HHS Public Access

Author manuscript

J Biomed Inform. Author manuscript; available in PMC 2018 October 01.

Published in final edited form as:

J Biomed Inform. 2017 October ; 74: 33–45. doi:10.1016/j.jbi.2017.08.007.

Using Pathfinder Networks to Discover Alignment between Expert and Consumer Conceptual Knowledge from Online Vaccine Content

Muhammad Amith, MS[a], **Rachel Cunningham, MPH**[b], **Lara S. Savas, PhD**[a], **Julie Boom, MD**[b], **Roger Schvaneveldt, PhD**[c,d], **Cui Tao, PhD**[a], and **Trevor Cohen, MBChB, PhD**[a,*]

[a]University of Texas Health Science Center; Houston, TX

[b]Texas Children's Hospital, Houston, TX

[c]Arizona State University, Tempe Arizona

[d]New Mexico State University, Las Cruces, NM

Abstract

This study demonstrates the use of distributed vector representations and Pathfinder Network Scaling (PFNETS) to represent online vaccine content created by health experts and by laypeople. By analyzing a target audience's conceptualization of a topic, domain experts can develop targeted interventions to improve the basic health knowledge of consumers. The underlying assumption is that the content created by different groups reflects the mental organization of their knowledge. Applying automated text analysis to this content may elucidate differences between the knowledge structures of laypeople (heath consumers) and professionals (health experts). This paper utilizes vaccine information generated by laypeople and health experts to investigate the utility of this approach. We used an established technique from cognitive psychology, Pathfinder Network Scaling to infer the structure of the associational networks between concepts learned from online content using methods of distributional semantics. In doing so, we extend the original application of PFNETS to infer knowledge structures from individual participants, to infer the prevailing knowledge structures within communities of content authors. The resulting graphs reveal opportunities for public health and vaccination education experts to improve communication and intervention efforts directed towards health consumers. Our efforts demonstrate the feasibility of using an automated procedure to examine the manifestation of conceptual models within large bodies of free text, revealing evidence of conflicting understanding of vaccine concepts among health consumers as compared with health experts. Additionally, this study provides insight into

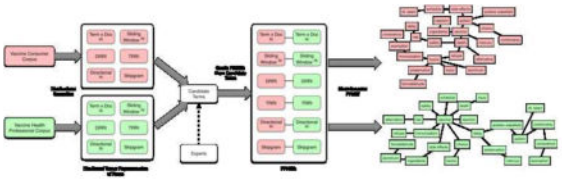
*Corresponding author trevor.cohen@uth.tmc.edu (Trevor Cohen, MBChB, PhD).

2010 MSC: 00-01, 99-00

URL: http://uth.tmc.edu (University of Texas Health Science Center), https://www.texaschildrens.org (Texas Children's Hospital), http://www.asu.edu (Arizona State University), https://www.nmsu.edu (New Mexico State University)

the differences between consumer and expert abstraction of domain knowledge, revealing vaccine-related knowledge gaps that suggest opportunities to improve provider-patient communication.

Graphical Abstract



Keywords

Distributional semantics; vaccination; consumer health; knowledge acquisition; mental models; natural language processing; social media; public health informatics; big data; semantic spaces; consumer informatics

1. Introduction

Misinformation about vaccination in the popular media has made it even more important for the public health community to provide accurate information in order to reduce parents' hesitancy to vaccinate their children. Generally, it is a challenge to develop effective health education messages for patients [1, 2, 3]. An aspect of health literacy, which is defined as the health consumer's ability to acquire and process health information [4], involves critical thinking about this information [5, 6]. When this information is insufficient, health consumers may seek other sources that may contain misleading information, or information that may be misunderstood.

Designing content with an understanding of the layperson's conceptual understanding has been shown to improve communication and education efforts [7, 8, 9, 10, 11, 12], and it has been argued that basing communication efforts on health consumers' mental models is a more effective way to engage them [13, 14]. For example, increased knowledge of vaccines may improve vaccine perception and uptake [15, 16, 17]. Due to socio-cultural differences, educational or training pedigree, expert and consumer representations of knowledge may differ [18]. This has implications in the health domain, and therefore, needs to be examined if health care-related artifacts are to be targeted for particular audiences [19]. Educational materials authored primarily by medical and health professionals should be developed with the consumer's level of understanding of the intervention domain in mind. In the field of cognitive psychology, such insight is gleaned from detailed studies of small numbers of subjects that use methods such as verbal protocol analysis [14], amongst other methods of knowledge elicitation [20]. However, the labor-intensive nature of these methods restricts their application to those small groups of subjects.

The purpose of this study is to understand the differences in knowledge organization that underlie the content produced by health professional and lay authors. When presented with a set of domain concepts, health professionals and health consumers will mentally organize

the concepts differently, in accordance with their knowledge of the domain [19, 21, 22]. We propose that by inferring this knowledge organization we will gain insight into communication gaps between health professionals and health consumers. This insight could provide guidance for educational interventions to address gaps in patient knowledge, and reveal opportunities to rectify inaccurate information at its origin.

For this purpose, it would be helpful to be able to automatically infer the structure of domain knowledge from large amounts of unstructured information authored by health consumers. Pathfinder Network Scaling (PFNETS) was originally developed in the domain of psychology. Given a set of entities from a domain of interest, PFNETS preserves the significant links between these entities, based on the pairwise distance between them, revealing an organizational structure. This approach has been used widely to characterize differences in knowledge structure (i.e. mental organization of knowledge) using estimates of semantic distance provided by humans [23], particularly in assessing differences in expertise between groups of individuals (e.g. students and teachers [24, 25]). The underlying theoretical framework is based upon the notion of connectionist representations, in which the main burden of representations rests on the relative strength of associations between concepts in a network. For a summary of work conducted up to 1990, we direct the interested reader to [25]. In this study, we used PFNETS as a tool to identify knowledge structures that reveal cognitive differences between health experts and health consumers, specifically their views on vaccination.

The novelty of this study lies in its use of PFNETS models produced from a large body of vaccine-related documents to represent authors' (health experts and health consumers) collective understanding of vaccine-related knowledge. Methods of distributional semantics [26, 27] provide the means to automate the derivation of Pathfinder networks by learning estimates of the relatedness between terms from their distribution across corpora of electronic text [23, 28, 29, 30, 31]. We posit that the knowledge structures inferred from the text authored by health experts and health consumers will reveal differences in their organization of vaccine-related domain knowledge.

Related Studies

Prior studies have applied methods of distributional semantics to analyze health-related content produced by consumers, with a focus on categorizing this content automatically by applying methods of supervised machine learning to the resulting distributional representations. Chen and his colleagues classified breast cancer portal pages into three different classes using Hyperspace Analogue to Language (HAL) [32, 33] distributional models with > 90% accuracy [34]. Myneni and her colleagues analyzed online social media content using distributional semantics and categories that emerged from qualitative coding to classify messages exchanged between users of an online social network for smoking cessation [35]. A recent study examined the use of convolutional neural network, with term vectors from a distributional model as their input layer, to classify topics from an online breast cancer community over time. The results of their study showed a relationship between the topic of online and the length of community participation [36]. The application of methods of distributional semantics in these studies was motivated by the idea that

information about the contextual use of words may be of value for machine learning classifiers.

Of particular relevance to our current efforts, McKeown and Sheehy utilized distributional semantics outside of classification tasks. They sought to discern the mental models of experts and laypeople by extrapolating knowledge structures for HIV transmission from an expert corpus (publications from government health institutions) and layperson corpora (curated documents from Reuters, and various in-house general sources and the British National Corpus [37]). The curated expert corpus was assembled from various government publications on the topic of HIV and AIDS. Latent Semantic Analysis (LSA) [38, 39] was applied to identify relationships between terms from these corpora. The closest relationships to a set of terms of interest were then compared to concepts in an preexisting, manually-curated expert model of HIV, by visualizing related terms as graphs constructed by retrieving the nearest neighboring terms to a cue term of interest, and the nearest neighbors of each of these terms [40]. The authors' hypothesis was that the model drawn from a domain-specific corpus would be better aligned with the preexisting expert model, and as anticipated the LSA-based graph, derived from the expert corpus by searching for nearest neighbors of the term "transmission", had some correspondence with the pre-existing expert model for HIV transmission. However, the graph derived from the British National Corpus consisted of terms that were unrelated to HIV, such as a cluster of words to do with motor mechanics. Overall, though the study did show inconsistencies between distributional models derived from different corpora, these differences may have been due to the different granularity of these corpora (topic-specific vs. general domain), rather than differences in their authors' understanding of HIV. For our study, and as we have described in the next section, we filtered our corpora based on vaccine-related documents in order to produce detailed knowledge structures that represent the vaccine domain exclusively.

