



# Finding Needles in the Haystack: Clinical Utility Score for Prioritisation (CUSP), an Automated Approach for Identifying Spontaneous Reports with the Highest Clinical Utility

Vijay Kara<sup>1</sup> · Greg Powell<sup>2</sup> · Olivia Mahaux<sup>3</sup> · Aparna Jayachandra<sup>4</sup> · Naashika Nyako<sup>1</sup> · Christopher Golds<sup>1</sup> · Andrew Bate<sup>1,5</sup>

Accepted: 11 June 2023 / Published online: 3 August 2023  
© GlaxoSmithKline Biologicals S.A 2023, corrected publication 2023

## Abstract

**Introduction** Spontaneous reporting of adverse events has increased steadily over the past decades, and although this trend has contributed to improving post-marketing surveillance pharmacovigilance activities, the consequent amount of data generated is challenging to manually review during assessment, with each individual report requiring review by pharmacovigilance experts. This highlights a clear need for alternative or complementary methodologies to help prioritise review.

**Objective** Here, we aimed to develop and test an automated methodology, the Clinical Utility Score for Prioritisation (CUSP), to assist pharmacovigilance experts in prioritising clinical assessment of safety data to improve the rapidity of case series review when case volumes are large.

**Methods** The CUSP method was tested on a reference dataset of individual case safety reports (ICSRs) associated to five drug-event pairs that led to labelling changes. The selected drug-event pairs were of varying characteristics across the portfolio of GSK's products.

**Results** The mean CUSP score for 'key cases' and 'cases of low utility' was 19.7 (median: 21; range: 7–27) and 17.3 (median: 19; range: 4–27), respectively. CUSP distribution for 'key cases' were skewed toward the higher range of scores compared with 'all cases'. The overall performance across each individual drug-event pair varied considerably, showing higher predictive power for 'key cases' for three of the drug-event pairs (average CUSP between these three: 22.8; range: 22.5–23.0) and lesser power for the remaining two (average CUSP between these two: 17.6; range: 14.5–20.7).

**Conclusion** Although several tools have been developed to assess ICSR completeness and regulatory utility, this is the first attempt to successfully develop an automated clinical utility scoring system that can support the prioritisation of ICSRs for clinical review.

## 1 Introduction

Spontaneous reporting remains critical for identifying emerging safety issues once a drug has been approved [1–3]. Whilst safety concerns can be detected from a wide range of sources, such as clinical studies, scientific literature and

✉ Vijay Kara  
vijay.x.kara@gsk.com

<sup>1</sup> GSK, 980 Great West Road, London TW8 9GS, UK

<sup>2</sup> GSK, Research Triangle, NC, USA

<sup>3</sup> GSK, Wavre, Belgium

<sup>4</sup> GSK, Bangalore, India

<sup>5</sup> Department of Non-Communicable Disease Epidemiology, LSHTM, London, UK

### Key Points

Techniques to systematically identify and distinguish ICSRs of higher clinical utility from lower utility ones will improve timeliness of case series review during safety assessment when case volumes are large.

We developed an automated methodology to assist pharmacovigilance experts in the prioritisation of ICSRs for clinical assessment of safety data within a safety database.

The method's performance demonstrated the feasibility of this approach, but also highlighted opportunities for further refinement.

spontaneous reports, the latter remain the most commonly cited data source supporting regulatory actions [4–6]. Spontaneous reporting is therefore an efficient way to monitor safety concerns. Whilst statistical methods can support screening large spontaneous reporting databases, qualitative review of individual case safety reports (ICSRs) forms a vital step following the initial identification supporting any subsequent validation or assessment activities. This process relies heavily on manual assessment by safety experts through the use of existing frameworks (e.g. Bradford Hill [7], Naranjo et al. [8]) which provide guidance in the assessment of causality.

Overall reporting of adverse events (AEs) has steadily increased over the past decades, resulting in a larger body of data which offers the potential to identify safety signals earlier and more accurately [9]. However, the processing of this sheer amount of data has appeared as an obstacle to the optimal detection of clinically significant AEs and adverse drug reactions, impeding manual review of safety reports and highlighting the need for alternative or complementary automated data review systems [4].

In addition, the review of spontaneous safety reports is complicated by qualitative issues: ICSRs are often incomplete, making the establishment of any causal relationships between a reported AE and a drug difficult. Although international guidelines recommend that all detected AEs must be reported with the aim of improving and better understanding product safety, follow-up is often unsuccessful and does not improve the overall information needed for assessment of a causal relationship. Indeed, a review of a random sample of 1000 AEs noted, across potentially important data elements used in causality assessment, an average of only 12.7 out of 26 were included in initial reports [10]. While follow-up should theoretically provide additional data and contribute to higher completeness of reports, studies suggest successful follow-up is relatively rare (only 13.7% of ICSRs received follow-up) [11] and generally fail to provide much additional data (ICSRs with a follow-up contained an average of 15.5 out of the 26 potentially important data elements) [10].

