Drug Safety (2022) 45:549–561
https://doi.org/10.1007/s40264-022-01162-7

ORIGINAL RESEARCH ARTICLE

Automated Drug Coding Using Artificial Intelligence: An Evaluation of WHODrug Koda on Adverse Event Reports

Eva-Lisa Meldau[1] · Shachi Bista[1] · Emma Rofors[1] · Lucie M. Gattepaille[1]

Accepted: 9 February 2022

Abstract

Introduction Coding medicinal products described on adverse event (AE) reports to specific entries in standardised drug dictionaries, such as WHODrug Global, is a time-consuming step in case processing activities despite its potential for automation. Many organisations are already partially automating drug coding using text-processing methods and synonym lists, however addressing challenges such as misspellings, abbreviations or ambiguous trade names requires more advanced methods. WHODrug Koda is a drug coding engine using text-processing algorithms, built-in coding rules and machine learning to code drug verbatims to WHODrug Global.

Objective Our aim was to evaluate the drug coding performance of WHODrug Koda on AE reports from VigiBase, the World Health Organization's global database of individual case safety reports, in terms of level of automation and coding quality.

Methods Koda was evaluated on 4.8 million drug entries from VigiBase. Automation level was computed as the proportion of drug entries automatically coded by Koda and was compared to a simple case-insensitive text-matching algorithm. Coding quality was evaluated in terms of coding accuracy, by comparing Koda's prediction to the WHODrug entries found on the AE reports in VigiBase. To better understand the cases in which Koda's coding results did not match with the WHODrug entries in VigiBase, a manual assessment of 600 samples of disagreeing encodings was performed by two teams of expert drug coders.

Results Compared with a simple direct-match baseline, Koda can increase the automation level from 61% to 89%, while providing high coding quality with an accuracy of 97%.

Conclusions Even though Koda was designed for use in clinical trials, Koda achieves automation level and coding quality for drug coding of AE reports comparable with the performance observed in a previous evaluation of Koda on clinical trial data. Koda can thus help organisations to automate their drug coding of AE reports to a large degree.

1 Introduction

Drug coding to standardised terminologies is a crucial data processing step to enable structuring drug information from various data sources such as Electronic Health Records (EHRs) systems, Electronic Data Capturing systems for clinical trials and spontaneous reporting databases for postmarketing surveillance of drugs [1–3]. In EHRs, mentions of medications can be found in clinical notes or coded directly to a drug dictionary [4]. Being able to accurately associate patients to drug exposures allows the statistical utilisation of the data for epidemiological studies in these real-world data sources [5]. The use of standardised drug

Key Points

WHODrug Koda is one of the first drug coding engines using artificial intelligence.

Originally developed for the use in clinical trials, Koda reaches equally good performance on adverse event reports.

Koda can automatically code large proportions of drugs, including ambiguous drug names, using its internal coding rules and additional information about the drug, such as route, indication and country.

Designed to code only when confident, Koda can identify challenging cases and leave these for manual coding while making helpful suggestions for a large proportion of inputs.

✉ Eva-Lisa Meldau
Eva-Lisa.Meldau@who-umc.org

[1] Uppsala Monitoring Centre, Uppsala, Sweden

terminology in EHR systems can also facilitate the exchange of information between different parts of healthcare [6, 7]. In clinical trials, coding concomitant medications uniformly can be challenging due to multicentric trials across several countries. Coding trade names of local markets into a global drug dictionary allows identification of the active ingredients involved and thus the harmonisation of inclusion and exclusion criteria across all drug markets. In postmarketing surveillance, coding reported medications to standardised terminologies is necessary to link the adverse events (AE) to the appropriate medicinal products and subsequently to identify safety signals for the associated active ingredients across all related products or groups of products sharing similar properties [2, 3].

In the three aforementioned contexts, data about medications may be first entered in the respective systems in free-text form before being handled manually or programmatically in a second standardisation step, where the verbatim descriptions of the medications are coded to a standardised terminology. Even with some degree of automation, this task can be time-consuming [8] and fully automating it is non-trivial. Verbatims might contain abbreviations, misspellings, or ambiguous trade names, or might not match directly to a record in the drug terminology for other reasons, and thus require either trained experts or support of advanced technical tools, either based on synonym dictionaries or on artificial intelligence (AI).

In this paper, we evaluate one such system, WHODrug Koda, an AI-powered drug coding engine that is designed to automatically code drug verbatims to entries in WHODrug Global [2], with the help of optional, user-provided additional drug information. Originally developed for the purpose of coding concomitant drugs in clinical trials, WHODrug Koda could be used in the context of coding drugs reported on the AE reports used in postmarketing surveillance. The purpose of this paper was therefore to evaluate WHODrug Koda in such a context by quantifying the algorithm's effectiveness and accuracy on reported drugs found in VigiBase, the WHO global database of individual case safety reports. In this article, individual case safety reports from VigiBase are referred to as AE reports and WHODrug Koda is referred to as Koda.

2 Background

2.1 Drug Coding

In Natural Language Processing (NLP), in general and especially medical NLP, the task of mapping a concept of interest given in free text (either recognised during Named Entity Recognition [NER] or given in a free-text data field) to a specific, unambiguous entry in a terminology of choice is a challenging task and still an open area of research [4]. During subsequent analysis, the concept may then be represented by the entry in the terminology. This allows grouping semantically identical or related concepts into a single term and limits the number of concepts in the analyses [4, 9]. Most data-mining pipelines based on free text include such a mapping step. For example, numerous studies based on clinical text describe a mining algorithm to extract clinically relevant information and map that information to appropriate terminologies (e.g. medical conditions mapped to SNOMED CT [10–12] or MedDRA [13, 14] codes, temporal expressions mapped to the ISO-TimeML standard [15, 16]).

Drug coding or drug mapping is an example of a standardisation task where the concepts to be mapped are medicinal products and the terminologies used are drug dictionaries such as RxNorm, the Anatomical Therapeutic Chemical (ATC) classification system, or WHODrug Global. Developed by the US National Library of Medicine (NLM), RxNorm has been created to address the need for interoperability across medical information systems. It provides standardised drug codes based on ingredient(s), dose form and strength(s), and links these codes to other drug terminologies, such as the National Drug File Reference Terminology (NDF-RT) or the National Drug Code (NDC) directory used by the US FDA [3]. RxNorm has been used for drug NER [17]. The ATC system is a hierarchical classification system developed and maintained by the WHO Collaborating Centre for Drug Statistics Methodology that classifies active drugs according to the organ or system on which they act, as well as their therapeutic, pharmacological and chemical properties. WHODrug Global [2] is a resource for drug coding in international databases. It has global coverage of trade names from 167 different drug markets and is developed and maintained by the Uppsala Monitoring Centre (UMC), the WHO Collaborating Centre for International Drug Monitoring. WHODrug Global contains information about medicinal products and active substances intended for human and medicinal use, both of conventional and natural origin. The records included in the dictionary are classified as *trade names* (the name under which a medicinal product is marketed), *generic records* (substance information) or *umbrella records*, which describe a drug category (e.g., *Hormonal contraceptives for systemic use* or *Cough and cold preparations*). In the remainder of the paper, we refer to WHODrug Global as WHODrug. All drug names in WHODrug have one or more assigned ATC codes, representing either an official code assigned by the WHO Collaborating Centre for Drug Statistics Methodology or additional UMC-assigned codes. UMC-assigned ATC codes are included due to the specific use cases of WHODrug, when official ATC codes are not sufficient for efficient analysis of medicinal products and their properties.

