Natural Language Processing (NLP) in Qualitative Public Health Research: A Proof of Concept Study

William Leeson¹, Adam Resnick¹, Daniel Alexander¹, and John Rovers²

Abstract
Qualitative data-analysis methods provide thick, rich descriptions of subjects’ thoughts, feelings, and lived experiences but may be time-consuming, labor-intensive, or prone to bias. Natural language processing (NLP) is a machine learning technique from computer science that uses algorithms to analyze textual data. NLP allows processing of large amounts of data almost instantaneously. As researchers become conversant with NLP, it is becoming more frequently employed outside of computer science and shows promise as a tool to analyze qualitative data in public health. This is a proof of concept paper to evaluate the potential of NLP to analyze qualitative data. Specifically, we ask if NLP can support conventional qualitative analysis, and if so, what its role is. We compared a qualitative method of open coding with two forms of NLP, Topic Modeling, and Word2Vec to analyze transcripts from interviews conducted in rural Belize querying men about their health needs. All three methods returned a series of terms that captured ideas and concepts in subjects’ responses to interview questions. Open coding returned 5–10 words or short phrases for each question. Topic Modeling returned a series of word-probability pairs that quantified how well a word captured the topic of a response. Word2Vec returned a list of words for each interview question ordered by which words were predicted to best capture the meaning of the passage. For most interview questions, all three methods returned conceptually similar results. NLP may be a useful adjunct to qualitative analysis. NLP may be performed after data have undergone open coding as a check on the accuracy of the codes. Alternatively, researchers can perform NLP prior to open coding and use the results to guide their creation of their codebook.

Keywords
machine learning, natural language processing, open coding, topic modeling, Word2Vec, public health research

Background
Qualitative methods are a core research tool in a number of fields, including the social sciences, history, education, and public health. Typically, the data in qualitative research are words. Researchers frequently record interviews with research subjects, transcribe the recordings, and perform qualitative analysis on the transcripts. This methodology provides rich, thick descriptions of research subjects’ beliefs, attitudes, and lived experiences and allows the researcher to explore a subject with greater rigor and depth than is often possible using quantitative methods. It may, in fact, be a better way of understanding human experiences (Miles, 1994; Rahman, 2017).

Advantages aside, qualitative methods also have their disadvantages. Collecting, interpreting, and analyzing qualitative data are both time-consuming and labor-intensive and usually require a team of analysts. In order to abstract relevant information from interviews, researchers use a process known as coding to identify and mark passages using short text labels (known variously as codes or nodes) that mark key ideas from which overarching themes are abstracted (Saldana, 2016). However, it can take weeks using these methods to analyze a complete corpus of interviews (Alshenqeti, 2014). Disagreements among analysts must be reconciled, which leads to concerns that qualitative analysis may not be replicable, may be prone to disciplinary predilections, or may be predisposed to bias.

¹ College of Arts & Sciences, Drake University, Des Moines, IA, USA
² College of Pharmacy & Health Sciences, Drake University, Des Moines, IA, USA

Corresponding Author:
John Rovers, College of Pharmacy & Health Sciences, Drake University, 2507 University Avenue, Des Moines, IA 50311, USA.
Email: john.rovers@drake.edu
Natural language processing (NLP) uses algorithmic approaches rooted in statistical techniques to ascertain semantic meaning from textual data. Early applications of NLP were somewhat deterministic in the sense they were based on developing an extensive unchanging and predetermined rule book (e.g., rules mapping between the French and English meanings of a specific phrase) used to identify simple patterns to complete tasks such as language translation. Recent approaches to NLP are more dynamic in the sense that they employ machine learning. A subfield of NLP is natural language generation or the creation of human-understandable text from qualitative or quantitative inputs, contrasting the more general extraction of information from qualitative data represented by NLP (Reiter & Dale, 2000).

While machine learning algorithms also begin with human-defined rules for a given task, these algorithms are designed to improve their performance automatically without human intervention by defining additional rules for the task. Further, these machine learning algorithms are generally applied to other problems in addition to NLP when the problem is too complex for direct human analysis. Some examples include computer vision, speech recognition, or problems that adapt and change over time, such as product recommendations for shoppers or spam filters.

