Sep;59(3):2763–79. doi:10.1007/s10579-025-09832-5
180. Gurulingappa H, Rajput AM, Roberts A, Fluck J, Hofmann-Apitius M, Toldo L. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of Biomedical Informatics*. 2012 Oct;45(5):885–92. doi:10.1016/j.jbi.2012.04.008
 181. Kontsioti E, Maskell S, Dutta B, Pirmohamed M. A reference set of clinically relevant adverse drug-drug interactions. *Sci Data*. 2022 Mar 4;9(1):72. doi:10.1038/s41597-022-01159-y
 182. Wongpakaran N, Wongpakaran T, Wedding D, Gwet KL. A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. *BMC Med Res Methodol*. 2013 Apr 29;13:61. doi:10.1186/1471-2288-13-61 PubMed PMID: 23627889; PubMed Central PMCID: PMC3643869.
 183. Busa G, Burlina A, Damuzzo V, Chiumente M, Palozzo AC. Comorbidity, Polytherapy, and Drug Interactions in a Neurological Context: An Example of a Multidisciplinary Approach to Promote the Rational Use of Drugs. *J Pharm Pract*. 2018 Feb;31(1):58–65. doi:10.1177/0897190017699138 PubMed PMID: 29278992.
 184. Kovačević M, Vezmar Kovačević S, Miljković B, Radovanović S, Stevanović P. The prevalence and preventability of potentially relevant drug-drug interactions in patients admitted for cardiovascular diseases: A cross-sectional study. *Int J Clin Pract*. 2017 Oct;71(10):e13005. doi:10.1111/ijcp.13005
 185. Innes GK, Ogden CL, Crensil V, Concato J, Fakhouri TH. Prescription Medication Use Among Older Adults in the US. *JAMA Intern Med*. 2024 Sep 1;184(9):1121–3. doi:10.1001/jamainternmed.2024.2781 PubMed PMID: 38949837; PubMed Central PMCID: PMC11217884.
 186. Chikatipalli R, Kuttiappan A, Kumar S, R V, Kujur P, N A A, et al. Prescribing trends and rational drug use patterns in cardiovascular patients: A cross-sectional observational study. *Bioinformatics*. 2024;20(11):1582–7. doi:10.6026/9732063002001582 PubMed PMID: 40162464; PubMed Central PMCID: PMC11953553.
 187. Pewsner D, Battaglia M, Minder C, Marx A, Bucher HC, Egger M. Ruling a diagnosis in or out with “SpPin” and “SnNOut”: a note of caution. *BMJ*. 2004 Jul 24;329(7459):209–13. doi:10.1136/bmj.329.7459.209 PubMed PMID: 15271832; PubMed Central PMCID: PMC487735.
 188. Muylle KM, Gentens K, Dupont AG, Cornu P. Evaluation of an optimized context-aware clinical decision support system for drug-drug interaction screening. *International Journal of Medical Informatics*. 2021 Apr;148:104393. doi:10.1016/j.ijmedinf.2021.104393
 189. Walsh T. Fuzzy gold standards: Approaches to handling an imperfect reference standard. *Journal of Dentistry*. 2018 Jul;74:S47–9. doi:10.1016/j.jdent.2018.04.022

190. Reitsma JB, Rutjes AWS, Khan KS, Coomarasamy A, Bossuyt PM. A review of solutions for diagnostic accuracy studies with an imperfect or missing reference standard. *Journal of Clinical Epidemiology*. 2009 Aug;62(8):797–806. doi:10.1016/j.jclinepi.2009.02.005
191. Roberts K, Demner-Fushman D, Tonning JM. Overview of the TAC 2017 Adverse Reaction Extraction from Drug Labels Track. Theory and Applications of Categories [Internet]. 2017 [cited 2026 Mar 4]. Available from: <https://www.semanticscholar.org/paper/Overview-of-the-TAC-2017-Adverse-Reaction-from-Drug-Roberts-Demner-Fushman/5b8a7b11b987ddeb865dbf3aaa7b745a86ea5bf0>
192. Ly T, Pamer C, Dang O, Brajovic S, Haider S, Botsis T, et al. Evaluation of Natural Language Processing (NLP) systems to annotate drug product labeling with MedDRA terminology. *Journal of Biomedical Informatics*. 2018 Jul;83:73–86. doi:10.1016/j.jbi.2018.05.019
193. Rutjes A, Reitsma J, Coomarasamy A, Khan K, Bossuyt P. Evaluation of diagnostic tests when there is no gold standard. A review of methods. *Health Technol Assess*. 2007 Dec;11(50). doi:10.3310/hta11500
194. Crupi L. Development of a Real-World Derived Gold Standard for High-severity Drug-Drug Interactions and Its Application to the Validation of an Automated Detection Tool. Unpublished Manuscript. 2026.
195. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *BMJ*. 2015 Oct 28;h5527. doi:10.1136/bmj.h5527
196. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies. *Ann Intern Med*. 2011 Oct 18;155(8):529–36. doi:10.7326/0003-4819-155-8-201110180-00009
197. Collins J, Huynh M. Estimation of diagnostic test accuracy without full verification: a review of latent class methods. *Statist Med*. 2014 Oct 30;33(24):4141–69. doi:10.1002/sim.6218
198. McAdam AJ. Discrepant analysis: how can we test a test? *J Clin Microbiol*. 2000 Jun;38(6):2027–9. doi:10.1128/JCM.38.6.2027-2029.2000 PubMed PMID: 10834948; PubMed Central PMCID: PMC86720.
199. Karch A, Koch A, Zapf A, Zerr I, Karch A. Partial verification bias and incorporation bias affected accuracy estimates of diagnostic studies for biomarkers that were part of an existing composite gold standard. *Journal of Clinical Epidemiology*. 2016 Oct;78:73–82. doi:10.1016/j.jclinepi.2016.03.022
200. Milea D, Najjar RP, Jiang Z, Ting D, Vasseneix C, Xu X, et al. Artificial Intelligence to Detect Papilledema from Ocular Fundus Photographs. *N Engl J Med*. 2020 Apr 30;382(18):1687–95. doi:10.1056/NEJMoa1917130

201. Cavallo J, Sher A, Chen D, Avondo J, Sammer M. Performance of an adult-trained AI tool for intracranial hemorrhage detection on head CT in children aged 6-17 years. *Pediatr Radiol*. 2026 Jan 31;56(3):649–57. doi:10.1007/s00247-026-06527-z