Bioinformatics, 36(4), 2020, 1226–1233
doi: 10.1093/bioinformatics/btz678
Advance Access Publication Date: 3 September 2019
Original Paper

Data and text mining

Applying citizen science to gene, drug and disease relationship extraction from biomedical abstracts

Ginger Tsueng *, Max Nanis, Jennifer T. Fouquier, Michael Mayers, Benjamin M. Good and Andrew I. Su

Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, CA 92037, USA

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on April 17, 2019; revised on August 5, 2019; editorial decision on August 25, 2019; accepted on August 29, 2019

Abstract

Motivation: Biomedical literature is growing at a rate that outpaces our ability to harness the knowledge contained therein. To mine valuable inferences from the large volume of literature, many researchers use information extraction algorithms to harvest information in biomedical texts. Information extraction is usually accomplished via a combination of manual expert curation and computational methods. Advances in computational methods usually depend on the time-consuming generation of gold standards by a limited number of expert curators. Citizen science is public participation in scientific research. We previously found that citizen scientists are willing and capable of performing named entity recognition of disease mentions in biomedical abstracts, but did not know if this was true with relationship extraction (RE).

Results: In this article, we introduce the Relationship Extraction Module of the web-based application Mark2Cure (M2C) and demonstrate that citizen scientists can perform RE. We confirm the importance of accurate named entity recognition on user performance of RE and identify design issues that impacted data quality. We find that the data generated by citizen scientists can be used to identify relationship types not currently available in the M2C Relationship Extraction Module. We compare the citizen science-generated data with algorithm-mined data and identify ways in which the two approaches may complement one another. We also discuss opportunities for future improvement of this system, as well as the potential synergies between citizen science, manual biocuration and natural language processing.

Availability and implementation: Mark2Cure platform: https://mark2cure.org; Mark2Cure source code: https://github.com/sulab/mark2cure; and data and analysis code for this article: https://github.com/gtsueng/M2C_rel_nb.

Contact: gtsueng@scripps.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Biomedical literature is growing at rate of over a million new articles per year in PubMed (https://www.ncbi.nlm.nih.gov/books/NBK3827/) and represents a treasure trove of knowledge that forms the foundation for the design of future experiments. Making that knowledge more accessible and computable could save researchers time, effort and resources (Yang *et al.*, 2016; Zhu *et al.*, 2013). Researchers have effectively mined slices of the biomedical literature to identify potential treatments for Raynaud's syndrome (Swanson, 1986), drug candidates for Alzheimer's disease (Li *et al.*, 2009) and potential mechanisms of ovarian oncogenesis (Urzúa *et al.*, 2010). Given the large potential to make valuable inferences and the large volume of literature, many researchers have turned to information extraction algorithms to harvest information in biomedical texts and improve the value of existing data resources (Murray-Rust, 2017; Pletscher-Frankild *et al.*, 2015). Information extraction as a process can be divided into a few sub tasks: (i) Named Entity Recognition (NER), (ii) Entity Linking (EL) and (iii) Relationship Extraction (RE).

NER entails identifying specific types of entities within biomedical text [e.g. NGLY1(entity) is a gene(entity type)]. Once identified, the NER term must be linked to an appropriate entry in a known database to provide semantic context (e.g. NGLY1 can be linked to gene 55768 in the NCBI Gene database). The process of linking NER annotations to known databases to provide context and generate semantic annotations is known as EL or normalization (Jovanović and Bagheri, 2017; Morgan *et al.*, 2008). After NER and EL, the relationships between the semantic annotations are extracted (Relationship Extraction/RE)—e.g. [NGLY1] (gene) [mutations cause](relationship) [congenital disorder of deglycosylation](disease).

Algorithms for NER and EL have steadily improved thanks to the BioCreative challenges and availability of gold standard corpora (Morgan *et al.*, 2008; Wei *et al.*, 2015). Tools such as EXTRACT (Pafilis *et al.*, 2016) and those included in the PubTator suite (Wei *et al.*, 2013) are sufficient for use in facilitating manual biocuration efforts. Furthermore, NER and semantic annotation algorithms have expanded beyond the concept types originally explored by the BioCreative challenges and now include post-translational modifications (Sun *et al.*, 2017), Gene Ontology terms (Ruch, 2016), metadata (Panahiazar *et al.*, 2017), adverse effects (Cañada *et al.*, 2017) and more (Tseytlin *et al.*, 2016). Semantic annotation algorithms such as SemRep have been used to generate SemMedDB, a PubMed-scale repository of subject–predicate–object triples (Kilicoglu *et al.*, 2012).

Because of its dependency on NER, EL, the complexity of the task and the limited availability of training corpora and ontologies with relationship annotations, automated approaches for RE have yet to reach the performance levels of NER and EL (Wei *et al.*, 2016). To overcome those limitations, researchers have focused on improving NER and EL methods (Xing *et al.*, 2018), different learning and modeling approaches (Peng *et al.*, 2018) and expanding training datasets via mining of knowledge bases (Zhou *et al.*, 2018) or crowdsourcing via paid microtask platforms (Li *et al.*, 2016; Lossio-Ventura *et al.*, 2018). Crowdsourcing through paid microtask platforms to expand the training datasets has proven to have great potential, but questions regarding scalability prompted us to investigate citizen science as a potential avenue for crowdsourcing RE.

Citizen science is a form of crowdsourcing in which nonprofessional scientists voluntarily engage in different degrees of data collection, analysis and/or dissemination of a scientific project (Haklay, 2013). The scalability of citizen science has enabled researchers to collect, process and analyze unprecedented volumes of data leading to advances in conservation and environmental science (McKinley *et al.*, 2017; Schmiedel, 2016), astronomy (Banfield *et al.*, 2016; Kuchner *et al.*, 2016; Straub, 2016), biomedical research (Candido dos Reis *et al.*, 2015; Kim *et al.*, 2014; Luengo-Oroz *et al.*, 2012) and more (Palermo *et al.*, 2017; Williams *et al.*, 2014). Crowdsourcing and citizen science has previously been applied toward NER of disease mentions via a platform called Mark2Cure (M2C). It was found that in aggregate, annotations submitted by trained citizen scientists were on par with expert annotators (Good *et al.*, 2015; Tsueng *et al.*, 2016). Based on this finding, citizen scientists may serve as an additional check for annotations generated by computer algorithms, and address quality issues introduced by NER and EL tools. The problem of insufficient gold standard corpora for RE tasks can also be addressed by crowdsourcing (Aroyo and Welty, 2013; Burger *et al.*, 2014). In aggregate, nonexperts recruited via a microtask platform could perform relationship annotation on par with expert curation (Dumitrache *et al.*, 2015) provided that the task is appropriately designed (Khare *et al.*, 2015).

