Automating case definitions using literature-based reasoning

T. Botsis[1, 2]; R. Ball[1]

[1]Office of Biostatistics and Epidemiology, Center for Biologics Evaluation and Research (CBER), Food and Drug Administration (FDA), Rockville, MD; [2]Department of Computer Science, University of Tromsø, Tromsø, Norway;

Keywords

Case definition, safety surveillance, semantic networks, literature-based reasoning, anaphylaxis, similarity

Summary

Background: Establishing a Case Definition (CDef) is a first step in many epidemiological, clinical, surveillance, and research activities. The application of CDefs still relies on manual steps and this is a major source of inefficiency in surveillance and research.

Objective: Describe the need and propose an approach for automating the useful representation of CDefs for medical conditions.

Methods: We translated the existing Brighton Collaboration CDef for anaphylaxis by mostly relying on the identification of synonyms for the criteria of the CDef using the NLM MetaMap tool. We also generated a CDef for the same condition using all the related PubMed abstracts, processing them with a text mining tool, and further treating the synonyms with the above strategy. The co-occurrence of the anaphylaxis and any other medical term within the same sentence of the abstracts supported the construction of a large semantic network. The 'islands' algorithm reduced the network and revealed its densest region including the nodes that were used to represent the key criteria of the CDef. We evaluated the ability of the "translated" and the "generated" CDef to classify a set of 6034 H1N1 reports for anaphylaxis using two similarity approaches and comparing them with our previous semi-automated classification approach.

Results: Overall classification performance across approaches to producing CDefs was similar, with the generated CDef and vector space model with cosine similarity having the highest accuracy (0.825±0.003) and the semi-automated approach and vector space model with cosine similarity having the highest recall (0.809±0.042). Precision was low for all approaches.

Conclusion: The useful representation of CDefs is a complicated task but potentially offers substantial gains in efficiency to support safety and clinical surveillance.

Correspondence to:
Taxiarchis Botsis, PhD, MS
Office of Biostatistics and Epidemiology, CBER, FDA
Woodmont Office Complex 1, Rm 306N
1401 Rockville Pike
Rockville, MD 20852
Tel. +1 301 827 5405
E-mail: Taxiarchis.Botsis@fda.hhs.gov

Appl Clin Inform 2013; 4: 515–527
DOI: 10.4338/ACI-2013-04-RA-0028
received: April 24, 2013
accepted: October 8, 2013
published: October 30, 2013
Citation: Botsis T, Ball R. Automating case definitions using literature-based reasoning. Appl Clin Inf 2013; 4: 515–527
http://dx.doi.org/10.4338/ACI-2013-04-RA-0028

1. Introduction

One of the first steps in any clinical or epidemiological investigation is defining the outcome of interest; this is often done by establishing a case definition [1]. According to Merril,

> "A Case Definition involves standard clinical criteria that are used to establish whether a person has a particular disease. Applying a standard case definition will guarantee that every case is consistently diagnosed, no matter when and/or where the diagnosis occurs".

Case definitions (CDefs) are often used for public health surveillance. For example, the World Health Organization (WHO) develops CDefs for the surveillance of various conditions, such as pertussis [2]. The Centers for Disease Control (CDC) and Prevention perform the same task and maintain an annually updated database of CDefs for infectious and non-infectious conditions [3–5]. Surveillance for adverse events after vaccination has been facilitated by the Brighton Collaboration's (BC) development of standardized CDefs for adverse events following immunizations (AEFIs); these CDefs are endorsed by WHO [6]. Often, the CDef development is assigned to special working groups that accomplish this task in labor-intensive steps, such as systematic literature reviews and frequent expert meetings [7–9].

The discussion about theory building from case studies and the search for cross-case patterns in the literature was initiated a few decades ago [10, 11], but the existing technology to support those efforts was limited at that time. Even now there is little published literature on the systematic application of medical informatics approaches to case definition development, and what does exist tends to focus on applications to specific conditions in a particular context. A few standalone efforts have focused on the use of administrative data and ICD (International Classification of Diseases) or other codes for the development and validation of automated CDefs to identify relevant patient cases in Electronic Medical Records (EMR) [12–14]. Existing CDefs have also been applied to EMR data, such as the ICD-based Centers for Disease Control and Prevention CDef for non-fatal head trauma [15] and tailored CDefs for diabetes and asthma in Datalink and PEAL Network projects, respectively [16]. Furthermore, machine learning approaches have been used to distinguish cases from non-cases in EMRs [17]. Kohl et al have also shown that the application of six BC CDefs was very successful in detecting confirmed cases of reported clinical events [18]. In a previous study, we combined the BC CDef for anaphylaxis [19] with a dedicated text mining algorithm to classify a large set of spontaneous reports submitted to the FDA's and CDC's Vaccine Adverse Event Reporting System (VAERS) [20]. Despite its high performance, the development of a dedicated rule-based solution is labor-intensive.