Our study is similar in some respects to the aforementioned study, but diverges, in attempting to derive conceptual knowledge from topic-specific content authored by health professionals and health consumers, utilizing different distributional semantic methods, and the use of PFNETS to infer the structure of the relationships between a pre-selected set of terms. As a well-established method for this purpose, PFNETS presents a desirable alternative solution to the nearest-neighbor approach used by McKeown and Sheely to infer knowledge structures from online content, with the potential to aid timely assessment and intervention [41, 15].

2. Materials

Professional ("Health Expert") Corpus

Using the recommended websites from [42], online resources were identified to generate a corpus of expert-validated vaccine content. While authored by domain experts, this web content is intended for a health consumer audience. An ad-hoc, Java-based application [43] was developed using the JSoup [44] and HTMLUnit libraries [45] to parse and extract content from these trusted web resources. The content was then exported in plain text format.

In addition to this content, PDF documents were available. These PDF documents were downloaded and exported as plain text files. This endeavor resulted in a corpus containing 1,213 documents, with a total unique word count of 25,082.

Westbury USENET Corpus ("Consumer Corpus")

The Westbury USENET Corpus [46] was chosen as the source for the vaccine-related consumer corpus for our study. This corpus consists of mailing list messages exchanged between users from various USENET groups that have been curated by researchers from the University of Alberta. In order to extract vaccine-relevant content, we developed standalone software that uses the Apache Spark library [47] to extract (from a 37 GB plain text file) and filter messages that contain the strings *vaccine*, *vaccination*, *immunogen* and *vaccinate*, using regular expression matching [3]. This software generated a Lucene index (~3 GB) for the vaccine-related messages, which was used in subsequent steps involving the open source Semantic Vectors software package [48, 49, 50]. The health consumer corpus of vaccine-related posts contained 508,685 terms from 136,798 messages/documents out of ~33 million messages.

Subject Concepts

Collaborating domain experts from the Immunization Project at Texas Children's Hospital furnished a set of terms that seemed likely to reveal conceptual discrepancies between the corpora. These terms were generated through discussion and review of the language utilized by well-known anti-vaccine websites, with the aim to include terms expressing anti-vaccine sentiment and vaccine hesitancy. It includes the names of controversial figures such as Dr. Robert Sears (a known pediatrician propagating alternative vaccine schedules) and Andrew Wakefield (a researcher who publicized false claims in the late 1990s of the MMR vaccine's connection with autism) as well as terms such as *death* and *autism*. The terms for vaccine concepts are: *vaccine*, *toxins*, *injury*, *ingredients*, *autism*, *safety*, *exemption*, *conscience*, *delay*, *refusal*, *dr. sears*, *andrew wakefield*, *schedule*, *alternative*, *reaction*, *side effects*, *death*, *mercury*, *risk*, *choice*, *immunization*, *controversy*, *preservative*, *aluminum*, *formaldehyde*.

3. Method

Pathfinder Network Scaling

Developed by Roger Schvaneveldt and his colleagues [51], PFNETS uses a shortest path approach to discover significant links by filtering out edges with a distance exceeding that of alternate paths. For example, if $distance(A, C)$ were more than $distance(A, B) + distance(B, C)$ then the edge ($A \leftrightarrow C$) would be eliminated. The distances assigned to edges between concepts are calculated using a distance metric determined by the parameter "r," derived from Minkowski distance calculation. The algorithm is further constrained by the parameter "q", which determines the maximum length (in edges) of alternative pathways to be considered. When sparse Pathfinder Networks (PFNETs) are desired, r and q parameters are generally assigned the values ∞ and $n-1$ (where n is the number of nodes), and these are the values we utilized to generate a readily interpretable network graph for this study. In this

[3]This was accomplished on an 8-core 64 GB RAM Mac Pro machine and completed in ~30 minutes

case, these parameters yield minimal spanning trees (MSTs) because the cosine data contain unique MSTs. In general, PFNET(∞, $n - 1$) includes tied links (if any) that would yield alternative MSTs. For detailed treatment of PFNETs see [23].

Distributed Vector Representations

The *distributional hypothesis*, attributed to Zellig Harris, states that the meaning of a term can be inferred from the contexts in which it exists [52]. So by analyzing a term's contexts, such as its neighboring terms or the documents where it resides, one may infer its meaning in relationship to other terms, and the concepts they represent - similar concepts should occur in similar contexts. This principle can be applied to derive estimates of the semantic similarity between terms from a large body of free text, by estimating the similarity between the terms' respective contexts.

Geometric approaches to this problem derive vector representations of terms, such that terms occurring in similar contexts will have similar vectors [26, 27]. The resulting vector representations are known as *word spaces*, or *semantic spaces*. The initial representation is a term-by-context matrix, which may contain raw frequencies or statistical transformations of these frequencies. An example is Latent Semantic Analysis (LSA), which treats each document in a corpus as an independent context (with its own column vector). The dimensionality of the matrix is reduced using Singular Value Decomposition (SVD), a linear algebra technique that yields improvements in performance on synonym test evaluations [39, 38]. We utilized the Semantic Vectors package [48, 49, 50] that employs Random Indexing [53] as a scalable alternative to SVD [54].

Beside the document level approach, another spatial modeling approach is to estimate relatedness using neighboring terms. This is often referred to as a *sliding window* approach, and an example is the Hyperspace Analogue to Language (HAL)[32], which distinguishes between the context before and after a term. The size of the sliding window is a parameter of the model. Random Indexing (RI) can also be applied to generate a reduced-dimensional approximation of sliding window models [54], and has been adapted to encode the positions of words within a sliding window using permutations of sparse random vectors to indicate position (fore example, before or after) in relation to the targeted word [55]. Information about each context a term occurs in, whether this concerns document-level occurrence statistics or neighboring words within a sliding window, is encoded in a context vector that is superposed to generate semantic vector representation that encodes the collective contexts a particular term has occurred in. When position before or after a term is encoded using sliding window RI with permutations, the net result is a vector space as reduced-dimensional approximation of the term-by-term-by-position matrix used in the original HAL implementation. Reflective Random Indexing (RRI) is a recent variant of term-document RI developed to encode relationships between terms that do not co-occur directly [56], an aspect of distributional modeling that was not addressed by the original implementation of RI. RRI extends RI by initializing random vectors from either terms in a corpus (Term-based Reflective Random Indexing or TRRI) or documents (Document-based Reflective Indexing or DRRI), and iteratively generating term and document vectors from a corpus in

succession, with improvements in the ability to detect implicit relationships between terms [56].

The continuous bag of words (CBOW) and skip-gram architectures, developed by Mikolov and colleagues [57], are currently popular, sliding window-based approaches for the generation of term vector representations. These representations are generated by utilizing neural networks that learn predict terms in proximity to an observed term across a training corpus. Specifically, CBOW predicts a target term based on surrounding context, and the skip-gram architecture predicts the surrounding context based on the target term. Of the two, the skip-gram architecture has been recommended for use with smaller datasets [58, 59], such as our professional corpus (Section 2).

Term-by-document approaches, such as LSA models, tend to emphasize relationships between terms that occur sequentially in text (*syntagmatic relations*, as proposed by Swiss linguist Ferdinand de Saussure [60]), while sliding window approaches tend to emphasize relationships between terms that could be substituted for one another, such as sugar and glucose (*paradigmatic relations*, in de Saussure's terms [60]), though with wider sliding windows this distinction is blurred [61]. The data provided by these models can be used to produce Pathfinder Networks (PFNETs) that can visually represent relationships between terms or concepts (denoted by edges in the graph) suggested by their distribution across a large corpus of text [62]. When a particular community, such as either lay or professional authors, have produced a corpus, it is possible that PFNETs based on a distributional model of a corpus will reveal this community's understanding of the concepts it contains [63].

