

Original Paper

# Quantifying the Severity of Adverse Drug Reactions Using Social Media: Network Analysis

Adam Lavertu<sup>1,2</sup>, PhD; Tymor Hamamsy<sup>3</sup>, MS; Russ B Altman<sup>2,4,5</sup>, MD, PhD

<sup>1</sup>Biomedical Informatics Training Program, Stanford University, Stanford, CA, United States

<sup>2</sup>Department of Biomedical Data Science, Stanford University, Stanford, CA, United States

<sup>3</sup>Center for Data Science, New York University, New York, NY, United States

<sup>4</sup>Department of Bioengineering, Stanford University, Stanford, CA, United States

<sup>5</sup>Department of Genetics, Stanford University, Stanford, CA, United States

**Corresponding Author:**

Russ B Altman, MD, PhD

Department of Bioengineering

Stanford University

Shriram Center Room 213

443 Via Ortega MC 4245

Stanford, CA, 94305

United States

Phone: 1 650 725 0659

Email: [rbaltman@stanford.edu](mailto:rbaltman@stanford.edu)

## Abstract

**Background:** Adverse drug reactions (ADRs) affect the health of hundreds of thousands of individuals annually in the United States, with associated costs of hundreds of billions of dollars. The monitoring and analysis of the severity of ADRs is limited by the current qualitative and categorical systems of severity classification. Previous efforts have generated quantitative estimates for a subset of ADRs but were limited in scope because of the time and costs associated with the efforts.

**Objective:** The aim of this study is to increase the number of ADRs for which there are quantitative severity estimates while improving the quality of these severity estimates.

**Methods:** We present a semisupervised approach that estimates ADR severity by using social media word embeddings to construct a lexical network of ADRs and perform label propagation. We used this method to estimate the severity of 28,113 ADRs, representing 12,198 unique ADR concepts from the Medical Dictionary for Regulatory Activities.

**Results:** Our Severity of Adverse Events Derived from Reddit (SAEDR) scores have good correlations with real-world outcomes. The SAEDR scores had Spearman correlations of 0.595, 0.633, and  $-0.748$  for death, serious outcome, and no outcome, respectively, with ADR case outcomes in the Food and Drug Administration Adverse Event Reporting System. We investigated different methods for defining initial seed term sets and evaluated their impact on the severity estimates. We analyzed severity distributions for ADRs based on their appearance in boxed warning drug label sections, as well as for ADRs with sex-specific associations. We found that ADRs discovered in the postmarketing period had significantly greater severity than those discovered during the clinical trial ( $P < .001$ ). We created quantitative drug-risk profile (DRIP) scores for 968 drugs that had a Spearman correlation of 0.377 with drugs ranked by the Food and Drug Administration Adverse Event Reporting System cases resulting in death, where the given drug was the primary suspect.

**Conclusions:** Our SAEDR and DRIP scores are well correlated with the real-world outcomes of the entities they represent and have demonstrated utility in pharmacovigilance research. We make the SAEDR scores for 12,198 ADRs and the DRIP scores for 968 drugs publicly available to enable more quantitative analysis of pharmacovigilance data.

(*J Med Internet Res* 2021;23(10):e27714) doi: [10.2196/27714](https://doi.org/10.2196/27714)

**KEYWORDS**

social media for health; pharmacovigilance; adverse drug reactions; machine learning; network analysis; word embeddings; drug safety; social media

## Introduction

### Background

Adverse drug reactions (ADRs) are among the leading causes of mortality and morbidity in the United States, affecting hundreds of thousands of people and costing more than US \$500 billion every year in the United States alone [1,2]. An ADR is characterized as “an appreciably harmful or unpleasant reaction resulting from an intervention related to the use of a medicinal product” [3]. A drug’s ADRs are primarily derived from clinical trial data and augmented through postmarketing surveillance [4]. The ADR labeling process for each drug is based on the frequency of the ADRs in the treated populations and the severity of the outcomes associated with each ADR. The severity of an ADR is traditionally classified into one of three categories characterized by where it appears on the label: boxed warning, warnings and precautions, or adverse reactions, listed in decreasing order of associated severity [5]. Boxed warnings refer to “serious warnings, particularly those that lead to death or serious injury.” Under the existing system, death is severe enough to warrant a boxed warning, but so is restlessness; therefore, it would be useful for prescribers, patients, and researchers to have systems for recognizing that death is more severe than restlessness [6]. Similarly, warnings, precautions, and adverse reactions include a diverse spectrum of ADRs, and it can be difficult for patients, prescribers, and researchers to compare the risk profiles of different drugs.

The existing categorical definitions of ADR severity limit the ability of researchers and regulators to apply quantitative methods to regulatory and pharmacovigilance efforts. For instance, tracking the regulatory performance of drug safety efforts is primarily achieved through a categorical analysis of ADR case outcomes and drug label changes. The creation of a quantitative numerical scale that provides a relative severity score for each ADR would enable these analyses to leverage the powerful tools of quantitative decision and utility theory.

The pioneering study by Tallarida et al [7] examined the ability to create a continuous ADR severity scale. A total of 53 physicians were interviewed and asked to estimate the acceptable probabilities in a set of scenarios specifying risk-benefit trade-offs. These probabilities were then used to define a set of equations and the expected benefit to the patient in each scenario. Solving the system of equations resulted in numerical severity weights that were relative to the lowest ADR severity. The effort took approximately 45 minutes per interview and resulted in relative scores for seven ordinal categories of ADR severity.

The study by Gottlieb et al [8] expanded this ADR severity ranking task using the crowdsourcing platform Amazon Mechanical Turk. Individual workers were presented with a pair of ADRs and a link to more information on these ADRs and asked to select the ADR that they perceived to be more severe. Over 146 person-days, 2589 workers produced severity comparisons for 126,512 ADR pairs composed from 2929 unique ADRs. A linear programming algorithm was used to create a unified ADR severity ranking. To validate the ranking, the ADR severities were correlated with ADR case outcomes

from the Food and Drug Administration Adverse Event Reporting System (FAERS). The highest Spearman correlation,  $\rho=0.53$ , was associated with the proportion of cases resulting in death. This effort drastically increased the number of ADRs included in the severity ranking but was costly in terms of human time, labor, and money.

Recent work in natural language processing has resulted in numerous methods for creating vectorized word representations (also known as *embeddings*) that capture semantic meaning in a dense numerical representation [9,10]. These methods rely on the distributional hypothesis that word meaning is captured by the contexts in which a word appears [11]. The practical implication of the distributional hypothesis is that training a model to predict word context (ie, co-occurring word pairs) results in model weights that capture the meaning of the word. These model weights can then be used as a numerical representation of a given word.

Word embeddings learned on social media data sets have been deployed for pharmacovigilance previously but not for the purpose of exploring ADR severity [12]. By using a social media corpus generated by the general public, especially on pseudoanonymous social media platforms such as Reddit, researchers can capture meanings that reflect people’s unfiltered experiences of, and opinions about, health and disease [13,14]. Investigating the utility, benefits, and challenges of different social media platforms and methods for pharmacovigilance has been an area of active research [15]. Previous research on Twitter data annotated samples of tweets for personal medication intake versus individuals simply mentioning a drug [16,17]. Together, the 2 studies found that approximately 40% of the tweets mentioning a drug indicated that the individual tweeting was possibly personally taking that medication. Here, we focused on ADRs, but it is likely that the individuals discussing ADRs on social media have more direct experience with the ADRs, either by being directly affected by the ADRs or being informed about the ADR experience by a close relation, than individuals selected from a pool of crowdworkers. Word embeddings trained on a corpus generated by the general public can then be leveraged as a metric for public opinion in a similar fashion to previous crowdsourced approaches. We used the RedMed embeddings trained on more than 580 million health-enriched Reddit comments from more than 10 million users [18]. These numbers dwarf the number of votes gathered in a typical crowdsourcing experiment and are therefore potentially more indicative of a representative population’s perception of ADRs than traditional survey-based methods.

### Objective

In this study, we used publicly available word embeddings and a network-based label propagation method to estimate the severity of 12,198 ADR concepts from Medical Dictionary for Regulatory Activities (MedDRA) terminology. The resulting Severity of Adverse Events Derived from Reddit (SAEDR) scores were validated against human rankings as well as FAERS case outcomes. We used System Organ Classes (SOCs) and other groupings within the MedDRA to examine how the SAEDR severity scores differ at various levels of abstraction and within ADR categories. We used the SAEDR scores to

compare the severity among ADRs in different sections of drug labels, ADRs with disproportionate rates between sexes, and ADRs discovered at different stages of drug development. We combined the SAEDR scores with frequency information from Side Effect Resource (SIDER)-derived drug labels to generate drug-specific aggregate side effect severity scores [19]. The SAEDR scores enabled new analyses that were not possible with the existing categorical classifications.