In this article, we describe the application of citizen science toward RE from biomedical text. Specifically, we (i) provide a brief overview of the RE module within the M2C platform, (ii) Evaluate the ability of citizen scientists to perform RE taking into consideration the limitations and ambiguities inherent in the system and (iii) Compare the citizen science-generated data with the automated results from SemMedDB to understand how the two may complement and enhance each other.

2 Materials and Methods

2.1 M2C relationship app design

The NER module within the M2C platform has previously been described (Tsueng *et al.*, 2016). Since the beta study, M2C has been expanded to investigate multiple entity (or concept) types and for RE of abstracts of interest for the NGLY1-deficiency rare disease community. In M2C, 'Entities' are referred to as 'Concepts' because initial user studies indicated that users found the term to be less intimidating, confusing and off-putting. Hence, NER entities may be referred to as concepts interchangeably. The RE App has a separate training module, task list, task interface and feedback screen. M2C is an open-source project, and code is available at https://github.com/SuLab/mark2cure.

For the RE App, training consists of a series of interactive modules developed after several iterations of testing and feedback from the M2C community. The first module introduces the user to the task interface. The second module introduces the user to the two fundamental rules of the task: (i) select relationship based only on what is in the abstract (no prior knowledge) and (ii) select most granular relationship without guessing. Users are expected to extract relationships as they are asserted in the abstract, but are not expected to evaluate the underlying biological truth of these assertions. The third module has three submodules introducing the user to the different kinds of relationships that they will need to classify (gene–disease relationships, gene–drug relationships and disease–drug relationships). Because there are no gold standards for this task, users are provided with visual feedback on how their selection aligned with that of the all the other users who have done the same task. Each task needs to be evaluated by multiple users to be considered complete. Screenshots of the user interface for a sample task and user feedback screen can be found in Supplementary Figure S1.

The M2C relationship app currently pulls concept annotations using the PubTator suite (Wei *et al.*, 2013). For each abstract, every combination of heterogeneous concept pairs is calculated and designated as a task. Concept pairs within the same concept type are not included because the relationship between concepts of the same concept type tend to be hypernymic relationships (e.g. 'is a') which algorithms are good at identifying (Rindflesch and Fiszman, 2003). For example, a gene entity such as 'Aladin' will be paired with a disease entity such as 'Alacrima' to form a heterotypic concept pair, but that gene entity would not be paired with another gene entity such as 'ACTH' since that would be a homotypic concept pair. Users are presented with a concept pair and the concepts are highlighted in the abstract to provide context. Based on the abstract text, the users are asked to identify the relationship (or lack of relationship) between those two concepts. Users also have the ability to tag either of the concepts as incorrectly identified/inappropriately annotated.

M2C is an ongoing project with active data collection and user submissions. The data analyzed for this study were collected between May 5, 2016 and November 22, 2017. This dataset consists of 4047 concept pairs pulled from 1058 abstracts annotated by 147 contributors; of which 1009 concept pairs from 234 abstracts were marked by at least 15 different contributors.

2.2 Analytical methods

2.2.1 Generation of a M2C-specific pseudo-gold standard for quality control

There were 1009 completed concept pair relationship (i.e. task) annotations at the time of analysis, and a 10% sample (~100 concept pairs) was desired for quality control (QC). About 120 PubMed ID (PMID) PMID-specific concept pairs were randomly selected and manually inspected to determine the expected response based on the rules and available options. These annotations comprise the QC annotation set.

2.2.2 Contribution distribution, accuracy and aggregation threshold determination

A relationship annotation task was considered complete once it had been reviewed by at least 15 different contributors. The contribution distribution was limited to just the set of completed task annotations. For this set, the number of RE task annotations that each user contributed was determined, sorted and plotted. For individual accuracy estimation, the set of each user's task annotations was compared with the QC annotation set. The accuracy was estimated based on the intersecting task annotations of the two sets and the median of all user accuracy estimations was calculated. For aggregate accuracy estimation, the user annotations for the concept pairs that were QC'd were pulled into a data frame for further analysis. Voter numbers (n) were set at values ranging from 1 (single vote) to 15 (maximum voters). For each concept pair, at each value of n, n

users that annotated that concept pair were randomly selected and the majority response for that concept pair was identified. If the results were tied, one of the tied responses was selected at random. For each value of n, the random selection and majority determination was performed 10 times (i.e. 10 iterations). The accuracy of the responses in the QC'd set was calculated per each level of n, and iteration. The median accuracy for each level of n was calculated along with the q25 and q75 quartiles.

2.2.3 Identification of missing relationship types and verification of nonrelationships

To identify missing relationship types, PMID-specific concept pairs annotated as 'has relationship' or 'other relationship' were aggregated to obtain the total number of users that marked each concept pair as having a nonspecific relationship. User agreement threshold (K) ranged from 1 (single voter, no agreement) to 15 (maximum agreement). At each level of K, up to 25 PMID-specific concept pairs were randomly selected for qualitative review. The concept pairs and the respective user counts were exported, randomly assigned a number and randomly sorted by that number. The user counts were then masked to prevent biasing and each PMID-specific concept pair was reviewed in-house. The same process was applied to PMID-specific concept pairs marked as having 'no relationship' or 'cannot be determined'.

2.2.4 Evaluating the effect of concept distance on accuracy

The abstracts were analyzed at the sentence level using the NLP Tool Kit (NLTK) sentence tokenizer (Bird *et al.*, 2009) to obtain an average per-sentence character count which can be used to estimate the concept distance at the sentence level. Only concepts with known identifiers were analyzed as it would be more difficult to determine the positional location of a term in an abstract when the identifier is missing. Since the appropriateness of a concept annotation should not be affected by concept distance, relation annotations for concepts considered correctly annotated (i.e 'not broken') were treated separately.