We have selected RI for our study for several reasons. RI scales comfortably to large corpora, and does not require retraining when new data are introduced. In contrast, with standard LSA algorithms, incorporating new data requires regenerating and decomposing a text-by-document matrix representing the entire corpus. As is the case with the CBOW and skip-gram architectures, RI is an "online" algorithm that incrementally process segments of text without the need to generate an explicit co-occurrence matrix. This provides performance benefits by permitting parallelized implementations, and also better approximates the "online" reading behavior of humans processing text, which is desirable from a cognitive perspective. In addition, RI is a relatively simple algorithm, which is appealing from a theoretical perspective on account of the small number of assumptions that must be introduced to include it as a component of a cognitive model of text comprehension. In addition, and as is the case with all of the algorithms we include in the paper, implementations are available within a common platform, Semantic Vectors [48, 49, 50]. Consequently, we can be confident any observed differences are due to differences in the underlying algorithm, rather than differences in package-specific pre-processing of text. Finally, RI permits both sliding window and term-by-document approaches permitting evaluation of the choice of context for modeling on the coherence of the resulting Pathfinder networks. In addition, we have included neural embeddings generated using the Skipgram-with-Negative-Sampling algorithm, on account of the strong correlation between pairwise similarities between the resulting term vectors and human estimates of semantic relatedness in recent experiments [57, 64, 65].

Figure 1 describes the preprocessing and generation of PFNETs. Two corpora were used, and for each we built a set of distinct word-space models. Five of these applied variants of RI: term-by-document RI, TRRI, DRRI, sliding window based RI, and a directional (HAL-like) model that encodes whether context terms occur before or after a target term in a sliding window. We also generated word representations using the skip-gram architecture [4]. For the last three (sliding-window RI, the HAL-like model and the skipgram embeddings), we selected a 10-word sliding window radius (i.e. ten terms to the left and right are considered), as this has been shown to provide a balance between synonymy and more general relationships in prior experiments [61]. For TRRI and DRRI, we employed two iterative training cycles. Lastly, all RI models employed the default seed length of 10 and term vector dimensionality of 200. Table 1 summarizes the model parameters.

The cosine metric between vector representations of words in each of the models was taken as an estimate of the relatedness between the expert-identified terms (Subject Concepts from Section 2), based on the corpus on which the model was trained (with compound terms, such as “Andrew Wakefield” the term vector was constructed as the vector average of the term vectors for the individual terms). The end result was a set of proximity ratings for each term pair in the set, which was the data used to create the PFNETs. Overall, we generated 12 sets of PFNETs (6 for the health consumer corpus and 6 for the health professional corpus) based on concepts provided by subject matter experts.

Figure 2 outlines the steps for analyzing the PFNETs. First a selection criterion was used to select the distributional model that produced the most consistent estimates, and then a comparison between the two PFNETs produced by this model was conducted. For the first step, we used the Pathfinder software package’s Coherence measure. The Coherence measures the transitive consistency between the pairwise comparisons within each proximity set. Previous PFNET studies have utilized Coherence to validate proximity data from participants [66, 67]. In this study, it was used as a selection criterion to select networks for more detailed analysis.

For the networks produced by the distributional model with greatest mean coherence across the health consumer and health professional sets, discrepancies that suggest differences between health consumer and health expert conceptualization of vaccine-related knowledge were identified. The similarity between the networks was estimated using the network similarity measure implemented by JPathfinder [68] (a publicly available software application for PFNET analysis). Network similarity varies from 0 to 1, indicating the proportion of the unique links in two network that are found in both networks. The similarity metric is shown in Equation 1.

$$S=\frac{C}{(L_1+L_2-C)} \quad (1)$$

[4]We used the Semantic Vectors implementation of the skipgram architecture, with default parameters from the Semantic Vectors package (iterations=1, negative samples=5, learning rate=0.025, threads=4, word vector dimensions=200)

where,

S = network similarity

C = number of links common to the networks

L_1 = number of links in first network

L_2 = number of links in second network

All aspects of this research were conducted using publicly available open source software. The building of Lucene indices, vector wordspace models, and proximity data files was performed using custom software developed in Java [69], leveraging the Semantic Vectors library's API. The latest version of Semantic Vectors at the time of this writing (pre-release version 5.9) provided support for skip-gram modeling [50]. We used JPathfinder to visualize and export the PFNETs. Table 2 outlines the specifications for the PFNETs.

4. Results and Discussion

Table 3 shows the Coherence metric for each of the proximity data from their respective models. The TRRI-based models exhibited higher mean Coherence than other model types at $\mu = 0.8585$. DRRI-based models for the health consumer and health professional corpus followed with $\mu = 0.8035$, along with the term-by-document RI models at $\mu = 0.7785$.

The term-by-document models' (Term x Doc RI, DRRI and TRRI) Coherence ratings were high compared to the sliding window models. However, among the sliding window models, the skip-gram models performed best overall. It is uncertain whether the Coherence might be improved with different parameter configurations of the current models, other distributional models, or by varying the size of the training corpora. This is a limitation of this work that we plan to address in future studies.

Using the proximity data from the TRRI model for the health consumer set and the health professional set, we compared the number of links that are shared between the two resulting Pathfinder Networks. Table 4 shows estimates of the similarity between these networks, and the data from which one of these estimates was derived. Each network contained 24 links with only three links that were shared. The Similarity was 0.067, where 1 denotes an identical network, and 0 denotes no shared links. The probability of 3 or more shared links by chance is 0.299, indicating that the similarity is not statistically significant. The three shared links were *vaccine* ↔ *death*, *autism* ↔ *andrew wakefield*, and *exemption* ↔ *conscience*. Aside from the three shared links, it appears that the knowledge of health consumers and health professionals is incongruent, as indicated by the low similarity between the resulting networks.

Furthermore, we analyzed the similarities between the networks derived from the various models (Table 5), and calculated the mean and median similarity of the similarity between the networks for health consumers (e.g. Consumer TRRI network to Other Consumer networks), health professionals (e.g. Professional DRRI network to Other Professional networks), and their combinations (e.g. Consumer TRRI to all Professional networks). The results of these calculations are provided in Table 6. Both the means and the medians of the

within-group (e.g. consumer-to-consumer) comparisons were greater than those of the consumer-to-professional comparisons. These differences were statistically significant (Mann-Whitney $U = 73.5$, Wilcoxon $W = 739.5$, $n_1 = 30$, $n_2 = 36$, $p = 0.00$ two-tailed). This suggests that difference in authorship has a greater influence on the resulting networks than choice of distributional model.

We examined the network structures of our two candidate models, derived from health consumer (Fig 3a) and health professional (Fig 3b) content. Individual concepts and links were studied, and, if the link between a pair of concepts was difficult to interpret, we inspected secondary extended links to elaborate on the meaning of neighboring links for contextual clarification.

4.1. Prominent concepts

Graph theory, the mathematical study of networks, provides a number of metrics that can be used to identify salient nodes in a network. For the PFNETs presented here, salient nodes correspond to salient concepts in the network. The degree of a node is the number of links connected to the node. The Maximum Degree node or nodes indicate concepts that enter into many relations with the other concepts. The eccentricity of a node is the maximum number of links between that node and all other nodes in the network. The Center of the network is the node or nodes of minimum eccentricity. The Median of a network is the node with the minimum mean number of links between the node and all other nodes in the network. The Maximum Degree nodes, the Center nodes, and the Median nodes for the health consumer (Fig. 3a) and health professional (Fig. 3b) PFNETs are shown in Table 7.

In the knowledge organization of health consumers, safety-related issues related to *death* and *toxins* appear to predominate. However, this is not the case for health professionals. One interpretation of the health expert perspective is that all of the terms (See Subject Concepts) relate to vaccine hesitancy, as indicated by the prominence of *vaccine* and *delay*. From the health consumer perspective, all of the terms concern dangers and harm (or some other negative description) of vaccines, highlighted by the prominence of *death* and *toxins*.

4.2. General Perception of Vaccine and Immunization

Because the central focus of this study is vaccination, we first reviewed the links associated with the term *vaccine* in both health consumer (from the health consumer corpus) and health professional PFNETs (from the health expert corpus). In the health consumer model, *vaccine* and *autism* are associated (Fig. 3a). However, in the health expert model (Fig. 3b), *autism* is linked with *andrew wakefield*, a controversial researcher known for disseminating discredited research findings concerning alleged side effects of vaccines, including autism [70]. While the relatedness of *autism* with *andrew wakefield* is an appropriate association, the association of *autism* with *vaccine* suggests a conceptual link that is inconsistent with current scientific consensus. *Vaccine* is further associated with terms such as *death* and *mercury* in the health consumer model. These links may suggest misconceptions about vaccines - ethylmercury is an ingredient of the vaccine preservative thiomerosal, which was removed from most but not all vaccines on account of safety concerns stemming from health sequelae observed after incidents of chemical contamination of food with methylmercury in

Japan and Iraq, but without evidence of harm from the vaccine preservative itself [71] - and highlight the prominence of deleterious consequences in online consumer discussion of vaccination.