Advantages of NLP include the large amount of textual data that can be processed, the speed of analysis using a trained model (where hundreds of pages of information can be processed in seconds), and the ability to ascertain subtle patterns that could otherwise be indiscernible to human analysts. Disadvantages include the vastness of data sets (known as corpora) needed to train some NLP models; the time needed to train some NLP models, which can take days; and the difficulty of translating the results of NLP model analysis into a human readable format (Mikolov, Chen, Corrado, & Dean, 2013).

As researchers become increasingly conversant with NLP, it is being used in a variety of research areas including the health sciences. It has been shown to be a useful tool in several studies of health-care quality (Baldwin, 2008; Ranard et al., 2016), including the analysis of the effects of allowing patients to read their own health records in plain English (Chen et al., 2018), the classification of patients by clinical condition (Uzuner, Goldstein, Luo, & Kohane, 2008), the extraction of information about cardiovascular function from large data sets (Nath, Albaghdadi, & Jonnalagadda, 2016), and the evaluation of the continuing education needs of practicing dentists based on comments posted to a listserv (Bekhuis, Kreinacke, Spallek, Song, & O’Donnell, 2011).

There is evidence that NLP may have value as a tool for qualitative research. Since qualitative methods suffer some of the difficulties noted above, NLP may mitigate some of these difficulties and be a useful tool in qualitative data analysis. Indeed, Crowston and colleagues have evaluated the potential of NLP to partially automate the analysis of qualitative data (Crowston, Allen, & Heckman, 2012; Crowston, Liu, & Allen, 2010). They compared an earlier rule-based method for machine generation of codes, which used the laborious human development of a rule book to a machine-learning-based automatic coding NLP method in the analysis of postings to a software discussion forum. While they concluded that NLP may be able to speed up the analysis of qualitative data, the rule-based method produced the strongest results. However, the machine learning model used in Crowston’s study is a simple linear classifier, which does not incorporate more recent NLP approaches such as Topic Modeling or Word2Vec that are the two methods we used. Therefore, we seek to compare traditional qualitative methods to more recent NLP implementations to see if modern NLP methods can generate codes that support, augment, or perhaps replace traditional conventional qualitative analysis.

Although previous studies suggest NLP may have value in qualitative research, there do not appear to be published direct comparisons between NLP and hand coding of data using traditional qualitative methods.

**Objective**

This is a proof of concept paper in which we evaluate the potential of NLP in the analysis of qualitative data in an ongoing public health project. Specifically, we ask if NLP can support conventional qualitative analysis, and if so, what its role is.

**Method**

In this study, we compared a traditional qualitative method of open coding (the first step in thematic analysis and other traditional qualitative methods) with two forms of NLP—(1) Topic Modeling and (2) Word2Vec—to analyze transcripts from interviews with men in the Toledo District of rural Belize. We sought to determine if the results returned by either or both NLP method(s) were conceptually similar to those returned by open coding. For purposes of this study, conceptually similar is defined as all four authors reviewing the results of all three analyses and coming to an overall agreement as to whether the results were similar in meaning. For example, if open coding returned a code of “Behavior–Nutrition” and either or both NLP method(s) returned results referring to eating, food, or diet, or nutrition, we judged these to be conceptually similar.

**Data Sources**

Author John Rovers (J. R.) interviewed 56 men living in the Toledo District of Southern Belize to identify their health needs as part of a needs assessment study. All subjects were aged 18 years of age or older and provided verbal, informed assent to participate. Most interviews were performed in English, although in a few interviews, J. R.’s questions were translated in real time from English into Q’eqchi, with responses translated back into English. Semistructured interviews were performed and recorded on a Zoom H1n digital recorder in .wav format and converted to an MP3 format using the Audacity software program Version 2.1.2. The MP3 files were sent to...
in a corpus. These topics are associated with a probability into a group of topics. A topic is a theme that occurs frequently algorithms to summarize large amounts of texts, called a corpus, generate the most appropriate topics. When applied to the corpus, the LDA model produces a set of topics with the probability they are represented by a word (Blei, Ng, & Jordan, 2003). The approach represents a nongenerative approach to NLP as topics will be composed of words or phrases already present in the corpus (Blei et al., 2003). Further, the approach is purely statistical. It is nonpragmatic, nonsyntactic, and nonsemantic as the surrounding context of words, grammar, and meaning of words are not analyzed or interpreted and, solely, the prevalence of words in the corpus is considered.