## Methods

### Data Sources and Preparation

#### ADR Terms and Phrases

We sourced our ADR phrases and their synonyms from the *lowest-level terms* within the terminology of the MedDRA, version 22 [20]. We filtered terms based on their semantic types within the Unified Medical Language System Metathesaurus [21] and retained terms of the following semantic types: *Disease or Syndrome* (T047), *Finding* (T033), *Neoplastic Process* (T191), *Injury or Poisoning* (T037), *Pathologic Function* (T046), *Sign or Symptom* (T184), *Mental or Behavioral Dysfunction* (T048), and *Congenital Abnormality* (T019).

#### Word-Embedding Model

We used the word-embedding model from the RedMed project [18]. This model was selected because it was directly optimized for medical term similarity and was trained on a corpus of Reddit comments generated by the public and preprocessed to maximize the inclusion of ADR terms. ADRs that were a phrase such as “abdominal pain” were represented using the average embedding of all terms within the phrase. In previous work, not presented here, we found that the use of average embeddings, for example, “abdominal” + “pain”/2, produced better results than the use of embedded phrases, for example, “abdominal pain.”

#### Gottlieb Severity Data

We used the crowdsourced severity estimates included within the supplement of the study by Gottlieb et al [8], which included 2929 ADRs with rank scores.

### FAERS Data

All adverse drug event data were downloaded in JSON format from the openFDA website, which includes data from both the Legacy Adverse Event Reporting System and the FAERS. Data were retrieved on August 8, 2020, and included adverse event case reports up to June 30, 2020. Drug reactions were normalized to the MedDRA, version 22. We filtered to unique case IDs using the provided duplicated flag to remove duplicate reports and case reports that did not originate in the United States. Case outcomes were normalized according to the following schema: “Death”: {“Death”}, “Serious Outcome”: {“Life-Threatening,” “Hospitalization,” “Other Serious,” “Required Intervention”}, and “Disability”: {“Congenital Anomaly,” “Disability”}. We created an additional outcome category denoted *No Outcome* for cases with no reported outcome. We felt the need to create this category because the FAERS only allows for the reporting of serious case outcomes, and there is information in the absence of a serious outcome being reported. The outcome proportions for each ADR were calculated by dividing the number of cases reporting that ADR for a particular outcome by the total number of cases reporting that ADR.

### FAERS Severity Rankings

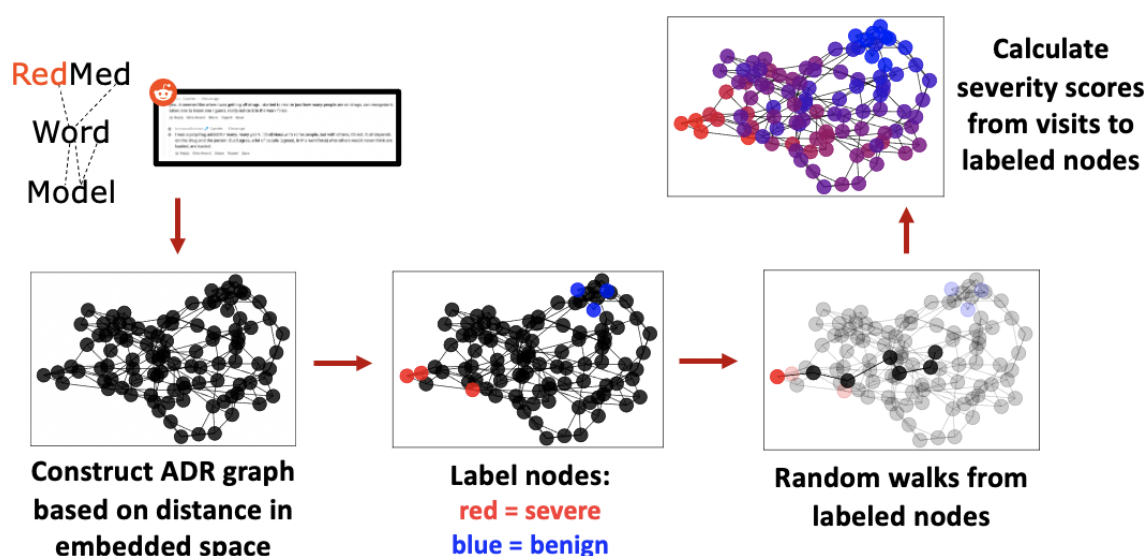
For the ranking of ADRs based on the FAERS data, we ranked ADRs based on their marginal likelihood to be included in a case with death or serious outcome as the reported case outcome. This was calculated by dividing the number of cases with the ADR that resulted in death or serious outcome by the number of cases without the ADR that resulted in these outcomes.

### Semisupervised Severity Propagation

#### Overview

Given a word-embedding model and a set of initial, potentially noisy, seed-word labels for the severe and benign categories, we sought to propagate severity information over the rest of the vocabulary, similar to the sentiment propagation methods described in the study by Hamilton et al [22]. A graphical overview of this method is presented in Figure 1.

**Figure 1.** Overview of the network method for estimating ADR severity from word embeddings. Word embeddings for 28,113 ADR terms and phrases were extracted from the RedMed word-embedding model. A network is constructed based on nearest-neighbors in the embedding space. A subset of nodes is labeled as severe or benign ADRs, and random walks from these labeled nodes are conducted. The Severity of Adverse Events Derived from Reddit scores for each ADR are calculated based on the relative number of encounters in random walks initiated from severe versus benign nodes. ADR: adverse drug reaction.



### Lexical Network Creation

The network was constructed by connecting words (nodes) with edges to their *k*-nearest neighbors based on the cosine distance between their embeddings. The corresponding edge weight was set to the cosine similarity of the two words (ie, the edge weight increases with term similarity).

### Seed Selection From Terminology

An initial set of ranked seed terms was labeled. Biomedical terminologies and ontologies often contain many terms that are highly similar to each other. To reduce the effect of lexical similarities on label propagation, we filtered the benign and severe seed terms based on the edit ratio among the terms in each list, as defined by the ratio function in the python-Levenshtein package [23]. Only terms that had an edit ratio >0.5 from a higher-ranking term were included in the list, resulting in a reduced list of lexically distinct seed terms. Both benign and severe seed term lists were truncated to the minimum seed term list size of both seed sets.

For instance, if the severe seed term list had [“death,” “cardiac arrest acute,” “cardiac arrest”] in rank order and the benign seed term list had [“dry skin,” “yawning,” “cold sweat”] in rank order, then, after lexical filtering, the severe seed term list would be [“death,” “cardiac arrest acute”], and the benign seed term list would remain unchanged. The final seed term lists would be truncated to the minimum seed term list size, resulting in a severe seed term list [“death,” “cardiac arrest acute”] and a benign seed term list [“dry skin,” “yawning”].

### Random Walks From Seed Nodes

Given a network and a set of seed nodes, we modified code from *node2vec* [24] to perform 5000 weighted random walks of length 200 from each seed node. We selected the number of random walks to ensure that all nodes within the graph were visited a nonzero number of times. We noted that this value

needs to be empirically determined and will likely change based on the size and structure of the network.

### SAEDR Score Calculation

The SAEDR score of a given node *u* is calculated using the following formula:

$$u_{\text{SAEDR}} = s_u / (s_u + b_u)$$

In the formula,  $s_u$  and  $b_u$  are the respective number of times that node *u* is encountered in a random walk from a severe node or a benign node. If node *u* was contained within one of our seed node sets, self-visits during the random walks were excluded from the calculation. To increase the robustness of these score estimates, we performed 10,000 iterations of bootstrap sampling [25] of the random walks, and the average of these bootstrap estimates was calculated. The SAEDR scores from multiple seed sets were averaged at the preferred term (PT) level, and the final combined scores were normalized to a zero to one range.

### Hyperparameter Tuning for Severity Propagation

We randomly split the crowdsourced severity data into training (1778/2369, 75.05%) and test (591/2369, 24.95%) sets, with 284 ADRs dropped because of mapping. We performed a grid search over the number of neighbors used to construct the lexical network {2, 5, 10, 15, 20, 25, 30} and the percentage of nodes to label for the severe and benign seed nodes {2, 5, 10, 15, 20, 25}. For example, an individual run would use the 25 nearest neighbors and 5% at each end of the training severity rank (ie, the top 5% most severe and bottom 5% least severe ADRs). We ultimately selected the parameters with the highest Spearman correlation with the training data.

### ADR Discovery Group Analysis

We analyzed the SAEDR score distributions of several different categories of ADRs.

### **ADRs Included in a Boxed Warning**

We downloaded the counts for the number of appearances of ADRs in the boxed warning section of drug labels from the supplement of the study by Wu et al [26]. ADRs that appear on a drug label based on their appearance in the SIDER but were not included in the list provided in the study by Wu et al [26] were considered to not have appeared in a boxed warning section [19].