Similarly, health professionals associated *vaccine* and *death* (Fig. 3b). The term *death* is linked to *injury*; however, it is difficult to ascertain any specific meaning from the perspective of a health expert conceptualization of vaccine, as it appears as the central organizing concept and is linked to many others. The term *death* in the health consumer model (Fig. 3a) is linked to *safety*, *refusal*, and *alternative* suggesting the expression of cautionary attitudes towards vaccines.

Immunization is synonymous with *vaccine*, and we found these two terms linked in the health professional model (Fig. 3b). In the health consumer model, we noted that *immunization* was linked to both *toxins* and *exemption* (Fig. 3a). Also, *refusal* is a term linked to *immunization* in the health professional model, and, again, it was difficult to determine meaning from a health professional/expert perspective. From the health consumer model, *refusal* appears to have more context as it linked to *death* and *toxins*, both linked to *vaccine* and *immunization*, respectively.

4.3. Vaccine constituents

Vaccine constituents are a specific concern that some health consumers have regarding vaccines and a factor in one's intention to vaccinate [72, 73, 74]. The term *ingredients* (associated with *vaccine*) and *reaction* are shown to be linked in the health consumer model (Fig. 3a). As discussed in the previous paragraph, *mercury* is an ingredient often associated with vaccines and misrepresented by health consumers. Robert Kennedy, Jr., notable for his anti-vaccination views, has published articles in the mainstream Rolling Stone magazine [75] and Slate, discussing the mercury-based preservative thimerosal, which he claims leads to autism. Figure 3a (the health consumer model) shows a link between *mercury* with *vaccine* (which also connects to *death* and *autism*). This suggests that both mercury and autism feature prominently in consumer association of vaccination, and that Kennedy's views prevail in these associations. In Figure 3b (the health professional model), *mercury* and *preservative* are associated, and this link is expressed with a connection to *delay* (and linked to *andrew wakefield*). The aluminum component in vaccines is also of concern to consumers [72, 73], and in the PFNET consumer conceptualization (Fig. 3a), it is associated with *toxins* (connected to *immunization* and *injury*). These links suggest an alarmist view of aluminum ingredients. In contrast, the health professional model has *aluminum* linked with *ingredients* and *formaldehyde*. Aluminum is a component of various vaccines, such as for hepatitis A and B, pneumococcus, *Haemophilus influenzae* type b (Hib) and diphtheria-tetanus-acellular pertussis (DTaP), to improve immune response [73, 72], which explains its link with *ingredients*. One of the controversial points that Dr. Sears advances as justification for vaccine delay pertains to overexposure to aluminum being harmful to infants [76].

4.4. Vaccine scheduling

As alluded to earlier, some health consumers are concerned about the frequency of vaccination and its overall impact on health. This can result in patients wishing to delay or

reduce the frequency of vaccination. We examined the term *schedule* on how it relates with other concepts. In the health professional model (Fig 3b), the term *schedule* is linked to *vaccine*, yet with the health consumer representation *schedule* is linked with both *side effects* and *dr. sears*, who, as described in the previous section, promotes a controversial alternative vaccine schedule, alleged to mitigate vaccine side effects [76]. Dr. Sears (Robert Sears) is a well-known physician who has appeared on mainstream television promoting this vaccination schedule [77], which provides a seemingly credible alternative to the CDC recommended immunization schedules for individuals who are fearful of vaccine side effects [78].

Immunization and *exemption* are two linked terms from the health consumer model (Fig. 3a). Many U.S. states allow individual exemptions from vaccines based on philosophical or personal objection. For more than a decade, vaccine exemptions have been a controversial and a topic debated by health consumers/patients and the health expert community. In contrast, in the health professional model, *exemption* is associated with *conscience*, which is extended with the term *controversy*. The two contrasting associations highlight dissimilar perspectives on vaccine-related exemption, which the health professional community views as controversial.

4.5. Vaccine reactions

A common factor among health consumers who hesitate or resist vaccination is the concern that there may be severe side effects. We saw that *death* and *vaccine* were associated in the health consumer model (Fig 3a), and other linked concepts were *autism* with *vaccine* and *side effects* with *autism*. The health professional model (Fig 3b), on the other hand, show *autism* linked with *controversy*, *andrew wakefield*, and *dr. sears*. Dr. Sears has authored a book alleging a relationship between vaccine and autism [79], and does not acknowledge studies that have failed to show a link between the Measles, Mumps, and Rubella (MMR) vaccine and autism [78].

4.6. Clustering by Nearest Neighbor (JPathfinder)

We further explored the proximity data by generating directed nearest neighbor network graphs with JPathfinder. In such graphs, arrows point from a node to the most semantically similar node in the set of nodes. Figure 4 shows nearest neighbor graphs for the health consumer model and the health professional model. Many of the links are shared with the respective Pathfinder networks presented earlier, but the nearest neighbor graphs reveal clusters of concepts that provide further information concerning the relatedness between terms in each model.

For example, in Figure 4a (the health consumer model), *dr. sears* and *schedule* compose their own cluster. *Vaccine* and *autism* are components of a coherent, inter-related cluster that also includes *mercury* and *andrew wakefield*. In contrast, in Figure 3b (health professional model) vaccine occurs in a separate cluster from these entities, several of which are included a cluster of terms related to controversial vaccine topics (*autism*, *dr. sears*, *andrew wakefield*, *controversy*). This is consistent with the marginal nature of these topics, in relation to current scientific consensus.

4.7. Summary

Our findings reveal aspects of health consumers' abstraction of vaccine knowledge that diverge from scientific consensus, especially in relation to the safety and efficacy of vaccines. While drawing definitive conclusions about the reasons for these differences is beyond the scope of the current paper, one hypothesis is that increased propagation of vaccine misinformation in various media avenues [80, 81] may influence the discourse on these topics that constitutes the corpus for our layperson models.

One limitation of this work was the challenge to determine precise meanings of some concept associations, for example, the terms *death* and *choice* within the health professional model. This suggests a need to broaden the conceptual territory evaluated. Expanding the range of concepts could shed additional light on health consumers' understanding of the domain, a possibility we plan to explore in future studies. In addition, we have not optimized the parameters used for our distributional models, and it may be the case that greater coherence can be obtained by exploring the parameter space further. Nonetheless, this study has revealed differences in knowledge conceptualization between health experts and health consumers, as reflected in the online content that they have authored. These differences could inform interventions, such as counseling or the design of materials, for health consumers. One example would be deliberately addressing any evidence for and against the merits of alternative vaccine schedules. Another might be addressing some of the strong associations of *vaccine* with *mercury* and *autism*, and the toxicity claim concerning aluminum adjuvants.

The health consumer corpus was extracted from publicly available mailing list documents provided by the University of Calgary, and only the vaccine-related messages were used. The Westbury USENET corpus contained messages between October 2005 and January 2011, so, there may be limitations with respect to concepts that have emerged since then. One future direction could be to gather more from the corpus, incorporate social media postings to expand the health consumer corpus, or focus on a specific source such as Facebook, Twitter, 4chan, or reddit or questionable websites with vaccine content, such as InfoWars or the HealthRanger. Here, we obtained interpretable results using relatively small corpora, which suggests that large datasets are not necessary for this approach.

Word clouds are a popular way to analyze online media [82], especially social media. Word clouds utilize the frequency with which terms occur to highlight frequently discussed concepts. However, it can be a challenge to interpret the meaning of these high-frequency terms, as the relationships between them are not shown. Topic models have been widely used to identify meaningful cluster of terms in text corpora [83]. Analysis of the degree to which a topic is represented may reveal the absence of knowledge of a topic within a community, which may present a target for intervention. In contrast, PFNETs reveal differences in the organization of knowledge of the same topic. Our experiences in the domain of vaccine adherence suggests the latter is a more pressing concern on account of the wide distribution of information contrary to scientific consensus through social media and other forums. However, an analysis of term frequency and the prevalence of topics within a corpus may provide information that is complementary to our approach. One potential application may be for the identification of terms to populate PFNETs. A limitation of our

current approach is its reliance upon the identification of a set of terms based on a suspicion of where discrepancies in understanding may be situated, limiting its scalability and applicability. Another potential solution to this problem might involve generation of distributional models of both corpora with the same initial seed vectors, and comparing them to identify terms that diverge in meaning from corpus-to-corpus.

