Drug Safety
https://doi.org/10.1007/s40264-020-01028-w

ORIGINAL RESEARCH ARTICLE

Developing Crowdsourced Training Data Sets for Pharmacovigilance Intelligent Automation

Alex Gartland[1] · Andrew Bate[2] · Jeffery L. Painter[3] · Tim A. Casperson[4] · Gregory Eugene Powell[5]

Accepted: 23 November 2020

Abstract

Introduction Machine learning offers an alluring solution to developing automated approaches to the increasing individual case safety report burden being placed upon pharmacovigilance. Leveraging crowdsourcing to annotate unstructured data may provide accurate, efficient, and contemporaneous training data sets in support of machine learning.

Objective The objective of this study was to evaluate whether crowdsourcing can be used to accurately and efficiently develop training data sets in support of pharmacovigilance automation.

Materials and Methods Pharmacovigilance experts created a reference dataset by reviewing 15,490 de-identified social media posts of narratives pertaining to 15 drugs and 22 medically relevant topics. A random sampling of posts from the reference dataset was published on Amazon Turk and its users (Turkers) were asked a series of questions about those same medical concepts. Accuracy, price elasticity, and time efficiency were evaluated.

Results Accuracy of crowdsourced curation exceeded 90% when compared to the reference dataset and was completed in about 5% of the time. There was an increase in time efficiency with higher pay, but there was no significant difference in accuracy. Additionally, having a social media post reviewed by more than one Turker (using a voting system) did not offer significant improvements in terms of accuracy.

Conclusions Crowdsourcing is an accurate and efficient method that can be used to develop training data sets in support of pharmacovigilance automation. More research is needed to better understand the breadth and depth of possible uses as well as strengths, limitations, and generalizability of results.

Key Points

Wider deployment of machine learning in pharmacovigilance requires further algorithmic evaluations, and appropriate contemporaneous test sets are lacking.

Crowdsourcing has become a frequently leveraged approach to a wide range of challenges, we present its application to the review of public domain data of potential use to safety.

We evaluated the crowdsourced approach and showed it to be a scalable, rapid, and effective approach for developed annotated social media data.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s40264-020-01028-w.

✉ Gregory Eugene Powell
gregory.e.powell@gsk.com

[1] College of Medicine, University of Central Florida, Orlando, FL, USA

[2] Safety and Medical Governance, GlaxoSmithKline, London, UK

[3] JiveCast, Raleigh, NC, USA

[4] North American Medical Affairs, GlaxoSmithKline, Research Triangle Park, NC, USA

[5] Pharma Safety, GlaxoSmithKline, 5 Moore Dr., Research Triangle Park, NC 27709, USA

Published online: 22 December 2020

1 Introduction

The volume of individual case safety reports (ICSRs) seems to be ever increasing [1]. While data from spontaneous reporting systems undoubtedly remain the cornerstone of post-marketing safety surveillance, there is a lack of evidence that the increasing volumes of such ICSRs result in an increased ability to find safety signals [2].

With technological advances across many sectors, we are increasingly seeing automation introduced into routine processes. This includes, for example, the many transactional activities that occur in the financial service sector such as invoice processing [3]. The ability to automate elements of ICSR intake is of paramount importance given the ever-increasing pharmacovigilance (PV) ICSR burden [4]. In addition to more efficient handling of reports, a potential benefit would include a more consistent approach in the handling of reports with a transparent audit trail. Automation offers the potential to allow PV medical expertise to be more focused on activity where clinical input is critical and therefore far more likely to impact the risk-benefit balance.

Machine learning (ML) approaches are an appealing solution to many of the challenges of automation. The ability to 'learn' rules through training data without explicit "a priori" rules are of particular use with unstructured data, such as ICSRs, where the amount of combinations of variants of data entry makes a few simple decision rules challenging to define. Machine learning algorithms need data to 'learn' from training data, results are then produced on a validation set and generalized performance is demonstrated through application to a separate 'test' set. Reliable annotated training/test data of sufficient volume and generalizability are therefore essential for ML. Continual feedback through external annotation on the data classification is required to enable dynamic improvements and adaptations as unstructured data changes over time. This is of particular importance as PV data are constantly changing (often in a non-random manner as therapeutic advances occur, healthcare system changes are introduced, Weber Effect), thus the need to continuously update algorithms to ensure contemporaneous automation is of paramount importance.

As automation using ML and artificial intelligence is explored in PV, a challenge is how to best train the algorithm [5]. This is time consuming as it requires manual clinical review and, given that training and test data sets themselves need to be updated over time, this is far from facile. While ICSRs are in public domain depositories, the narratives are not normally available, at least in full detail. Sharing of full ICSR data including clinical narratives is challenging given difficulties in effective anonymization of free-text narratives at scale. Other public domain narratives with potential information on narratives include social media data. Comfort et al. [6] in developing an automated approach ICSR classifier to classify valid ICSRs from a set of social digital media data, estimated that the task of human manual review of the unannotated social media posts would be 44,000 work hours. Abatemarco et al. [7] built a 14,000 strong annotated corpus of ICSRs (of an originally planned 20,000 ICSRs), and were explicit about the challenges in the time, effort, and cost of producing high-quality data for training and testing automation techniques. While organizations may potentially share training data, and/or develop simulated data for some of the training needs, developing scalable approaches to the development of accurate training or test data sets is a current major limitation in PV and will be of great importance to facilitating the use of automation in PV. As PV, therapies, and healthcare evolve, there will be a need for ongoing development of new training/validation/test data to ensure ML algorithms for automation remain trained and appropriate to ongoing routine use, thus the capability to rapidly produce new good training data sets will be ever more important.