To create these topic models, we used the Python tool kit Gensim (version 3.4.0), a package of code designed for topic modeling large portions of text (Rehurek & Sojka, 2010). To model the text files, all stop words that appear frequently but have little semantic value (e.g., “be,” “the,” “an,” “it”) were removed from the files (Stanford University, 2008). Finally, a model was created using the Gensim LDA model function to generate 10 topics for each question asked in the interviews. Each of these 10 topics in turn consist of 10 words that are candidates to represent the topic and the associated probabilities for each word. These probabilities represent the likelihood that the word appropriately encapsulates the meaning of the topic.

Word2Vec

Word2Vec is a connectionist application of NLP in which the semantic meaning of individual words is translated into numerical values, represented as vectors that possess a direction and a length, and represent a location in n-dimensional space relative to a point (a force acting on an object is an example of a vector in three-dimensional space). The method of Word2Vec we used was developed by Mikolov, Chen, Corrado, and Dean (2013) and uses shallow neural networks to transform massive data sets into high-quality word vector representations. The results are 300-dimensional word vectors that encode the semantic meaning of words determined using word contexts in sentences. Word2Vec uses a shallow neural network to simulate the human brain and derive the connected meaning of words between disparate portions of the analyzed passage (Kiefer, 2019). Further, the approach is pragmatic and semantic, as the surrounding context of a word is used to derive its semantic meaning, however, grammatical syntax is not considered.

Unlike open coding which is an abstractive process (in the sense that a respondent’s one to two major thoughts or ideas are abstracted into descriptors), Word Embedding codes are nongenerative and are extracted from a passage of a response and used to ascertain the functional meaning of the passage. Two characteristics of a useful code are: (1) the code must be

a professional transcription service, where they were converted to MS Word files. Qualitative analysis was performed on the MS Word files.

Open Coding

Open coding (Guest, 2012) was performed on the MS Word files by author J. R. Cleaning of the transcripts was performed by identifying up to two major concepts given by each respondent for a given question. The one to two major concepts, respondents’ demographic data, and any relevant analyst comments were entered into MS Excel question by question. The unit of analysis was by question rather than by individual response.

The Excel data were then imported into nVivo Version 9.0 and underwent open coding by J. R. to identify preliminary common concepts and ideas, which were summarized in a codebook of short descriptors, called nodes in nVivo. These codes were used to code subsequent transcripts. A parent node (e.g., Behavior) may also have one or more daughter nodes (e.g., Behavior–Nutrition) that describe the parent node in greater detail. The open coding was grounded in interpretive phenomenology since the analysis required interpretation of the lived experiences of the respondents and was not dependent upon a preexisting theoretical preconception. Interpretive phenomenology has also been shown to be valuable in previous studies of patients’ perspectives on their illness (Creswell, 2007; Smith, 2008; Smith & Osborn, 2015).

In a later stage of the project, the codebook and the output from the analysis in nVivo are reviewed in detail to identify overarching concepts. These concepts represent the themes identified in a thematic analysis (Guest, 2012) and will be described elsewhere.

Ideally, qualitative data are coded by a team of analysts so that disagreements about codes can be resolved. The coding in this study was performed by only a single analyst (author J. R.), which might have biased the analysis of the data. Although this was not ideal, it was largely unavoidable since the open coding was performed while author J. R. was on sabbatical leave in rural Belize and not in consistent contact with the other authors. In an attempt to overcome this concern, after the analysis was completed, four experts in Belizean health care reviewed author J. R.’s proposed codes and, after extensive discussion, the experts agreed the codes proposed were empirically reasonable. The four experts also reviewed the proposed themes from the thematic analysis (described elsewhere) and confirmed those identified by author J. R. By using four experts external to the project, the risk of errors or bias during either open coding or thematic analysis are minimized.

Topic Modeling

Topic Modeling is a statistical form of NLP that uses algorithms to summarize large amounts of texts, called a corpus, into a group of topics. A topic is a theme that occurs frequently in a corpus. These topics are associated with a probability measuring how strongly the topic is present in the corpus (Blei, 2012). A vocabulary, or dictionary, is used to evaluate how likely a word or phrase is to be used in a specific context (Lyra, 2017). Topic models use statistical methods to explain the similarity of topics. For this study, a latent Dirichlet allocation (LDA) model was used to examine the textual data and generate the most appropriate topics. When applied to the corpus, the LDA model produces a set of topics with the probability that they are represented by a word (Blei, Ng, & Jordan, 2003). The approach represents a nongenerative approach to NLP as topics will be composed of words or phrases already present in the corpus (Blei et al., 2003). Further, the approach is purely statistical. It is nonpragmatic, nonsyntactic, and nonsemantic as the surrounding context of words, grammar, and meaning of words are not analyzed or interpreted and, solely, the prevalence of words in the corpus is considered.
representative of the meaning of text which it is encoding, providing high-level “summative” meaning and (2) the code must have semantic depth; it must encode important semantic “essence-capturing” that belies the specific importance of a passage (Saldana, 2016).