2.2.5 Comparison with PMID-specific SemMedDB relationships

Because there are no expert-curated gold standard relationships extraction data available for the PMIDs covered in this experiment, a subset of the computationally derived open dataset, SemMedDB was used for comparative purposes. The PMIDs for the completed concept pairs were used to pull SemMedDB annotations specific to those PMIDs. For example on the PMID 16609705, the following M2C concept pair: 8086(AAAS gene), D000309 (adrenal insufficiency) was matched with the following IDs from SemMedDB C1422135, C0001623, respectively. SemMedDB concept annotations are linked to concept-unique identifiers (CUIs) from the Unified Medical Language System (UMLS). The UMLS integrates many biomedical vocabularies and standards (Bodenreider, 2004). Since UMLS supports many more concept types than M2C, only UMLS concept types that mapped to M2C concept types were included in the analysis. For example, SemMedDB concepts like Language (lang|T171) or Bird (bird|T012) would not be included; however, SemMedDB concepts like Disease or Syndrome (dysn|T047) and Sign or Symptom (sosy|T184) would both be included and mapped to the M2C Disease concept. In addition, the RE module in M2C is not yet connected to the previously described NER module (Tsueng *et al.*, 2016); hence, the concepts in the M2C relationship module are purely based on PubTator. Thus, the UMLS concept types were mapped to M2C concept types based on mappings used in the generation of the corpora used for the gene, disease and drug NER algorithms in PubTator with a few additional mappings to suit the expansion of the PubTator concept types (M2C Relation Extraction concept types) to the M2C NER concept types. The identifiers for relationships considered complete in M2C were converted to UMLS CUIs and used to filter an export of SemMedDB annotations for only relationships involving those CUIs. The semantic types of the SemMedDB subject and objects were mapped to M2C concept types for comparing relationships within the same concept type (e.g. gene–gene relationships) versus different concept types (e.g. gene–disease relationships). The types of concept pairs were compared between SemMedDB and M2C after concept pairs in which the majority of users marked a concept as incorrectly identified (i.e. 'broken') or unrelated were removed from the remaining set of PMIDs.

3 Results

3.1 Contribution distribution, accuracy and aggregation threshold determination

The relationships between 1009 concept pairs were annotated by at least 15 M2C volunteers. In total, we collected and analyzed 15 739 annotations from 147 volunteer contributors. As with other crowdsourcing systems (Cox *et al.*, 2015), we measured the distribution of effort by plotting out the contributions and calculating the Gini coefficient (Fig. 1A). The Gini coefficient was 0.73 which was comparable to what was observed in the disease mentions NER pilot study (gini = 0.716; Tsueng *et al.*, 2016) and slightly lower than those observed for several well-known online citizen science projects (gini range = 0.77–0.91; Sauermann and Franzoni, 2015).

To assess the accuracy of these relationship annotation results, we compared them to a manually curated subset of the full dataset. Based on this QC set, the median accuracy per user across all of their annotations was 0.61 (Fig. 1B), and was affected by NER issues (Supplementary Fig. S4). To assess how increasing the number of contributors affected the quality of the aggregate annotations, we simulated smaller numbers of annotators per document by randomly sampling from the QC set. When aggregating user responses and selecting the majority response, the accuracy increases along with the number of users n up until about six users. Beyond six users, the median accuracy does not further increase, suggesting that each relationship task should be annotated by a maximum of six users to maximize accuracy while minimizing redundancy of work done by the community (Fig. 1C).

A small subset of users were estimated to have very low accuracy despite contributing over 10 task annotations–warranting the need to verify that there were not unaccounted for issues with the data. Upon manual inspection of two of these outliers, we observed that these users did not mark any concepts as inappropriately annotated.

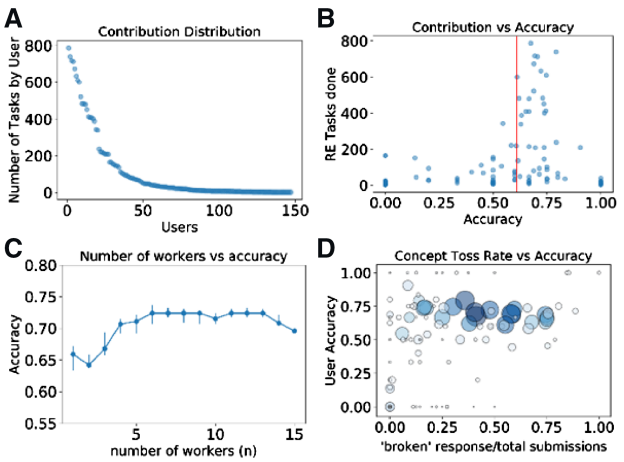

Fig. 1. (**A**) Contribution distribution of the RE task. (**B**) Each user's estimated accuracy (x) versus the number of tasks that user completed (y). The line illustrates the median accuracy. (**C**) Median accuracy with respect to the majority response of an aggregate of n users. The Q25 and Q75 quantiles are represented by the lower and upper error bars. The concept toss rate (i.e. concepts marked as broken or inappropriately annotated by the NER algorithm) versus accuracy of individual users (**D**). The size of the circles in (D) represent the number of total tasks that individual user contributed, whereas the intensity or value represents the number of that user's annotations which could be found in the QC set and, therefore, used for accuracy estimations

To investigate the effect of a user's reluctance to correct incorrectly identified concepts, we calculated each user's 'concept toss rate' (i.e. number of annotations they marked as incorrectly identified relative to number of annotations they submitted) and plotted their estimated accuracy relative to their toss rate (Fig. 1D). Because the accuracy estimates were based on a limited QC set, estimates of a user's performance is expected to be less representative then they had very little overlap with the QC set (less color) as compared with those who had greater overlap (more intense color). Nonetheless, a subset of users that had performed many tasks (medium circles) within the lowest range of accuracy (between 0 and 0.2) also had a low average rate of tossing out concepts highlighting the effects of NER quality on Relationship Annotation. Inquiries sent from a few users suggested that at least some of the error could be attributed to a lack of guidance on how to prioritize when dealing with multiple true states. For example, the 'WD' from an abstract discussing tryptophan-Aspartic Acid repeats (PMID 16609705) was incorrectly annotated by the NER algorithm as a disease (D006527/Wilson's Disease). The majority response in this case would be to mark the annotation as 'broken' or incorrect. However, inquiries with our users have indicated that our guidance here was lacking as a small subset of our users would reason that 'WD' is indeed a disease, even if it is not a disease in this particular abstract. Therefore, the two concepts would be treated as not having a relationship by this subset of users. Further clarification on how to prioritize multiple true responses could help to improve consistency and performance of RE across the M2C community.