### **ADRs With Disproportionate Reporting Between Sexes**

A set of ADRs shown to have sex differences was identified from the supplement of the study by Chandak et al [27]. We filtered out ADRs considered borderline (mean  $\log(\text{ROR}) < 0.4$  and mean  $\log(\text{ROR}) > -0.4$ ), as defined in the original study by Chandak et al [27]. Their study reported ADRs at the high-level group term (HLGT) level within the MedDRA, whereas our severity estimates were generated at the PT level. There are many PT ADRs per HLGT; therefore, all PTs within the specified HLGT terms were averaged to create a single SAEDR score for that HLGT. We did not use the drug information except to note distinct instances of sex-specific HLGTs (ie, HLGTs associated multiple times with different drugs each time).

### **ADRs Discovered at Different Stages of Drug Development**

ADRs discovered at the clinical trial stage were identified based on their inclusion in a drug label in the SIDER with a non-postmarketing ADR frequency [19]. Postmarketing ADRs were identified as those included on drug labels with postmarketing frequency information and no clinical trial frequency information. OFFSIDES and TWOSIDES ADRs were discovered from the FAERS data by looking at disproportionate reporting after correcting for the effects of demographics and other case information [28]. The OFFSIDES and TWOSIDES sets are the results of a study by Tatonetti et al [28] in which propensity score matching was used to identify matched cases and controls among the FAERS cases and to also identify potentially novel ADRs based on disproportional reporting metrics. This was done to discover ADRs that were associated with a single drug, OFFSIDES, as well as ADRs that were potentially the result of drug-drug interactions, TWOSIDES. We considered both OFFSIDES and TWOSIDES ADRs discovery-stage postmarketing ADRs because they have not yet been included in the drug labeling.

### **Drug-Risk Profile Scores**

The drug-risk profile (DRIP) scores were calculated based on our SAEDR score and the frequency of each ADR for a given drug. The DRIP score for a given drug is the sum of the severity multiplied by the frequency of each ADR on the drug label. Drug ADR frequency information was retrieved from the SIDER [19]. In situations where multiple ADR frequencies were indicated, 1000 frequency estimates were sampled from a uniform distribution with a lower bound at the minimum-reported frequency and an upper bound at the

maximum-reported frequency. When there was no frequency information, we sampled 1000 estimates from a uniform distribution over the interval of 0.001 to 0.01. The final DRIP score was the average score across 1000 samples.

## **Results**

### **Network Statistics**

We used the ADR terms and phrases from the MedDRA, version 22.0, lowest-level terms [20] as the initial lexicon for the lexical ADR network. Although the FAERS uses the MedDRA to encode ADRs, versioning differences over the years have resulted in some FAERS ADRs not being included in the MedDRA. We were able to generate embeddings for 92.35% (2450/2653) of the unique crowdsourced ADRs, 100% (14,045/14,045) of the unique FAERS ADRs, and 68.72% (28,113/40,905) of all filtered MedDRA terms. This resulted in a final lexical network with 28,113 nodes, representing 12,198 MedDRA PTs.

### **SAEDR Score Performance**

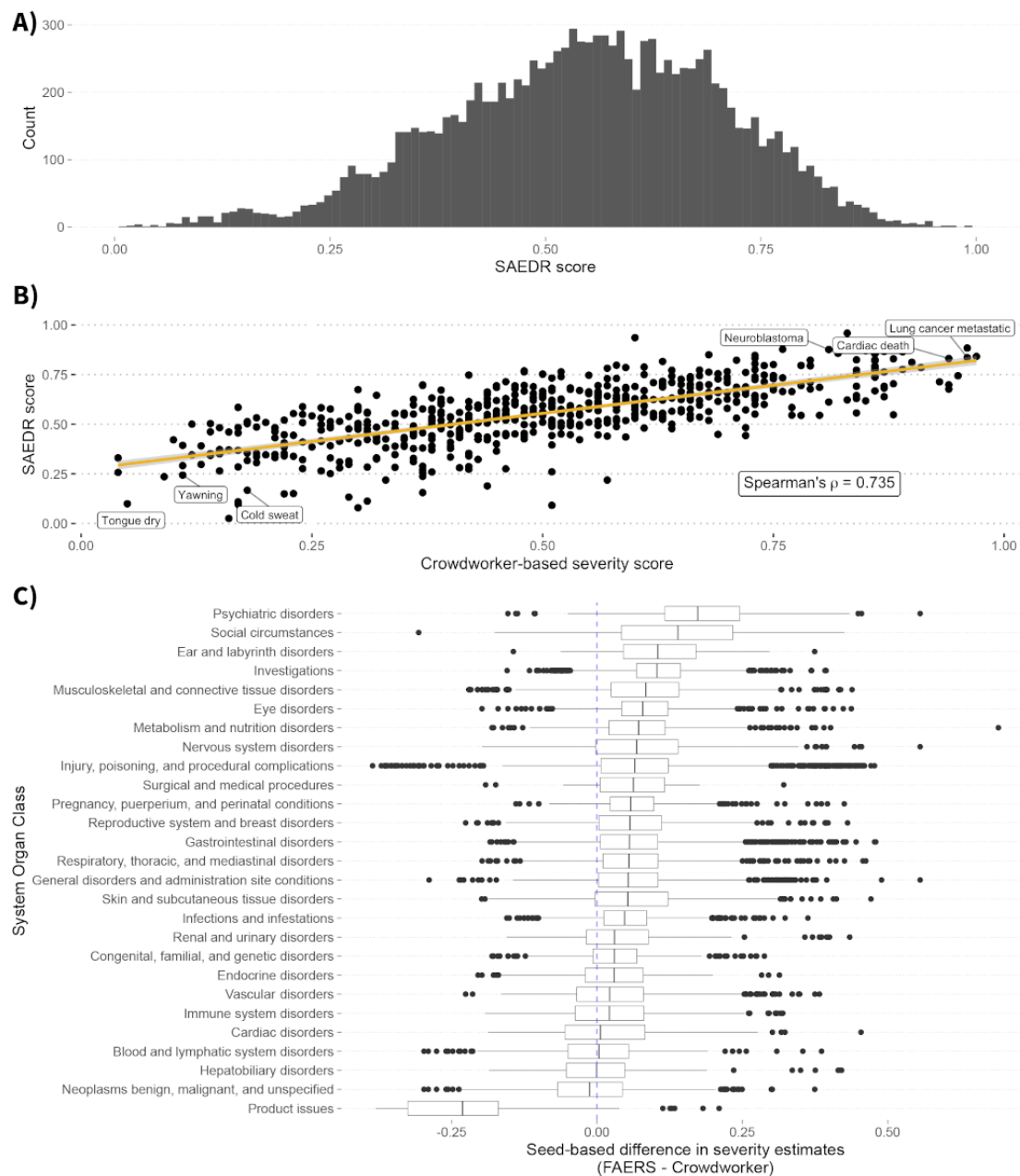
We compared our severity estimates using crowdworker-ranked seeds, FAERS-ranked seeds, and the average severity estimate across the two seed sets with two different ADR rankings: (1) a held-out test set of crowdworker-ranked ADRs ( $n=591$ ) and (2) ADRs ranked by case outcome statistics in the FAERS database.

The crowdworker-seeded severity estimates had the highest training performance with a 25 nearest-neighbors graph and using 10% of the most and least severe ADRs as seeds. The Spearman correlation with the crowdworker test set was 0.747 ( $P < .001$ ). The Spearman correlations for events with at least 100 reports in the FAERS were 0.595, 0.616, and  $-0.732$  for death, serious outcome, and no outcome, respectively, with  $P < .001$  for all (Figures S1 and S2 of [Multimedia Appendix 1](#)).

The FAERS-seeded severity estimates had the highest training performance with a 10 nearest-neighbors graph and using 10% of the most and least severe ADRs as seeds. The Spearman correlation with the crowdworker test set was 0.587. As no information from the crowdworker rankings was used to select these seeds, we can also report the Spearman correlation with the entire set of crowdworker ADRs: 0.765. The Spearman correlations for events with at least 100 reports in the FAERS were 0.509, 0.557, and  $-0.656$  for death, serious outcome, and no outcome, respectively, with  $P < .001$  for all (Figures S1 and S2 of [Multimedia Appendix 1](#)).

The SAEDR score—the average of the two severity estimates—had a Spearman correlation with the crowdworker test set of 0.735 ([Figure 2](#)). The Spearman correlations for events with at least 100 reports in the FAERS were 0.595, 0.633, and  $-0.748$  for death, serious outcome, and no outcome, respectively (Figures S1 and S2 of [Multimedia Appendix 1](#)). All these correlations were statistically significant ( $P < .001$ ).