For this reason, in addition to including an automated component to support assessment of increasing case volumes, new approaches should enable the rapid and effective identification of reports that have the highest potential to contribute, either alone or combined with other reports (e.g. case series), to safety signal assessment (i.e. the detection of AEs that may be causally linked with the drug of interest).

Because of the increasing volume of ICSRs and the need for manual review by safety stakeholders, systematic scoring methodologies assisting in the prioritisation of ICSRs, both at the drug-event level and ICSR level are quickly gaining popularity in the field of pharmacovigilance. These algorithms include methods based on case report completeness

(vigiGrade [12]), regulatory utility (Food and Drug Administration [FDA] [13]); prediction models based on algorithm from the Netherlands Pharmacovigilance Centre Lareb [14], or those developed for specific use cases (ClinDoc [15]; IMI WEB-RADR [16]), the purpose of which is to integrate qualitative attributes to support signal management activities.

The vigiGrade scoring system, developed to highlight quality issues in ICSR datasets by scoring the completeness of case reports, is currently the most widely used of these tools (predominantly in the field of pharmaceutical regulatory science), and both the vigiGrade and modified vigiGrade methods have previously been used successfully in several countries [12, 17–19]. However, a previous attempt to apply vigiGrade for prioritisation within a marketing authorisation holder (MAH) database failed to improve quantitative signal detection [20], suggesting that ‘completeness’ does not necessarily imply ‘utility’ (i.e. the potential for information to provide true drug safety insights) [13]. Similarly, regulatory utility scoring systems developed by Muñoz et al. [13] and Scholl et al. [14], while proven effective for their respective databases, require information that is not systematically included in MAH databases or differ in terms of regulatory requirements (e.g. expedited reporting) and therefore, may not be directly suited to the purpose of safety assessment in MAH datasets.

The goal of our research was to develop and test a systematic scoring methodology applied in an automated manner to assist pharmacovigilance experts when a case series review is required through prioritisation on the basis of clinical utility. In this study, clinical utility is defined as the likelihood of an ICSR containing information that will contribute to a meaningful clinical assessment of that case based on the presence of data within historically important fields to support a decision on the potential causal association between a drug-event pair.

With regard to safety data within safety databases, a major flaw of current approaches is that all cases are initially given the same weight and that manual review of ICSRs (sometimes > 1000) is required to isolate a small cluster of clinically meaningful and sufficiently well-documented cases that can support a causality assessment.

## 2 Methods

### 2.1 Methodology Presentation

We established a new method, called the Clinical Utility Score for Prioritisation (CUSP), built upon the literature for completeness and regulatory utility described above, and including a few notable additions relevant for overall clinical

assessment of causality not accounted for in other methods (e.g. duration of AE).

## 2.2 Variable Identification

Variables used in existing methods for the assessment of spontaneous report completeness (vigiGrade [12]; modified vigiGrade [17–19]), clinical quality scoring (ClinDoc [15]), and utility (FDA pharmacovigilance utility scoring [13]) were reviewed, evaluated and agreed upon by the consensus of a group of eight safety experts with a diversity of experiences who routinely conduct safety assessments. In addition, variables commonly used by safety staff, but not included in the aforementioned methodologies, were also considered. A comparison of variables included in CUSP and other pharmacovigilance scoring systems is presented in Supplementary Information S11.

A total of 30 variables of interest were deemed to contribute to the clinical utility of ICSRs and were included in the CUSP method (Table 1).

## 2.3 Variable Scoring

In the context of the CUSP methodology, variable fields were assigned a score of 1 if (1) they contained any information (regardless of the actual data entered) and (2) met variable-specific criteria if applicable (e.g. 'medical history' or 'indication' variable fields containing the information 'unknown', as in 'unknown medical history' or 'unknown indication', would be assigned a score of 0; the 'narrative length' field would be assigned a score of 1 if the length of the narrative was > 1 standard deviation above the average length of narratives across all cases). Inclusion/exclusion

criteria specific to CUSP variable fields are listed in Supplementary Information S12.

The CUSP score for a given ICSR is calculated as the sum of the scores of each variable field, with a minimum score of 2 (all ICSRs contain information relative to the 'suspect drug' and 'event' variable fields) and in theory a maximum score of 30.

## 2.4 ICSR Data Source and Mapping of Variables

Data used for the analysis were all ICSRs from GSK's internal safety database, made up of spontaneous reports up to the second quarter of 2021, where a GSK product was reported as the suspect drug.

The CUSP variables of interest were mapped to relevant data fields contained within GSK's internal safety database. This mapping was validated through an independent code review against the requirements.