3.2 Identification of missing relationship types

The relationship annotation options in M2C were based on higher-level relation properties from an ontology in development that was started in WebProtege but moved to Wikidata for more open discussions (https://www.wikidata.org/wiki/User: ProteinBoxBot#Task_permission_requests). With limited relationship options available in M2C, qualitative analysis of concept pairs annotated as 'has relationship' or 'other relationship' can provide insight into relationships missing from the currently available options in M2C. To identify missing relationship types, PMID-specific concept pairs annotated as 'has relationship' or 'other relationship' were aggregated to obtain the total number of users that marked each concept pair as having a nonspecific relationship. We sampled up to 25 PMID-specific concept pairs at each voter threshold (K), randomized the order of the samples, masked the number of users that marked it as having a nonspecific relationship and then manually inspected the relationship. If the relationship between the two concepts was an available option within the M2C system, it was binned as 'has available specific relationship'. If either of the concepts were inappropriately annotated, it was binned as 'concept broken'. Common relationships not available in the M2C system were binned together and new categories were created whenever the relationship did not fit in with previous categories.

As seen in Figure 2, there is a decreasing number of per-PMID concept pairs that were marked as having a better response option within M2C (pink/red bars) as the number of users that agreed with that assessment increased. Per-PMID concept pairs that were marked as 'has relationship' by a high number of users tended to genuinely have a relationship not captured by the system. However, setting the threshold too high increases the risk of missing interesting, nonobvious relationships that are otherwise not captured by M2C.

Interesting relationships that were missing from M2C's selection options included: Resistance/Insensitivity to the gene was associated with the disease; the disease conferred resistance to the drug; the drug was altered in the disease; the gene is a marker for inspecting samples involving the disease; the drug is used in the diagnosis of the disease; and a mutation in the gene causes a disease that was misdiagnosed as the disease according to the abstract (e.g. AAAS gene as it relates to Cerebral Palsy in an abstract for a case study where Allgrove Syndrome was misdiagnosed as Cerebral Palsy in the patient history). Some users also marked 'has relationship' when the text explicitly mentioned the investigation of a relationship between

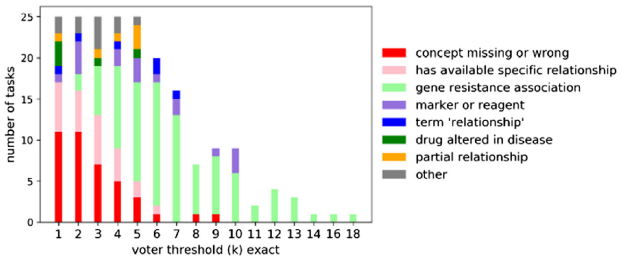

Fig. 2. Qualitative assessment of a sample of relationships marked as generic relationship/other relationship by total user counts. In red are 'has relation' annotations for a concept pair in which one of the concepts appear to be incorrect (i.e. broken). In pink are 'has relation' annotations for concept pairs which could be described with a more specific option. In green are concept pairs with a genuine relationship, not available as a selection option. In orange are concepts where the relationship in the text is partial or simultaneously true and false. All other colors indicate other ways in which the two concepts are discussed together (e.g. purple: gene is a marker used to study samples from patient with disease, etc.). To further explore these results, download an alternative interactive visualization of Figure 2, from https://git.io/fjomt and open it in a browser

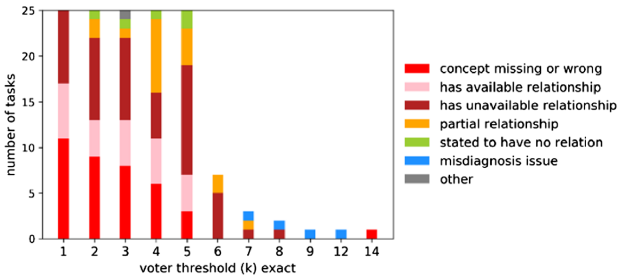

Fig. 3. Types of relationships marked as having 'no relationships'. Red indicates concept pairs in which one of the two are considered incorrectly annotated. Pink indicates that an appropriate relationship was available in the system even though the majority ruled there to be 'no relationship'. Dark Red are relationships not available in the system. Orange indicates that the text suggests an inconsistent or partial relationship. In green are concept pairs in which an attempt to establish a relationship failed (e.g. drug failed to treat a disease). Blue indicates that the relationship is between a gene or drug and a disease which was misdiagnosed in lieu of the actual disease. Other types of relationships are in gray. An alternative interactive visualization of Figure 3 can be downloaded from https://git.io/fjoqi

two concepts without actually revealing the relationship (i.e. 'we investigate the relationship between x and y', without stating the outcome of the investigation). Concept pairs which had an inconsistent relationship (such as in case studies) were marked by some users as 'has relationship' and 'has no relationship' by others (Fig. 3). The inconsistent relationship was either explicitly described in the text (e.g. 3 out of 11 patients had mutations in the gene), or implicitly described (e.g. 'We present an atypical case of Triple A syndrome without the expected ACTH-deficiency'). Since users in aggregate appear to be correctly identifying missing relationships as 'has relationship', we investigated the annotations marked as 'no relationship' to understand any rules/guidelines in the system in need of further clarification.

3.3 Verification of unrelated concepts and identification of rules in need of improvement

We applied the method used for Figure 2 to PMID-specific concept pairs that users marked as having 'no relationship'. Few PMID-specific concept pairs (each RE task) were marked by at least six users as having 'no relationship'; hence, sampling the RE tasks for qualitative analysis was only necessary for RE tasks with five or less users agreeing on the 'no relationship' response. At $K = 6$, the number of different RE tasks that was marked as having 'no relationship' drops to only seven; and drops again to less than half of that at $K = 7$. At $K = 8$ or above, we were unable to find RE tasks marked

by exactly K users for each level of K above 8, severely limiting the sample size (Fig. 3).