**Figure 2.** Comparison of SAEDR scores and crowdworker severity estimates. (A) Histogram of SAEDR scores for 12,198 adverse drug reactions (ADRs). (B) Crowdworker severity estimates (x-axis) versus SAEDR scores (y-axis) for a test set of 591 ADRs. The SAEDR scores showed a strong Spearman correlation,  $\rho=0.735$ , with the human crowdworker rankings. We noted that this correlation is greater than the interrater correlation,  $\rho=0.71$ , reported in the original crowdsourcing study by Gottlieb et al [8]. A select set of the least and most severe ADRs based on the SAEDR score has been annotated. (C) Differences between severity estimates seeded with FAERS rankings and those seeded with crowdworker rankings (x-axis) for different System Organ Class groups (y-axis). The dashed blue line indicates where the severity would be the same for both estimates. FAERS: Food and Drug Administration Adverse Event Reporting System; SAEDR: Severity of Adverse Events Derived from Reddit.



### Seed-Ranking Comparisons

We examined the effects of seeding on the severity estimates of individual ADRs. Most ADRs were not substantially changed by the use of different seeds. We aggregated ADR severity differences at the MedDRA SOC level and found that the FAERS seeds led to increased severity estimates for *Psychiatric Disorders* and *Social Circumstances* (Figure 2). The most drastic shift in values was the decrease in severity of *Product Issue*-related ADRs using FAERS seeds compared with their severity when using crowdworker-based seeding.

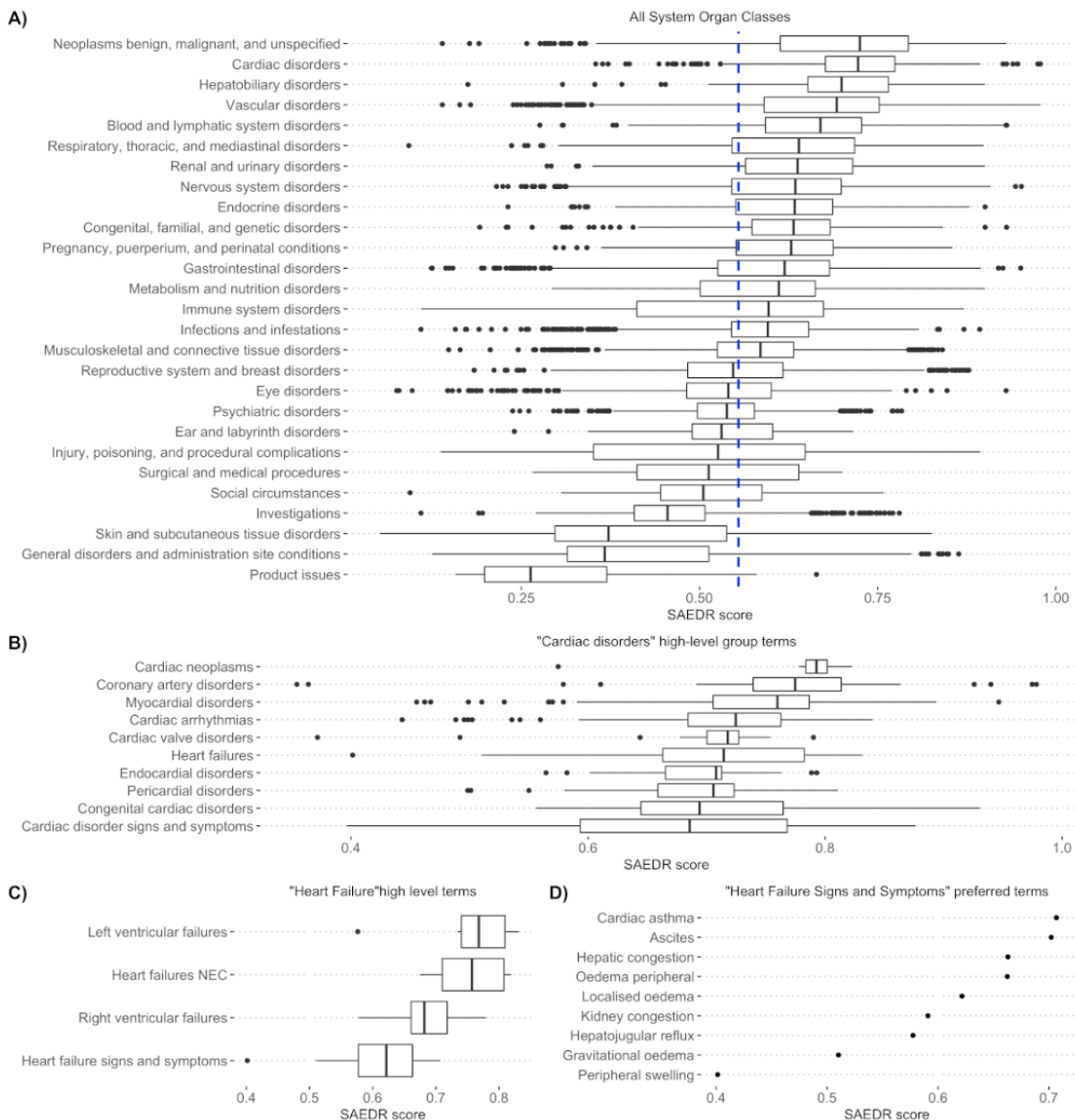
### SOC Severity Rankings

We compared the severity of ADRs based on the MedDRA SOC groupings as a qualitative evaluation (Figure 3). We found the ADRs related to cancer (neoplasms), cardiac, and liver (hepatobiliary) to be the most severe, with high SAEDR scores, whereas skin, general disorders, and product issues were considered the least severe, with low SAEDR scores. We performed a qualitative examination of ADRs at various other levels. We compared the severity distributions between the ADRs describing benign neoplasms and those describing malignant neoplasms (Figure S3 of Multimedia Appendix 1). A one-sided *t* test found malignant neoplasms to be significantly

more severe than benign neoplasms based on their SAEDR scores ( $P < .001$ ). We found a wide range in terms of severity within the different levels of the MedDRA ADR term groupings (Figure 3). We noted that among ADRs in the *Cardiac Disorders* HLGT, cardiac neoplasms have the highest SAEDR scores, whereas signs and symptoms of cardiac disorder have the lowest

relative SAEDR scores (Figure 3). When examining the level of individual ADRs within the *Heart Failure Signs and Symptoms* HLGT, we found that the ADRs that we would judge to be more severe were ranked higher than those that are more general and considered less severe.

**Figure 3.** Severity of adverse drug reactions (ADRs) at different group resolutions. (A) SAEDR score (x-axis) for 11,981 ADRs with System Organ Class (SOC) groups (y-axis). The dashed blue line indicates an overall median SAEDR score of 0.554. The SAEDR score distributions indicate that cancers and cardiac-related side effects are considered the most severe, with product issues being the least severe. (B) Severity distributions of the Medical Dictionary for Regulatory Activities high-level group terms (HLGTs) within the Cardiac Disorders SOC. There are large differences in severity within these term groups. (C) High-level terms within the Heart Failure HLGTs show a tighter distribution of severity, with "Heart Failure Signs and Symptoms" being the least severe within this group. (D) Individual side effects within the signs and symptoms of heart failure suggest cardiac asthma and ascites (accumulation of fluid in the peritoneal cavity) are the most severe symptoms, and peripheral swelling is the least severe symptom NEC: necrotizing enterocolitis; SAEDR: Severity of Adverse Events Derived from Reddit.