In postmarketing surveillance and clinical trials, drug coding involves coding the verbatim describing the active substance or product used by the patient to a drug terminology. Additional information, such as the strength, the dosage form or the indication of the product may be used to inform the coding decision [18]. Since the verbatim might be entered manually, it can contain abbreviations, misspellings, and the trade name, with or without the corresponding active ingredients, and might not be formatted consistently, depending on data entry conventions. Drug coding can also be challenging when the trade names used are ambiguous. *Aircort*, for example, is a trade name that is marketed in both Morocco and Italy; however, while it represents a product containing *Beclometasone dipropionate* in Morocco, *Aircort* represents a product containing *Budesonide* in Italy. Even within a single drug market, ambiguous trade names can be found: *Losec* is a trade name in Sweden marketed in different pharmaceutical formulations such as injections and tablets. Depending on the form, different variations of the active ingredient are used, for example the *omeprazole magnesium* variation is used in tablets and the *omeprazole sodium* variation is used in injection formulations. Addressing these challenges, which present varying degrees of complexity, may require the expertise of trained coders and makes drug coding a time-consuming task. In the context of safety case processing, based on a survey of pharmaceutical companies, the drug and AE coding part of the case processing is estimated to take 1–4 min per case [8]. Although drug coding appears to be performed manually to a large degree, in both the postmarketing surveillance and the clinical trials contexts, there have been some recent efforts to automate the task [19, 20].

Tools for drug coding through direct matches of the verbatim, with or without transformation, to drug terminologies can be referred to as *auto-encoders* and the process itself *auto-coding* or *auto-encoding* [2, 18]. Direct matches are coded automatically to the drug terminology of interest [18] and synonym lists are commonly created by organisations during manual coding to record and reapply coding decisions. However, the creation and maintenance of such synonym lists requires considerable amounts of manual work. In order to facilitate some of this work, some drug-coding systems proposed use reordering of tokens or different text-processing steps. Systems developed for RxNorm, for example, identify the different parts of the drug name, i.e., ingredient, product name, strength and form, and standardise to the appropriate format [1, 17, 21]. Neither synonym lists nor text-processing methods can, with certainty, select between ambiguous drug names.

More advanced, AI-based systems may help in dealing with ambiguities. To our knowledge, there are very few such systems in use. One study explored the use of deep neural networks to code non-standardised drug order texts in EHRs to ATC codes [22]. The authors evaluated their drug-coding system as a ranking system and reached a Mean Reciprocal Rank of 98% on their dataset comprising fewer than 1000 drug-order texts from one medical centre. In another study, Abatemarco and colleagues [20] proposed a deep neural network approach to WHODrug coding and other case processing tasks. They trained two different models towards this task, a drug entity recognition model that identifies the drug mentions in the case narrative and a subsequent classifier mapping the detected entities to WHODrug. They reported a top-5 accuracy score of 98% for the mapping task. However, a system that can only reliably predict the correct record in a top-5 list would require a human expert to select the correct record in all cases. Therefore, such a system is only supporting, instead of automating, the task. In contrast to drug coding, the task of drug name extraction or NER of drugs has been more frequently approached using machine-learning methods and, in recent years, deep neural networks have been applied for drug name extraction [20, 23, 24].

2.2 Drug Coding with WHODrug Koda

WHODrug Koda is an automated coding engine custom built by UMC. Koda is one of the first AI-based drug coding systems available for use.[1] The purpose of Koda is to assist coders in interpreting free-text drug information selecting the most appropriate drug name in WHODrug. Koda can scale-up coding capacity and support drug coders in their manual work.

For a given drug entry, Koda does one of the following: (1) select a WHODrug record with high certainty; (2) suggest a set of WHODrug records to choose from; or (3) leave the entry uncoded, in cases requiring human expertise. Inputs to Koda consist of a verbatim and optional fields about the route, indication and country in which the product was obtained. In the case of ambiguous drug names, Koda harnesses the additional information, if provided, to identify the correct WHODrug record. Koda also has the ability to select the most appropriate of the ATC codes assigned to a WHODrug record per intended use, meeting regulatory expectations for coding concomitant medications in clinical trials [26]. However, because ATC selection is usually not part of AE case processing, evaluating Koda for its ATC selection capabilities is out of scope in this study and we evaluated Koda on its WHODrug coding capabilities exclusively.

To select the correct record in WHODrug, Koda uses a combination of text-processing algorithms, built-in coding rules and machine learning. The machine learning module was trained on a dataset of millions of reported drugs in VigiBase, combined with annotations established by a team

[1] A trial web interface is available for WHODrug users [25].

of drug-coding experts. It is retrained by UMC on every new WHODrug version to ensure that Koda stays up to date with new additions and changes in the dictionary.

To ensure that Koda follows the most recent coding guidelines for coding concomitant drugs, it also uses built-in coding rules. The coding rules have been developed in collaboration with a reference group consisting of experienced coders from the industry and are aligned with accepted best practices (such as the WHODrug Best Practices [27]) and regulatory expectations [26]. Three coding rules can be toggled in Koda to enable user-specific coding conventions: the *preferred base* rule, the *generic* rule, and the *country* rule. Koda can fall back on these coding rules in cases where route and indication information cannot disambiguate the drug names. The *preferred base* coding rule allows Koda to select the base substance in cases where the trade name is ambiguous. For example, the trade name *Doxycyclin Al* contains different salt variations of *Doxycycline* depending on the pharmaceutical formulation. In cases where Koda is unable to code to one specific trade name, the preferred base substance *Doxycycline* will be selected. The *generic* rule allows Koda to select the generic record in the dictionary if a trade name is marketed with the same name as a generic record. For example, the generic record *Magnesium* will be selected even though there is a trade name with the same name but different active ingredients. The *country* rule allows Koda to utilise the country information to select the correct trade name in the dictionary.

An independent study performed at Novo Nordisk evaluated Koda for its coding performance for coding concomitant drugs in clinical trials [19]. The authors found that 79% of the concomitant drugs in their dataset could be coded to a WHODrug record with a high certainty by Koda. Koda could additionally make suggestions for 15% of the drugs, leaving 6% of the drugs uncoded. For 96% of the drugs coded with high certainty, Koda's prediction agreed with the existing coding done by Novo Nordisk internal coding practices. During manual evaluation of a sample of 181 disagreeing drug encodings, the assessors found the Koda encoding to be at least as acceptable and precise as the Novo Nordisk encoding for 90% of the drugs. The performance of Koda for drug coding on AE reports is as yet unknown. Evaluation of Koda for drug coding on AE reports is the purpose of this study.

3 Methods

3.1 Evaluation Data

Figure 1 shows the creation of the evaluation dataset. In a first step, we extracted all AE reports first received in VigiBase between 1 January 2020 and 31 December 2020 (included). The deduplicated version of VigiBase for this time period was used, where duplicate removal had been performed using the vigiMatch algorithm [28]. For each reported drug, we extracted the description of the drug—the original, verbatim description of the drug received from the primary source of the report, the indication (when provided), the route of administration (when provided), and the country where the drug was obtained (when available, otherwise inferred as the primary source country of the report) (Table 1). In this study, each instance of verbatim, route, indication and country for a given reported drug is referred to as a Koda input and represents one reported drug in our evaluation dataset.