This approach also opens the possibility of developing an analytical and visualization tool that allows for on-demand experimentation with health consumer-generated online media. This tool would be of value as a means to monitor web-based content [41, 15]. Such dynamic application of our methods would provide the means to monitor and understand misalignments of conceptual knowledge as they emerge, for vaccines and other public health concerns. This understanding could inform tailoring of educational interventions concerning vaccination and other consumer health topics to enhance the health-related decision-making of the general public.

The methods described may also be extended to the other problem domains. Other situations where we would anticipate finding discrepancies between consumer and expert knowledge organization include nutrition (for example, fad diets and supplements) and the risks of emerging methods for the consumption of nicotine (such as e-cigarettes and “vaping”). Beyond the healthcare domain, one might envision this method providing insight into the ways that groups with different political affiliations differ in their understanding of the world.

5. Conclusion

This paper describes an automated approach to analysis of large corpora of free text and visualization of the results as PFNET, providing insight into the knowledge organization of the authors. Open-source tools were used to generate proximity data for terms found in our corpora. The methods produced estimates of relatedness between the concepts and similarity measures between PFNET. Dissimilarity was found between PFNET derived from health consumers and health expert texts, and qualitative evaluation of specific concept associations indicated differences in perspectives on various vaccine-related issues, particularly those relating to controversial concepts including autism, vaccine ingredients, Andrew Wakefield, and efficacy of vaccine, with implications for the design of interventions to encourage vaccination.

Acknowledgments

This research is supported by the National Library Of Medicine of the National Institutes of Health under Award Number R01LM011829 and R01LM011563. The authors also gratefully acknowledge the support from the UTHealth Innovation for Cancer Prevention Research Training Program (Cancer Prevention and Research Institute of Texas grant # RP140103). The authors thank Carol Kakalec Kohn, MS, ELS(D) for editorial assistance.

References

1. Emmett CL, Montgomery AA, Peters TJ, Fahey T. Three-year follow-up of a factorial randomised controlled trial of two decision aids for newly diagnosed hypertensive patients. British journal of general practice. 2005; 55(516):551–553. [PubMed: 16004744]

2. World Health Organization and others. Behaviour change strategies and health: the role of health systems. 2008

3. Sheridan SL, Viera AJ, Krantz MJ, Ice CL, Steinman LE, Peters KE, Kopin LA, Lungelow D. The effect of giving global coronary risk information to adults: a systematic review. Archives of internal medicine. 2010; 170(3):230–239. [PubMed: 20142567]

4. Selden, CR., Zorn, M., Ratzan, SC., Parker, RM. Health literacy. Bethesda (MD): National Library of Medicine;

5. Nutbeam D. Health literacy as a public health goal: a challenge for contemporary health education and communication strategies into the 21st century. Health promotion international. 2000; 15(3): 259–267.

6. Nutbeam D. Defining and measuring health literacy: what can we learn from literacy studies? International Journal of Public Health. 2009; 54(5):303–305. [PubMed: 19641847]

7. MacGregor DG, Slovic P, Morgan MG. Perception of risks from electromagnetic fields: a psychometric evaluation of a risk-communication approach. Risk analysis. 1994; 14(5):815–828. [PubMed: 7800866]

8. Morgan MG, Florig HK, Nair I, Cortés C, Marsh K, Pavlosky K. Lay understanding of low-frequency electric and magnetic fields. Bioelectromagnetics. 1990; 11(4):313–335. [PubMed: 2285416]

9. Read D, Morgan MG. The efficacy of different methods for informing the public about the range dependency of magnetic fields from high voltage power lines. Risk analysis. 1998; 18(5):603–610. [PubMed: 9853395]

10. Downs JS, Murray PJ, de Bruin WB, Penrose J, Palmgren C, Fischhoff B. Interactive video behavioral intervention to reduce adolescent females' std risk: A randomized controlled trial. Social science & medicine. 2004; 59(8):1561–1572. [PubMed: 15279915]

11. Downs JS, de Bruin WB, Fischhoff B. Parents' vaccination comprehension and decisions. Vaccine. 2008; 26(12):1595–1607. [PubMed: 18295940]

12. Morgan, MG. Risk communication: A mental models approach. Cambridge University Press; 2002.

13. Marteau TM, Weinman J. Self-regulation and the behavioural response to dna risk information: a theoretical analysis and framework for future research. Social science & medicine. 2006; 62(6): 1360–1368. [PubMed: 16162383]

14. Sivaramakrishnan M, Patel VL. Reasoning about childhood nutritional deficiencies by mothers in rural india: A cognitive analysis. Social science & medicine. 1993; 37(7):937–952. [PubMed: 8211312]

15. Habel MA, Liddon N, Stryker JE. The HPV vaccine: a content analysis of online news stories. Journal of women's health. 2009; 18(3):401–407.

16. Myers, MG., Pineda, D. Misinformation about vaccines. In: Barrett, AD., Stanberry, LR., editors. Vaccines for biodefense and emerging and neglected diseases. Academic Press; 2009. p. 255-270.

17. Scherer LD, Shaffer VA, Patel N, Zikmund-Fisher BJ. Can the vaccine adverse event reporting system be used to increase vaccine acceptance and trust? Vaccine. 2016; 34(21):2424–2429. [PubMed: 27049120]

18. Schvaneveldt RW, Durso FT, Goldsmith TE, Breen TJ, Cooke NM, Tucker RG, De Maio JC. Measuring the structure of expertise. International journal of man-machine studies. 1985; 23(6): 699–728.

19. Zhang J. Representations of health concepts: a cognitive perspective. Journal of Biomedical Informatics. 2002; 35(1):17–24. [PubMed: 12415723]

20. Cooke NJ. Knowledge elicitation. Handbook of applied cognition. 1999:479–510.

21. Patel V, Arocha J, Kushniruk A. Emr: re-engineering the organization of health information. J Biomed Inform. 2002; 35:8–16. [PubMed: 12415722]

22. Patel VL, Arocha JF, Kaufman DR. A primer on aspects of cognition for medical informatics. Journal of the American Medical Informatics Association. 2001; 8(4):324–343. [PubMed: 11418539]