For the qualitatively inspected tasks–at each value of K that was below six, a better relationship (generic or specific) existed 20% of the time. At each value of K of six or above, a better (generic or specific relationship) was not available. As K increases from 1 to 5, the number of tasks in which a concept should have been marked as inappropriate decreased from 11 to 3. At $K \geq 6$, the number of tasks in which a concept should have been marked as inappropriate was mostly 0. RE tasks for which a specific relationship exists (but the option is not available) averaged to make up about a third of the annotations marked as having no relationship when K was <6. Instances where a drug failed to treat a disease and, therefore, can truly be counted as having 'no relationship' was not observed when at $K < 5$ due to the limited sample availability. Misdiagnoses accounted for some of the instances of 'no relationship' for RE tasks with $K \geq 6$ (see Supplementary Fig. S2 for the breakdown and variety of 'no relationship'). Based on these results, clarification and improved guidance is needed for the treatment of abstracts that cover case studies or clinical trials (in which the relationship may be inconsistent/partial) and for instances where a relationship was suspected, but found to be untrue such as in a misdiagnosis or failed drug trial.

3.4 Effects of concept distance on relationship identification

Many semantic and NER text mining algorithms perform optimally when used in conjunction with text analyzed at the sentence level (Lou *et al.*, 2017; Muzaffa *et al.*, 2015; Zhu *et al.*, 2018). Limiting the M2C RE task to concepts that share the same sentence could reduce the amount of text contributors would need to read and reduce the amount of text displayed in mobile devices. However, such limitations would also result in losing the option to identify relationships between concepts in the text that do not appear in the same sentence. To evaluate the pros and cons of restricting the task to the sentence-based concept pairs, we looked for relationship annotations of concepts that were not in the same sentence and we inspected the effects of the distance between concept pairs in a task on accuracy.

As seen in Figure 4, most concept pairs were less than a sentence apart. Each green dot represents a concept pair. Its location on the x-axis represents the number of sentences apart the concept pair appears to be, while the y-axis represents the estimated accuracy when inspected with different numbers of voters (n). Multiple synonymous mentions of concepts increase the likelihood of concept pairs to be located within shorter distances than farther ones, nonetheless there were still plenty of relationship identified between concept pairs that were estimated to be two or more sentences apart. In cases where the concept distance was not expected to affect the relationship (i.e. one of the concepts is considered inappropriately annotated), there were still more concept pairs located closely together in space than farther apart. Estimated accuracy appeared to be more affected by the number of voters (n) than by the estimated minimum distance between the concept pairs. This was particularly visible in the difference observed at $n = 15$ between concept pairs with a relationship and concept pairs marked as broken (Supplementary Fig. S3). For concept pairs considered broken, the accuracy at $n = 15$ decreases due to the inclusion of annotations from users uncomfortable with discarding concepts as broken in all runs. The concept pairs were further subdivided between those marked as 'unrelated' in the QC set to determine if unrelated concept pairs were more likely to be located farther apart. This was not found to be the case (Supplementary Fig. S3). Not every annotation from Pubtator was linked to an identifier. Annotations lacking identifiers could be synonymous with other annotations lacking identifiers within an abstract, making it difficult to calculate the minimum distance between two concepts if there are multiple annotations lacking identifiers.

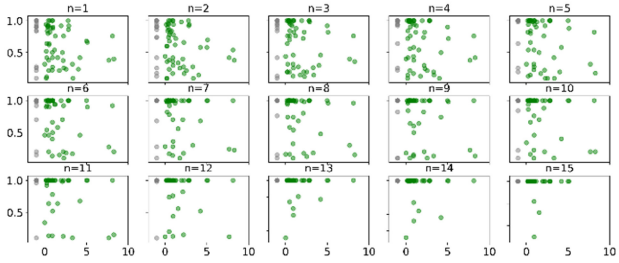

Fig. 4. The minimum estimated number of sentences between two concepts (x-axis) versus the average accuracy of the majority response (y-axis) at different voter numbers (n) for concept pairs which were not considered inappropriate. At distance less than 0 are concepts for which no identifier was available and the minimum distance was not calculated

3.5 Comparison with PMID-specific SemMedDB relationships

Although concept distance is one factor which distinguishes M2C from automated methods that analyze text at the sentence level, we wanted to investigate how the relationships annotated in M2C compared with those annotated via automated methods. We pulled subject–predicate–object triples from SemMedDB and restricted the SemMedDB data to just the abstracts in common with the M2C dataset. Some differences in the relations between SemMedDB and M2C were immediately visible. SemMedDB employs sentence-level analysis and mines the relationships from these sentences regardless of concept type resulting in a very different set of relationship annotations as compared with M2C.

In contrast, M2C users are only presented with concept pairs which are different in type (heterotype), no matter where they may appear in the abstract. For example, users may be asked about the relationship between a gene term and a disease term, but never about the relationship between a disease term and a different disease term. In addition, SemMedDB has many more entity types than M2C, and to do a more detailed comparison, we restricted our comparison of the two to the entities that were found in common.

As seen in Figure 5, the majority of relationships in SemMedDB for the abstracts that have at least one subject entity and one object entity in common with M2C are relationships within the same type (homotype) of concepts. Only four abstracts contained concept pairs that appeared to have a relationship in both SemMedDB and M2C.

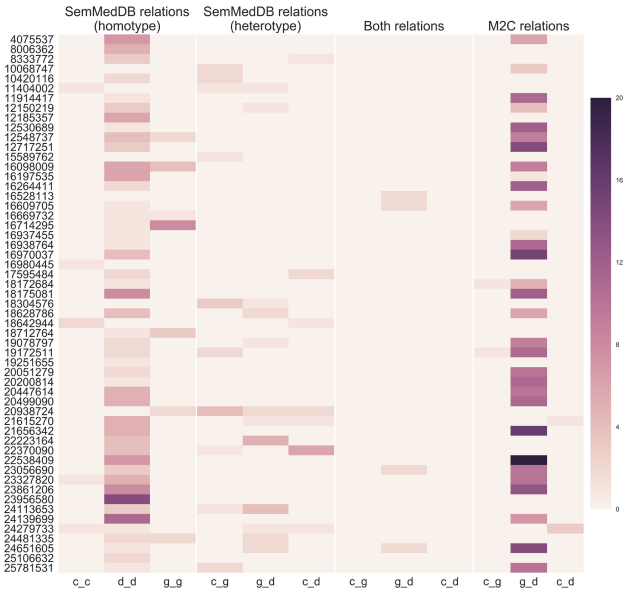

Fig. 5. Heatmap illustrating the number of relationships identified in each abstract via SemMedDB, both SemMedDB and M2C (both relations), and just M2C. M2C only allows relationships between different types of concepts (heterotypic relationships) such as gene–disease (g_d), drug–disease c_d) and drug–gene (c_g) relationships. In contrast, SemMedDB mines for all relationships including relationships within the same types of concepts (homotypic relationships) such as disease–disease (d_d), gene–gene (g_g) and drug–drug (c_c) relationships

Most of the relationships in M2C appear to be gene–disease relationships, but this is due to the removal of concept pairs which were marked as inappropriately annotated. Given the differences in relationship types available in SemMedDB and M2C, we wanted to explore the potential complementarity between the two. For example, gene–disease relationships which are likely to be underrepresented in SemMedDB due to the sentence-level analysis could appear with greater frequency in M2C. Similarly, broad/nonspecific relationships identified in M2C could be explained by multinode/indirect (e.g. A relates to B, B relates to C in lieu of A relates to C) relationships or other relationships available in SemMedDB but not in M2C.