To develop automated approaches, it is necessary to start with training data sets relevant to PV. One such approach, crowdsourcing, was examined as a potential solution in transforming a large amount of unstructured data, using discussions of adverse events on social media as an exemplar, to structured accurately classified data that can subsequently be used to develop scalable contemporaneous algorithms to support automation.

Crowdsourcing is defined as the practice of obtaining needed services, ideas, or content by soliciting contributions from a large group of people, especially from the online community, rather than from traditional employees or suppliers [8]. The use of crowdsourcing for health-related activities and across the medicine development lifecycle is still an emerging area, but preliminary results have been promising [9–11]. For example, the study by MacLean et al. concluded that crowd-labeled data is a scalable and effective technique for automatically identifying medical terms in patient-authored text [12]. There has only been limited exploration of crowdsourcing in PV [13]. This study looked to investigate whether crowdsourcing executed on social media data could be an effective tool in the development of training data sets that would then ultimately be used for PV automation algorithms.

2 Methods

2.1 Reference Dataset

Our team began evaluating the usefulness of social media posts for PV by examining commercially available, anonymized Facebook and Twitter posts from Epidemico that contained mentions of 15 products of interest over a 1-year period (1 September, 2013–31 August, 2014). The 15 products were chosen as a diverse representation of prescription and over-the-counter drugs. Initially, over 212,000 posts were identified for inclusion in the study. We looked for alternative methods to reduce the volume of these data so that it could be reviewed by our curation team (experienced safety physicians and scientists) in a reasonable amount of time. Language-detection algorithms were applied to identify only the posts written in English. In addition, steps were taken to remove duplicate posts and to reduce potential noise due to spam or advertising. These methods reduced the dataset to 81,300 posts, which were still too numerous to review within the timeline of the project. Approximately 19% of posts (15,490 posts) were selected by random sampling to be read and reviewed by our curation team.

Prior to reviewing the data, 22 medical topics relevant to PV were identified, to test whether these types of topics could be mined from social media data. The wide variety of topics included the type of person who was posting (e.g., patient, family member, healthcare provider, or friend), socio-economic indicators (e.g., occupation, level of education), and contributing health factors (e.g., pregnancy, smoking, or alcohol use). Additional items of interest included more complex topics such as drug indication/dose, adverse events, and efficacy. These 22 topics were evaluated across the set of all 15 drug products.

The curation process was completed using a software tool called Insight Explorer, which had been built by researchers at GSK to help manage the curation process [14]. The team comprised 13 curators who worked over a 5-month period (November 2014 to March 2015) to create our reference dataset. The software helped to track the total time spent to review the 15,490 posts and reported that it took 333 hours averaging ~ 1.3 min per post.

2.2 Experimental Design

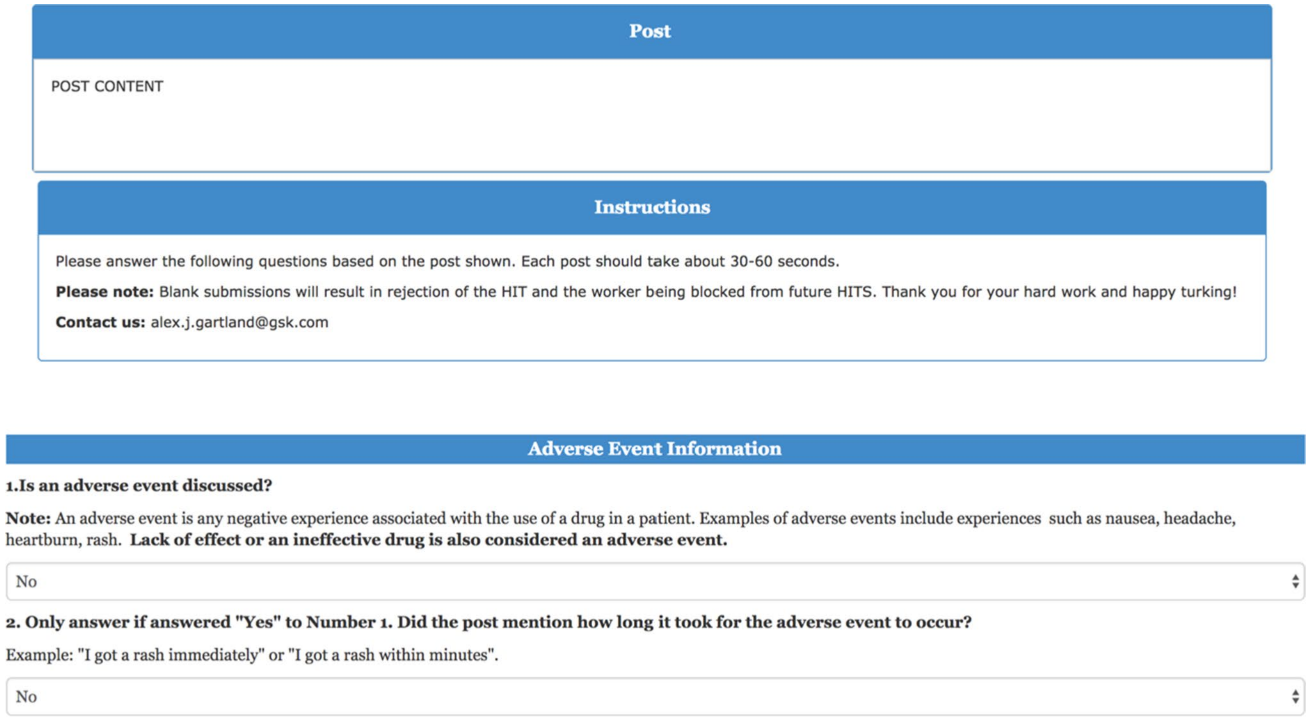
Crowdsourcing is the process of dividing a large project or task into small assignments and outsourcing them to a large number of people (i.e., a crowd), typically using the Internet [14]. Amazon's "Mechanical Turk" (MTurk) is one of the more popular crowdsourcing platforms and was the one chosen for this study [15]. Mechanical Turk has been extensively used for research in other fields including social science for several years [16–19]. The workers, whose identities are hidden from the requestor, are known as "Turkers." They complete assignments known as Human Intelligence Tasks (HITs). A HIT is a task that cannot easily be completed by a computer, requiring a real person to read, understand, and respond to one or more questions [20].