A key concept from the field of case-based reasoning, which includes the idea of using existing cases to create knowledge about a particular topic, and build upon it, is that of similarity. Similarity has been developed to provide a rigorous framework for measuring the semantic similarity of cases of interest based on particular features [21] or concepts [22]. Some work has been done on combining case-based reasoning and machine learning to identify features that are most useful for predicting outcomes of interest [23–25] and extracting associations from literature and clinical documents [26, 27] but such approaches have not been directly applied to automating CDef development [28]. Similarity in case-based reasoning is consistent with the standard epidemiological notion of a CDef introduced above [1] and can serve as the basis for automating the development and updating of CDefs. The definition of "similar" depends on the purpose for which the CDef is being developed. Typically, CDefs are developed by experts to achieve a certain goal; for example, a common use of a CDef is to ensure that similar cases are enrolled in a clinical trial to study the effect of a drug on a particular medical condition [29] or identify conditions of public health importance in a surveillance system. In this similarity framework, a case definition consists of the features (e.g. signs, symptoms and laboratory values) and describes their relationship to one another as well as their contribution to validly and reliably predicting an outcome.

We illustrated this approach when we recently demonstrated how features extracted by a text mining system previously developed, namely Vaccine *a*dverse *e*vent Text Mining (V*ae*TM), can be used within a similarity framework to automatically assess whether the reports meet criteria for clas-

sification as possible cases of anaphylaxis using the same CDef [30]. To make this process more efficient and generalizable, we used a semi-automated strategy (▶Figure 1A) that included:

i the specification of the key words for each criterion of the BC CDef for anaphylaxis;
ii the synonym identification for those key words using the Unified Medical Language System (UMLS) Metathesaurus; and
iii the manual curation of the lists of key words and synonyms.

The automation of the whole process, to include not just the application of CDefs but their development and update, could result in a more efficient process. Here, we outline what we consider to be the main problems in translating existing CDefs into automated algorithms and automating the generation of CDefs from the medical literature. We illustrate some of the key issues by extending our previous work for the use of CDefs in the area of post-market medical product safety surveillance [20, 30], and provide some preliminary results exploring possible solutions. Even though we focus on the development of a CDef that is of primary interest for safety surveillance, we believe that this approach can find application to the broader spectrum of CDefs in other settings.

2. Methods

2.1. Translating Existing Case Definitions into Machine Readable Algorithms

In our previous work [18], two steps required human intervention, namely step (ii) – the synonym identification for those key words using the Unified Medical Language System (UMLS) Metathesaurus – and, step (iii) – the manual curation of the lists of key words and synonyms. In the current work, the steps related to the manual identification of synonyms are substituted by a fully automated process (▶Figure 1B). The BC CDef criteria for anaphylaxis are processed with the NLM MetaMap tool by selecting the "word sense disambiguation" option and a score threshold of 850 to extract the top mappings to the UMLS Metathesaurus [31, 32]. Subsequently, among the various terms that appear in the MetaMap output with their semantic type, we select those terms that appear under the following semantic types: "Disease or Syndrome", "Disease/Finding", "Finding", "Sign or Symptom", "Pathologic Function", "Neoplastic Process", "Mental or Behavioral Dysfunction" and "Injury or Poisoning", because these types mostly include medical terms that could serve our purposes. The selected medical terms and their synonyms synthesize the "translated" CDef.

2.2. Automatically Generating Case Definitions from Medical Literature

We used published biomedical literature and principles of knowledge discovery through the mining of large corpora. Following the corpora creation, the key terms (and their synonyms) related to the condition of interest were specified in an automated fashion. Various text mining strategies have been applied before to extract relationships from biomedical literature, such as the gene-disease and disease-treatment relationships [33]. Similar approaches have built semantic networks to explore the semantic relationships between the nodes of the network that represent the key terms in the corpus, such as regulatory gene-protein networks [34]. A number of literature-based discovery systems have combined semantic with graph-based methodologies [35–38]. The term co-occurrence in a sentence has been studied before by constructing dependency trees [39, 40] or performing semantic role labeling using natural language processing and other techniques [41–43].

To construct the corpus we selected abstracts including the term for "anaphylaxis" and applied a text mining technique to extract medical terms. We used a graph-based method to create the semantic network structure to represent the relationships between the terms existing in the same document and, particularly, the terms co-occurring in the same sentence since they are semantically related. We relied on the term co-occurrence considering that the term for the condition would be reported in the same sentence with other important medical terms related to the condition. Co-occurrence would then form the relationships that would be represented as edges in the semantic network. It was expected that the final network would be a large structure with many connections and

nodes. We therefore applied an algorithm that reduces the network and identifies the densest region(s) within it. The reduced network structure is used to define the CDef for the condition under study (▶Figure 1C).

To identify the most densely connected nodes in this network we applied the islands algorithm using the triangular weights of the edges (TW) [44]. The TW for any given edge is equal to the number of triangles this particular edge participates to, while the islands algorithm creates a subnetwork including a pre-specified number of nodes above a certain TW threshold [45]. As we showed before, this combined approach filters out weak connections and allows patterns to stand out [46].

We executed a PubMed query on July 16, 2012 to retrieve all publications that included the word "anaphylaxis" in the title/abstract. Subsequently, the V*ae*TM system processed the free text and extracted the corresponding features [30]. Only the non-negated medical terms under DIAGNOSIS, CAUSE_OF_DEATH, SECOND_LEVEL_DIAGNOSIS and SYMPTOM features were retained. Based on our experience, negations are not significant in the case of anaphylaxis but might be important for other conditions and should be included in the generation process. The anaphylaxis-related terms (e.g. anaphylactic reaction) were replaced by the term "anaphylaxis" to create a single representation of the core concept in the CDef and along with their co-occurring terms that co-appeared in the same sentences of the corpus formed a list (hereafter, "final list"). A unique sentence id was included in the "final list" to denote the term co-occurrence, i.e. the co-existence of "anaphylaxis" and other medical terms in the same sentence. Thus, a sentence id in the final list was necessarily linked to at least one "anaphylaxis" term and one or more other medical terms.