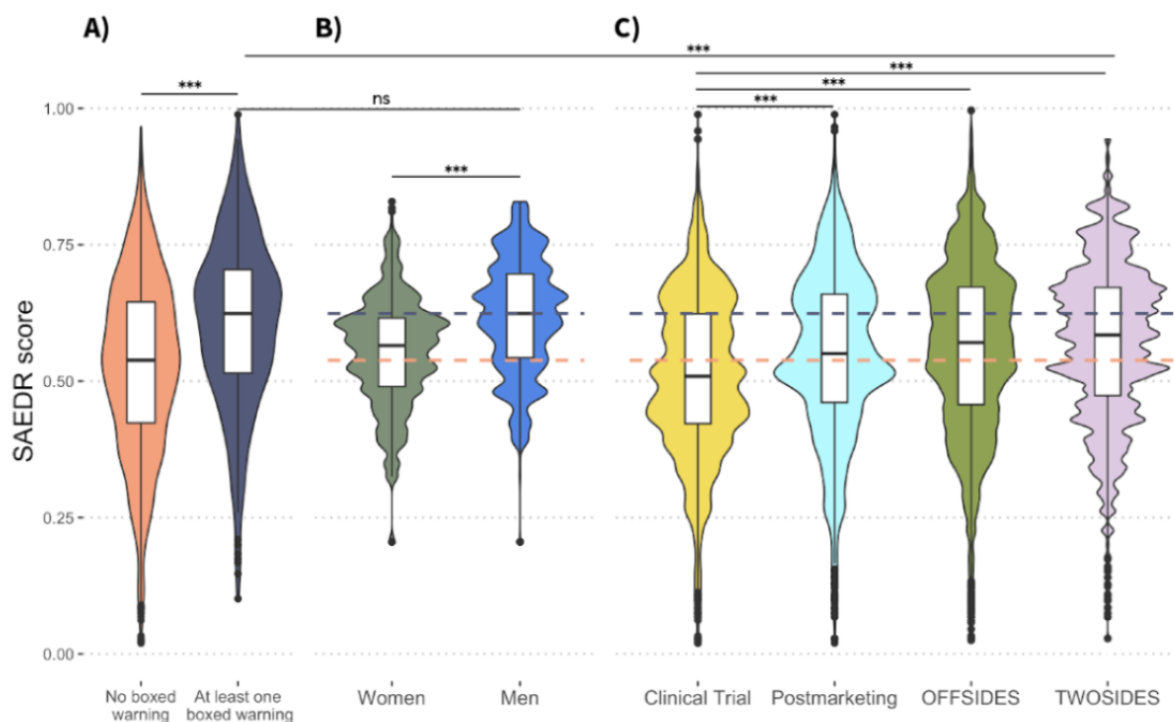


### Severity of ADRs Grouped by Labeling Section, Sex Effects, and Time Point of Discovery

On the basis of our estimate, we calculated that the ADRs that have never been included in a boxed warning had a median SAEDR score of 0.538, whereas the ADRs that had been

included in at least one boxed warning had a median SAEDR score of 0.624. We found that among the ADRs that have appeared on a drug label, those that have been included in at least one boxed warning were significantly more severe ( $P < .001$ ) than the ADRs that have never been included in a boxed warning section based on a one-sided  $t$  test (Figure 4).

**Figure 4.** Differences in adverse drug reaction (ADR) severity between ADR groupings and discovery periods: ADR groups (x-axis) versus SAEDR scores (y-axis). The grey dashed line indicates the median severity, 0.624, of the ADRs that have been included in a boxed warning. The orange boxed line indicates the median severity, 0.538, of the ADRs that appear on a drug label but have not been included in a boxed warning. (A) The ADRs that were listed as a black box warning at least once (n=356) were significantly more severe than those that have not appeared in a black box warning (n=3305). (B) The ADRs that are disproportionately reported for men (n=56,405) are significantly more severe than those disproportionately reported for women (n=50,801). There was no significant difference (ns) in severity between the ADRs included in black box warnings and those disproportionately reported for men. (C) The ADRs discovered in the postmarketing period (n=11,506) are significantly more severe (\*\*\*) than those discovered in the clinical trials (n=35,450). The ADRs identified in the postmarketing period through OFFSIDES (n=350,631) and the postmarketing polypharmacy ADRs identified through TWOSIDES (n=4,210,513) are significantly more severe than those discovered in the clinical trials. The severity of all postmarketing ADR groups is significantly less than the severity of the ADRs that have appeared in a black box warning. SAEDR: Severity of Adverse Events Derived from Reddit.



We found that the ADR HLGTS that are disproportionately reported for men were significantly more severe ( $P < .001$ ) than those disproportionately reported for women (Figure 4). The HLGTS disproportionately experienced by men were not significantly different from the severity of the ADRs that have been included in at least one boxed warning,  $P = .13$ , based on a one-sided  $t$  test.

We examined the severity of the ADRs that were found during the clinical trial versus the severity of those discovered in the postmarketing period based on the SIDER label annotations. We found that the ADRs discovered in the postmarketing period had significantly higher severity ( $P < .001$ ) than the ADRs discovered in the clinical trials based on a one-sided  $t$  test (Figure 4). We compared the severity of the clinical trial ADRs with the severity of those discovered in the postmarketing period using FAERS data. The study by Tatonetti et al [28] identified two different sets of ADRs: OFFSIDES is a set of ADRs disproportionately reported for a drug, whereas TWOSIDES is a set of ADRs disproportionately reported for a pair of drugs being taken concurrently, after correcting for case demographics and other information. The severity distribution of ADRs in both OFFSIDES and TWOSIDES was significantly higher ( $P < .001$ ) than the severity distribution discovered in the clinical trials based on a one-sided  $t$  test (Figure 4). We found that the ADRs associated with polypharmacy, through TWOSIDES,

had the highest severity of the postmarketing ADRs. All postmarketing ADRs were significantly lower in severity than the boxed warning ADRs based on a one-sided  $t$  test.

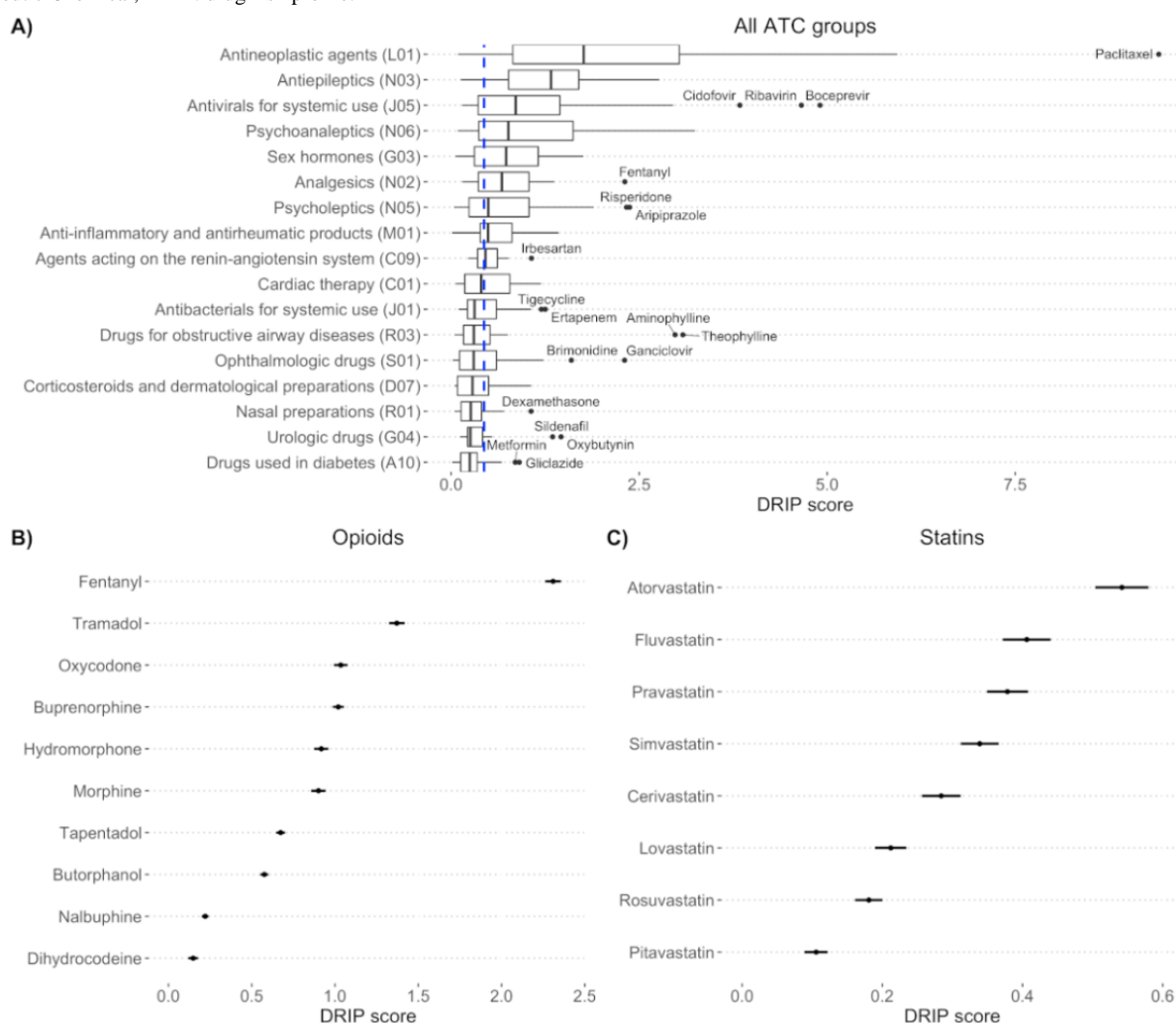
### DRIP Score Analysis

We calculated the DRIP scores for 968 drugs using SIDER label data, with a resulting median DRIP score of 0.439. The Spearman correlation between the drugs ranked by the proportion of FAERS cases, with that drug as the primary suspect that resulted in death, and the drugs ranked by our DRIP scores was 0.377,  $P < .001$ .