For this scientific methodological study, the same 22 medical concepts (see Table 1 in Section 3.1 for the list) being analyzed by the curation team were also given to the Turkers, using instructions written at a level that could be easily understood. The project was divided into two phases. The first phase tested the impact of price on the accuracy and efficiency of Turker reviews, when compared to the experts. This was done by (1) having a post reviewed by either one or multiple Turkers and (2) varying the price paid per post reviewed. In phase II, the questionnaire, reward per HIT, and number of reviewers per post were adjusted based on the results of phase I. Additionally, the accuracy and time efficiency of the Turkers were evaluated on a larger, more representative scale by increasing the number of posts reviewed from 500 to 5000.

The MTurk system allowed us to create a new project with relative ease. For our study, we developed a single survey form to use as our template, which contained the medical concepts to be evaluated by the Turkers (see Fig. 1). The MTurk system generated unique HITs based on this template after we uploaded the dataset containing our randomly sampled, anonymized social media posts. Once all HITs were completed, MTurk handled the collation of results for further analysis.

MTurk allows the requester to specify how many times each HIT should be reviewed, making it simple to test whether multiple reviews add value to the process. Additional criteria specifying which Turkers can complete a HIT may be enforced. For this study, we selected workers who reside in the USA to increase the likelihood of English language proficiency.

The Turkers were presented with an interface that displayed the social media post content to be reviewed and the 22 questions to identify the medical topics (such as, "Is more than one medication mentioned within the post?" or "Is illicit drug use mentioned within the post?") that were the same as those identified by the GSK curation team (see Fig. 1). All but one of the attributes (poster type) were modified to binary outcomes (Yes/No) for the Turkers to answer. This reduced the need for Turkers to have extensive training in pharmacology and related medical disciplines and made comparison to the reference dataset more straightforward.



Post

POST CONTENT

Instructions

Please answer the following questions based on the post shown. Each post should take about 30-60 seconds.

Please note: Blank submissions will result in rejection of the HIT and the worker being blocked from future HITS. Thank you for your hard work and happy turking!

Contact us: alex.j.gartland@gsk.com

Adverse Event Information

1.Is an adverse event discussed?

Note: An adverse event is any negative experience associated with the use of a drug in a patient. Examples of adverse events include experiences such as nausea, headache, heartburn, rash. **Lack of effect or an ineffective drug is also considered an adverse event.**

No

2. Only answer if answered "Yes" to Number 1. Did the post mention how long it took for the adverse event to occur?

Example: "I got a rash immediately" or "I got a rash within minutes".

No

Fig. 1 Amazon's "Mechanical Turk" (MTurk) user interface. A Turker interface was designed to mimic the social media curation interface, Insight Explorer, that was used by the expert safety scientists in the creation of the gold standard data set [11]. For each Human Intelligence Task (HIT), a single post would appear in the box labeled "Post" followed by overall instructions and detailed directions on how to answer each of the 22 questions corresponding to the 22 medical topics of interest

2.3 Phase I

In phase I of the study, 500 posts were randomly sampled from the reference dataset and uploaded to MTurk. This phase tested the influence of price elasticity upon Turker efficiency and accuracy. Additionally, this phase was used to understand whether there was a difference in accuracy when comparing (1) having multiple Turkers review a single post and aggregating their results through a voting system or (2) using only a single Turker to review a post. The sample data to be reviewed were published at various reward levels per HIT.