As in the translation of the CDef (▶Figure 1B), the "final list" of terms was processed by the NLM MetaMap tool selecting the "term processing" and "word sense disambiguation" options as well as the same score threshold. The terms falling under the same aforementioned NLM MetaMap semantic types replaced their mapping terms in the "final list" to include the "normalized" terms from the MetaMap output and better facilitate the generation of the CDef. To trim this list and recognize the most related to anaphylaxis terms, we created a network with the nodes and the edges representing the terms and their co-occurrence (as denoted by the sentence id) in the sentences, respectively. Thus, edges were created between the anaphylaxis and other medical term(s) nodes as well as between the medical terms nodes.

All the steps that were followed for the generation of the CDef are illustrated in ▶Figure 1C. The network was built and reduced with Pajek 2.02, a tool for large network analysis developed at the University of Ljubljana (Ljubljana, Slovenia); ORA 2.2.8b, a dynamic meta-network assessment and analysis tool developed at Carnegie Mellon (Pittsburgh, PA), was also used to create the visualization of the reduced network.

2.3. Use case: Automated case definitions and case classification for post-market vaccine safety surveillance

We evaluated both the "translated" and the "generated" CDef by investigating their ability to support the classification of reports and comparing their performance with our previous results that were based on the semi-automated approach [30]. We also tested whether the subnetwork resulting from the steps described in section 2.2 was equivalent to the formal CDef for anaphylaxis and its nodes represented the corresponding key medical terms that synthesized the "generated" CDef. We used the same set of 6034 VAERS reports for H1N1 vaccine as before [20, 30]; this set was randomly split into five subsets that were used to evaluate the three approaches in a 5-fold cross validation process. In our previous semi-automated work, we processed the symptom text of these reports with V*ae*TM and mapped the output to the criteria of the BC CDef for anaphylaxis using a manually curated list of synonyms [30]. To create this list we identified the synonyms to the BC criteria using the MetamorphoSys Unified Medical Language System (UMLS) [47] as well as the Medical Dictionary for Regulatory Activities (MedDRA) [48] browsers. One of the authors (RB) reviewed the initial list and confirmed the correct mapping of each term to the appropriate criterion of the BC case definition. In the current study, the same V*ae*TM output was mapped to the terms of the "translated" and "generated" CDefs.

Following the identification of the CDef terms in the reports per se, we applied two methods to quantify the similarity of each report to each CDef. First, we used the vector space model and repre-

sented both the reports and the CDef as vectors. These vectors included the key terms of the CDef originating from either the fully- or the semi-automated approach. The vector components were weighted by applying one of the weighting schemes suggested by the SMART notation [49]. Using the numerical representation of the vectors, we calculated the cosine similarity of the two vectors [49]. Second, we used the information theoretic similarity framework proposed by Lin [50] and applied to a document network [51] and medical cases in electronic medical records [21]. We thus measured the similarity between each report and the CDef as the ratio between the common information the two objects share and the total information needed to fully describe them. For each of the similarity approaches and CDef, the reference standard (i.e. medical expert's classification) and a 5-fold cross validation were used to calculate the standard text classification metrics of recall, precision, accuracy and F-measure.

Subsequently, the average values over the five assessments of the cross-validation were calculated.

3. Results

3.1 "Translated" CDef

We used the criteria of the BC CDef for anaphylaxis, which is summarized in the ▶Supplementary File Appendix 1, and processed them with the NLM MetaMap tool. This processing resulted in the identification of sixty-eight synonyms for the criteria and formed the "translated" CDef.

3.2 "Generated" CDef

The search of PubMed for "anaphylaxis" returned 10300 entries with 7757 of them including an abstract. This subset (N = 7757) was split into sentences that were further filtered to identify those that included the term (or a synonym) of anaphylaxis. These sentences (N = 14067) formed the corpus that was used for the development of the automated CDef for anaphylaxis. Based on the "final list" that was described above we created a network of 967 nodes. The application of the triangle weight and islands algorithm resulted in a reduced network of forty-eight nodes that represented the key terms in the "generated" CDef for anaphylaxis plus the anaphylaxis node that appeared in the center of the reduced network topology (▶Figure 2). Most of these terms also appear in the major and minor criteria of the corresponding BC CDef [19].

3.3. Use case

As shown in ▶Table 1 there was no clear winner across both similarity measures and all performance metrics. Surprisingly, the "generated" CDef performed better compared to our previous semi-automated approach in terms of average accuracy suggesting its effectiveness in the correct identification of the true cases (positive and negative for anaphylaxis). On the other hand, the "translated" CDef showed the lowest performance in general. Noteworthy, the "semi-automated" had the best recall and all three CDefs had poor precision due to the considerable number of false positives. Moreover, the use of the two similarity approaches did not result in any considerable differences in the classification of reports.

Since we are only testing one example in the use case, additional examples would need to be evaluated to determine whether characteristics of a particular condition might affect performance of the method.