To explore the potential complementarity of the data from the two systems, the quality-checked concept pairs were aggregated in terms of the majority response given per PMID in M2C. The concept pairs were also checked for a direct relationship (i.e. a single SemMedDB triple with both concepts) and for co-occurrence of the concepts in the pair (potential indirect relationships) in SemMedDB (i.e. SemMedDB triples with only one concept in each triple) after filtering for only relationships within the quality-checked PMIDs. As seen in Figure 6, concept pairs with specific relationships appear in many more PMIDs in M2C than in SemMedDB. This difference is largely attributable to NER differences between PubTator and SemMedDB.

For example, the concept pair 8086_x_D000309' (AAAS gene × Adrenal Insufficiency) was observed in 28 abstracts (Supplementary Tables S1 and S2), but only six of those abstracts were both entities part of relationships in SemMedDB. In some cases, these were due to differences in the way text was annotated by PubTator versus SemMedDB (e.g. PubTator marking mentions of 'adrenal insufficiency' as mentions of 'Triple A syndrome' or 'Allgrove syndrome'). In other cases (like '8086_x_D009461': AAAS gene × neurologic dysfunction), mentions observed in PubTator were missed altogether in SemMedDB. If each of the concepts in a concept pair were found in an abstract in SemMedDB, the relationship between the two concepts was not annotated unless the two concepts were in the same sentence as seen in the case of 'D003981_x_D012640' (Diazoxide × Seizures). This concept pair was observed in PMID: 11916319 by both PubTator/M2C and SemMedDB; however, because the two concepts were found in different sentences, a direct relationship between the two concepts is not annotated in SemMedDB.

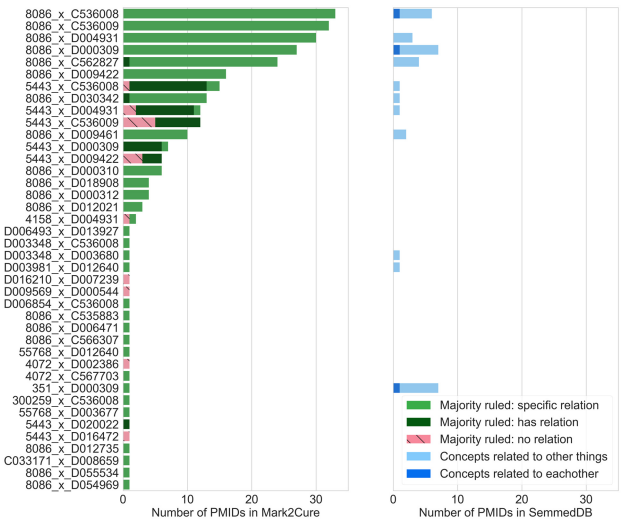

Fig. 6. In spite of the abundance of a specific relationship across multiple PMIDs (as identified in M2C), SemMedDB may miss many of these instances in the same PMIDs. On the left are the number of abstracts in which the majority of users ruled that the concepts had a specific relationship, dark bars is the number of abstracts in which the majority of users ruled that the concepts had a nonspecific relationship, and hatched bars are the number of abstracts which the majority of users ruled that the concepts were unrelated. On the right are the same concept pairs pulled from SemMedDB restricted to the same set of abstracts as they appear in relationship with other concepts (light) or in a direct relationship with one another (dark)

The frequency of relationships in concept pairs arising from M2C may be useful for selecting important relationships in SemMedDB. Data from M2C can be used to identify relationships missed from SemMedDB due to sentence-level analysis. Furthermore, SemMedDB may be useful for clarifying relationships between concept pairs that users consistently marked as having an unspecified relationship if the concepts in M2C can successfully be mapped to those in SemMedDB.

4 Discussion

Improvements in automated RE from biomedical text have been hampered by limited gold standard corpora and dependencies on named entity recognition and entity linking. In spite of these limitations, RE data generated from automated methods like SemRep have been useful for identifying potential prostate cancer drugs (Zhang *et al.*, 2014a), identifying potential drug–drug interactions (Zhang *et al.*, 2014b) and identifying the molecular effects of drugs (Fathiamini *et al.*, 2016) when augmented with other methods or data sources to improve their quality. Generating datasets for improving RE algorithms will help to improve their value as a tool for researchers.

Towards that goal, we found that citizen scientists were willing and capable of performing RE, and that the relationships they extracted from the full abstract were different than those obtained via automated methods like SemMedDB. System-specific restrictions such as sentence-level analysis in SemMedDB and inclusion of only heterogeneous concept pairs in M2C contributed to these differences. Aggregate task performance by the citizen science community was affected by three primary issues: (i) NER quality issues, (ii) training and platform issues and (iii) issues with the documents. Consistent with the literature on information extraction, the quality of the RE data was affected by the NER quality issues (Li *et al.*, 2016; Xing *et al.*, 2018) and users who were uncomfortable or unwilling to discard concepts had lower performance results. NER quality issues also effectively decreased user throughput on actual RE since many RE tasks ended up being NER quality checking task.

User discomfort or inability to discard poor NER concepts further suggests that there are design and training issues. In aggregate, the users generally selected the most appropriate response in spite of the NER issues and limited (and sometimes ambiguous) choice options in the RE task. Also consistent with the literature, design issues (Gabriele and Pölz, 2016; Kosmala *et al.*, 2016) such as the lack of guidance on prioritizing multiple true responses still affected the performance of the citizen scientists, in spite of multiple iterations in the development of the training and platform in an effort to ensure high-quality data. Because annotation guidelines can vary greatly from corpus to corpus, we expect that adherence to the annotation guidelines to affect performance more greatly than the educational/occupational backgrounds of individual citizen scientists.