To test our voting system method, each post was reviewed by three unique Turkers in this phase. The quality of completed HITs was reviewed to ensure that the Turker had not left any questions blank and had spent an adequate amount of time reviewing the post ($>$ 10 s). Tasks that failed this quality test were rejected and republished until successfully completed.

2.4 Phase Two

In phase II of the study, 5000 posts were randomly sampled from the reference dataset and uploaded to MTurk. Phase I results, which are discussed in detail later, showed that neither a reward per assignment nor having a post reviewed by multiple Turkers had a statistically significant effect on overall match to the reference dataset. In phase I, two questions (regarding adverse events and poster type) had an agreement rate below 80%. For phase II, the wording around those attributes was revised to help Turkers better understand how to answer those questions appropriately, the specific changes were:

- In phase I, a lack of effect was listed as a possible adverse event but was not specifically clarified for Turkers. In phase II, the wording was changed to say "Lack of effect or an ineffective drug is also considered an adverse event".
- In phase I, poster type was defined as the relationship between the author and person receiving the medication. However, in phase II, each of the seven poster types was specifically defined as to aid the Turker.

In phase II, each social media post was reviewed by just one Turker. The reward level for phase II was raised, which had an added benefit by making our HITs attractive enough to be completed in a timely manner. As with phase I, all results were checked for quality assurance; if they failed, those results were rejected and republished until successfully completed.

2.5 Data Analysis

In phase I, a majority voting algorithm was implemented to compare Turker results to the reference dataset. For each post, all three Turker responses were compared across all 22 attributes. If more than one Turker voted yes for a single attribute, then the plurality ruled for a "Yes" response to that question. Only the "poster type" attribute was computed differently because it was not a binary outcome; the default response was set to "Unknown" unless two or more Turkers agreed on a single classification, such as "HCP" or "Patient". The plurality vote for each post was then compared to the reference dataset response for match and accuracy.

A matrix of all Turker responses was created with columns set to the attribute type and row entries populated with a (1) for a match and (0) otherwise. In phase I, there were 500 social media posts tested, each with 22 attributes, for a total of 11,000 entries in the matrix. A simple analysis was then performed to find the overall percent match to the reference dataset. A similar process was followed for phase II, except that there was no plurality vote required because each of the 5000 posts was reviewed only once. This resulted in a matrix of 110,000 entries, and the same analysis performed for phase I was computed against these results.

The total time for Turkers to complete the HITs was computed to the nearest hour by determining the duration between when the first and last HITs were completed. Finally, we computed the percentage of false-positive and false-negative errors for the phase II results.

3 Results

3.1 Phase I Results

Accuracy for the phase I batches (1500 posts across three batches) was computed to be 92.8% (with a range of 92.6–92.9), remaining steady across varying reward amounts. Turkers struggled the most with correctly identifying (1) whether there was an adverse event (82–83% accuracy) and (2) the poster type attribute (71–73% accuracy).

Table 1 Phase I summary statistics: agreement to a reference data set

Question name	Number of posts	Matched	Yes/no	Match %
AE information				
Proto-AE	1500	1237	474/1500	82.47%
Time to onset	1500	1461	21/1500	97.40%
Outcome	1500	1402	57/1500	93.47%
Poster type	1500	1087	Multiple categories	72.47%
Post mentions				
PII*	1500	1465	18/1500	97.67%
Concomitant Medications	1500	1355	291/1500	90.33%
Occupation	1500	1481	18/1500	98.73%
Education	1500	1500	0/1500	100.00%
Smoking	1500	1464	54/1500	97.60%
Alcohol use	1500	1486	15/1500	99.07%
Illicit drug use	1500	1490	12/1500	99.33%
Medical history	1500	1407	93/1500	93.80%
Pregnancy	1500	1490	18/1500	99.33%
Health services Information	1500	1436	108/1500	95.73%
Seeking information	1500	1430	108/1500	95.33%
Drug abuse	1500	1484	12/1500	98.93%
Product complaint	1500	1436	45/1500	95.73%
Product information				
Route	1500	1268	261/1500	84.53%
Formulation	1500	1238	375/1500	82.53%
Dosing	1500	1415	108/1500	94.33%
Indication	1500	1265	639/1500	84.33%
Benefit discussed	1500	1325	261/1500	88.33%
Total questions	33,000	30,622	3813/33,000	92.79%

AE adverse event, *PII* Personally identifiable information

This decrease in accuracy is likely owing to the increasing complexity of these concepts. For example, the average person may not realize “lack of effect” is considered an adverse event for regulatory reporting purposes. Further details and a concept-by-concept breakdown of the results from phase I can be found in Table 1. The time to complete the review of the 1500 posts (no difference seen across the batches, data not shown) was approximately 441 h.

3.2 Phase II Results

Phase II attempted to measure overall accuracy when using a larger sample size of posts (5000) that would be more representative of a real-world scenario, and also measured the duration to complete the HIT assignments at a higher reward level. It also tested whether improving the wording of the two questions about adverse events and poster type would help increase the accuracy of matching those attributes against the reference dataset.