## 2.5 Reference Dataset for Evaluation

A reference dataset was created to evaluate the performance of the CUSP methodology. The reference dataset was built from selected safety signals previously evaluated and completed by GSK safety experts, where a manual review of ICSRs had been undertaken prior to the implementation of the CUSP score to support the assessment of causality (i.e. case classification was independent of ICSR CUSP score). Of these signal evaluations, only those meeting the following criteria were included:

- Conducted prior to December 2020 (data lock point)

**Table 1** CUSP variables of interest, categorised according to information type

Patient general information	Patient medical history	Drug information	Event information	Miscellaneous information
Age	Historical drug	Suspect drug	Event	HCP report
Gender	Medical history	Number of suspect drugs	Number of events	Dechallenge
Weight		Indication	DME	Rechallenge
		Dose	Event onset date	Verbatim narrative
		Concentration	Event duration	Narrative length
		Dosing regimen	Time to onset	
		Route of administration	Results of test and procedure	
		Start/stop	Treatment product	
		Lot number	Outcome of AE	
		Disposition of drug		
		Concomitant medications		

AE adverse event, CUSP Clinical Utility Score for Prioritisation, DME designated medical event, HCP healthcare professional

- Regulatory action (e.g. labelling change) was recommended beyond the continuation of routine pharmacovigilance
- Data from the GSK safety database was the main data source evaluated
- Represented a range of drugs and events with varying attributes.

Of the resulting evaluations, five drug-event pairs with varying characteristics across GSK's product portfolio were selected to test the CUSP method. Events with a short time to onset (TTO) and those manifesting over longer periods of exposure were included in the analyses. For each evaluation, we noted the total number of ICSRs that had previously undergone manual review. Of those, we reported the total number of ICSRs that were determined to be of poor quality and/or lacking the sufficient clarity to be assessed, the total number of ICSRs that were of sufficient quality for assessment, and those that could be considered as 'key cases'. A 'key case' is a case or group of cases most suggestive of a causal relationship [21].

The five safety evaluations included in the reference dataset contained a total of 773 ICSRs, of which 33 were considered 'key cases' (positive control), 344 of poor quality (negative control; henceforth referred to as 'cases of low utility') (Table 2), and 396 reports containing alternative explanations (medical history and/or concurrent medications). These datasets ('key cases' and 'cases of low utility') were thereafter used along with 'all cases' within the safety database to assess distribution of CUSP scores, performance of CUSP methodology, and contribution of individual variable fields to the score.

### 3 Results

#### 3.1 Overall Performance of the CUSP Methodology for All Data

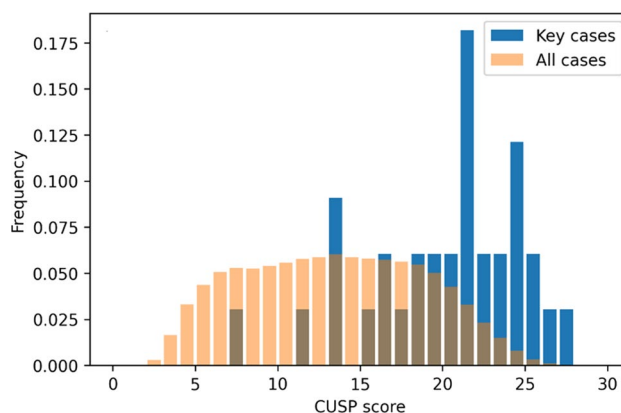
The mean CUSP scores for 'key cases' and 'cases of low utility' were 19.7 (median: 21; range: 7–27) and 17.3 (median: 19; range: 4–27), respectively. The distribution of CUSP scores for 'key cases' were clearly skewed toward the higher range of scores as compared with 'all cases' (Fig. 1).

The number and proportion of 'key cases' were assessed with regard to ranges of CUSP score thresholds (e.g. top 95%, top 85%, etc.). The overall distribution of CUSP scores according to percentiles is presented in Table 3.

As an example of using CUSP thresholds as cutoff points for case retrieval and prioritisation for subsequent manual evaluation by safety experts: if only ICSRs in the

**Table 2** Number of 'key cases' and 'cases of low utility' extracted from the reference dataset (for each drug-event pair and in total)

Drug-event pair	Number of 'key cases' (positive control)	Number of 'cases of low utility' (negative control)
Drug-Event #1	5	16
Drug-Event #2	7	195
Drug-Event #3	10	77
Drug-Event #4	2	56
Drug-Event #5	9	0
Total	33	344



**Fig. 1** Distribution of CUSP scores in 'key cases' subset and full reference dataset ('all cases'). CUSP Clinical Utility Score for Prioritisation

top 20% of all CUSP scores were selected (i.e. ICSRs with CUSP scores in the 80th percentile), 72.7% ( $n = 24$ ) of 'key cases' would be retrieved versus only 58.1% ( $n = 200$ ) of 'cases of low utility'. In other words, 41.9% ( $n = 144$ ) of 'cases of low utility' would not be prioritised for subsequent manual evaluation by safety experts versus only 27.3% ( $n = 9$ ) of 'key cases'.