As Koda does not handle non-Latin characters, we excluded all Koda inputs containing at least two non-Latin characters[2] (see step 2 in Fig. 1). In step 3, we also excluded reported drugs that, at the extraction date, had not been mapped in VigiBase to any WHODrug record (e.g., invalid drug information, drugs awaiting manual coding). The remaining Koda inputs constituted our final evaluation dataset, all written using the Latin alphabet and associated with a valid WHODrug record. Newline characters were removed from all free-text fields. No additional pre-processing was applied. As our evaluation dataset is extracted from VigiBase, each Koda input has a corresponding VigiBase encoding, which is what the drug has been mapped to in VigiBase. This VigiBase encoding is considered a gold-standard label and is used to evaluate the Koda prediction. Importantly, none of the reports included in the evaluation dataset and their associated gold-standard labels were used to train Koda.

3.2 Construction of the Gold Standard

The WHODrug records created by UMC's internal coding processes and stored in VigiBase formed our gold-standard labels and were used to evaluate the correctness of Koda encodings. During UMC's coding processes, reported drugs on AE reports in VigiBase are coded by an automated process that is independent of Koda and consists of directly matching verbatims to WHODrug in combination with text-processing algorithms, as well as a compiled synonym list. When this is not successful, the more challenging cases are manually coded by an expert who investigates the information on the report to make the coding decision. The coding experts follow UMC's internal coding rules, which are based on WHODrug Best Practices [27], to assure high quality and consistency of the coded data. They also feed coding decisions back to the automated process by updating the synonym list to reuse the coding decisions on similar contexts in the future.

[2] We chose to cut-off at two characters because some single characters such as unit specifiers (such as mg [U+338E]) could have been poorly encoded, but the text may nevertheless contain enough information for Koda.

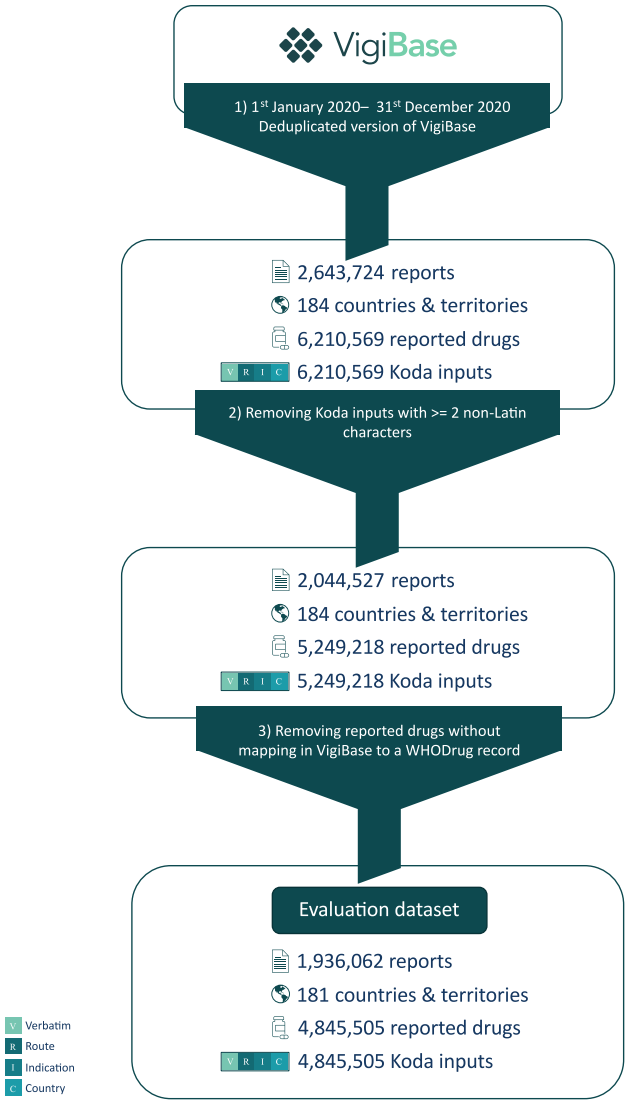

Fig. 1 Steps for the extraction of the evaluation dataset and the general data statistics after each step

A significant proportion of the manual coding step involves resolution of *ambiguous drug names*, also referred to as non-unique trade names. They share the same drug name in WHODrug but include different active ingredients and thus correspond to different records in the dictionary [27]. To differentiate ambiguous drug names in WHODrug, in the B3-format[3] of the dictionary, trade names are appended with an alphabetically ordered list of active ingredients. Ambiguous drug names appear in WHODrug if:

(a) a product is marketed under the same trade name with different active ingredients; *ASTAFEN [KETOTIFEN FUMARATE]* is marketed in Turkey and *ASTAFEN [PARACETAMOL]* is marketed in the Republic of Korea;
(b) a product is marketed under the same trade name with different salt variations of the same moiety, possibly in different pharmaceutical forms; *ACIFRE [OMEPRAZOLE]* includes the preferred base in the capsules form and *ACIFRE [OMEPRAZOLE SODIUM]* includes the salt variation in the vials form;
(c) the active ingredients of a product have been modified without a change to the trade name, resulting in two records reflecting both the old product and the new product; *ACTON [CORTICOTROPIN]* is not marketed any longer; however, a new product *ACTON [PARACETAMOL]* is currently on the market with the same name but different active ingredients;
(d) a product's trade name has the same name as a generic record in WHODrug but different active ingredients; *CALCIUM* is the generic record that represents the substance information in WHODrug; *CALCIUM [ASCORBIC ACID;CALCIUM;COLECALCIFEROL]* is also a trade name in Egypt with a different content;
(e) a trade name is the same as an umbrella term; the trade name *PROBIOTICS [BIFIDOBACTERIUM LONGUM; LACTOBACILLUS ACIDOPHILUS;LACTOBACILLUS RHAMNOSUS]* has the same name as an umbrella collective term *PROBIOTICS [UMBRELLA TERM]*.

[3] WHODrug Global is provided in two formats, the B3-format and the more detailed C3-format [27].

Table 1 Description of Koda input fields

Field name	Field type	Possibly missing	Koda input requirement
Verbatim	Free text	No	Mandatory
Route of administration	Structured (70 values possible)	Yes	Optional
Indication	Free text	Yes	Optional
Country	Structured (251 values[a])	No	Optional

[a]ISO 3166-1 alpha-3 country code or unknown

To measure Koda's performance against ambiguous drug names, we defined a Koda input to reference an ambiguous drug name if the gold-standard WHODrug record fulfilled any of the five above-mentioned criteria in the WHODrug March 2021 release.

3.3 Models

This study used the March 2021 release of Koda, with the *preferred base*, *generic* and *country* rules turned on. It coded drug entries to the March 2021 release of WHODrug (B3-format), selecting records at *high certainty*, as *suggested,* or leaving them *uncoded*.

To give Koda a baseline comparator and to show how much simple automation can be used to solve the task of drug coding on AE reports, we compared Koda's coding results to the encodings created by a case-insensitive text-matching algorithm, where verbatims were searched against all drug names in the March 2021 release of WHODrug in B3-format. This baseline algorithm coded to the record for which the drug name in WHODrug matches the verbatim exactly, disregarding letter case differences. No preprocessing was performed on the verbatim to keep the model as simple as possible. Based on the text-matching approach described above, ambiguous drug names were only matched when the verbatim provided the list of active ingredients in the correct format, or, alternatively, if they corresponded exactly to a *generic record*. This behaviour corresponded to a coding convention allowing ambiguous drug names to be coded to a *generic record* if it exists.