23. Schvaneveldt, RW. Pathfinder associative networks: Studies in knowledge organization. Ablex Publishing; 1990.

24. Housner LD, Gomez R, Griffey DC. A pathfinder analysis of pedagogical knowledge structures: A follow-up investigation. Research quarterly for Exercise and Sport. 1993; 64(3):291–299. [PubMed: 8235050]
25. Goldsmith, TE., Johnson, PJ. Pathfinder associative networks: Studies in knowledge organization. Ablex Publishing; 1990. Ch. A structural assessment of classroom learning
26. Cohen T, Widdows D. Empirical distributional semantics: methods and biomedical applications. Journal of biomedical informatics. 2009; 42(2):390–405. [PubMed: 19232399]
27. Turney PD, Pantel P, et al. From frequency to meaning: Vector space models of semantics. Journal of artificial intelligence research. 2010; 37(1):141–188.
28. Cohen T. Exploring medline space with random indexing and pathfinder networks. AMIA. 2008
29. Schvaneveldt R, Cohen T, Whitfield GK. Paths to discovery, Expertise and Skills Acquisition: The Impact of William G Chase. 2013:147–177.
30. Schvaneveldt, RW., Cohen, TA. Computer-Based Diagnostics and Systematic Analysis of Knowledge. Springer; 2010. Abductive reasoning and similarity: Some computational tools; p. 189-211.
31. Cohen T, Whitfield GK, Schvaneveldt RW, Mukund K, Rindflesch T. Epiphanet: an interactive tool to support biomedical discoveries. Journal of biomedical discovery and collaboration. 2010; 5:21–49. [PubMed: 20859853]
32. Lund CB, Kevin. Modelling parsing constraints with high-dimensional context space. Language and cognitive processes. 1997; 12(2–3):177–210.
33. Burgess C, Livesay K, Lund K. Explorations in context space: Words, sentences, discourse. Discourse Processes. 1998; 25(2–3):211–257.
34. Chen G, Warren J, Riddle P. Semantic Space models for classification of consumer webpages on metadata attributes. Journal of Biomedical Informatics. 2010; 43(5):725–735. [PubMed: 20601122]
35. Myneni S, Cobb NK, Cohen T. Finding meaning in social media: content-based social network analysis of quitnet to identify new opportunities for health promotion. MedInfo. 2013:807–811.
36. Zhang, S., Grave, E., Sklar, E., Elhadad, N. Longitudinal Analysis of Discussion Topics in an Online Breast Cancer Community using Convolutional Neural Networks. arXiv preprint arXiv: 1603.08458. URL https://arxiv.org/abs/1603.08458
37. Oxford University Computing Services on behalf of the BNC Consortium. The British National Corpus, version 3 (BNC XML Edition). URL http://www.natcorp.ox.ac.uk/
38. Landauer TK, Dumais ST. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological review. 1997; 104(2):211.
39. Landauer TK, Foltz PW, Laham D. An introduction to latent semantic analysis. Discourse processes. 1998; 25(2–3):259–284.
40. McKeown, G., Sheehy, N. Visualising Textual Knowledge about Risks to Aid Risk Communication. Proceedings of the 2005 Human-Computer Interaction Conference;
41. Tozzi AE, Buonuomo PS, Ciofi degli Atti ML, Carloni E, Meloni M, Gamba F. Comparison of Quality of Internet Pages on Human Papillomavirus Immunization in Italian and in English. Journal of Adolescent Health. 2010; 46(1):83–89. [PubMed: 20123262]
42. Healy CM, Pickering LK. How to communicate with vaccine-hesitant parents. Pediatrics. 2011; 127(Supplement 1):S127–S133. arXiv:http://pediatrics.aappublications.org/content/127/Supplement_1/S127.full.pdf+html. [PubMed: 21502238]
43. Amith, M. webscraper. 2015. URL https://bitbucket.org/MFAmith/webscraper
44. Hedley, J. JSoup: Java html parser. 2010. URL http://jsoup.org/
45. Bowler, M. HtmlUnit. 2002. URL http://htmlunit.sourceforge.net/
46. Shaoul, C., Westbury, C. A usenet corpus. University of Alberta; Canada: 2005–2009.
47. Apache Spark - Lightning-fast cluster computing. 2013. URL http://spark.apache.org/
48. Widdows D, Ferraro K. Semantic vectors: a scalable open source package and online technology management application. LREC. 2008

49. Widdows, D., Cohen, T. The semantic vectors package: New algorithms and public tools for distributional semantics. Semantic Computing (ICSC), 2010 IEEE Fourth International Conference on; IEEE; 2010. p. 9-15.
50. Semantic Vectors. Github; Sep. 2016 URL https://github.com/semanticvectors/semanticvectors
51. Schvaneveldt RW, Durso FT, Dearholt DW. Network structures in proximity data. The psychology of learning and motivation. 1989; 24:249–284.
52. Harris ZS. Distributional structure. Word.
53. Kanerva, P., Kristofersson, J., Holst, A. Random indexing of text samples for latent semantic analysis. Proceedings of the 22nd annual conference of the cognitive science society; Citeseer. 2000.
54. Sahlgren, M. An introduction to random indexing. Methods and applications of semantic indexing workshop at the 7th international conference on terminology and knowledge engineering, TKE; 2005.
55. Sahlgren, M., Holst, A., Kanerva, P. Permutations as a means to encode order in word space. Proc. 30th Annual Meeting of the Cognitive Science Society (CogSci'08);
56. Cohen T, Schvaneveldt R, Widdows D. Reflective random indexing and indirect inference: A scalable method for discovery of implicit connections. Journal of biomedical informatics. 2010; 43(2):240–256. [PubMed: 19761870]
57. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.
58. Mikolov, T., Le, QV., Sutskever, I. Exploiting similarities among languages for machine translation. 2013. arXiv preprint arXiv:1309.4168Bibtex: mikolov exploitingURL http://arxiv.org/abs/1309.4168
59. Mikolov, T. de-obfuscated python + question, word2vec-toolkit - Google Group. Oct. 2013 URL https://groups.google.com/forum/#!searchin/word2vec-toolkit/c-bow/word2vec-toolkit/NLvYXU99cAM/E5ld8LcDxlAJ
60. Saussure, Fd. Course in general linguistics. Harris, R., translator. London: Duckworth;
61. Sahlgren, M. PhD Dissertation. The word-space model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces.
62. Widdows, D., Cederberg, S., Dorow, B. Text, Speech and Dialogue. Springer; 2002. Visualisation techniques for analysing meaning; p. 107-114.
63. Kandula S, Zeng-Treitler Q. Exploring relations among semantic groups: a comparison of concept co-occurrence in biomedical sources. Studies in Health Technology and Informatics. 2010; 160(Pt 2):995–999. [PubMed: 20841833]
64. Levy O, Goldberg Y, Dagan I. Improving distributional similarity with lessons learned from word embeddings. Transactions of the Association for Computational Linguistics. 2015; 3:211–225.
65. Baroni M, Dinu G, Kruszewski G. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. ACL. 2014; (1):238–247.
66. Rowe LJ, Schvaneveldt RW, Bennett W Jr. Measuring pilot knowledge in training: The pathfinder network scaling technique. Tech rep, DTIC Document. 2007
67. Dunlap JC, Grabinger S. Using pathfinder networks to examine structural knowledge.
68. Schvaneveldt, R. JPathFinder. 2015. URL http://interlinkinc.net/
69. Amith, M. semvec4vaccinev2. 2016. URL goo.gl/lor18C
70. Plotkin S, Gerber JS, Offit PA. Vaccines and autism: a tale of shifting hypotheses. Clinical Infectious Diseases. 2009; 48(4):456–461. [PubMed: 19128068]
71. Baker JP. Mercury, vaccines, and autism: one controversy, three histories. American Journal of Public Health. 2008; 98(2):244–253. URL http://ajph.aphapublications.org/doi/abs/10.2105/ajph.2007.113159. [PubMed: 18172138]
72. Offit PA, Jew RK. Addressing parents' concerns: Do vaccines contain harmful preservatives, adjuvants, additives, or residuals? Pediatrics. 2003; 112(6):1394–1397. arXiv:http://pediatrics.aappublications.org/content/112/6/1394.full.pdf. URL http://pediatrics.aappublications.org/content/112/6/1394. [PubMed: 14654615]

73. Fernandez S. Aluminum in Vaccines: Addressing Parents' Concerns. Pediatric Annals. 2016; 45(7):e231–e233. URL http://www.healio.com/doiresolver?doi=10.3928/00904481-20160606-01. DOI: 10.3928/00904481-20160606-01 [PubMed: 27403668]
74. Kocourkova A, Honegr J, Kuca K, Danova J. Vaccine ingredients: Components that influence vaccine efficacy. Mini reviews in medicinal chemistry.
75. Kennedy RF Jr. Deadly immunity: Robert f. kennedy jr. investigates the government cover-up of a mercury/autism scandal. Rolling Stone. 2005; 977:978.
76. Sears, RW. Little, Brown. 2011. The vaccine book: making the right decision for your child.
77. Poland GA, Jacobson RM. The clinicians guide to the anti-vaccinationists galaxy. Human immunology. 2012; 73(8):859–866. [PubMed: 22504410]
78. Offit PA, Moser CA. The problem with dr bob's alternative vaccine schedule. Pediatrics. 2009; 123(1):e164–e169. [PubMed: 19117838]
79. Sears, RW. Autism book. New York: Little Brown;
80. Newport, F. Tech rep. Gallup; Mar. 2015 U.S., Percentage Saying Vaccines Are Vital Dips Slightly.
81. Kata A. Anti-vaccine activists, Web 2.0, and the postmodern paradigm An overview of tactics and tropes used online by the anti-vaccination movement. Vaccine. 2012; 30(25):3778–3789. bibtex: kata anti-vaccine 2012 URL http://www.sciencedirect.com/science/article/pii/S0264410X11019086. DOI: 10.1016/j.vaccine.2011.11.112 [PubMed: 22172504]
82. Smith G. Tagging: people-powered metadata for the social web. New Riders. 2007
83. Blei DM. Probabilistic topic models. Communications of the ACM. 2012; 55(4):77–84.

Appendix A. Proximity Data

Table A.8

Links from proximity data. High semantic similarity (cosine values) indicate stronger association between concepts. Links with shaded backgrounds are from health professional models. [S] denotes the shared links.