By investigating concept pairs that had high levels of agreement for different responses in different abstracts, we identified areas in need of refinement in terms of available relationship options and modeling; which led us to identify issues with the documents themselves. The M2C community extracted relationships from abstracts surrounding a unique symptom of NGLY1-deficiency (alacrima), resulting in a dataset that was homogenous and narrow in scope, limiting the types of missing and nonrelationships that could be identified. Furthermore, this document set had a number of case studies in which the relationship between a pair of concepts could be simultaneously true and false. Based on our findings, we expect to be able to raise the community performance on the RE task by providing more guidance on reviewing the concepts (NER) and prioritizing responses in situations involving multiple 'true' responses and/or simultaneous 'true and false' responses. Incorporating additional training or evaluations sets may also help improve volunteer performance and the downstream data analysis in future iterations of the project.

Citizen science is a potential avenue for generating new training datasets for improving automated RE tools, but should not be considered a cheaper version of crowdsourcing. Citizen science projects

require significant investments in training and community management. Resources must be allocated toward incremental user learning and sustained engagement to ensure the success of both the project and its participants.

Although preliminary, we demonstrate that citizen scientists can contribute different types of relationship annotations across three different types of concepts. In contrast, many RE efforts focus on a specific type of RE (Cañada *et al.*, 2017; Collier *et al.*, 2015; Li *et al.*, 2016; Xing *et al.*, 2018). This difference makes it difficult to draw comparisons on contributor performance, but opens up interesting avenues for exploration in RE. In particular, it would be interesting to evaluate the results from this approach with those from nonspecific (Mintz *et al.*, 2009) and medically tailored (Wang and Fan, 2014) RE algorithms, or those from crowdsourced efforts involved in active (Liu *et al.*, 2016) or semi-supervised (Angeli *et al.*, 2014) learning.

Acknowledgements

This article was made possible by citizen scientists. The importance of their contributions both within the M2C system and outside (especially their questions, feedback, suggestions) cannot be emphasized enough, and we are extremely grateful for the privilege of working with such wonderful contributors and collaborators. Contributors who consented to publishing their names can be found here: https://mark2cure.org/blog/relation-paper-contributors/. We would also like to thank the Mights, the Esteniks, the Jennisons and other members of the NGLY1 rare disease community for their inspirational activities and words of support.

Funding

This work was supported by the US National Institute of Health [grant number U54GM114833 to A.I.S.]. This work was also supported by the Scripps Translational Science Institute, an NIH-NCATS Clinical and Translational Science Award [grant number CTSA; 5 UL1 RR025774].

Conflict of Interest: None declared.

References

Angeli,G. *et al.* (2014) Combining distant and partial supervision for relation extraction. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics. Doha, Qatar, pp. 1556–1567.

Aroyo,L. and Welty,C. (2013) Crowd truth: harnessing disagreement in crowdsourcing a relation extraction gold standard. In: *ACM Web Science Conference*. ACM, New York, NY, USA.

Banfield,J. *et al.* (2016) Radio galaxy zoo: discovery of a poor cluster through a giant wide-angle tail radio galaxy. *Mon. Not. R. Astron. Soc.*, **460**, 2376–2384.

Bird,S. *et al.* (2009) *Natural Language Processing with Python*. O'Reilly Media Inc., Sebastopol, CA.

Bodenreider,O. (2004) The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.*, 32, D267–70.

Burger,J. *et al.* (2014) Hybrid curation of gene–mutation relations combining automated extraction and crowdsourcing. *Database*, **2014**, 1–13.

Cañada,A. *et al.* (2017) LimTox: a web tool for applied text mining of adverse event and toxicity associations of compounds, drugs and genes. *Nucleic Acids Res.*, **45**, W484–W489.

Candido dos Reis,F. *et al.* (2015) Crowdsourcing the general public for large scale molecular pathology studies in cancer. *Ebiomedicine*, **2**, 681–689.

Collier,N. *et al.* (2015) PhenoMiner: from text to a database of phenotypes associated with OMIM diseases. *Database*, **2015**, bav104.

Cox,J. *et al.* (2015) Defining and measuring success in online citizen science: a case study of zooniverse projects. *Comput. Sci. Eng.*, **17**, 28–41.

Dumitrache,A. *et al.* (2015) Achieving expert-level annotation quality with CrowdTruth: the case of medical relation extraction. In: Song,D. *et al.* (eds.) *International Workshop on Biomedical Data Mining, Modeling, and Semantic Integration: A Promising Approach to Solving Unmet Medical Needs*. CEUR Workshop Proceedings (CEUR-WS.org), Bethlehem.

Fathiamini,S. *et al.* (2016) Automated identification of molecular effects of drugs (AIMED). *J. Am. Med. Inform. Assoc.*, **23**, 758–765.

Gabriele,W. and Pölz,E. (2016) Data quality in citizen science projects: challenges and solutions. *Front. Environ. Sci.*, **4** doi: 10.3389/conf.FENVS.2016.01.00011.

Good,B. *et al.* (2015) Microtask crowdsourcing for disease mention annotation in PubMed abstracts. *Pac. Symp. Biocomput*, 282–293.

Haklay,M. (2013) Citizen science and volunteered geographic information: overview and typology of participation. In: Sui,D. *et al.* (eds.) *Crowdsourcing Geographic Knowledge*. Springer, Dordrecht, pp. 105–122.

Jovanović,J. and Bagheri,E. (2017) Semantic annotation in biomedicine: the current landscape. *J. Biomed. Semantics*, **8**. doi: 10.1186/s13326-017-0153-x.

Khare,R. *et al.* (2015) Scaling drug indication curation through crowdsourcing. *Database*, **2015**, 1–10.

Kilicoglu,H. *et al.* (2012) SemMedDB: a PubMed-scale repository of biomedical semantic predications. *Bioinformatics*, **28**, 3158–3160.

Kim,J. *et al.* (2014) Space–time wiring specificity supports direction selectivity in the retina. *Nature*, **509**, 331–336.

Kosmala,M. *et al.* (2016) Assessing data quality in citizen science. *Front. Ecol. Environ.*, **14**, 551–560.

Kuchner,M. *et al.* (2016) Disk detective: discovery of new circumstellar disk candidates through citizen science. *Astrophys. J.*, **830**, 84.