4. Discussion

To the best of our knowledge this is the first effort that attempts to discuss and investigate the automated representation of CDefs based on the concept of information retrieval and knowledge discovery from the biomedical literature. We "translated" and "generated" a CDef aiming at the algorithmic representation for a particular condition that could facilitate the automated identification of

potential cases for anaphylaxis. The selected use case and our classification results showed that the proposed methodology, i.e. the combination of a dedicated text mining tool (i.e. V*ae*TM) with existing resources (i.e. NLM MetaMap) allows the implementation of this idea. The performance of the "generated" CDef might be attributed to the large corpus that supported the development of the CDef based on a larger pool of medical terms compared to the BC criteria. The suggested methodology might facilitate not only the automated development but also the continuous update of CDefs. The latter is particularly important for uncommon events that have not been fully investigated and are subject to periodic updates based on the new findings.

Given its preliminary nature our study has certain limitations. We have intentionally used fairly simple information retrieval algorithms, which are widely available. These algorithms cannot replicate the sophisticated cognitive processes of experts. It could be also argued that the proposed methodology does not reproduce all the characteristics of a BC CDef. For example, the BC CDefs define various levels of diagnostic certainty [52] that are not determined by our suggested methodology. It might be possible to identify highly linked nodes in a network (e.g. nodes representing gastrointestinal and the dermatological medical terms) that form distinct regions in the anaphylaxis network (▶Figure 2), and use them to replicate the levels of diagnostic certainty in a subsequent analysis. In any case, the retrieved information combines both the research findings and clinical knowledge that are some of the critical components in the decision making process. More sophisticated emulation of human cognition or application of more rigorous approaches to similarity might improve results.

We recognize that the use of abstracts instead of full texts might be also a restricting factor for the complete retrieval of the available information. It should be noted though that the power of the summarized information available in abstracts has been previously demonstrated through the use of a corpus including MEDLINE titles and abstracts only. This effort was very successful in identifying unknown protein-protein interactions based on the conjunction of two protein names in the same abstract [53]. Furthermore, the synthesis of a corpus of full texts would be time-consuming and expensive and the benefits of doing so would have to be demonstrated compared with the use of abstracts alone. We should also clarify that we used the entire corpus of abstracts to develop our CDef because we had an existing expert labeled set of cases for validation purposes.

The anaphylaxis subnetwork included nodes that represent inflections of the same term, e.g. "allergic reactions" and "reactions, allergic" for "allergic reaction"; merging them into a single node might have resulted in a more concise CDef and should be further investigated, e.g. by incorporating a distributional semantics approach in our network analysis technique [54]. However, this should be treated automatically – any manual curation would violate the main goal in our methodology, i.e. the elimination of any non-automated steps. This is definitely a general issue related to the identification as well as the representation of synonym terms by single entities irrespective of the technique used for the definition of semantically related concepts.

To fully evaluate this approach additional development of certain key aspects will be necessary. These are the determination of the key medical concepts to represent a case definition and the identification of the synonyms to represent those concepts as well as the inclusion of additional features, such as laboratory values, in the construction of a CDef. Approaches to automatic synonym generation have been described in the literature but there is no agreed upon, standard approach. For example, the multi-word terms of a gastrointestinal terminology were split into single words and all their potential combinations were mapped to the existing UMLS terms to identify the list of synonyms [55, 56]. This process was not fully automated and involved manual curation of the final list. In a recent literature-based discovery study, the synonym issue was treated by a more automated strategy using an existing dictionary to identify gene synonyms [57]. NLM MetaMap tool is a more complete solution that is a fully automated and maps any medical term to the UMLS Metathesaurus; it also resolves some of the disambiguation issues [31, 32] that are critical in biomedical text processing [58, 59]. The inclusion of abnormal laboratory values is not critical when building a CDef for anaphylaxis but might be important for other conditions. The proposed methodology should be then extended to extract the laboratory values and determine their abnormality. This might require substantive work using natural language processing techniques but the overall approach described above would not need to be altered.

We have also elected to build a semantic network around the condition of interest rather than pursue a different approach, e.g. a machine learning strategy, assuming that co-occurrence might better support the identification of the key terms in a CDef. The next challenging step in the semantic network process is to filter out the weak and retain the strong condition-to-term(s) relationships; those that really matter for a particular condition and could practically synthesize the corresponding CDef. The triangle formation is particularly important in semantic networks, since it is strongly correlated with the content similarity nodes that participate in the triangle [60]. This is applicable to our proposed methodology, given that the actual conceptual unit (sentence rather than the whole abstract) that defines the term co-occurrence is likely to include semantically related terms. While we used the triangle approach, the identification of the optimal algorithm is an important area of further research. Machine learning algorithms and statistical approaches to key concept identification [29, 34], are also worthwhile to explore.

It may be the case that the approach we described here works best for already developed case definitions used for safety surveillance where the need for rapid screening of large numbers of reports allows for the use of relatively simple informatics approaches. If fully automated, this process could be generalized to other published CDefs. We note that automation of the synonym list also allows periodic update of the case definition, assuming the fundamental structure and medical concepts in the CDef have not changed. We believe that the paradigm of automated CDef development introduced here might be more widely applied to other applications. This broader paradigm might be thought of as an instantiation of distributed cognition or Literature-Based Reasoning (LiBRe). LiBRe is the intersection of the case-based reasoning (CBR) that capitalizes on past experience to solve current problems [28] and the Literature-Based Discovery (LBD) that identifies unknown correlations between key terms in non-interacting literatures [61]. LiBRe could also borrow from both areas in the sense that not only utilizes existing knowledge for a particular topic but also attempts to identify the most connected terms in a corpus of published studies.