We analyzed DRIP score distributions by the Anatomical Therapeutic Chemical (ATC) classification system for the subset of drugs with an ATC designation (Figure 5). Antineoplastic and antiepileptic drugs were the drug classes with the highest DRIP scores, indicating that the drugs in these classes have severe ADR profiles. Drugs used in the management and treatment of diabetes and urologic issues had the lowest DRIP scores, indicating that these drugs are relatively safe and have less-severe side effects. We examined individual drugs within the opioid class and found that fentanyl had a markedly higher DRIP score than other drugs in the same class (Figure 5). Drugs within the statin class were primarily below the overall median DRIP score (0.439), with only atorvastatin being markedly above the median DRIP score (Figure 5).



**Figure 5.** DRIP scores calculated using side effect severity and frequency. (A) DRIP scores (x-axis) for ATC classification system groups (y-axis) ordered by median severity with outliers labeled. The dashed blue line indicates the median DRIP score of 0.439 across all drugs. (B) DRIP scores (x-axis) for opioid class drugs (y-axis), as denoted by the N02A ATC group. Fentanyl, an extremely potent synthetic opioid, is ranked first as having the most severe side effect. (C) DRIP scores (x-axis) for statins (y-axis), as denoted by the C10AA ATC group. Atorvastatin, the strongest statin in broad use, is ranked to have the most severe side effect, whereas lower-efficacy statins are ranked as having less-severe side effects. ATC: Anatomical Therapeutic Chemical; DRIP: drug-risk profile.



## Discussion

### Principal Findings

In this study, we present the SAEDR scores and quantitative estimates of the severity of 28,113 MedDRA ADR terms. The SAEDR scores have strong correlations with both crowdworker-based severity estimates and real-world ADR case outcome statistics. We generated these estimates using a network-based label propagation approach that required only a small percentage of the terms to be labeled. This study demonstrated the feasibility of using distantly supervised techniques such as label propagation to generate quantitative values for medical concepts.

Our approach enabled us to increase the number of ADRs with quantitative severity estimates from 2929 to 28,113, an almost tenfold increase, while minimizing the human time involved in generating the estimates. Our severity estimates can be updated routinely based on the recomputation of the embeddings and reseeding of the reference severe or benign ADRs. We noted

that our severity estimates had a higher correlation with the aggregate crowdworker rankings ( $\rho=0.735$ ) than the crowdworker rankings had among the three replicates in the original study ( $\rho=0.710$ ). In addition, our estimates had a greater correlation with real-world outcomes such as FAERS cases resulting in death ( $\rho=0.595$ ) than the crowdworker rankings ( $\rho=0.53$ ). Thus, our severity estimates track with both human judgments of ADR severity and real-world outcomes of consequence to patient health.

We found that our SAEDR scores had a strong negative correlation,  $\rho=-0.748$ , with no case outcome being reported in the FAERS. We assigned a *No Outcome* label to the FAERS cases for which none of the available but relatively severe case outcomes had been reported. We took this as an indication that the FAERS cases lacking outcomes may be a result of cases having an outcome below the level of seriousness that is reportable in the FAERS. FAERS case outcome completeness could potentially be improved by creating a lower-severity category for case outcome documentation. We also noted that

the SAEDR scores have a stronger correlation with serious outcomes in FAERS cases than with death outcomes. This can be partially explained by the fact that many more ADRs result in serious outcomes than in death, leading to poor differentiation of severity based on death for many ADRs where death is a rare or unobserved outcome.

In an effort to compare the impacts of using different methods to select the initial severe and benign ADR seed terms for label propagation, we compared crowdworker and real-world outcomes-based seed selections. We found that both methods resulted in consistently high severity scores for ADRs in the highest-severity SOC groups such as *Neoplasms* and *Cardiac Disorders*. We observed the largest differences in the *Product Issues* SOC group. Real-world outcomes-based seeds increased the severity of psychiatric disorders and social circumstances and decreased the severity of product issues. The importance and impact of psychiatric disorders may not be apparent to those not in the medical profession. Thus, it is possible that the crowdworkers did not perceive mental illness to be as severe as physical illnesses, leading to a bias in the rankings. The use of real-world outcomes-based rankings to select seeds led to more psychiatric disorders being included in the seed terms for the severity propagation. Similarly, the shifts in social circumstances and product issues are likely due to their limited presence or absence within the original crowdworker rankings because only seven ADRs in the social circumstances category were included in the 2929 ADRs ranked by crowdworkers, and no *Product Issues* ADRs were included in that ranking. This absence resulted in no ADRs from these groups being included in the initial seed set of labeled ADRs for the crowdworker-seeded method run, likely limiting the ability of the label propagation to accurately estimate the severity of these ADRs.

We found that the overall severity ordering of the SOC classes based on our SAEDR scores made intuitive sense. We were surprised by the ability of the label propagation approach to adjust the severity estimates based on particular key terms. For instance, the ADR terms for cancer (*Neoplasm*) ADRs are relatively similar on a lexical basis, but individual modifiers such as malignant and benign resulted in significantly different SAEDR scores for the respective groups of ADRs (Figure S3 of [Multimedia Appendix 1](#)). Similarly, examining groups of ADRs at different MedDRA hierarchy levels demonstrated relative severity estimates that were sensible to us, such as *Heart Failure Signs and Symptoms* having lower severity than the different types of actual heart failure.

We found a significantly higher estimate of severity among the ADRs that had been included in a boxed warning section than among those that had not. This offers further validation of the SAEDR scores because they are in agreement with past regulatory and drug labeling decisions. Our comparison of HLGTS with disproportionate rates between sexes revealed that male-associated ADRs were more severe. Notably, the SAEDR scores of the male-associated ADRs were not significantly different from those of the ADRs that have been included in a boxed warning. This is in agreement with previous work that highlighted the relatively higher severity of ADRs experienced by men [29,30].

We found that ADRs discovered in the postmarketing period are generally more severe than those discovered in the clinical trials based on our SAEDR scores. This indicates that postmarketing surveillance and ongoing regulatory discussions of risk-benefit trade-offs for particular drugs are necessary to keep the public safe [31]. These findings are not new because a recent study examining internal Food and Drug Administration data on drug labeling changes found that 35% of the ADRs added to drug labels were added to the boxed warnings and warnings and precautions sections [32]. Of the postmarketing ADRs, those involving more than one drug, as identified by TWOSIDES, were among those with the highest severity, indicating the increased risks associated with polypharmacy. This finding highlights the need for further research into the safety of polypharmacy because approximately half of the individuals prescribed a prescription drug are prescribed more than one concurrent medication [33].

One of the original aims of the pioneering study by Tallarida et al [7] was to create quantitative risk scores for individual drugs. Our DRIP scores are an attempt to do so, and we combined ADR severity with frequency information to generate a numerical estimate of a drug's risk profile. We were able to generate DRIP scores for 968 drugs using ADR frequency data from the SIDER. Our DRIP scores have a modest Spearman correlation,  $\rho=0.377$ , with FAERS cases that resulted in death, where that drug was the primary suspect. Generally, correlations greater than 0.8 are considered strong, but this generally assumes that the comparison is with a gold standard ground truth. Because of a lack of an objective gold standard of DRIPs, we compared with FAERS case outcomes for the drugs in question, although the case outcomes are imperfect because of reporting bias and other issues. In addition, the DRIP scores were created using estimates of ADR frequencies from the SIDER that are primarily derived from clinical trial data. These ADR frequency estimates are affected by the dosage and underlying health of the clinical trial population, both of which are only sometimes reported. High-quality frequency estimates for postmarketing ADRs are essentially nonexistent, and many challenges exist to accurately estimating their frequency [29]. These issues limit the ability of our DRIP scores to accurately quantify risk. However, despite the limitations of the benchmark data and the ADR frequency information, the DRIP scores still demonstrate a modest correlation with real-world outcomes. We interpret this as an indication that our DRIP scores track with real-world outcomes and capture signals related to a drug's safety profile that could be useful for downstream application. Further evidence of their validity is provided by the qualitative evaluation of the DRIP score distributions for the ATC groups. The relative distributions of the ATC groups make intuitive sense, with drugs used to treat cancer—antineoplastic drugs—having the most severe side effect profiles, whereas drugs used to treat diseases that are known to have safe and effective treatment options, such as diabetes, are ranked among the safest drug categories.

We present ADR severity estimates that track with both real-world outcomes and human perception, but these estimates are still limited. We used word embeddings trained on Reddit data, but other social media data could improve the severity

prediction model. We used RedMed because the model was publicly available and contained many ADR terms of interest to this study. However, Reddit as a social media platform is skewed toward men and young people, who may discuss ADRs in different ways from the general population. Learning new embeddings from social media corpora generated by the patient groups most at risk for a given set of ADR experiences or from biomedical literature might improve model performance and address issues with demographic model biases.