In phase II, the Turkers achieved an overall match of 91.8% compared to the reference dataset (similar to the 92.8% overall accuracy results in phase I). In addition, all assigned HITs were completed and accepted within 33 hours. This was a dramatic reduction in time compared to phase I.

We can extrapolate that the time to complete all 15,490 posts would be less than a week, or about 5% of the total amount of time it took the experts to create the reference dataset. Further details of the results from phase II can be found in Table 2. Furthermore, to test that the classification was not biased toward either a yes or no answer, the percentages of false positives and false negatives were computed for all the binary variables (excluding “poster type”) and are shown in Table 3.

4 Discussion

To test the accuracy and efficiency of crowdsourcing to effectively and rapidly develop a PV training set through the review and annotate social media data, crowdsourced curation using MTurk was compared to a reference dataset of posts reviewed by a trained curation team (experienced safety physicians and scientists). A random sample of 5000 posts was uploaded to MTurk, where Turkers were asked to identify the same 22 attributes the curation team had been asked to detect. Curation of these 5000 posts was completed in 33 hours and had an overall agreement with the reference dataset of 91.8%.

4.1 Limitations

Although our results show the viability of using crowdsourcing for a review of social media posts, there are several potential limitations. First, social media data were used for this study, generalizability to other data sources is unknown. For this analysis, we included anonymized Facebook and Twitter social media data provided by an external vendor, Epidemico, who has conducted research into the use of social media for drug safety for many years, including various regulatory collaborations with the US Food and Drug Administration and in the public private Innovative Medicines Initiative-funded project on social media WEBRADR [21–25]. The landscape of available social media data is ever changing [26, 27], not only will new data sources emerge but some existing data sources are likely to change their terms of use over time and or make substantive changes to their data platform [28, 29]. The method we outlined in this paper will likely be generalizable for many sources of consumer data of relevance or potential relevance to PV, and perhaps beyond to for example, EMRs, particularly if crowdsourcing participation was limited to some healthcare/medical qualification. However, one needs to understand the terms of use for each data source as well as carefully evaluate data characteristics (e.g., consumer vs HCP data) and the context of use (e.g., drug safety) when deciding on the appropriateness of crowdsourced training data sets to ensure appropriate legal, ethical, and scientific use of the data. It should be noted our project underwent extensive internal reviews by our safety, privacy, and legal groups to ensure appropriate use, governance, and oversight.

Additionally, only basic medical/safety insights were evaluated (adverse events, medication usage, dose), thus the ability to identify more complicated insights (causality assessment, drug interactions) has yet to be determined. However, most social media posts come from non-medically trained individuals, thus the Turkers are likely to be somewhat representative of people who post on social media. The generalizability of our approach beyond general social media and simplistic medical/safety insights required further investigation. Additional “fine tuning” of crowdsourced training data sets by domain experts may be required for some activities (e.g., adding that a lack of effect should be considered an adverse event). However, crowdsourcing the initial training data set with subsequent refinement by domain experts should still take significantly less time than having domain experts create the entire data set from scratch.

Another limitation is some opaqueness in how the Turk review was conducted. There is no way to know the individual who is performing the work, whether they correctly entered their demographic information, and whether they are actually putting thought into answering questions or just clicking on buttons for monetary motivation. One safeguard with crowdsourcing is that you have the benefit of large numbers of testers, which should minimize the risks. As a result, one should consider crowdsourcing a

Table 2 Phase II summary statistics

Question name	Number of posts	Matched	Match %
AE information			
Proto-AE	5000	4177	83.5%
Time to onset	5000	4823	96.5%
Outcome	5000	4542	90.8%
Poster type	5000	3481	69.6%
Post mentions			
PII	5000	4906	98.1%
Concomitant medications	5000	4461	89.2%
Occupation	5000	4907	98.1%
Education	5000	4963	99.3%
Smoking	5000	4874	97.5%
Alcohol use	5000	4973	99.5%
Illicit drug use	5000	4949	99.0%
Medical history	5000	4800	96.0%
Pregnancy	5000	4968	99.4%
Health services information	5000	4728	94.6%
Seeking information	5000	4711	94.2%
Drug abuse	5000	4891	97.8%
Product complaint	5000	4694	93.9%
Product information			
Route	5000	4181	83.6%
Formulation	5000	3911	78.2%
Dosing	5000	4719	94.4%
Indication	5000	4031	80.6%
Benefit discussed	5000	4291	85.8%
Total questions	110,000	100,981	91.8%

AE adverse event, *PII* Personally identifiable information

large data set as well as placing a limit on the number of posts a single person can review (for our pilot, we capped the maximum amount a single person could have earned in an effort to minimize single-source bias/confounding). There may be value in routine activity to ensure perceived trust in outputs of having two or more people annotating the same post. One also has the ability to assess overall performance by comparing the verbatim free-text post to the Turk review output. One should nevertheless be aware of the trade-off between speed and cost and some unavoidable uncertainty with respect to the accuracy of the work.