The generation of CDefs is a complicated task and multiple techniques must be combined at all stages: from the composition of the appropriate corpus to the identification of the key terms and concepts that are expected to synthesize the CDef. We showed that some of those techniques could be used for the automated translation of existing CDefs for medical product safety surveillance.

Acknowledgements
This project was supported in part by the appointment of Taxiarchis Botsis to the Research Participation Program at the Center for Biologics Evaluation and Research administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the U.S. Food and Drug Administration. The authors would like to thank Michael D. Nguyen for the insightful review of the manuscript and his constructive comments.

Competing Interest
None declared.

Protection of Human and Animal Subjects
Human and/or animal subjects were not included in the project.

Informal Communication Disclaimer
Our contributions are an informal communication and represent our own best judgment. These comments do not bind or obligate FDA.

Contributorship Statement
TB and RB equally contributed to framing the idea, designing the methodology, and writing the paper.

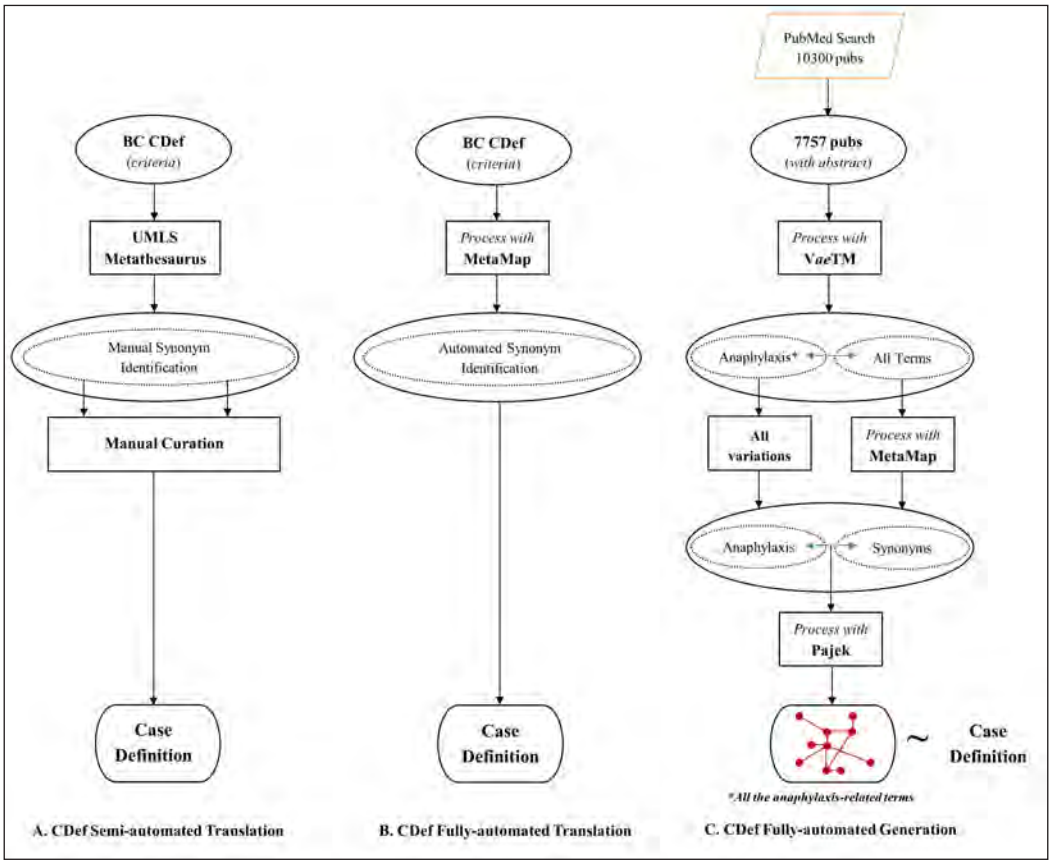

Fig. 1 Flow chart illustrating: **A)** the semi-automated translation of the case definition for anaphylaxis; **B)** the fully-automated translation of the case definition for anaphylaxis; **C)** the automated generation of the case definition for anaphylaxis.