3.4 Evaluation Metrics

Automation level was measured by the proportion of Koda inputs coded by Koda at *high certainty.* The proportion was compared with the baseline model described in the paragraph above, where a successful match to a WHODrug drug name was considered *coded*.

Coding quality indicates how good Koda was at selecting *or* suggesting the correct drug record for a given input. This was measured in terms of accuracy as well as precision, recall and *F*1 score, comparing Koda's prediction against the gold-standard drug encoding present in our evaluation dataset. The accuracy was computed overall and per Koda confidence level (*high certainty*, *suggested* or *uncoded*). For the *high-certainty* cases, we directly compared the drug encoding in the gold standard with the drug encoding predicted by Koda. A positive match was obtained when the gold-standard encoding and Koda encoding were identical, while a mismatch was obtained in any other case. A similar strategy was used to compare Koda's *suggested* encodings to the gold standard. A positive match was obtained when one of the *suggested* encodings was identical to the gold standard at the detail level provided by the B3-format. The coding quality was then computed as the proportion of positive matches between the outputs from Koda and the gold standard. Precision, recall and *F*1-score were computed for all Koda inputs as the macro average across all WHODrug records present in the gold standard and as a weighted average of the metrics per WHODrug record, weighing each record by its prevalence in the gold standard. In all precision, recall and *F*1 computations, any *suggested* or *uncoded* Koda prediction counts as a false negative for the associated gold-standard WHODrug record.

3.5 Manual Evaluation of Mismatches to the Gold Standard

To better understand Koda's coding quality, we manually evaluated cases in which Koda failed to predict the gold-standard encodings. For this, we randomly sampled 200 *high-certainty* mismatches, 200 *suggested* encodings where no suggestion matched the gold standard, and 200 Koda inputs left uncoded by Koda. We consequently categorised each mismatch and summarised the results on Koda's confidence levels. The assessment was done by five drug coding experts divided into two independent teams. Inter-annotator agreement for each confidence level was measured between the two teams using Cohen's kappa statistic to judge the reliability of the assessments and as an indirect measure for the difficulty of the task.

For the *high-certainty* samples, we considered four categories of mismatches: (1) Koda's encoding was more precise; (2) the gold-standard encoding was more precise; (3) both were acceptable; and (4) both were incorrect.

For the *suggested* samples, the categories were (1) the gold-standard encoding was more precise than any of Koda's predictions; (2) at least one Koda suggestion was more precise; (3) both the gold-standard encoding and at least one Koda suggestion were acceptable; and (4) none of the encodings was acceptable.

Finally, for the uncoded samples, we considered the three following categories: (1) the gold-standard encoding was correct; (2) the gold-standard encoding needed manual review based on the available information from the Koda input; or (3) the gold-standard encoding was not correct.

3.6 Effect of Route, Indication and Country Fields

Koda was designed to not only automate simple coding decisions but to also enable the coding of ambiguous drug names. Fields such as indication, route, and country, while optional, are assumed to provide additional context to Koda when the verbatim alone is not enough to resolve these ambiguous drug names. To test this hypothesis, we performed a masking experiment on a sample of Koda inputs in the evaluation dataset. Due to computational limitations, it was not possible to perform this experiment for all Koda inputs from the gold standard that had been coded to an ambiguous drug name. We therefore randomly sampled 2500 Koda inputs from the set of unique Koda inputs where all fields were non-empty and where the gold-standard label was an ambiguous drug name (ambiguous, as defined in Sect. 3.2). We refer to these 2500 unique ambiguous Koda inputs as the Verbatim-Route-Indication-Country (VRIC) dataset.

From the VRIC dataset, five additional synthetic datasets were derived through masking combinations of the optional

fields: (1) the V dataset, retaining only the verbatim field while route, indication and country fields are masked; (2) the VC dataset, where the verbatim and country were kept but the route and indication were masked; (3) the VRC dataset, where only the indication was masked; (4) the VIC dataset, where only the route was masked; and finally (5) the VRI dataset, where only the country field was masked. Together, these five datasets were run through Koda with identical coding rule configuration. Country was treated slightly differently from the route and indication fields because the country field is linked to the country coding rule of Koda. Country information is also typically present on AE reports but not necessarily available in the clinical trial context.

In our analysis, we used the VRIC dataset with all fields present as a reference dataset, which we compared against the masked versions to see how the additional fields affected Koda's automation level and the WHODrug record selection.

4 Results

4.1 Evaluation Data Overview

After step 1 in our dataset preparation as depicted in Fig. 1, we had extracted 2,643,724 reports that had been received in VigiBase in 2020 and after duplicate removal. These covered 184 countries and territories and contained 6,210,569 reported drugs, with an average of two reported drugs per report.

Our final evaluation dataset contains 1,936,062 reports from 181 countries and territories and 4,845,505 Koda inputs, all written using the Latin alphabet and associated with a valid WHODrug record. The 4,845,505 Koda inputs can be grouped into 907,153 *unique* Koda inputs.

The top-five countries represented on the reports, as well as the reporter qualifications, are presented in Tables 2 and 3, respectively. The drugs in the dataset have been coded to 63,010 distinct WHODrug records. The top-five reported WHODrug records and the most common ATC level 2 codes associated with reported drugs are presented in Tables 4 and 5, respectively. Statistics about the number of times a WHODrug record has been reported and the number of unique inputs associated with them is presented in Table 6.

The presence of a route or indication field in the dataset does not imply that they are informative. Of all Koda inputs in the dataset, 46% have an informative indication[4] and 51% have an informative route[5]. Only three Koda inputs in this dataset have country reported as *unknown*.

[4] Indication present and not containing the phrase 'unknown indication'.

[5] Route present and not 'unknown', 'other', or derived from verbatim.

Table 2 Top-5 report countries in the evaluation dataset

Country	%
USA	43
Korea	12
Germany	6
UK	6
France	4
Rest of the world	29

Table 3 Top-5 report qualification in the evaluation dataset

Reporter qualification	%
Consumer	39
Physician	24
Other health professionals	19
Pharmacist	13
Lawyer	3
Unknown	2

Table 4 Top-5 reported drugs in the evaluation dataset

Drug	%
Other anti-acne preparations for topical use [umbrella term]	2.0
Adapalene	1.2
Revlimid	1.0
Humira	1.0
Zantac	0.9

4.2 Automation Level and Coding Quality

On the evaluation dataset, Koda shows an automation level of 89%. This is significantly higher than our direct-match baseline, which is automatically coding 61% of the drug entries in this dataset. For an additional 6% of inputs, Koda makes one or more suggestions, while leaving 5% of inputs uncoded (Fig. 2a). Figure 2b shows the coding quality of Koda's *high-certainty* encodings. Koda codes at high certainty with an accuracy of 97%; this means that for 97% of the cases, Koda's *high-certainty* predictions agree with the gold-standard encoding. Similarly, in Koda's *suggested* predictions, 76% of the *suggested* encodings contain the gold-standard WHODrug record (Fig. 2c).