Li,J. *et al.* (2009) Building disease-specific drug-protein connectivity maps from molecular interaction networks and PubMed abstracts. *PLoS Comput. Biol.*, **5**, e1000450.

Li,T. *et al.* (2016) A crowdsourcing workflow for extracting chemical-induced disease relations from free text. *Database*, **2016**, baw051.

Liu,A. *et al.* (2016) Effective crowd annotation for relation extraction. In: *Proceedings of NAACL-HLT 2016, San Diego, California*, pp. 897–906.

Lossio-Ventura,J. *et al.* (2018) OC-2-KB: integrating crowdsourcing into an obesity and cancer knowledge base curation system. *BMC Med. Inform. Decis. Mak.*, **18**, *55*.

Lou,Y. *et al.* (2017) A transition-based joint model for disease named entity recognition and normalization. *Bioinformatics*, **33**, 2363–2371.

Luengo-Oroz,M. *et al.* (2012) Crowdsourcing malaria parasite quantification: an online game for analyzing images of infected thick blood smears. *J. Med. Internet. Res.*, **14**, e167.

McKinley,D. *et al.* (2017) Citizen science can improve conservation science, natural resource management, and environmental protection. *Biol. Conserv.*, **208**, 15–28.

Mintz,M. *et al.* (2009). Distant supervision for relation extraction without labeled data. In: *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP*. Suntec, Singapore, pp. 1003–1011.

Morgan,A. *et al.* (2008) Overview of BioCreative II gene normalization. *Genome Biol.*, **9** (**Suppl. 2**), S3.

Murray-Rust,P. (2017) ContentMine: mining the scientific literature. In: *OSC*. SlideShare, Cambridge, MA, doi: 10.17863/CAM.12526.

Muzaffar,A. *et al.* (2015) A relation extraction framework for biomedical text using hybrid feature set. *Comput. Math. Methods Med.*, **2015**, 1–12.

Pafilis,E. *et al.* (2016) EXTRACT: interactive extraction of environment metadata and term suggestion for metagenomic sample annotation. *Database*, **2016**, baw005.

Palermo,E. *et al.* (2017) A natural user interface to integrate citizen science and physical exercise. *PLoS One*, **12**, e0172587.

Panahiazar,M. *et al.* (2017) Predicting biomedical metadata in CEDAR: a study of Gene Expression Omnibus (GEO). *J. Biomed. Inform.*, **72**, 132–139.

Peng,Y. *et al.* (2018) Extracting chemical–protein relations with ensembles of SVM and deep learning models. *Database*, **2018**, 1–9.

Pletscher-Frankild,S. *et al.* (2015) DISEASES: text mining and data integration of disease–gene associations. *Methods*, **74**, 83–89.

Rindflesch,T., and Fiszman,M. (2003) The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *J. Biomed. Inform.*, **36**, 462–477.

Ruch,P. (2016) Text mining to support gene ontology curation and vice versa. *Methods Mol. Biol.*, **1446**, 69–84.

Sauermann,H., and Franzoni,C. (2015) Crowd science user contribution patterns and their implications. *Proc. Natl. Acad. Sci. USA*, **112**, 679–684.

Schmiedel,U. *et al.* (2016) Contributions of paraecologists and parataxonomists to research, conservation, and social development. *Conserv. Biol.*, **30**, 506–519.

Straub,M.C.P. (2016) Giving citizen scientists a chance: a study of volunteer-led scientific discovery. *Citiz. Sci.*, **1**, 1–10.

Sun,D. *et al.* (2017) MPTM: a tool for mining protein post-translational modifications from literature. *J. Bioinform. Comput. Biol.*, **15**, 1740005.

Swanson,D. (1986) Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspect. Biol. Med.*, **30**, 7–18.

Tseytlin,E. *et al*. (2016) NOBLE—flexible concept recognition for large-scale biomedical natural language processing. *BMC Bioinform*., **17**, 1–15.

Tsueng,G. *et al*. (2016) Citizen science for mining the biomedical literature. *Citiz. Sci*., **1**, 14.

Urzúa,U. *et al*. (2010) Tumor and reproductive traits are linked by RNA metabolism genes in the mouse ovary: a transcriptome-phenotype association analysis. *BMC Genomics*, **11** (**Suppl. 5**), S1.

Wang,C. and Fan,J. (2014) Medical relation extraction with manifold models. In: Toutanova,K. and Wu,H. (eds.) *52nd Proc. Conf. Assoc. Comput. Linguist. Meet.* Vol. **1**: Long Papers. Association for Computational Linguistics (ACL), Baltimore, MD, USA, pp. 828–838.

Wei,C. *et al*. (2013) PubTator: a web-based text mining tool for assisting biocuration. *Nucleic Acids Res*., **41**, W518–W522.

Wei,C. *et al*. (2015) GNormPlus: an integrative approach for tagging genes, gene families, and protein domains. *Biomed. Res. Int*., **2015**, 1.

Wei,C. *et al*. (2016) Assessing the state of the art in biomedical relation extraction: overview of the BioCreative V chemical-disease relation (CDR) task. *Database*, **2016**, 1–8.

Williams,A. *et al*. (2014) A computational pipeline for crowdsourced transcriptions of Ancient Greek papyrus fragments. In: *2014 IEEE International Conference on Big Data (Big Data)*, pp. 100–105.

Xing,W. *et al*. (2018) A gene–phenotype relationship extraction pipeline from the biomedical literature using a representation learning approach. *Bioinformatics*, **34**, i386–i394.

Yang,H. *et al*. (2016) Literature-based discovery of new candidates for drug repurposing. *Brief. Bioinform*., **18**, 488–497.

Zhang,R. *et al*. (2014a) Exploiting literature-derived knowledge and semantics to identify potential prostate cancer drugs. *Cancer Inform*., **13** (**Suppl. 1**), 103–111.

Zhang,R. *et al*. (2014b) Using semantic predications to uncover drug–drug interactions in clinical data. *J. Biomed. Inform*., **49**, 134–147.

Zhou,H. *et al*. (2018) Chemical-induced disease relation extraction with dependency information and prior knowledge. *J. Biomed. Inform*., **84**, 171–178.

Zhu,F. *et al*. (2013) Biomedical text mining and its applications in cancer research. *J. Biomed. Inform*., **46**, 200–211.

Zhu,Q. *et al*. (2018) GRAM-CNN: a deep learning approach with local context for named entity recognition in biomedical text. *Bioinformatics*, **34**, 1547–1554.