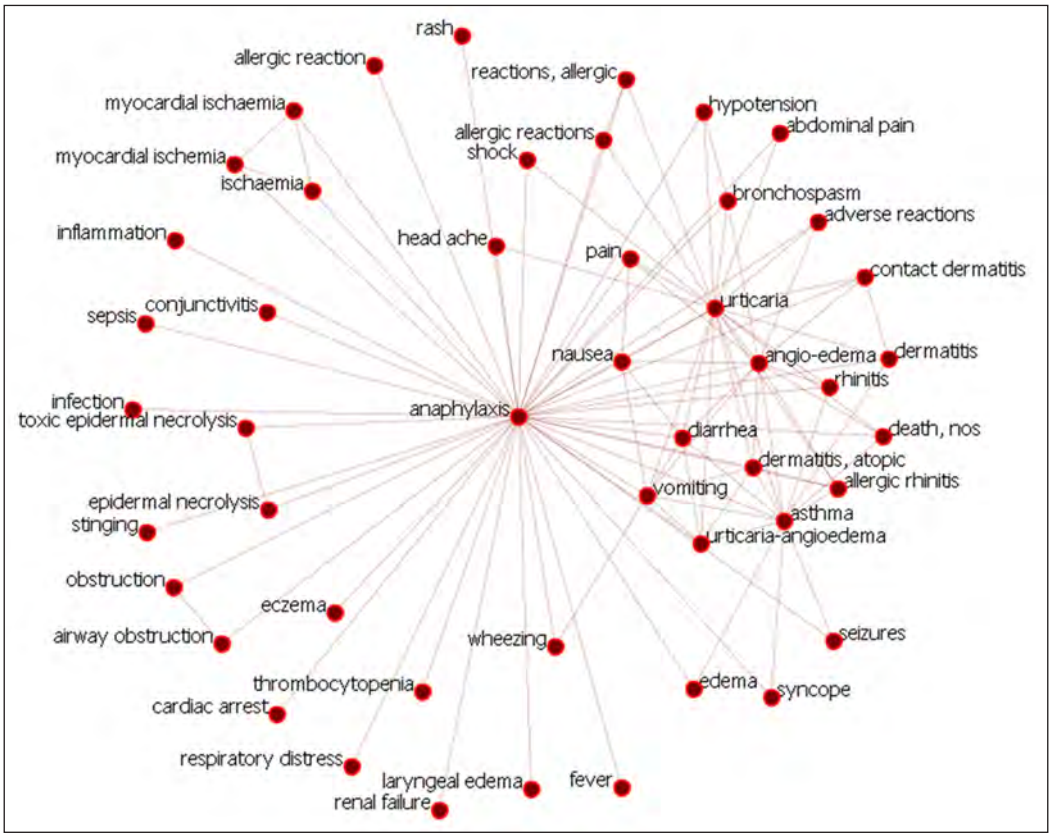

Fig. 2 The subnetwork (or island) based on the triangular weights; it represents the automated case definition and its 48 nodes the corresponding key terms.

Table 1 Average recall, precision and accuracy and their associated standard errors (SE) of the classification accomplished by the three Case Definitions (CDefs) for anaphylaxis using the vector space model and the information theoretic similarity over the 5-folds of the cross validation; also, the corresponding F-measure values are provided.

		recall±SE	precision±SE	accuracy±SE	F-measure
Vector Space Model w/cosine similarity	Translated CDef	0.664±0.059	0.103±0.008	0.731±0.022	0.173±0.010
	Semi-automated approach	0.809±0.042	0.106±0.006	0.701±0.027	0.186±0.009
	Generated CDef	0.667±0.045	0.141±0.008	0.825±0.004	0.232±0.012
Information-theoretic Similarity	Translated CDef	0.629±0.057	0.116±0.009	0.768±0.022	0.190±0.012
	Semi-automated approach	0.791±0.031	0.127±0.005	0.777±0.003	0.218±0.008
	Generated CDef	0.707±0.050	0.115±0.006	0.772±0.003	0.197±0.010