The demographics of the Turkers are not well understood and likely to change non-randomly over time. Some studies have shown that most of the people are based in the USA, but their geographic location, age, and educational experience are unknown [15], as with other convenience samples, this can be very useful, but one needs to consider generalizability and use of the outputs carefully [30]. This limitation was mitigated to an extent by the use of a well-stablished, extensively studied crowdsourcing platform that is likely to be more stable to changes over time than newer, smaller, or less well-known platforms; additionally, there is extensive literature investigating the generalizability of this particular platform, for example [17, 31]. Nevertheless, one must be cognizant that cultural differences that may lead to unanticipated bias being introduced into the work could potentially limit the generalizability of the results; although it is not clear the extent to which this would be problematic in the use of outputs in training sets for developing ML solutions.

Table 3 False positives and false negatives: phase II

	Count out of 110,000	Percentage (%)	Posts
Correct match	100,981	91.8	5000
False positive	3665	3.3	5000
False negative	3824	3.5	5000

False-positive and false-negative computations exclude poster type because this is not a binary outcome

Accuracy of annotation and minimizing the gaming that might occur with crowdsourcing is of paramount importance to ensuring the best training data sets possible [32]. Our study demonstrated that having each post reviewed by three individuals (voting system) in phase I yielded similar accuracy as compared to having each post reviewed by a single individual in phase II (92.9% vs 91.8%). Some studies have shown similar results whereas others have shown greater variability in overall accuracy and differences between group and individual reviews, such as Good et al. who found the greatest improvement in performance was going from two to three annotators per post [33–39]. There are some inherent limitations of having posts reviewed by a single individual. We attempted to reduce the single source bias in our study by limiting the number of posts a single individual could review, nevertheless the distribution of work by individuals showed most individuals reviewed a very small number of posts whereas a small group of individuals reviewed a disproportionately larger group of posts. As a result, some bias may have still been introduced into our results. Activities that may improve accuracy and minimize bias may include, but are not limited to, using 'Master Turkers' (Turkers with significant positive feedback), limiting the number of posts that can be reviewed by a single person, using qualification tests to help with diversity, as well as using various techniques to help prevent/detect potential problems (e.g., wording questions in a manner that prevents simply clicking on the default answer, identifying outliers with respect to time to complete task).

4.2 Moral, Ethical, and Legal Considerations

The Amazon MTurk platform was used specifically for this scientific methodological study, and no promotional activity was conducted as part of the engagement. It should be noted that ethical questions have arisen with the use of crowdsourced platforms in general and specific to Amazon MTurk since this study was initiated, see for example [40, 41]. We are committed to conducting studies of the highest ethical standards. Therefore, for any future studies looking at this issue, further assessment of differential ethical approaches across crowdsourcing platforms is warranted.

In addition to these limitations, there are also a number of other ethical considerations. Publicly available de-identified posts for their social listening activities were used; however, posting these on a public site like mturk.com may draw undue attention to the posts. In theory, a Turker could copy the post verbatim and search for it on the Internet in an attempt to identify the original source of the post. If successful, the Turker might then have access to the original author's username and could potentially even respond to the post. If the approach herein were pursued further, then additional anonymization of the posts might be considered.

A final consideration is around setting the rate of pay. The perception of creating an Internet "sweatshop" is a real risk with any funded crowdsourcing approach [42]. Our approach to mitigating this risk was to benchmark pay in the much smaller phase I trial based on an initial set of assumptions and then adjust pay for the larger phase II trial based on empiric results to ensure an appropriate level of compensation. Additionally, we limited the maximum amount a single individual could earn in aggregate. However, it should be noted that a study by Cohen et al. showed that increasing the rate of pay for each completed task did not lead to a linear increase in the total amount of pay, as the more they paid per activity the longer the Turkers took to complete the task [43]. Clearly ensuring clarity on the purpose of the foundational research and developing test sets that could be shared across many organizations for foundational necessary methodological work for drug safety would seem less of a concern than potential commercial applications. Using established, widely known, tested crowdsourcing applications with appropriate processes and safeguards in place clearly reduces the risk. As there is more experience with a specific application, for example, the time needed to conduct certain tasks appropriately, further safeguards can be included to ensure that participants are remunerated appropriately. Although crowdsource workers may not feel that they are being taken advantage of, ensuring participants are appropriately recompensed in all future applications is critical [44, 45].

Finally, there is very little guidance on the use of technologies such as crowdsourcing. In the meantime, the best approach is for stakeholders is to share lessons learned and begin to develop industry-wide best practices on the use of crowdsourcing.

5 Conclusions

Crowdsourcing is an accurate and efficient method that can be used to develop training data sets in support of PV automation. More research is needed to better understand the breadth and depth of possible uses as well as the strengths, limitations, and generalizability of results.

Acknowledgements The authors acknowledge the following current and former GSK safety scientists and physicians for their help in developing the reference dataset used in this study: Laurie Anderson, Heidi Bell, James Blowers, Linda Clayton, Amy Curry, Lynn Henson, Sandra McPhail, Lorrie Schifano, Soraya Shaikh, Eric Smith, Michele Thomas, and Nancy Yuen. The authors thank Arooj Akhtar and Richard Le for their work on this project, including assisting with the study design and planning the analysis. The authors also thank Epidemico for the provision of the social media used in this analysis, and the anonymized Turkers who conducted the annotations. The opinions and conclusions in this study are not necessarily those of GSK, nor the vendor Epidemico.