#### 3.2 Variables of Interest Contributing to the CUSP

The proportions of reports including specific CUSP variables varied across the reference dataset (Fig. 2; a table detailing these data is presented in Supplementary Information SI3). Differences were most pronounced for the designated medical event (DME; 73% for 'key cases' versus 16% for 'all cases'), TTO (94% for 'key cases' versus 47% for 'all cases') and long narrative (70% for 'key cases' versus 26% for 'all cases') case attributes.

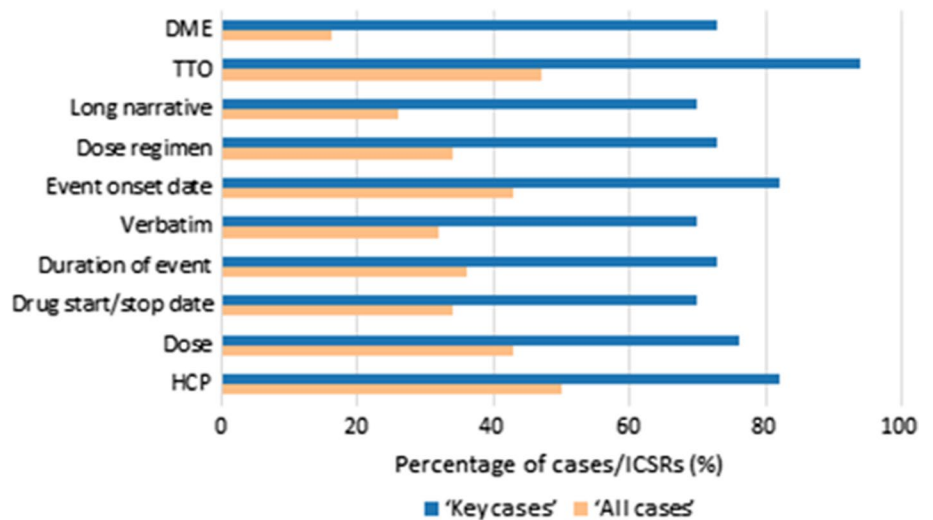
**Table 3** Performance of the CUSP score based on reference dataset.

		Percentiles from overall CUSP score distribution ('all cases')					Total
		95	90	85	80	15 <sup>a</sup>	
'Key cases' (N = 33)	Number of retrieved cases (n)	12	20	22	24	33	33
	Proportion of retrieved cases (%)	36.4	60.6	66.7	72.7	100	100
'Cases of low utility' (N = 344)	Number of retrieved cases (n)	104	160	183	200	332	344
	Proportion of retrieved cases (%)	30.2	46.5	53.2	58.1	96.5	100

CUSP Clinical Utility Score for Prioritisation, N cases included in given datasets, n cases retrieved from given datasets depending on cutoff points (percentiles) applied

<sup>a</sup>The 15th percentile was the lowest threshold at which all 'key cases' were retrieved

**Fig. 2** Presence (%) of selected CUSP variables in the safety database ('all cases') and 'key cases' subset, listed by decreasing difference (top 10 highest) between the 'key cases' subset and the full reference dataset. CUSP Clinical Utility Score for Prioritisation, DME designated medical event, HCP healthcare professional, ICSRs individual case safety reports, TTO time to onset



### 3.3 Performance of CUSP for Specific Drug-Event Pairs

The overall performance across each individual drug-event pair varied greatly, with 'key cases' retrieval ranging from 20% to 60% when applying an 80th percentile CUSP score threshold (Table 4). Three of the drug-event pairs (#1, #2 and #4) demonstrated a discrimination of 'key cases' compared with the data for the drug-event pair of at least 50% (at the 80th percentile), whereas performance for the other two drug-event pairs (#3 and #5) was < 50% (Table 4; Fig. 3). The average CUSP score for drug-event pairs #1, #2 and #4 was 22.8 (range: 22.5–23.0), while the average CUSP score for drug-event pairs #3 and #5 was 17.6 (range: 14.5–20.7).