SE: Standard Error; CDef: Case Definition

References

1. Merrill R. Introduction to Epidemiology. 5th ed. Jones & Bartlett Learning; 2010.
2. Ghanaie RM, Karimi A, Sadeghi H, Esteghamti A, Falah F, Armin S, Fahimzad A, Shamshiri A, Kahbazi M, Shiva F. Sensitivity and specificity of the World Health Organization pertussis clinical case definition. International Journal of Infectious Diseases 2010; 14(12): e1072-e1075.
3. CDC. National Notifiable Diseases Surveillance System (NNDSS). December 7, 2012. Available from: http://wwwn.cdc.gov/nndss/.
4. Koo D, Wharton M, Birkhead G. Case Definitions for Infectious Conditions Under Public Health Surveillance. MMWR Recomm Rep 1997; 46(RR-10): 1–64.
5. Wharton M, Chorba TL, Vogt RL, Morse DL, Buehler JW. Case definitions for public health surveillance. MMWR Recomm Rep 1990; 39(RR-13): 1–43.
6. Bonhoeffer J, Kohl K, Chen R, Duclos P, Heijbel H, Heininger U, Jefferson T, Loupi E. The Brighton Collaboration: addressing the need for standardized case definitions of adverse events following immunization (AEFI). Vaccine 2002; 21(3): 298–302.
7. Ball R, Halsey N, Braun MM, Moulton LH, Gale AD, Rammohan K, Wiznitzer M, Johnson R, Salive ME. Development of case definitions for acute encephalopathy, encephalitis, and multiple sclerosis reports to the Vaccine Adverse Event Reporting System. Journal of Clinical Epidemiology 2002; 55(8): 819–824.
8. Berry SH, Bogart LM, Pham C, KARIN LIU, Nyberg L, Stoto M, Suttorp M, Clemens JQ. Development, validation and testing of an epidemiological case definition of interstitial cystitis/painful bladder syndrome. The Journal of Urology 2010; 183(5): 1848–1852.
9. Bines JE, Ivanoff B, Justice F, Mulholland K. Clinical case definition for the diagnosis of acute intussusception. Journal of Pediatric Gastroenterology and Nutrition 2004; 39(5): 511–518.
10. Eisenhardt KM. Building theories from case study research. Academy of Management Review 1989; 532–550.
11. Hullermeier E. Case-based approximate reasoning. 44 ed. Springer; 2007.
12. Cunningham A, Stein CM, Chung CP, Daugherty JR, Smalley WE, Ray WA. An automated database case definition for serious bleeding related to oral anticoagulant use. Pharmacoepidemiology and Drug Safety 2011; 20(6): 560–566.
13. Leslie WD, Lix LM, Yogendran MS. Validation of a case definition for osteoporosis disease surveillance. Osteoporosis International 2011; 22(1): 37–46.
14. Reid AY, et al. Development and validation of a case definition for epilepsy for use with administrative health data. Epilepsy Res 2012; 102(3): 173–179.
15. Parks S, Sugerman D, Xu L, Coronado V. Characteristics of non-fatal abusive head trauma among children in the USA, 2003–2008: application of the CDC operational case definition to national hospital inpatient data. Injury Prevention 2012; 18(6): 392–398.
16. Desai JR, Wu P, Nichols GA, Lieu TA, O' Connor PJ. Diabetes and Asthma Case identification, validation, and representativeness when using electronic health data to construct registries for comparative effectiveness and epidemiologic research. Medical Care 2012; 50: S30-S35.
17. Afzal Z, Schuemie MJ, van Blijderveen JC, Sen EF, Sturkenboom MC, Kors JA. Improving sensitivity of machine learning methods for automated case identification from free-text electronic medical records. BMC Medical Informatics and Decision Making 2013; 13(1): 1–11.
18. Kohl KS, Magnus M, Ball R, Halsey N, Shadomy S, Farley TA. Applicability, reliability, sensitivity, and specificity of six Brighton Collaboration standardized case definitions for adverse events following immunization. Vaccine 2008; 26(50): 6349–6360.
19. Ruggeberg JU, Gold MS, Bayas JM, Blum MD, Bonhoeffer J, Friedlander S, de Souza BG, Heininger U, Imoukhuede B, Khamesipour A. Anaphylaxis: case definition and guidelines for data collection, analysis, and presentation of immunization safety data. Vaccine 2007; 25(31): 5675.
20. Botsis T, Nguyen MD, Woo EJ, Markatou M, Ball R. Text mining for the Vaccine Adverse Event Reporting System: medical text classification using informative feature selection. Journal of the American Medical Informatics Association 2011; 18(5): 631–638.
21. Cao H, Melton GB, Markatou M, Hripcsak G. Use abstracted patient-specific features to assist an information-theoretic measurement to assess similarity between medical cases. Journal of Biomedical Informatics 2008; 41(6): 882–888.
22. Batet M, Sánchez D, Valls A. An ontology-based measure to compute semantic similarity in biomedicine. Journal of Biomedical Informatics 2011; 44(1): 118–125.
23. Begum S, Ahmed MU, Funk P, Xiong N, Von Scheele B. A case-based decision support system for individual stress diagnosis using fuzzy similarity matching. Computational Intelligence 2009; 25(3): 180–195.