Declarations

Funding No external funding was received for conduct of this study or the preparation of this article.

Conflict of interest Greg Powell and Andrew Bate are current full-time employees and stockholders of GlaxoSmithKline (GSK), Jeffery Painter and Tim Casperson were full-time employees and stockholders of GSK at the time the study was conducted, and Alex Gartland was a student at the University of North Carolina at Chapel Hill and an intern at GSK when the study was completed.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

Data availability Social media data were provided by Epidemico.

Code availability Not applicable.

Author contributions GP, JP, TC, and AG jointly planned, analyzed the results, drafted, and edited the article. AB analyzed the results and drafted and edited the article. All authors read and approved the final version.

References

1. 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. Pharm Med. 2019;33(6):499–510.

2. Bate A, Hornbuckle K, Juhaeri J, Motsko SP, Reynolds RF. Hypothesis-free signal detection in healthcare databases: finding its value for pharmacovigilance. Ther Adv Drug Saf. 2019;5(10):2042098619864744.
3. Li Y, Muthiah M, Routh A, Dorai C. Cognitive computing in action to enhance invoice processing withcustomized language translation. In: Proceedings of the 2017 IEEE international conference on cognitive computing (ICCC), 25−30 June 2017, Honolulu;2019. p. 136–139.
4. Ghosh R, Kempf D, Pufko A, Martinez LFB, Davis CM, Sethi S. Automation opportunities in pharmacovigilance: an industry survey. Pharm Med. 2020;34(1):7–18.
5. Lewis DJ, McCallum JF. Utilizing advanced technologies to augment pharmacovigilance systems: challenges and opportunities. Ther Innov Regul Sci. 2020;54(4):888–99.
6. Comfort S, Perera S, Hudson Z, Dorrell D, Meireis S, Nagarajan M, et al. Sorting through the safety data haystack: using machine learning to identify individual case safety reports in social-digital media. Drug Saf. 2018;41(6):579–90.
7. Abatemarco D, Perera S, Bao SH, Desai S, Assuncao B, Tetarenko N, et al. Training augmented intelligent capabilities for pharmacovigilance: applying deep-learning approaches to individual case safety report processing. Pharm Med. 2018;32(6):391–401.
8. Merriam Webster Dictionary. https://www.merriam-webster.com/dictionary/crowdsourcing. Accessed 18 Oct 2017.
9. Khare R, Burger JD, Aberdeen JS, et al. Scaling drug indication curation through crowdsourcing. Database (Oxford). 2015;2015:bav016.
10. Khare R, Good BM, Leaman R, et al. Crowdsourcing in biomedicine: challenges and opportunities. Brief Bioinform. 2016;17(1):23–32.
11. Bentzien J, Bharadwaj R, Thompson DC. Crowdsourcing in pharma: a strategic framework. Drug Discov Today. 2015;20(7):874–83.
12. MacLean DL, Heer J. Identifying medical terms in patient-authored text: a crowdsourcing-based approach. J Am Med Inform Assoc. 2013;20:1120–7.
13. Bate A, Beckmann J, Dodoo A, Härmark L, Hartigan-Go K, Hegerius A, et al. Developing a crowdsourcing approach and tool for pharmacovigilance education material delivery. Drug Saf. 2017;40(3):191–9.
14. Casperson TA, Painter JL, Dietrich J. Strategies for distributed curation of social media data for safety and pharmacovigilance. In: Proceedings of the international conference on data mining (MDIN 2016); 27 July 2016; Las Vegas (NV).
15. Ross J, Irani I, Silberman M, et al. Who are the crowdworkers? Shifting demographics in Amazon Mechanical Turk. In: ACM CHI conference, April 2010, Atlanta; 2010. p. 2863–2872.
16. Mason W, Suri S. Conducting behavioral research on Amazon's Mechanical Turk. Behav Res Methods. 2012;44:1–23.
17. Paolacci G, Chandler J, Ipeirotis PG. Running experiments on Amazon Mechanical Turk. Judgm Decis Mak. 2010;5(5):411–9.
18. Crump MJ, McDonnell JV, Gureckis TM. Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. PLoS ONE. 2013;8(3):e57410.
19. Cheung JH, Burns DK, Sinclair RR, Sliter M. Amazon Mechanical Turk in organizational psychology: an evaluation and practical recommendations. J Business Psychol. 2017;32(4):347–61.
20. Introduction to Amazon Mechanical Turk. Amazon Mechanical Turk developer guide. Amazon Web Services; 2018. https://docs.aws.amazon.com/AWSMechTurk/latest/AWSMechanicalTurkGettingStartedGuide/SvcIntro.html. Accessed 12 Feb 2020.
21. US FDA. Enhancing tobacco surveillance through online monitoring. https://www.fda.gov/tobacco-products/research/enhancing-tobacco-surveillance-through-online-monitoring. Accessed 30 Nov 2020.
22. Pierce CE, Bouri K, Pamer C, et al. Evaluation of Facebook and Twitter monitoring to detect safety signals for medical products: an analysis of recent FDA safety alerts. Drug Saf. 2017;40(4):317–31. https://doi.org/10.1007/s40264-016-0491-0.
23. Policy & Medicine. FDA releases MedWatcher reporting for healthcare providers, patients and caregivers. https://www.policymed.com/2014/01/fda-releases-medwatcher-reporting-for-healthcare-providers-patients-and-caregivers.html. Accessed 30 Nov 2020.
24. van Stekelenborg J, Ellenius J, Maskell S, et al. Recommendations for the use of social media in pharmacovigilance: lessons from IMI WEB-RADR. Drug Saf. 2019;42:1393–407. https://doi.org/10.1007/s40264-019-00858-7.
25. Pierce CE, de Vries ST, Bodin-Parssinen S, Härmark L, Tregunno P, Lewis DJ, et al. Recommendations on the use of mobile applications for the collection and communication of pharmaceutical product safety information: lessons from IMI WEB-RADR. Drug Saf. 2019;42(4):477–89. https://doi.org/10.1007/s40264-019-00813-6.
26. Öztamur D, Karakadılarilhan IS. Exploring the role of social media for SMEs: as a new marketing strategy tool for the firm performance perspective. Proc Soc Behav Sci. 2014;150:511–50.
27. Dey L, Haque SM, Khurdiya A, Shroff G. Acquiring competitive intelligence from social media. In: Proceedings of the 2011 joint workshop on multilingual OCR and analytics for noisy unstructured text data; 2011: p. 3.
28. Facebook. An update on our plans to restrict data access on Facebook. https://about.fb.com/news/2018/04/restricting-data-access/. Accessed 30 Nov 2020.
29. Rosen A. Tweeting made easier. https://blog.twitter.com/en_us/topics/product/2017/tweetingmadeeasier.html. Accessed 30 Nov 2020.
30. Landers RN, Behrend TS. An inconvenient truth: arbitrary distinctions between organizational, Mechanical Turk, and other convenience samples. Ind Organ Psychol. 2015;8(2):142–64.
31. Ipeirotis PG. Demographics of mechanical turk (March 2010). NYU Working Paper No. CEDER-10-01, Available at SSRN: https://ssrn.com/abstract=1585030
32. Suri S, Goldstein DG, Mason WA. Honesty in an online labor market. Proceedings of the 3rd Human Computation Workshop (HCOMP); August 2011; San Francisco (CA).
33. Ipeirotis PG, Provost F, Wang J. Quality management on Amazon Mechanical Turk. In: HCOMP '10: Proceedings of the ACM SIGKDD Workshop on Human Computation. Washington, DC; 2010. p. 64–67. https://doi.org/10.1145/1837885.1837906
34. Bentley FR, Daskalova N, White B. CHI EA '17: Proceedings of the 2017 CHI Conference extended abstracts on human factors in computing systems; 2017; pp. 1092–9.
35. Buhrmester M, Kwang T, Gosling S. Amazon's Mechanical Turk: a new source of inexpensive, yet high-quality, data? Perspect Psychol Sci. 2011;1:3–5.
36. Nowak S, Rüger S. How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In: Proceedings of the international conference on multimedia information retrieval; 2010; pp. 557–66.
37. Hsueh PY, Melville P, Sindhwani V. Data quality from crowdsourcing: a study of annotation selection criteria. In: Proceedings of the NAACL HLT 2009 workshop on active learning for natural language processing; 2019: p. 27–35.
38. Zubiaga A, Liakata M, Procter R, Bontcheva K, Tolmie P. Crowdsourcing the annotation of rumourous conversations in social media. In: Proceedings of the 24th international conference on World Wide Web; 2015; pp. 347–53.
39. Good BM, Nanis M, Wu C, Su AI. Microtask crowdsourcing for disease mention annotation in PubMed abstracts. Pacific