FAERS case outcome proportions as a benchmark for severity are limited because of the outcomes being reported at the case level by the FAERS. Because of this case-level aggregation, the ADRs in the severe category may have inflated severity. For instance, symptoms of cardiac arrest may often be reported as separate ADRs for a FAERS case of cardiac arrest, resulting in symptoms that seem to be more severe than they are when occurring independently of the severe ADR with which they are associated.

Our results indicate that men reported more severe ADRs based on sex-specific ADRs identified in the study by Chandak et al [27]. Other research efforts have reported similar findings based on ADR-reporting databases [29], but there are potential sources of bias that could have affected this result. The data underlying the sex-specific ADRs identified in the study by Chandak et al [27] were derived from the FAERS. The study method focused on the analysis of a sex-balanced cohort of case reports, with cohort creation through propensity score matching of individuals. Propensity score matching can correct for factors included in the model, but it is unable to correct for underlying biases in reporting. It has been documented that there are higher rates of ADR reporting for female patients than for male patients, whereas male patients tend to have more severe outcomes reported [30]. It is likely that male patients are enriched for severe ADRs because of this reporting bias, and ADR reports should not be conflated with all ADRs experienced by individuals because of the low rate of ADR reporting [34].

Another limitation is that sex-specific ADRs were only reported at the HLT level, resulting in SAEDR scores that were an average of all PTs within the HLT. Figure 3 highlights the large range in terms of severity, even within a particular HLT. It is possible that an analysis of sex-specific ADRs at the PT level would result in a different finding. Overall, although ADRs disproportionately reported in men are more severe than those disproportionately reported in women, more research is needed to reach a complete understanding of sex-specific differences in ADR severity.

Although we captured some nuances in ADR severity, such as benign versus malignant neoplasms, we only created SAEDR scores for the ADR terms contained in the MedDRA. However, while exploring the RedMed word embeddings, we observed that individual ADR terms were often contained in phrases indicating modified severity. For instance, a report of stomach pain might be modified with an adjective such as excruciating or mild. Although the MedDRA does not contain an intensity scale component for its terms, capturing this information from patients might further enable risk-benefit trade-off calculations. The approaches for general word sentiment assignment, similar to the one used in the study by Hamilton et al [22], could be repurposed here to assign severity to ADR modifier terms.

## Conclusions

In summary, we demonstrated that lexical networks and label propagation can be used quantitatively to estimate the severity of medical conditions. We showed the distributions of ADR severity among different groups of conditions, different groups of patients, and different ADR discovery time points. We combined our SAEDR scores with available ADR frequency data to generate quantitative DRIP scores and examined the distribution of the resulting DRIP scores. Our results (and future improved estimates) enable new quantitative analyses within the field of pharmacovigilance. To this end, we provide the complete set of SAEDR scores and DRIP scores in [Multimedia Appendix 2](#) and [Multimedia Appendix 3](#), respectively.

---

## Acknowledgments

The authors would like to thank the editor and the reviewers for their helpful and thoughtful feedback during the review process. Some of the computations for this project were performed on the Sherlock cluster. The authors would like to thank Stanford University and the Stanford Research Computing Center for providing computational resources and support that contributed to the research results. AL is supported by the National Science Foundation (NSF) Graduate Research Fellowship (DGE 1656518). TH is supported by the National Institutes of Health (NIH; R01DK103358); Simons Foundation and NSF (IOS-1546218, R35GM122515, and NSF CBET 1728858); and the NIH (R01AI130945). RBA is supported by the NIH (GM102365 and HG010615) and the Chan Zuckerberg Biohub.

---

## Conflicts of Interest

None declared.

---

## Multimedia Appendix 1

Supplementary figures and associated captions.  
[\[PTX File , 885 KB-Multimedia Appendix 1\]](#)

---

## Multimedia Appendix 2

Severity of Adverse Events Derived from Reddit scores for 12,198 adverse drug reactions.

[\[ZIP File \(Zip Archive\), 391 KB-Multimedia Appendix 2\]](#)

### Multimedia Appendix 3

Drug-risk profile scores.

[\[ZIP File \(Zip Archive\), 30 KB-Multimedia Appendix 3\]](#)

### References

1. Bourgeois F, Shannon M, Valim C, Mandl KD. Adverse drug events in the outpatient setting: an 11-year national analysis. *Pharmacoepidemiol Drug Saf* 2010 Sep;19(9):901-910 [FREE Full text] [doi: [10.1002/pds.1984](https://doi.org/10.1002/pds.1984)] [Medline: [20623513](https://pubmed.ncbi.nlm.nih.gov/20623513/)]
2. Watanabe JH, McInnis T, Hirsch JD. Cost of prescription drug-related morbidity and mortality. *Ann Pharmacother* 2018 Sep;52(9):829-837. [doi: [10.1177/1060028018765159](https://doi.org/10.1177/1060028018765159)] [Medline: [29577766](https://pubmed.ncbi.nlm.nih.gov/29577766/)]
3. Aronson JK, Ferner RE. Clarification of terminology in drug safety. *Drug Saf* 2005;28(10):851-870. [doi: [10.2165/00002018-200528100-00003](https://doi.org/10.2165/00002018-200528100-00003)] [Medline: [16180936](https://pubmed.ncbi.nlm.nih.gov/16180936/)]
4. Ahmad SR. Adverse drug event monitoring at the Food and Drug Administration. *J Gen Intern Med* 2003 Jan;18(1):57-60 [FREE Full text] [doi: [10.1046/j.1525-1497.2003.20130.x](https://doi.org/10.1046/j.1525-1497.2003.20130.x)] [Medline: [12534765](https://pubmed.ncbi.nlm.nih.gov/12534765/)]
5. Gunderson F, Gunderson H, Ferguson Jr E. Code of Federal Regulations. In: *Food Standards and Definitions in the United States*. Cambridge, Massachusetts, United States: Academic Press; 1963:147-154.
6. Shaffer ML, Watterberg KL. Joint distribution approaches to simultaneously quantifying benefit and risk. *BMC Med Res Methodol* 2006 Oct 12;6:48 [FREE Full text] [doi: [10.1186/1471-2288-6-48](https://doi.org/10.1186/1471-2288-6-48)] [Medline: [17038184](https://pubmed.ncbi.nlm.nih.gov/17038184/)]
7. Tallarida RJ, Murray RB, Eiben C. A scale for assessing the severity of diseases and adverse drug reactions. Application to drug benefit and risk. *Clin Pharmacol Ther* 1979 Apr;25(4):381-390. [doi: [10.1002/cpt1979254381](https://doi.org/10.1002/cpt1979254381)] [Medline: [428184](https://pubmed.ncbi.nlm.nih.gov/428184/)]
8. Gottlieb A, Hoehndorf R, Dumontier M, Altman RB. Ranking adverse drug reactions with crowdsourcing. *J Med Internet Res* 2015;17(3):e80 [FREE Full text] [doi: [10.2196/jmir.3962](https://doi.org/10.2196/jmir.3962)] [Medline: [25800813](https://pubmed.ncbi.nlm.nih.gov/25800813/)]
9. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. *arXiv - Computer Science, Computation and Language*. URL: <http://arxiv.org/abs/1301.3781> [accessed 2021-07-19]
10. Pennington J, Socher R, Manning C. Glove: Global vectors for word representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2014 Presented at: *Conference on Empirical Methods in Natural Language Processing (EMNLP)*; October, 2014; Doha, Qatar p. 1532-1543. [doi: [10.3115/v1/d14-1162](https://doi.org/10.3115/v1/d14-1162)]
11. Firth J. A Synopsis of Linguistic Theory 1930-1955, Selected Papers of JR Firth 1952 - 1959. In: *Studies in Linguistic Analysis*. Oxford, United Kingdom: Philological Society; 1957:168-205.
12. Nikfarjam A, Sarker A, O'Connor K, Ginn R, Gonzalez G. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *J Am Med Inform Assoc* 2015 May;22(3):671-681 [FREE Full text] [doi: [10.1093/jamia/ocu041](https://doi.org/10.1093/jamia/ocu041)] [Medline: [25755127](https://pubmed.ncbi.nlm.nih.gov/25755127/)]
13. Hanson CL, Cannon B, Burton S, Giraud-Carrier C. An exploration of social circles and prescription drug abuse through Twitter. *J Med Internet Res* 2013;15(9):e189 [FREE Full text] [doi: [10.2196/jmir.2741](https://doi.org/10.2196/jmir.2741)] [Medline: [24014109](https://pubmed.ncbi.nlm.nih.gov/24014109/)]
14. Yan L, Yan X, Tan Y, Sun S. Shared Minds: How patients use collaborative information sharing via social media platforms. *Prod Oper Manag* 2018 Jun 26;28(1):9-26. [doi: [10.1111/poms.12895](https://doi.org/10.1111/poms.12895)]
15. Sarker A, Ginn R, Nikfarjam A, O'Connor K, Smith K, Jayaraman S, et al. Utilizing social media data for pharmacovigilance: A review. *J Biomed Inform* 2015 Apr;54:202-212 [FREE Full text] [doi: [10.1016/j.jbi.2015.02.004](https://doi.org/10.1016/j.jbi.2015.02.004)] [Medline: [25720841](https://pubmed.ncbi.nlm.nih.gov/25720841/)]
16. Alvaro N, Conway M, Doan S, Lofi C, Overington J, Collier N. Crowdsourcing Twitter annotations to identify first-hand experiences of prescription drug use. *J Biomed Inform* 2015 Dec;58:280-287 [FREE Full text] [doi: [10.1016/j.jbi.2015.11.004](https://doi.org/10.1016/j.jbi.2015.11.004)] [Medline: [26556646](https://pubmed.ncbi.nlm.nih.gov/26556646/)]
17. Klein A, Sarker A, Rouhizadeh M, O'Connor K, Gonzalez G. Detecting personal medication intake in Twitter: An annotated corpus and baseline classification system. *BioNLP* 2017:136-142. [doi: [10.18653/v1/W17-2316](https://doi.org/10.18653/v1/W17-2316)]
18. Lavertu A, Altman RB. RedMed: Extending drug lexicons for social media applications. *J Biomed Inform* 2019 Nov;99:103307 [FREE Full text] [doi: [10.1016/j.jbi.2019.103307](https://doi.org/10.1016/j.jbi.2019.103307)] [Medline: [31627020](https://pubmed.ncbi.nlm.nih.gov/31627020/)]
19. Kuhn M, Letunic I, Jensen LJ, Bork P. The SIDER database of drugs and side effects. *Nucleic Acids Res* 2015 Oct 19:1075-1079 [FREE Full text] [doi: [10.1093/nar/gkv1075](https://doi.org/10.1093/nar/gkv1075)] [Medline: [26481350](https://pubmed.ncbi.nlm.nih.gov/26481350/)]
20. Brown EG, Wood L, Wood S. The medical dictionary for regulatory activities (MedDRA). *Drug Saf* 1999 Feb;20(2):109-117 [FREE Full text] [doi: [10.2165/00002018-199920020-00002](https://doi.org/10.2165/00002018-199920020-00002)] [Medline: [10082069](https://pubmed.ncbi.nlm.nih.gov/10082069/)]
21. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004 Jan 1;32(Database issue):267-270 [FREE Full text] [doi: [10.1093/nar/gkh061](https://doi.org/10.1093/nar/gkh061)] [Medline: [14681409](https://pubmed.ncbi.nlm.nih.gov/14681409/)]
22. Hamilton WL, Clark K, Leskovec J, Jurafsky D. Inducing domain-specific sentiment lexicons from unlabeled corpora. *Proc Conf Empir Methods Nat Lang Process* 2016 Nov;2016:595-605 [FREE Full text] [doi: [10.18653/v1/D16-1057](https://doi.org/10.18653/v1/D16-1057)] [Medline: [28660257](https://pubmed.ncbi.nlm.nih.gov/28660257/)]
23. Määttä E, Haapala A. Python-Levenshtein. Python Software Foundation. URL: <https://pypi.org/project/python-Levenshtein/> [accessed 2020-05-25]