24. Huang ML, Hung YH, Lee WM, Li RK, Wang TH. Usage of case-based reasoning, neural network and adaptive neuro-fuzzy inference system classification techniques in breast cancer dataset classification diagnosis. Journal of Medical Systems 2012; 36(2): 407–414.
25. van den Branden M, Wiratunga N, Burton D, Craw S. Integrating case-based reasoning with an electronic patient record system. Artificial Intelligence in Medicine 2011; 51(2): 117–123.
26. Chen ES, Hripcsak G, Xu H, Markatou M, Friedman C. Automated acquisition of diseaseGÇôdrug knowledge from biomedical and clinical documents: an initial study. Journal of the American Medical Informatics Association 2008; 15(1): 87–98.
27. Markatou M, Don PK, Hu J, Wang F, Sun J, Sorrentino R, Ebadollahi S. Case-based reasoning in comparative effectiveness research. IBM Journal of Research and Development 2012; 56(5): 4–1.
28. Bichindaritz I, Marling C. Case-based reasoning in the health sciences: What's next? Artificial Intelligence in Medicine 2006; 36(2): 127–135.
29. Letang E, Naniche D, Bower M, Miro JM. Kaposi sarcoma-associated immune reconstitution inflammatory syndrome: In need of a specific case definition. Clinical Infectious Diseases 2012; 55(1): 157–158.
30. Botsis T, Buttolph T, Nguyen MD, Winiecki S, Woo EJ, Ball R. Vaccine adverse event text mining system for extracting features from vaccine safety reports. Journal of the American Medical Informatics Association 2012; 19(6): 1011–1018.
31. Aronson AR. Metamap: Mapping text to the UMLS metathesaurus. Bethesda, MD: NLM, NIH, DHHS 2006.
32. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. Journal of the American Medical Informatics Association 2010; 17(3): 229–236.
33. Bundschus M, Dejori M, Stetter M, Tresp V, Kriegel HP. Extraction of semantic biomedical relations from text using conditional random fields. BMC Bioinformatics 2008; 9(1): 207.
34. Spacic J, Jensen LJ, Ouzounova R, Rojas I, Bork P. Extraction of regulatory gene/protein networks from Medline. Bioinformatics 2006; 22(6): 645–650.
35. Cameron D, Bodenreider O, Yalamanchili H, Danh T, Vallabhaneni S, Thirunarayan K, Sheth AP, Rindflesch TC. A graph-based recovery and decomposition of swanson's hypothesis using semantic predications. Journal of Biomedical Informations 2013; 46(2): 238–251.
36. Miller CM, Rindflesch TC, Fiszman M, Hristovski D, Shin D, Rosemblat G, Zhang H, Strohl KP. A closed literature-based discovery technique finds a mechanistic link between hypogonadism and diminished sleep quality in aging men. Sleep 2012; 35(2): 279.
37. Wilkowski B, Fiszman M, Miller CM, Hristovski D, Arabandi S, Rosemblat G, Rindflesch TC. Graph-Based Methods for Discovery Browsing with Semantic Predications. American Medical Informatics Association Annual Meeting 2011. p. 1514.
38. Yetisgen-Yildiz M, Pratt W. A new evaluation methodology for literature-based discovery systems. Journal of Biomedical Informatics 2009; 42(4): 633.
39. Coulet A, Shah NH, Garten Y, Musen M, Altman RB. Using text to build semantic networks for pharmacogenomics. Journal of Biomedical Informatics 2010; 43(6): 1009–1019.
40. Fundel K, Kuffner R, Zimmer R. RelEx-Relation extraction using dependency parse trees. Bioinformatics 2007; 23(3): 365–371.
41. Barnickel T, Weston J, Collobert R, Mewes HW, Stümpflen V. Large scale application of neural network based semantic role labeling for automated relation extraction from biomedical texts. PLoS One 2009; 4(7): e6393.
42. Bethard S, Lu Z, Martin JH, Hunter L. Semantic role labeling for protein transport predicates. BMC Bioinformatics 2008; 9(1): 277.
43. Kogan Y, Collier N, Pakhomov S, Krauthammer M. Towards semantic role labeling & IE in the medical literature. American Medical Informatics Association Annual Meeting 2005. p. 410.
44. Zaversnik M, Batagelj V. Islands. Sunbelt XXIV, Portoroz, Slovenia.
45. De Nooy W, Mrvar A, Batagelj V. Exploratory social network analysis with Pajek. 34 ed. Cambridge Univ Press; 2011.
46. Ball R, Botsis T. Can Network Analysis Improve Pattern Recognition Among Adverse Events Following Immunization Reported to VAERS&quest. Clinical Pharmacology & Therapeutics 2011; 90(2): 271–278.
47. NLM. UMLS® Reference Manual. September 2009. Available from http://www.ncbi.nlm.nih.gov/books/NBK9676/.
48. Brown EG, Wood L, Wood S. The medical dictionary for regulatory activities (MedDRA). Drug Safety 1999 ; 20(2): 109–117.
49. Manning CD, Raghavan P, Schutze H. Introduction to information retrieval. 1 ed. Cambridge University Press Cambridge; 2008.
50. Lin D. An information-theoretic definition of similarity. ICML 1998. p. 296–304.

51. Aslam JA, Frost M. An information-theoretic measure for document similarity. SIGIR 2003. p. 449–450.
52. Kohl KS, Bonhoeffer J, Braun MM, Chen RT, Duclos P, Heijbel H, Heininger U, Loupi E. The Brighton Collaboration: Creating a global standard for case definitions (and guidelines) for adverse events following immunization. Advances in Patient Safety 2005; 2: 87–102.
53. van Haagen HH, 't Hoen P, Bovo AB, de Morree A, van Mulligen EM, Chichester C, Kors JA, den Dunnen JT, van Ommen GJB, van der Maarel SM. Novel protein-protein interactions inferred from literature context. PLoS One 2009; 4(11): e7894.
54. Cohen T, Widdows D. Empirical distributional semantics: methods and biomedical applications. Journal of Biomedical Informatics 2009; 42(2): 390–405.
55. Huang K, Geller J, Halper M, Cimino JJ. Piecewise synonyms for enhanced UMLS source terminology integration. American Medical Informatics Association Annual Meeting 2007. p. 339.
56. Huang KC, Geller J, Halper M, Perl Y, Xu J. Using WordNet synonym substitution to enhance UMLS source integration. Artificial Intelligence in Medicine 2009; 46(2): 97–109.
57. Ozgur A, Xiang Z, Radev DR, He Y. Literature-based discovery of IFN and vaccine-Mediated Gene Interaction Networks. Journal of Biomedicine and Biotechnology 2010; 2010: 426479.
58. Schuemie MJ, Kors JA, Mons B. Word sense disambiguation in the biomedical domain: an overview. Journal of Computational Biology 2005; 12(5): 554–565.
59. Xu H, Markatou M, Dimova R, Liu H, Friedman C. Machine learning and word sense disambiguation in the biomedical domain: design and evaluation issues. BMC Bioinformatics 2006; 7(1): 334.
60. Cheng XQ, Ren FX, Zhou S, Hu MB. Triangular clustering in document networks. New Journal of Physics 2009; 11(3): 033019.
61. Swanson DR. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspectives in Biology and Medicine 1986; 30(1): 7.