Manual inspection of a sample of inputs showed that Koda's coding quality is affected by verbatims that contain non-trivial misspellings, and verbatims with much additional information such as unrelated text, or strength without units. Koda also struggles on verbatims describing combination products—products containing more than one ingredient—that are listed in various formats. Collective terms such as *vitamins* or *antibiotics*, which should often be coded to an umbrella term, also seem to be particularly challenging for

Table 5 Top-5 reported ATC codes in the evaluation dataset

ATC level 2	%
V91 Homeopathic preparation	1.2
V90 Unspecified herbal and traditional medicine	0.5
R01 Nasal preparations	0.5
D03 Preparations for treatment of wounds and ulcers	0.2
N05 Psycholeptics	0.1

ATC Anatomical Therapeutic Chemical

Koda to code. In a few cases, Koda's performance was negatively affected by the indication information, mainly when the reported indication was rare or formulated in an unusual way, and therefore very different from the training dataset.

We observed that 61% of the Koda inputs for which Koda provided suggestions had been coded manually in the gold standard. Similarly, 98% of the inputs Koda left *uncoded* had been coded manually in the gold standard. Many of the uncoded Koda inputs were identical or referred to a very specific brand of over-the-counter products, showing that only very few unique Koda inputs are left fully uncoded by Koda. Even when coding as *suggested*, Koda suggests only a single WHODrug record for most inputs (Fig. 3), and, for more than 99% of inputs, Koda suggests six or fewer WHODrug records (Table 7).

The evaluation result of Koda's coding quality across all Koda inputs, including those that Koda chooses to leave as *suggested* encodings or *uncoded* requiring human input, is presented in Table 8. Compared with the direct-match baseline with an accuracy of 60.4%, Koda increases accuracy to 86.0%. We find that Koda reaches a macro average *F*1 score of 87.6% and weighted average *F*1 score of 88.2%, which is significantly higher than the direct-match baseline (64.4% and 66.1%, respectively).

4.3 Mismatches to the Gold Standard

Manual evaluation of the 200 sampled Koda predictions coded at *high certainty* that did not match the gold standard showed that Koda's predictions were as good as or more precise than the gold standard in over 90% of the samples (Fig. 4), with an inter-annotator score of 0.83 (Cohen's

Table 6 Descriptive statistics of WHODrug records in the evaluation dataset

	Number of unique Koda inputs per WHODrug record	Number of unique verbatims reported per WHODrug record	Number of unique routes reported per WHODrug record (total = 67)	Number of unique indications reported per WHODrug record (total = 26,379)
Minimum	1	1	0	0
25th percentile	1	1	1	1
Median	3	1	1	1
75th percentile	11	2	2	4
Maximum	98,033	281	34	1487

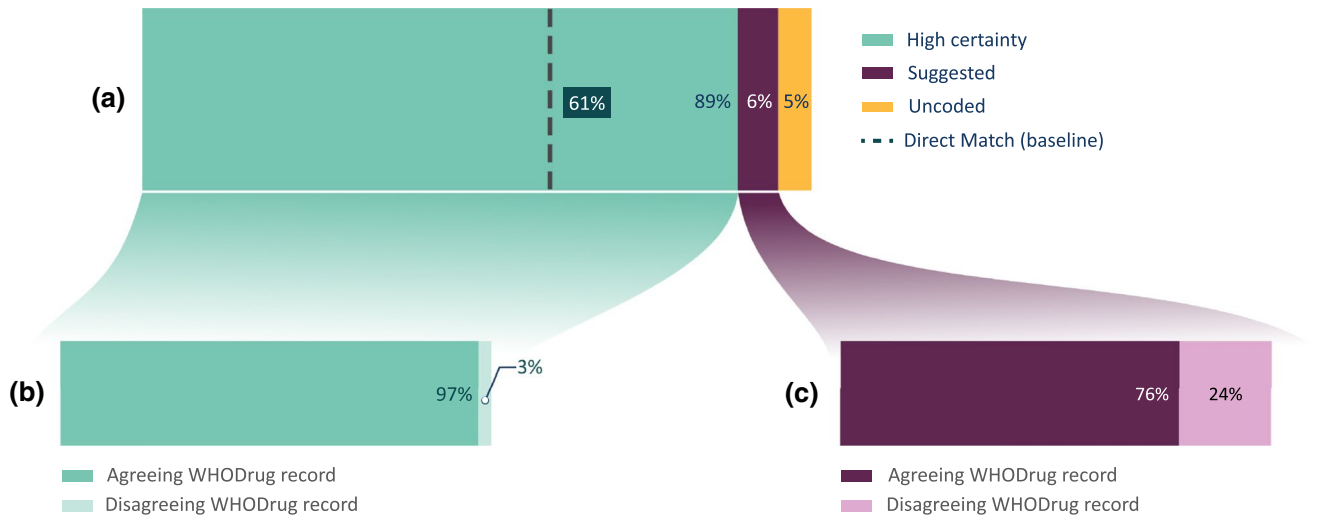

Fig. 2 **a** Koda's automation level showing percentages of high certainty, suggested encodings and uncoded Koda inputs compared with the direct-match baseline. **b** Agreement between high-certainty Koda encodings and gold standard. **c** Agreement between Koda suggestions and gold standard

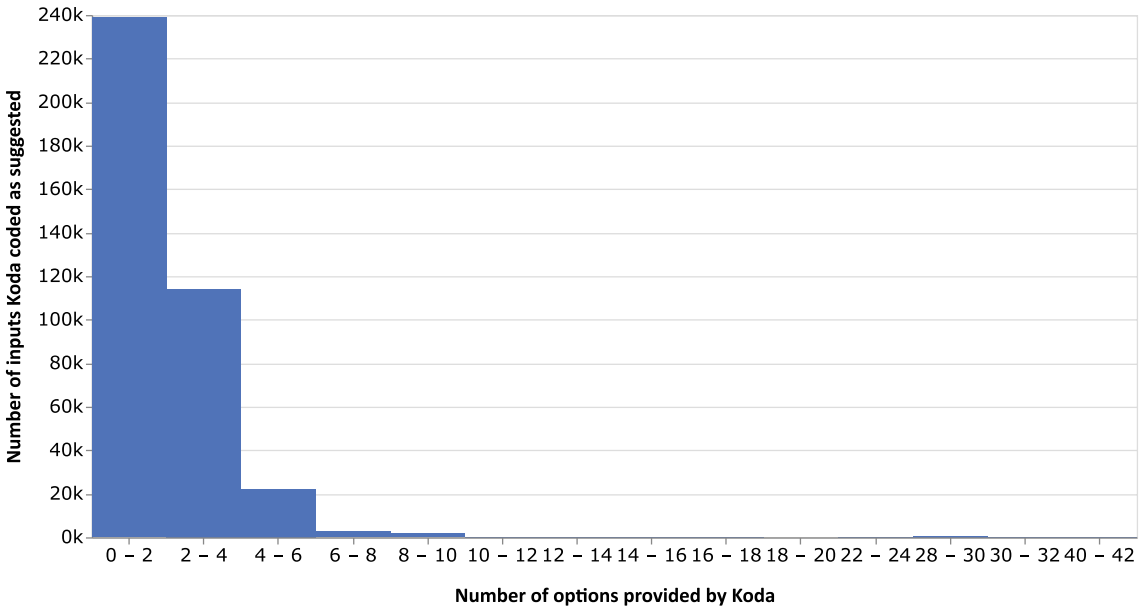

Fig. 3 Number of options provided by Koda when coding as suggested

kappa), which suggests strong agreement between the teams [29].