## 4 Discussion

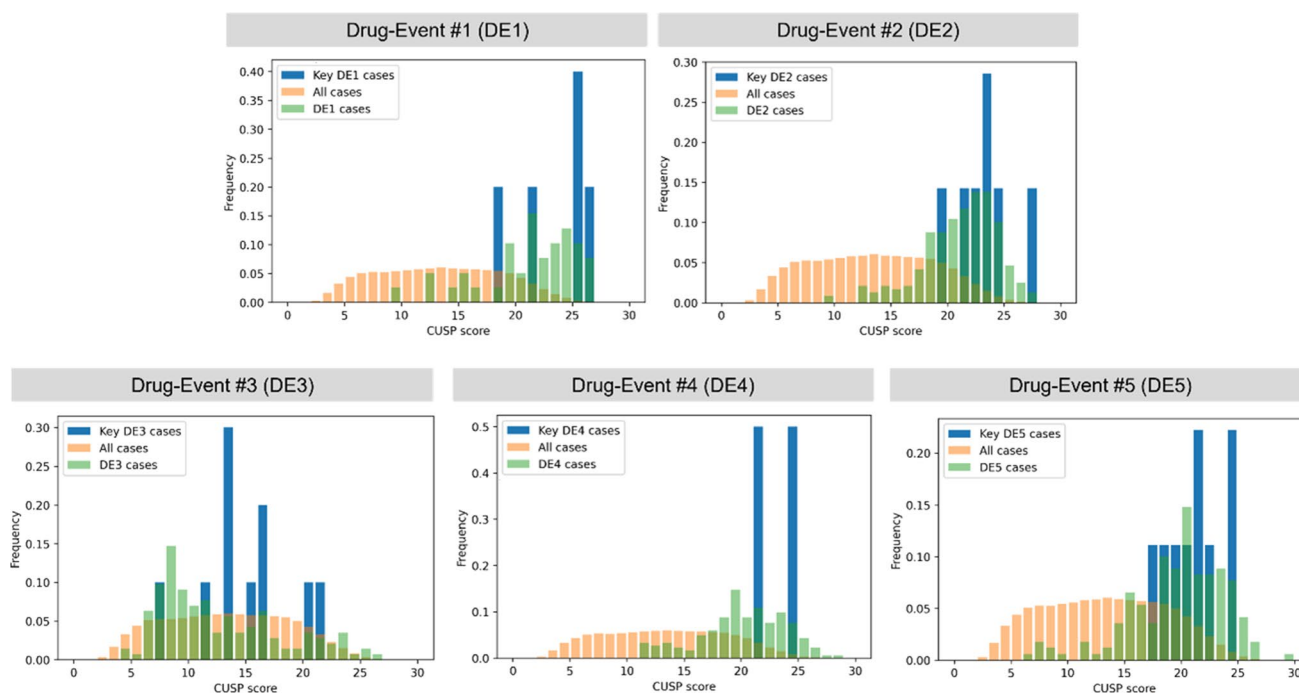
Building upon prior work done for completeness and regulatory utility assessment, as well as internal expertise, our present research aimed to develop and test CUSP, an automated methodology to support the clinical assessment of

**Table 4** Proportion (%) of 'key cases' in 80th percentile (defined from 'all cases' of given drug-event pairs)

Drug-event pairs	'Key cases' at 80th percentile (%)
Drug-Event #1	60
Drug-Event #2	57
Drug-Event #3	20
Drug-Event #4	50
Drug-Event #5	22

safety data within a case series by ranking ICSRs on the basis of clinical utility.

The CUSP method for the overall dataset of five drug-event pairs was able to retrieve 72.7% of 'key cases' at the 80th percentile, producing higher average CUSP score for 'key cases' than 'cases of low utility', and score distributions skewing towards higher CUSP scores for the 'key cases' stratum compared with the 'cases of low utility' stratum. Furthermore, the use of thresholds/cutoff points based on percentiles of CUSP scores for the reference dataset promoted



**Fig. 3** Distributions of CUSP scores for drug-event pairs. *CUSP* Clinical Utility Score for Prioritisation

a preferential retrieval of ‘key cases’ (72.7% retrieval with an 80th percentile cutoff point) compared with ‘cases of low utility’ (58.1% retrieval with an 80th percentile cutoff point). This highlights the potential of the scoring system to discriminate between cases of higher and lower clinical utility and assist in prioritisation of reports.

Although questions could be raised concerning that not all ‘key cases’ were always ranked highly using the CUSP method, it is important to frame this result in the context of real-world pharmacovigilance activities. A small number of cases reports (as low as one, if a strong and well-documented report is isolated) can be enough to justify regulatory action [22, 23], however, this is generally rare. Typically, a safety assessment is supported by numerous cases, such that the absence of some ‘key cases’ is unlikely to change the overall assessment of the drug-event pair. In practice however, the application of a CUSP threshold is arbitrary, due to the variability in performance across drug-event pairs, and all cases can be reviewed when needed. The CUSP offers a systematic approach for prioritisation which will often, but not always, reduce time, as a sufficient body of evidence for considering next steps will often surface to the top of a sorted list for clinical review.