Link		Semantic Similarity
choice	controversy	0.99754
vaccine	mercury	0.99443
vaccine	autism	0.99125
autism	choice	0.98826
delay	risk	0.98249
autism	side effects	0.97496
vaccine	risk	0.96787
ingredients	reaction	0.96623
vaccine	immunization	0.96549
vaccine	ingredients	0.96524
vaccine	schedule	0.94231
alternative	risk	0.94176
vaccine	death	0.94002 [S_1]
safety	risk	0.93127
vaccine	reaction	0.9277
vaccine	delay	0.91741
injury	death	0.91525
vaccine	choice	0.90658
vaccine	death	0.90132 [S_1]

Link		Semantic Similarity
exemption	conscience	0.88088 [S_3]
vaccine	safety	0.87345
mercury	preservative	0.87239
vaccine	side effects	0.87082
autism	andrew wakefield	0.85667 [S_2]
safety	death	0.84273
toxins	choice	0.8236
ingredients	aluminum	0.76991
ingredients	side effects	0.76686
aluminum	formaldehyde	0.7632
autism	controversy	0.76309
autism	dr. sears	0.75828
delay	preservative	0.71412
autism	andrew wakefield	0.70008 [S_2]
refusal	immunization	0.67507
delay	andrew wakefield	0.6726
dr. sears	schedule	0.66473
toxins	refusal	0.63559
toxins	preservative	0.62951
schedule	side effects	0.62948
toxins	immunization	0.61651
exemption	immunization	0.54379
conscience	controversy	0.54193
alternative	death	0.50042
refusal	death	0.49126
toxins	injury	0.3695
preservative	formaldehyde	0.20115
toxins	aluminum	0.19739
exemption	conscience	0.15035 [S_3]

HIGHLIGHTS

1. We introduce a method using Pathfinder networks derived from distributional models of expert and consumer authored corpora to characterize differences in understanding of the vaccine domain.
2. The resulting Pathfinder networks reveal difference in organization of knowledge on topics related to vaccine safety and adherence.
3. Distributional models using entire documents as a context produced more coherent Pathfinder networks than those using sliding windows as a context.

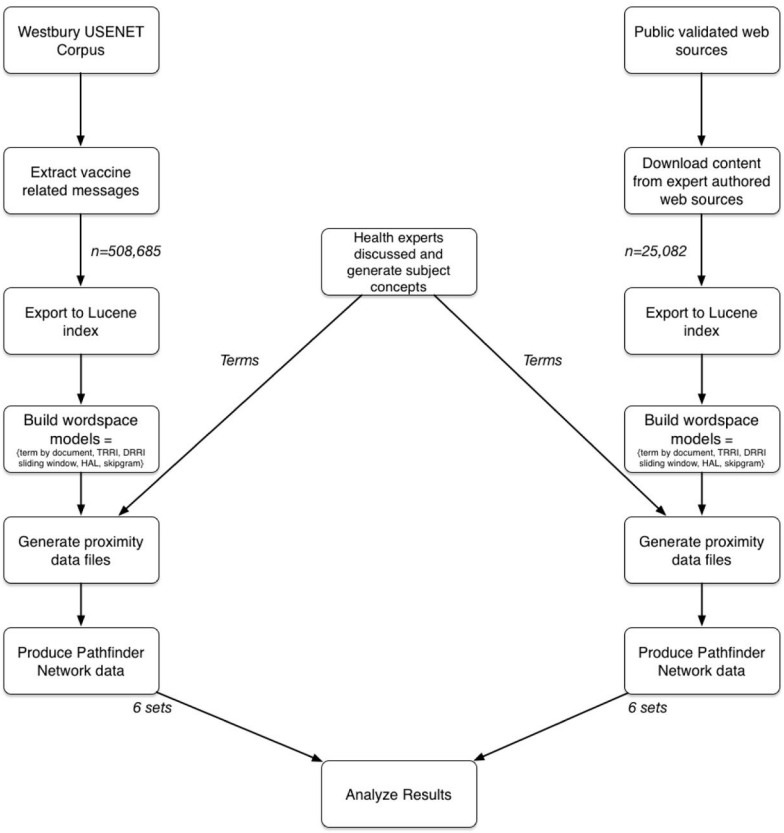


Figure 1.
Process of generating the PFNET from corpus and subject concepts.

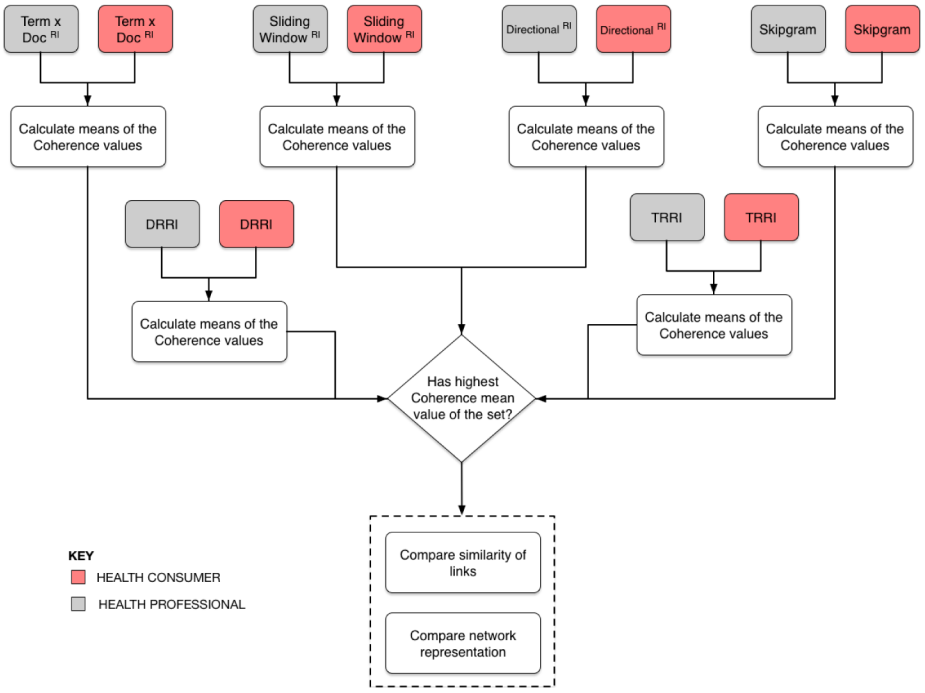
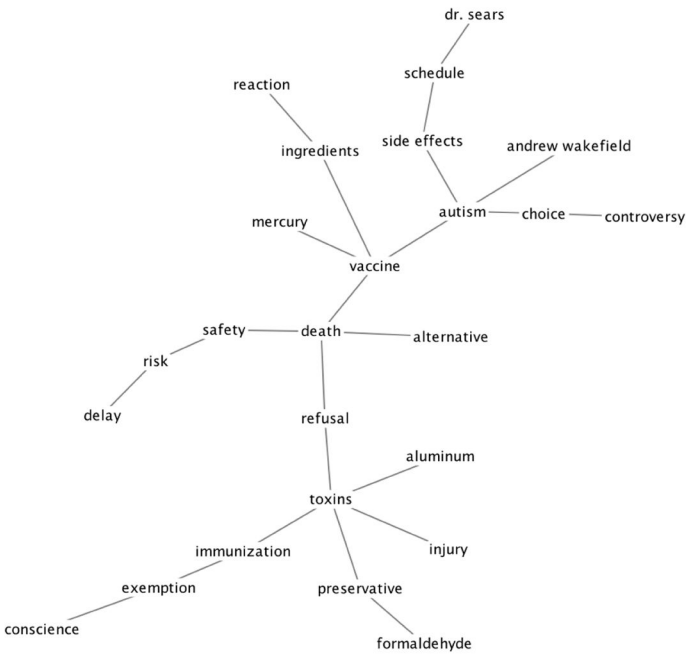
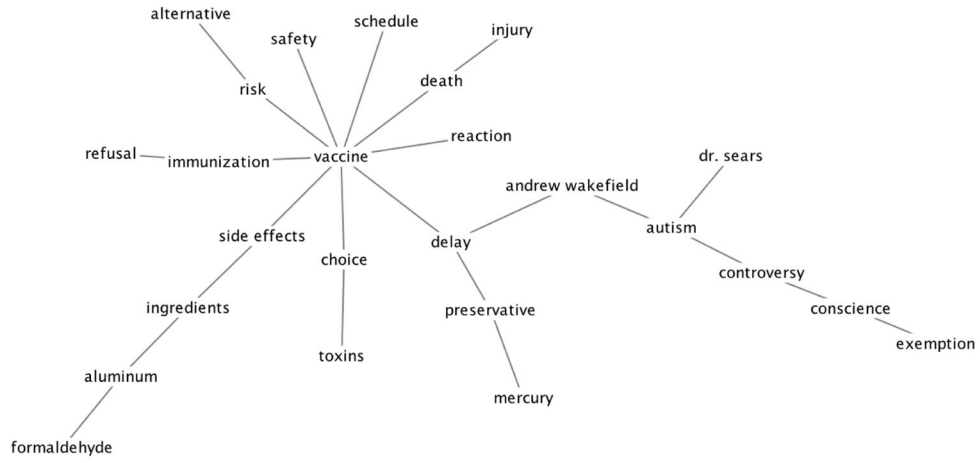


Figure 2.