Similarly, in the 200 samples for which Koda suggested one or several WHODrug records that did not contain the gold-standard WHODrug record, the two teams found that approximately 50% of Koda's suggestions were as good as or better than the gold standard (Fig. 4). The gold standard was considered more correct in 48% of the mismatches by Team 1 and 44% by Team 2. None of the encodings were judged acceptable in 1.5% and 7.5% of mismatches according to Team 1 and Team 2, respectively. The inter-annotator agreement score for this evaluation was 0.81, which can also be considered strong agreement [29].

Of the 200 mismatches that Koda left uncoded, the gold-standard encoding was found to either be incorrect or require manual review in more than 80% of the samples. The Cohen Kappa inter-annotator score for these samples was 0.89, indicating strong agreement between the two teams [29].

4.4 Effect of Route, Indication and Country Fields

None of the Koda inputs in the VRIC dataset with confidence level *suggested* transitioned to *high certainty* when one or more of the optional fields were masked. We thus provide the results of the effect of the optional fields only for the 2124 entries coded by Koda at *high certainty* in the VRIC dataset. For each of the dataset variations (VRC, VIC, VC, VRI and V), we report the changes in encoding and Koda's confidence in Table 9.

Table 7 Centre and dispersion of the number of options for suggested encodings

Min	Median	75th pctl	95th pctl	99th pctl	Max
1	1	2	4	6	42

Min minimum; *max* maximum; *pctl* percentile

As can be seen in Table 9, Koda can still resolve ambiguous drug names even when additional fields are removed. Addition of the country field seems to have a larger effect, compared with route and indication, on increasing the automation level from *suggested* to *high certainty*: 129 entries were demoted to *suggested* when the country field was removed (Table 9, row VRI), while only 9 were demoted to *suggested* when the route and indication fields were removed (Table 9, row VC). The combined effect of removing all additional fields had the greatest impact, with 246 entries being demoted to *suggested* (Table 9, row V). No records were demoted to uncoded for any of the masked datasets.

Most entries stayed at *high certainty* when additional fields were removed. However, upon removal of these additional fields, a small proportion of these *high-certainty* coded Koda inputs were coded to a different WHODrug record. When route and indications were removed (Table 9, row VC), 105 entries at *high certainty* were coded to a different WHODrug record, while 81 entries at *high certainty* were

Table 8 Macro average precision, recall and *F*1 score across all WHODrug records in the gold standard, and the average of these metrics per WHODrug record weighted by the classes' prevalence in the gold standard for WHODrug Koda and our direct-match baseline

	Accuracy	Macro average			Weighted average		
		Precision	Recall	*F*1	Precision	Recall	*F*1
Koda	86.0%	91.0%	86.7%	87.6%	94.9%	86.0%	88.2%
Direct-match baseline	60.4%	71.9%	62.1%	64.4%	88.6%	60.4%	66.1%

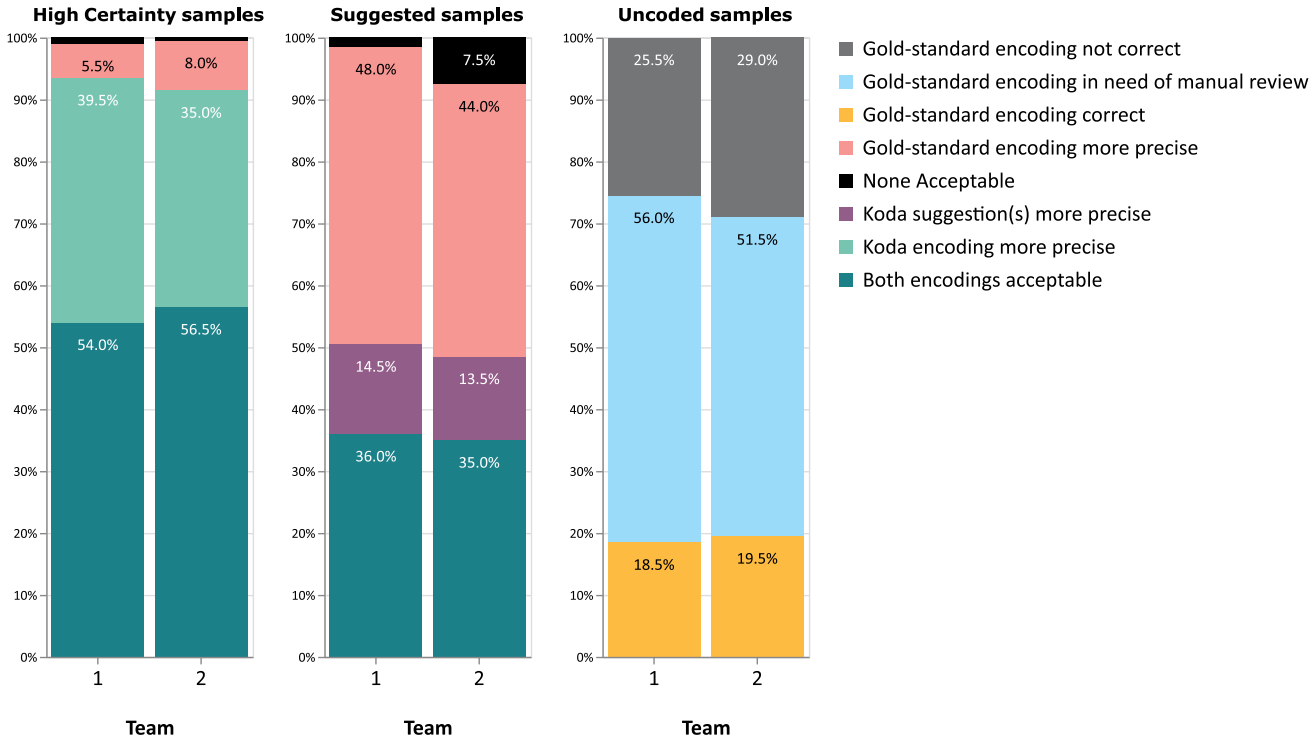

Fig. 4 Manual assessment results for Koda's three confidence levels

coded to a different WHODrug record when all fields were masked (Table 9, row V). Removing only route (Table 9, row VIC) led to more changes in the chosen WHODrug record than removing only indication (Table 9, row VRC), while the removal of only the country field (Table 9, row VRI) did not change the WHODrug record chosen for this dataset.

Of the 105 Koda inputs whose WHODrug record changed when route and indication were masked (Table 9, row VC), 86 were changed to an entry with the same active moiety; 61 of 86 were coded to the preferred base substance. Similarly for the case when all fields were masked (Table 9, row V), 78 were changed to a WHODrug record with the same active moiety, of which 59 were coded to the preferred base substance. These results show that Koda can select a sensible, but less specific, WHODrug record even when route and indication fields are missing.