Symposium on Biocomputing Co-Chairs, 3–7 January 2014, Fairmont Orchid, Big Island of Hawaii; 2014, p. 282–293.

40. Bourhis P, Demartini G, Elbassuoni S, Hoareau E, Rao HR. Ethical challenges in the future of work. Data Eng. 2019;55:55–64.
41. Adda G, Cohen KB. Amazon Mechanical Turk: gold mine or coal mine. Comput Lingustics. 2017;37(2):2–10.
42. Newsweek. The internet creates a new kind of sweatshop. https://www.newsweek.com/internet-creates-new-kind-sweatshop-75751. Accessed 1 Dec 2020.
43. Cohen KB, Fort K, Adda G, et al. Ethical issues in corpus linguistics and annotation: pay per HIT does not affect hourly rate for linguistic resource development on Amazon Mechanical Turk. LREC Int Conf Lang Resour Eval. 2016;W40:8–12.
44. Busarovs A. Ethical aspects of crowdsourcing, or is it a modern form of exploitation. Int J Econ Business Admin. 2017;1(1):3–14.
45. Wertheimer A. Exploitation. In: Zalta EN, editor. The Stanford Encyclopaedia of Philosophy (Fall Edition). 2018. http://plato.stanford.edu/archives/fall2008/entries/exploitation/. Accessed 30 Nov 2020.