24. Grover A, Leskovec J. node2vec: Scalable feature learning for networks. KDD 2016 Aug;2016:855-864 [[FREE Full text](#)] [doi: [10.1145/2939672.2939754](https://doi.org/10.1145/2939672.2939754)] [Medline: [27853626](https://pubmed.ncbi.nlm.nih.gov/27853626/)]
25. Efron B. Bootstrap Methods: Another look at the jackknife. Ann Statist 1979 Jan 1;7(1):1-26. [doi: [10.1214/aos/1176344552](https://doi.org/10.1214/aos/1176344552)]
26. Wu L, Ingle T, Liu Z, Zhao-Wong A, Harris S, Thakkar S, et al. Study of serious adverse drug reactions using FDA-approved drug labeling and MedDRA. BMC Bioinformatics 2019 Mar 14;20(Suppl 2):97 [[FREE Full text](#)] [doi: [10.1186/s12859-019-2628-5](https://doi.org/10.1186/s12859-019-2628-5)] [Medline: [30871458](https://pubmed.ncbi.nlm.nih.gov/30871458/)]
27. Chandak P, Tatonetti NP. Using machine learning to identify adverse drug effects posing increased risk to women. Patterns (N Y) 2020 Oct 09;1(7):100108 [[FREE Full text](#)] [doi: [10.1016/j.patter.2020.100108](https://doi.org/10.1016/j.patter.2020.100108)] [Medline: [33179017](https://pubmed.ncbi.nlm.nih.gov/33179017/)]
28. Tatonetti NP, Ye PP, Daneshjou R, Altman RB. Data-driven prediction of drug effects and interactions. Sci Transl Med 2012 Mar 14;4(125):125ra31 [[FREE Full text](#)] [doi: [10.1126/scitranslmed.3003377](https://doi.org/10.1126/scitranslmed.3003377)] [Medline: [22422992](https://pubmed.ncbi.nlm.nih.gov/22422992/)]
29. Watson S, Caster O, Rochon PA, den Ruijter H. Reported adverse drug reactions in women and men: Aggregated evidence from globally collected individual case reports during half a century. EClinicalMedicine 2019 Dec;17:100188 [[FREE Full text](#)] [doi: [10.1016/j.eclinm.2019.10.001](https://doi.org/10.1016/j.eclinm.2019.10.001)] [Medline: [31891132](https://pubmed.ncbi.nlm.nih.gov/31891132/)]
30. Holm L, Ekman E, Jorsäter Blomgren K. Influence of age, sex and seriousness on reporting of adverse drug reactions in Sweden. Pharmacoepidemiol Drug Saf 2017 Mar;26(3):335-343. [doi: [10.1002/pds.4155](https://doi.org/10.1002/pds.4155)] [Medline: [28071845](https://pubmed.ncbi.nlm.nih.gov/28071845/)]
31. Lavertu A, Vora B, Giacomini KM, Altman R, Rensi S. A new era in pharmacovigilance: Toward real-world data and digital monitoring. Clin Pharmacol Ther 2021 May 25;109(5):1197-1202 [[FREE Full text](#)] [doi: [10.1002/cpt.2172](https://doi.org/10.1002/cpt.2172)] [Medline: [33492663](https://pubmed.ncbi.nlm.nih.gov/33492663/)]
32. Pinnow E, Amr S, Bentzen SM, Brajovic S, Hungerford L, St George DM, et al. Postmarket safety outcomes for New Molecular Entity (NME) drugs approved by the Food and Drug Administration between 2002 and 2014. Clin Pharmacol Ther 2018 Aug;104(2):390-400. [doi: [10.1002/cpt.944](https://doi.org/10.1002/cpt.944)] [Medline: [29266187](https://pubmed.ncbi.nlm.nih.gov/29266187/)]
33. Quinn KJ, Shah NH. A dataset quantifying polypharmacy in the United States. Sci Data 2017 Oct 31;4:170167 [[FREE Full text](#)] [doi: [10.1038/sdata.2017.167](https://doi.org/10.1038/sdata.2017.167)] [Medline: [29087369](https://pubmed.ncbi.nlm.nih.gov/29087369/)]
34. Hazell L, Shakir SA. Under-reporting of adverse drug reactions : a systematic review. Drug Saf 2006;29(5):385-396. [doi: [10.2165/00002018-200629050-00003](https://doi.org/10.2165/00002018-200629050-00003)] [Medline: [16689555](https://pubmed.ncbi.nlm.nih.gov/16689555/)]

## Abbreviations

**ADR:** adverse drug reaction

**ATC:** Anatomical Therapeutic Chemical classification system

**DRIP:** drug-risk profile

**FAERS:** Food and Drug Administration Adverse Event Reporting System

**HLGT:** high-level group term

**MedDRA:** Medical Dictionary for Regulatory Activities

**PT:** preferred term

**SAEDR:** Severity of Adverse Events Derived from Reddit

**SIDER:** Side Effect Resource

**SOC:** System Organ Class

*Edited by R Kukafka; submitted 03.02.21; peer-reviewed by A Sarker, S Matsuda; comments to author 03.04.21; revised version received 25.05.21; accepted 14.06.21; published 21.10.21*

*Please cite as:*

Lavertu A, Hamamsy T, Altman RB

Quantifying the Severity of Adverse Drug Reactions Using Social Media: Network Analysis

J Med Internet Res 2021;23(10):e27714

URL: <https://www.jmir.org/2021/10/e27714>

doi: [10.2196/27714](https://doi.org/10.2196/27714)

PMID:

©Adam Lavertu, Tymor Hamamsy, Russ B Altman. Originally published in the Journal of Medical Internet Research (<https://www.jmir.org>), 21.10.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.jmir.org/>, as well as this copyright and license information must be included.