Process of analyzing PFNETs data based on the Coherence rating. ^{RI} for Random Indexing variant

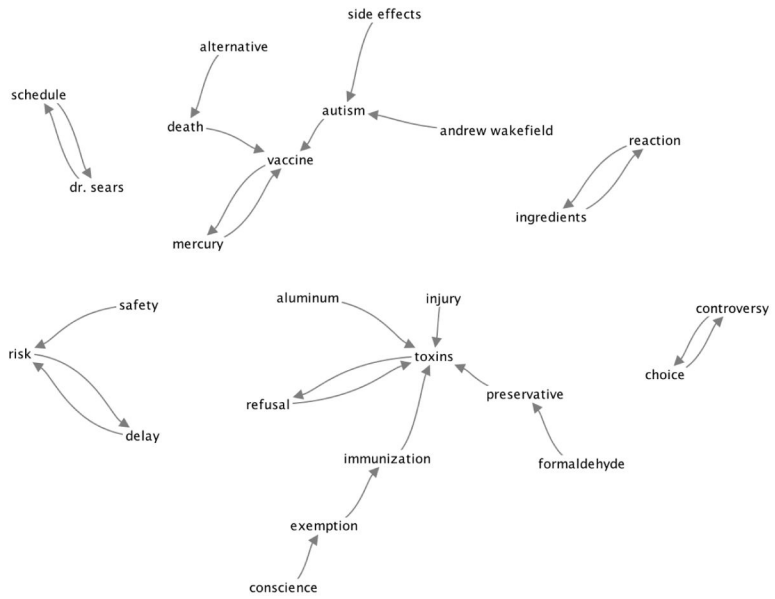


(a) PFNET from health consumer model (TRRI model)

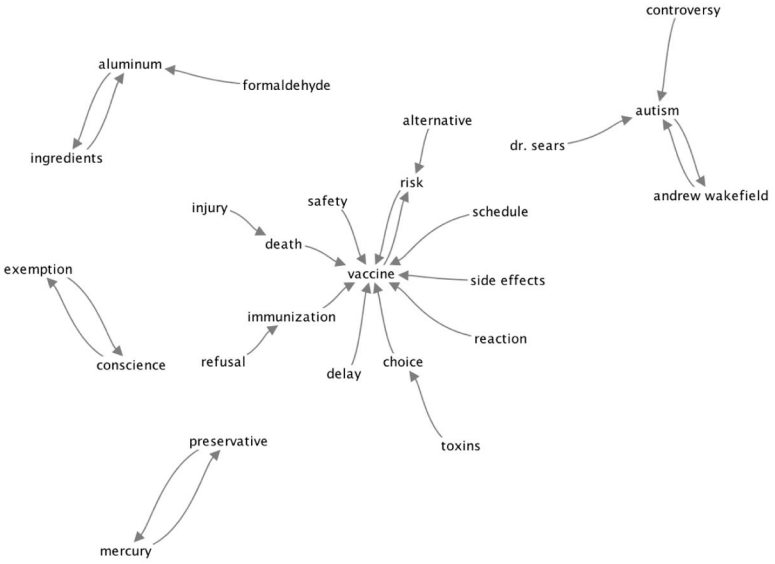


(b) PFNET from health professional model (TRRI model)

Figure 3.
PFNETs for health consumer and health professional model



(a) Nearest neighbor graphs from health consumer model (TRRI model)



(b) Nearest neighbor graphs from health professional model (TRRI model)

Figure 4.
Nearest neighbor graphs for health consumer and health professional model

Table 1

Parameters for models. *(10)* denotes window size of 10. ^{RI} for Random Indexing variant.

Model	Seed length	Iterative cycles	Sliding-window radius	Dimensionality
Term x Doc^{RI}	10	NA	NA	200
DRRI	10	2	NA	200
TRRI	10	2	NA	200
Sliding Window (10)^{RI}	10	NA	10	200
Directional (10)^{RI}	10	NA	10	200
Skip-gram (10)	NA	1	10	200

Table 2

Parameters for PFNETs.

Parameters	Value
Type of data	similarity
Number of nodes	50
Order of data values	coord
Dimension size	200
Distance computation	Cosine (standard)
r	∞
q	$n-1$

Table 3

Coherence ratings for proximity data (^{RI} Random Indexing variant) from the health consumer and health professional models. Highlighted row denotes selected model.

Proximity Data	Consumer Model	Professional Model	Mean (set)
Term x Doc^{RI}	0.772	0.785	0.7785
DRRI	0.717	0.89	0.8035
TRRI	0.822	0.895	**0.8585**
Sliding Window (10)^{RI}	0.183	0.671	0.427
Directional (10)^{RI}	−0.059	0.505	0.223
Skip-gram (10)	0.748	0.616	0.682
Mean (corpus)	0.752	0.82	0.786
Median (corpus)	0.74	0.773	

Table 4

Similarity of links between health professional and health consumer PFNETs.

Number of Links (health consumer)	24
Number of Links (health professional)	24
Common Links	3
Similarity	0.067
Probability of 3 or more links in common by chance	0.299

Table 5

Pairwise similarity between models. Column and row labels with black backgrounds refer to health professional models. *Dir* is shorthand for directional and *SWin* is shorthand for sliding-window.

	PROFESSIONAL						CONSUMER					
	Term x Doc	**DRRI**	**TRRI**	**SWin**	**Dir**	**Skip-gram**	**Term x Doc**	**DRRI**	**TRRI**	**SWin**	**Dir**	**Skip-gram**
Term x Doc	1	0.333	0.455	0.263	0.333	0.091	0.116	0.021	0.067	0.091	0.021	0.091
DRRI	0.333	1	0.297	0.231	0.333	0.143	0.067	0.021	0.067	0.043	0.067	0.116
TRRI	0.455	0.297	1	0.297	0.297	0.143	0.067	0.043	0.067	0.067	0.043	0.067
SWin	0.263	0.231	0.297	1	0.333	0.116	0.091	0.021	0.043	0.067	0.067	0.067
Dir	0.333	0.333	0.297	0.333	1	0.091	0.116	0.067	0.067	0.043	0.021	0.116
Skip-gram	0.091	0.143	0.143	0.116	0.091	1	0.021	0.043	0.043	0.043	0.091	0.091
Term x Doc	0.116	0.067	0.067	0.091	0.116	0.021	1	0.143	0.2	0.2	0.116	0.116
DRRI	0.021	0.021	0.043	0.021	0.067	0.043	0.143	1	0.371	0.091	0.2	0.116
TRRI	0.067	0.067	0.067	0.043	0.067	0.043	0.2	0.371	1	0.2	0.143	0.091
SWin	0.091	0.043	0.067	0.067	0.043	0.043	0.2	0.091	0.2	1	0.091	0.021
Dir	0.021	0.067	0.043	0.067	0.021	0.091	0.116	0.2	0.143	0.091	1	0.2
Skip-gram	0.091	0.116	0.067	0.067	0.116	0.091	0.116	0.116	0.091	0.021	0.2	1

Table 6

Mean and median of pairwise similarities among models.

	Mean	Median
Consumer to Other Consumer	0.153	0.143
Professional to Other Professional	0.250	0.297
Consumer to Professional	0.063	0.067

Table 7

Prominent concepts based on properties of the PFNETs.

	PFNET	
	Consumer	**Professional**
Maximum Degree	toxins	vaccine
Center	death	delay
Median	death	vaccine