Another analysis was carried out to evaluate the relative contribution of variables to the CUSP score. Since variables considered for the CUSP method were selected with the aim of facilitating the establishment of possible links between a drug and an event, it is unsurprising that high-scoring ICSRs therefore contain a considerable amount

of information directly related to the suspected drug and reported event. In particular, DME, TTO and long narratives were frequently present in high CUSP scoring reports, and would therefore be likely to contribute to the identification of ‘key cases’. This too is unsurprising if we consider the nature of these attributes. The DME list includes designated important medical events appearing on the European Medicines Agency important medical events list, which details serious and usually drug-related medical conditions using standardised terminology [24]. In addition, TTO supports the establishment of a temporal relationship between the drug and event. Finally, ICSRs containing longer narratives are hypothesised to contain more clinically rich information that may help understand the chronology of the AE.

Assessment of the method for the five individual drug-event pairs included in the reference dataset revealed variations of performance: CUSP performed well (of at least 50%, at the 80th percentile) for three of the pairs, enabling retrieval of 50–60% of ‘key cases’ compared with 20–22% for the two other pairs. It is also worth noting that the average CUSP score for ‘key cases’ obtained for the three drug-event pairs for which the method performed better was higher (22.8) than the scores obtained for the reference dataset (12.9 for ‘all cases’, 19.7 for ‘key cases’, and 17.3 for ‘cases of low utility’). This further highlights the fact that the current iteration of the CUSP method is likely to benefit the pharmacovigilance activities surrounding some drug or event types more than others.

Variables that are suggestive of a given ICSR clinical utility may differ considerably depending on the safety issue(s) being assessed [13]. For instance, the CUSP method is expected to perform less well for events which are considered less clinically complex and/or require very few attributes to diagnose (e.g. hypersensitivity reactions). In the example of hypersensitivity reactions, some CUSP variables are likely to be less relevant (e.g. concomitant medications) than others (e.g. TTO, dechallenge, rechallenge). In addition, CUSP scoring may be impacted for events which have highly complex diagnostic criteria (e.g. psychiatric disorders), where a healthcare professional may not know the clinically relevant information as the event manifests over a much longer period of time and is multifactorial in nature. Therefore, the healthcare professional is more likely to include information that may not be relevant in the assessment of the event. Put together, these results suggest that the CUSP method is an effective ranking classifier, but that bespoke adjustments a priori for specific drug- or vaccine-event pairs may be an area of opportunity. However, the rationale of such adjustments should be carefully documented to maintain reproducibility and consistency of approach across drug-event pairs when appropriate.

Important limitations of the current study include the relatively small sample size of its reference dataset (33 positive and 344 negative controls) and the reduced number of drug-event pairs considered for the analyses. Although selected drug-event pairs displayed varying attributes, a lot of drug-event pairs with different characteristics were not covered here, and CUSP performance may vary substantially. As previously discussed by Muñoz et al. [13], our work further highlights the difficulty in defining key variables as a major challenge to the large-scale use of measures of utility for safety assessments [13]. The CUSP score only considers whether a variable is present or absent, not whether the said variable conveys clinically relevant and useful information (e.g. is rechallenge available versus confirming a positive rechallenge). Although this binary classification might appear as simplistic, the lack of rationale and evidence to support a non-binary approach (e.g. how to apply variable weighting between 0 and 1 in a clinically appropriate and reproducible way taking into consideration the heterogeneity of drug AEs through a semi-automated solution) currently precludes the use of more sophisticated variable weighting for the CUSP method. However, with further experience using the method in the field and the concurrent build-up of evidence, a weighted approach may be considered in the future. In addition, the CUSP score did not consider a time-based variable both for potential time savings utilising a CUSP-based methodology or in an assessment to determine if retrieval of 'key cases' occurs sequentially to support decision making. The CUSP scores can potentially be further

refined to provide a higher level of accuracy through the accumulation of additional data.

Results presented here support the effectiveness and potential of the CUSP method, however, there remains room for further improvement. This first iteration of the CUSP method used a relatively simple binary scoring system based on the presence or absence of data, regardless of the drug-event association being assessed. A modified version of CUSP limiting variables to only those of relevance to the drug and/or event being assessed could improve the performance of the method. Although this may require a priori input from safety experts, the burden of manual work involved would be largely offset by the drastically improved reviewing performance, and this methodology could be used to improve the performance of drug-event pairs #3 and #5. It is worth noting, though, that manual interventions, even small, may be a source of bias or error (e.g. wrong weighting of individual variables) and great caution should be exerted. More broadly, the need for any situation-specific modifications to the assessment system should be the object of a robust framework to ensure consistency and adequate documentation of deviations.

Other directions for improvement are the development of a regression model where each variable is weighted differently, or the systematic weighting of variables according to their importance to the causality assessment (e.g. rechallenge, which is of high importance, would carry more weight than AE treatment, which is less important). The challenge of such approaches is to reduce the risk of overfitting, particularly when the available data are sparse. In the longer term, artificial intelligence and machine learning may come into play to further improve approaches using the CUSP method, for example, through the systematic identification and prioritisation of clinically rich reports at earliest possible time (i.e. as soon as a safety report enters the system) or the automated focus of follow-up efforts on these reports.