5 Discussion

Our evaluation revealed Koda's performance for drug coding on AE reports to be on par with its performance on drug coding of concomitant drugs in clinical trials, on both automation level and coding quality [19]. Furthermore, Koda could automatically code significantly more than a simple text-matching baseline and managed to automate the coding of inputs with ambiguous drug names to a large extent. Ambiguous drug names were resolved mostly by Koda's internal coding rules, while the country field marginally increased Koda's confidence level, and the route and indication fields maintained Koda's confidence level but marginally

Table 9 Result of masking experiment, masking various fields

	Verbatim	Route	Indication	Country	High Certainty		Suggested
					Same WHODrug record	Changed WHODrug record	
VRIC	Provided	Provided	Provided	Provided	2,124 (100.0%)	-	-
VRC	Provided	Provided	Masked	Provided	2,089 (98.3%)	27 (1.3%)	8 (0.4%)
VIC	Provided	Masked	Provided	Provided	2,036 (95.9%)	87 (4.1%)	1 (0.0%)
VC	Provided	Masked	Masked	Provided	2,010 (94.6%)	105 (5.0%)	9 (0.4%)
VRI	Provided	Provided	Provided	Masked	1,995 (93.9%)	0 (0.0%)	129 (6.1%)
V	Provided	Masked	Masked	Masked	1,797 (84.6%)	81 (3.8%)	246 (11.6%)

influenced the choice of WHODrug record. Records of trade names are more likely to be coded to generic records of the base substance in the absence of route and indication fields, leading to correct but less precise encodings. Built-in coding rules are thus an important component of Koda's intelligence compared with standard text processing-based systems.

Not only could Koda automate the coding of a large proportion of cases, but also for approximately half of the non-automated cases, Koda suggested one or multiple WHODrug records, significantly reducing the search space of possible records. These suggestions were found to be sensible for three of four inputs, and manual evaluation of mismatches with the gold standard revealed that Koda's effective coding quality is likely even higher. From a coder's perspective, Koda can reduce manual coding effort to a minimum in many cases, including non-trivial ones. Koda cannot however support all coding decisions, with 5% of all inputs from the evaluation dataset remaining uncoded.

Even though route and indication are optional, their presence improved the record selection. Since indication is provided in free text, it allows greater room for error due to variations in its content. Machine learning can help in cases where information is entered in free text, but it requires high-coverage training datasets. Even though a well-trained model should generalise to new, unseen input, it may fail if the data are too different from what was seen during training, which was also seen in this study. Therefore retraining and continuous evaluation of the model is crucial.

In this study, we evaluated Koda for its ability to code drugs on AE reports, while in a previous study, Koda had been evaluated for its ability to code concomitant drugs in clinical trials. In the context of EHRs, drug coding may be applied to free-text drug fields or as the mapping step following the NER of drugs. Evaluating Koda's coding capabilities on drug mentions in EHRs would be an interesting future study. When coding drug verbatims that were extracted from narratives in EHRs, Koda's performance might be negatively affected since they might differ in nature and since additional drug information might be challenging to extract from the narratives.

There are several limitations to consider when interpreting the results of our study. First, we have only evaluated Koda using its default configuration with all coding rules turned on. The default configuration partly reflects UMC's internal coding conventions and is thus much in line with how the gold standard was developed. A different configuration may affect the results negatively.

Second, the baseline in this study was chosen for its simplicity and does not make use of synonym lists that organisations may develop as part of their automated coding processes. The addition of a synonym list to the baseline would likely increase its automation level. Constructing such a synonym list is a non-trivial, manual task that often requires domain-specific knowledge and maintenance. A comparison with more intelligent, automated drug coding systems for WHODrug would be of interest, however, to the best of our knowledge, no such systems are publicly available.

Third, in the manual evaluation of the mismatches to the gold standard, assessors were provided the same information as Koda, namely verbatim, route, indication and country, instead of the full AE report, which is used in practice during manual drug coding. As a result, there may be cases where the gold-standard encoding was in fact more correct based on the information derived from the entire AE report, but was judged to be as good or worse than Koda's prediction. This approach was chosen to be fair to Koda, which can only use the verbatim, route, indication and country information to produce its predictions.

Furthermore, our results should be interpreted considering the difficulty of the drug coding task. Many Koda predictions mismatching with the gold standard were nevertheless deemed acceptable and since several assessments made by the two evaluation teams were different, we can deduce that there is a certain amount of imprecision to drug coding, which makes evaluation of drug coding systems harder. Additionally, differences between the WHODrug versions used by Koda or the gold standard may affect the accuracy of both Koda and the baseline method when compared with the gold standard.

Moreover, while Koda is an independent system developed and maintained by an independent team and sharing no modules or code with UMC's automated drug coding processes for VigiBase reports, there is a possible influence on Koda's development from the automated processes used

during the creation of the gold standard in the form of transfer of knowledge and learnings.

Finally, during the creation of our gold-standard dataset, all reported drugs considered invalid during manual coding were excluded. Koda's ability to recognise these inputs as invalid and leave them uncoded would be worth evaluating during future studies but was outside the scope of this evaluation of Koda.

6 Conclusion

Drug coding to a standardised dictionary, such as WHODrug, is an essential step in case processing activities for pharmacovigilance; however, it is a non-trivial task due to the varying data quality for the reported drugs, variation in submitted information, as well as presence of ambiguous drug names. Auto-encoders based on text-processing algorithms and synonym lists commonly used during drug coding can usually only automate the coding of drug names that have a single record in the drug dictionary. Resolving ambiguous drug names is however harder and perhaps represents the fundamental challenge for automation in drug coding. It requires additional decision making in the form of coding conventions to choose the correct record among a list of possible candidates. Automating or supporting this process could save a significant amount of time for drug coders.

WHODrug Koda is a product developed for the specific purpose of automatically coding free-text drug information of concomitant drugs in clinical trials to WHODrug. Evaluation of Koda's coding performance on AE reports was the focus of this study. In our dataset of reports from VigiBase, Koda achieved an automation level gain of 46% compared with the simple baseline, requiring no human assistance. While Koda is designed as a human-in-the-loop coding system, its high automation level minimises manual interaction. Human assistance is only required in particularly difficult cases and is in many cases supported by sensible coding suggestions provided by Koda. Koda appears to handle ambiguous trade names very well and at high quality with the help of additional information, even when additional information is not provided, by following configurable coding rules.

Koda's ability to automatically code reported drugs on AE reports at a high confidence level (including drugs with ambiguous names) and suggest WHODrug records in cases identified as more challenging appears to be a novelty. To the best of our knowledge, there are no such systems in use whose performance have been systematically studied.

Even though Koda was designed for concomitant drug coding in clinical trials, it achieves high automation level and coding quality for drug coding of AE reports in VigiBase. Koda can thus be a valuable tool for automating and supporting coding practices during case processing for pharmacovigilance while ensuring high data quality.

Acknowledgements The authors sincerely thank the WHODrug Terminology specialists Anna Frisk, Jenny Adamsson, Jenny Klint, Michaela Lindh, and Susanna Johansson for the thorough assessment of our data during the manual evaluation study. They would also like to thank Klas Östlund for his support in running their dataset through Koda and in understanding Koda's inner workings. The authors would further like to thank Jessica Nilsson for helping them understand UMC's internal coding processes and her support during data extraction and Jim W. Barrett for proofreading the manuscript. The authors are indebted to the national centres that make up the WHO Programme for International Drug Monitoring (PIDM) and contribute reports to VigiBase. However, the opinions and conclusions of this study are not necessarily those of the various centres or of the WHO.

Declarations

Funding There was no specific funding for this study.

Conflicts of interest/competing interests The UMC is a non-profit organisation and the WHO Collaborating Centre for International Drug Monitoring. UMC sells WHODrug Koda as part of its product offerings to pharmaceutical companies. Eva-Lisa Meldau, Shachi Bista, Emma Rofors and Lucie Gattepaille declare being employed by UMC at the time of writing but have no other conflicts of interest that are directly relevant to the contents of this article.

Availability of data and material The datasets generated and analysed during the current study are not publicly available due to agreements between contributors of data to the database used (VigiBase) and the custodian of this database. National centres (mainly national drug regulatory authorities) constituting the WHO PIDM contribute data to VigiBase and the UMC is the custodian in its capacity as WHO Collaborating Centre for International Drug Monitoring. Some subsets of the data may be available from the corresponding author on reasonable request.

Code availability WHODrug Koda is proprietary software provided by the UMC (an independent, non-profit foundation) and neither the software nor the code for this evaluation are available for public release.

Authors' contributions All authors designed and conceptualised the manuscript, interpreted the data, drafted the manuscript, and revised the manuscript. Additionally, EM and SB generated and analysed the data, LG supervised the study, and ER provided support to the manual evaluation teams. All authors read and approved the final version.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

References

1. Peters L, Kapusnik-Uner JE, Bodenreider O. Methods for managing variation in clinical drug names. AMIA Annu Symp Proc. 2010;2010:637–41.
2. Lagerlund O, Strese S, Fladvad M, Lindquist M. WHODrug: a global, validated and updated dictionary for medicinal information. Ther Innov Regul Sci. 2020;54(5):1116–22. https://doi.org/10.1007/s43441-020-00130-6.
3. Nelson SJ, Zeng K, Kilbourne J, Powell T, Moore R. Normalized names for clinical drugs: RxNorm at 6 years. J Am Med Inf Assoc. 2011;18(4):441–8. https://doi.org/10.1136/amiajnl-2011-000116.
4. Kreimeyer K, Foster M, Pandey A, et al. Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. J Biomed Inf. 2017;73:14–29. https://doi.org/10.1016/j.jbi.2017.07.012.
5. Cowie MR, Blomster JI, Curtis LH, et al. Electronic health records to facilitate clinical research. Clin Res Cardiol. 2017;106(1):1–9. https://doi.org/10.1007/s00392-016-1025-6.
6. Mandel JC, Kreda DA, Mandl KD, Kohane IS, Ramoni RB. SMART on FHIR: a standards-based, interoperable apps platform for electronic health records. J Am Med Inf Assoc. 2016;23(5):899–908. https://doi.org/10.1093/jamia/ocv189.
7. Begoyan A. An overview of interoperability standards for electronic health records. Society For Design And Process Science; 2007.
8. Ghosh R, Kempf D, Pufko A, Barrios Martinez LF, Davis CM, Sethi S. Automation opportunities in pharmacovigilance: an industry survey. Pharm Med. 2020;34(1):7–18. https://doi.org/10.1007/s40290-019-00320-0.
9. Hwang M, Jeong D-H, Jung H, Sung W-K, Shin J, Kim P. A term normalization method for better performance of terminology construction. In: Rutkowski L, Korytkowski M, Scherer R, Tadeusiewicz R, Zadeh LA, Zurada JM, editors. Artificial intelligence and soft computing. Vol 7267 Lecture Notes in Computer Science. Berlin: Springer, Berlin Heidelberg; 2012. p. 682–90. https://doi.org/10.1007/978-3-642-29347-4_79.
10. Allones JL, Martinez D, Taboada M. Automated mapping of clinical terms into SNOMED-CT. An application to codify procedures in pathology. J Med Syst. 2014;38(10):134. https://doi.org/10.1007/s10916-014-0134-x.
11. Stenzhorn H, Pacheco EJ, Nohama P, Schulz S. Automatic mapping of clinical documentation to SNOMED CT. Stud Health Technol Inf. 2009;150:228–32.
12. Patrick J, Wang Y, Budd P. An automated system for conversion of clinical notes into SNOMED clinical terminology. In: Proceedings of the Fifth Australasian Symposium on ACSW Frontiers. Vol 68. Australian Computer Society, Inc.; 2007. pp. 219-226.
13. Combi C, Zorzi M, Pozzani G, Moretti U, Arzenton E. From narrative descriptions to MedDRA: automagically encoding adverse drug reactions. J Biomed Inf. 2018;84:184–99. https://doi.org/10.1016/j.jbi.2018.07.001.
14. Gattepaille LM, Hedfors Vidlin S, Bergvall T, Pierce CE, Ellenius J. Prospective evaluation of adverse event recognition systems in twitter: results from the Web-RADR Project. Drug Saf. 2020;43(8):797–808. https://doi.org/10.1007/s40264-020-00942-3.
15. Pustejovsky J, Lee K, Bunt H, Romary L. ISO-TimeML: An International Standard for Semantic Annotation. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). Valletta, Malta: European Language Resources Association (ELRA); 2010.
16. Sun W, Rumshisky A, Uzuner O. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. J Am Med Inf Assoc. 2013;20(5):806–13. https://doi.org/10.1136/amiajnl-2013-001628.
17. Sohn S, Clark C, Halgrim SR, Murphy SP, Chute CG, Liu H. MedXN: an open source medication extraction and normalization tool for clinical text. J Am Med Inf Assoc. 2014;21(5):858–65. https://doi.org/10.1136/amiajnl-2013-002190.
18. Babre D. Medical coding in clinical trials. Perspect Clin Res. 2010;1(1):29–32.
19. Herrgard S, Gil C, Holst I, et al. Assessment of machine learning methods in coding of concomitant medications in clinical trials. In: ML13 (Phuse Connect US 2020). Orlando, FL: Novo Nordisk; 2020. https://phuse.s3.eu-central-1.amazonaws.com/Archive/2020/Connect/US/Virtual/PAP_ML13.pdf.
20. Abatemarco D, Perera S, Bao SH, et al. Training augmented intelligent capabilities for pharmacovigilance: applying deep-learning approaches to individual case safety report processing. Pharm Med. 2018;32:391–401. https://doi.org/10.1007/s40290-018-0251-9.
21. Peters L, Kapusnik-Uner JE, Nguyen T, Bodenreider O. An approximate matching method for clinical drug names. AMIA Annu Symp Proc. 2011;2011:1117–26.
22. Raiskin Y, Eickhoff C, Beeler PE. Categorization of free-text drug orders using character-level recurrent neural networks. Int J Med Inf. 2019;129:20–8. https://doi.org/10.1016/j.ijmedinf.2019.05.020.
23. Uzuner Ö, Solti I, Cadag E. Extracting medication information from clinical text. J Am Med Inf Assoc. 2010;17(5):514–8. https://doi.org/10.1136/jamia.2010.003947.
24. Jauregi Unanue I, Zare Borzeshi E, Piccardi M. Recurrent neural networks with specialized word embeddings for health-domain named-entity recognition. J Biomed Inform. 2017;76:102–9. https://doi.org/10.1016/j.jbi.2017.11.007.
25. Koda information. https://www.who-umc.org/whodrug/whodrug-portfolio/whodrug-koda/.
26. Study Data Technical Conformance Guide—Technical Specifications Document. https://www.fda.gov/media/151717/download. Accessed 30 Aug 2021.
27. WHODrug Best Practices Version 6.0. https://www.who-umc.org/media/164209/whodrug-best-practices-vers-6-revised-final.pdf. Accessed 25 Nov 2021.
28. Norén GN, Orre R, Bate A, Edwards IR. Duplicate detection in adverse drug reaction surveillance. Data Min Knowl Discov. 2007;14:305–28. https://doi.org/10.1007/s10618-006-0052-8.
29. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med. 2012;22(3):276–82.