The CUSP methodology has use cases across multiple pharmacovigilance processes (e.g. supporting the identification of potentially clinically important cases at case intake through to supporting prioritisation during safety assessment). The CUSP methodology is an important step toward allowing safety stakeholders to focus efforts on higher quality data. The CUSP method is an effective ranking classifier, which supports a structured review of the data prioritising clinically rich cases. CUSP scoring can not only identify ICSRs with potential for increased clinical utility to support causality assessment, but can also be used to identify what variables are more important to help focus follow-up efforts. In addition, CUSP has the potential to support in-stream qualitative reviews aiming to identify if changes in a particular ICSR have the potential to improve clinical assessment [22, 23]. Further work will endeavour to explore these

potential use cases; currently it does not support the signal detection process, nor does it infer clinical insights without appropriate pharmacovigilance expertise.

## 5 Conclusion

Our study demonstrates that the CUSP methodology has the potential to enable the systematic identification of ICSRs with higher clinical utility. Although other tools have been developed to assess ICSR completeness and regulatory utility, this is the first attempt to successfully develop an automatable clinical utility scoring system that can support the prioritisation of ICSRs during case series review. The results presented here demonstrate the feasibility of automating the identification of ICSRs with greater clinical utility to support pharmacovigilance experts in prioritising clinical review. The efficiency of such an approach could allow safety staff to reach a decision faster and therefore focus on activities with higher added value (e.g. risk mitigation) rather than reviewing the subset of ICSRs that offer very little value to clinical assessment.

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s40264-023-01327-y>.

**Acknowledgements** The authors would like to thank the GSK staff involved in the study. The authors would like to thank Akkodis Belgium for writing support, editorial assistance, and manuscript coordination, on behalf of GSK.

## Declarations

**Conflicts of interests** All authors are employed by GSK and declare financial and non-financial relationships and activities. VK, GP, OM, AJ, NN, CG and AB also hold shares/stocks/stock options in the company.

**Funding** This work was supported by GlaxoSmithKline Biologicals S.A., which was the funding source. GlaxoSmithKline Biologicals S.A. was involved in all stages of the study conduct and analysis and paid for all costs associated with the development and publishing of this manuscript.

**Author contributions** VK, GP, AJ, NN, CG and AB were involved in the conception or the design of the study. VK, OM, AJ and NN participated in the collection or generation of the study data. VK, GP and OM performed the study. All authors were involved in the analysis or interpretation of the data. All authors reviewed and commented on drafts of the manuscript for intellectual content and gave final approval to submit for publication. All authors attest they meet the ICMJE criteria for authorship.

**Data availability** The datasets generated and/or analysed during the current study are not publicly available as the study was based on data from the GSK safety database. The data contains PII and confidential/sensitive patient related Information. They could be available from the corresponding author on reasonable request.

**Ethics approval** Not applicable.

**Consent to participate** Not applicable.

**Consent for publication** Not applicable.

**Code availability** Not applicable.

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

## References

1. Alomar M, Tawfiq AM, Hassan N, Palaian S. Post marketing surveillance of suspected adverse drug reactions through spontaneous reporting: current status, challenges and the future. *Ther Adv Drug Saf.* 2020;11:2042098620938595. <https://doi.org/10.1177/2042098620938595>.
2. Bate A, Evans SJW. Quantitative signal detection using spontaneous ADR reporting. *Pharmacoepidemiol Drug Saf.* 2009;18(6):427–36. <https://doi.org/10.1002/pds.1742>.
3. Bate A, Stegmann JU. Safety of medicines and vaccines—building next generation capability. *Trends Pharmacol Sci.* 2021;42(12):1051–63. <https://doi.org/10.1016/j.tips.2021.09.007>.
4. Stergiopoulos S, Fehrle M, Caubel P, Tan L, Jebson L. Adverse drug reaction case safety practices in large biopharmaceutical organizations from 2007 to 2017: an industry survey. *Pharmaceut Med.* 2019;33(6):499–510. <https://doi.org/10.1007/s40290-019-00307-x>.
5. Clarke A, Deeks JJ, Shakir SAW. An assessment of the publicly disseminated evidence of safety used in decisions to withdraw medicinal products from the UK and US markets. *Drug Saf.* 2006;29(2):175–81. <https://doi.org/10.2165/00002018-200629020-00008>.
6. Lane S, Lynn E, Shakir S. Investigation assessing the publicly available evidence supporting postmarketing withdrawals, revocations and suspensions of marketing authorisations in the EU since 2012. *BMJ Open.* 2018;8(1):e019759. <https://doi.org/10.1136/bmjopen-2017-019759>.
7. Hill AB. The environment and disease: association or causation? *Proc R Soc Med.* 1965;58(5):295–300.
8. Naranjo CA, Busto U, Sellers EM, Sandor P, Ruiz I, Roberts EA, et al. A method for estimating the probability of adverse drug reactions. *Clin Pharmacol Ther.* 1981;30(2):239–45. <https://doi.org/10.1038/clpt.1981.154>.
9. Caster O, Aoki Y, Gattepaille LM, Grundmark B. Disproportionality analysis for pharmacovigilance signal detection in small databases or subsets: recommendations for limiting false-positive associations. *Drug Saf.* 2020;43(5):479–87. <https://doi.org/10.1007/s40264-020-00911-w>.



10. Kara V, Powell G, Merico E, Kaur N, Bate A. An assessment of the clinical completeness of spontaneous adverse event reports pre and post-follow-up. *Drug Saf.* 2021;44(12):1451–2.
11. Kara V, Powell G, Merico E, Kaur N, Bate A. Impact of follow-up activities on spontaneous reports. *Drug Saf.* 2021;44(12):1451–2.
12. Bergvall T, Noren GN, Lindquist M. *vigiGrade*: a tool to identify well-documented individual case reports and highlight systematic data quality issues. *Drug Saf.* 2014;37(1):65–77. <https://doi.org/10.1007/s40264-013-0131-x>.
13. Muñoz MA, Dal Pan GJ, Wei YJJ, Delcher C, Xiao H, Kortepeter CM, et al. Towards automating adverse event review: a prediction model for case report utility. *Drug Saf.* 2020;43(4):329–38. <https://doi.org/10.1007/s40264-019-00897-0>.
14. Scholl JHG, van Hunsel FPAM, Hak E, van Puijtenbroek EP. A prediction model-based algorithm for computer-assisted database screening of adverse drug reactions in the Netherlands. *Pharmacoepidemiol Drug Saf.* 2018;27(2):199–205. <https://doi.org/10.1002/pds.4364>.
15. Oosterhuis I, Rolfes L, Ekhart C, Muller-Hansma A, Härmark L. First experiences with a tool to measure the level of clinical information present in adverse drug reaction reports. *Expert Opin Drug Saf.* 2018;17(2):111–5. <https://doi.org/10.1080/14740338.2018.1400008>.
16. Oosterhuis I, Taavola H, Tregunno PM, Mas P, Gama S, Newbould V, et al. Characteristics, quality and contribution to signal detection of spontaneous reports of adverse drug reactions via the WEB-RADR mobile application: a descriptive cross-sectional study. *Drug Saf.* 2018;41(10):969–78. <https://doi.org/10.1007/s40264-018-0679-6>.
17. Kumar R, Kumar P, Kalaiselvan V, Kaur I, Singh GN. Best practices for improving the quality of individual case safety reports in pharmacovigilance. *Ther Innov Regul Sci.* 2016;50(4):464–71. <https://doi.org/10.1177/2168479016634766>.
18. Obieze AI, Oreagba IA, Osakwe AI, Yusuf AA. Quality assessment of individual case safety reports in the Nigerian national pharmacovigilance centre database. *Pharmacoepidemiol Drug Saf.* 2017;26:319.
19. Rodriguez Y, Condori Y, Hernandez M. Assessing the quality of the individual case safety reports received at the pharmacovigilance center of Peru using *vigiGrade*. *Pharmacoepidemiol Drug Saf.* 2018;27:254.
20. Lee I, Jokinen JD, Crawford SY, Calip GS, Kilpatrick RD, Lee TA. Exploring completeness of adverse event reports as a tool for signal detection in pharmacovigilance. *Ther Innov Regul Sci.* 2021;55(1):142–51. <https://doi.org/10.1007/s43441-020-00199-z>.
21. Edwards IR, Lindquist M, Wiholm BE, Napke E. Quality criteria for early signals of possible adverse drug reactions. *Lancet.* 1990;336(8708):156–8. [https://doi.org/10.1016/0140-6736\(90\)91669-2](https://doi.org/10.1016/0140-6736(90)91669-2).
22. Aronson JK, Hauben M. Anecdotes that provide definitive evidence. *BMJ.* 2006;333(7581):1267–9. <https://doi.org/10.1136/bmj.39036.666389.94>.
23. Meyboom RHB, Egberts ACG, Edwards IR, Hekster YA, de Koning FHP, Gribnau FWJ. Principles of signal detection in pharmacovigilance. *Drug Saf.* 1997;16(6):355–65. <https://doi.org/10.2165/00002018-199716060-00002>.
24. European Medicines Agency. Important medical event terms list (MedDRA version 25.0). 2022. <https://www.ema.europa.eu/en/human-regulatory/post-authorisation/pharmacovigilance/signal-management>. Accessed 28 Jul 2023