A simple statistic to indicate how representative a word is of a sentence is the average angle between the word’s vector representation and the vector representations of all other words in the sentence. Common words used in varied contexts will likely be closest to the greatest number of other words’ vectors. Therefore, words with the most specific meaning will be furthest away from the greatest number of other word vectors.

Word significance is correlated with word vector magnitude (Schakel & Wilson, 2015). Thus, word vector magnitude is used as a simple criterion for the semantic depth of a word.

With criteria formed for the two key components of high-quality codes, representativeness (inverse average angular distance) and depth (word vector magnitude), we used the metric

$$\text{score} = (\text{angular vector distance})^{-1} \times \text{magnitude},$$

to score each word in a passage, with a high score indicating a useful code. The \(j\) parameter is tuned for each application of the algorithm. A passage with words that are complex and diffuse in meaning will require a lower \(j\) value (as vectors are more naturally dispersed in space). For this corpus of interviews, a \(j\) value of 3 was determined to produce optimal codes.

Our rationale for choosing these particular NLP methods was twofold: first, each is a relatively recent development within NLP and, second, each uses a different approach—statistical (Topic Modeling) versus neural/connectionist (Word2Vec)—providing a broader sampling of NLP methodologies than if we used a single method.

**Ethical Review**

This study was approved by the Drake University Institutional Review Board (File # 2016-17023). The Belizean Ministry of Health reviewed the research protocol and provided a letter noting the study was in compliance with trial protocols and international codes of ethical standards. After the study was described to them, study procedures reviewed, and risks and benefits explained, all interviewees gave informed verbal assent before being interviewed.

**Results**

The most four to five common results for both NLP methods and open coding as well as the interview questions are shown in Table 1. Interpretations of the results follow the table.

**Open Coding**

After the transcripts underwent open coding, nVivo was used to create a series of charts that showed how often any given code was used in the coding of a given question. An example is shown below in Figure 1. The results show that the most common responses to the question “What do you believe health is?” were coded at the following nodes:

- Health is a function of Behavior \((n = 50)\).
- Health is a function of the daughter node of Behavior–Nutrition \((n = 25)\).
- Health is a general, nonspecific good \((n = 18)\).
- Health is a function of the daughter node of Behavior–Activity \((n = 14)\).
- Health is being able to work \((n = 9)\).

Similar output was generated for all 13 interview questions, and the results are shown in Table 1.

**Topic Modeling**

The results of the model can be interpreted by looking at each word-probability pair. For example, the question “How healthy do you feel” yielded the following response:

\[
\begin{align*}
0.230^{*}\text{“healthy”} & \quad + \quad 0.121^{*}\text{“good”} & \quad + \quad 0.071^{*}\text{“eat”} \\
0.051^{*}\text{“find”} & \quad + \quad 0.049^{*}\text{“something”} & \quad + \quad 0.046^{*}\text{“got”} \\
0.029^{*}\text{“morning”} & \quad + \quad 0.010^{*}\text{“used”} & \quad + \quad 0.008^{*}\text{“little”} \\
& \quad + \quad 0.008^{*}\text{“told”}
\end{align*}
\]

which can be interpreted as concluding that the topic has a 23% chance of being represented by “healthy,” a 12.1% chance of being represented by “good,” a 7.1% chance of being represented by “eat,” and so on. This means that throughout the interviews, one of the topics of discussion was likely to be about feeling healthy and good. It is also likely that among multiple responses, the theme of health was connected to eating or the morning.

**Word2Vec**

Some examples of machine-generated codes are shown in Table 2. For the response given by Respondent 1, after the vector representation of each word in the sentence was computed, each word was scored on the criteria of a combination of representativeness (average angle distance) and depth (vector magnitude). Though this long passage of text had words such as “well” and “whatever” that did not contribute semantic meaning, the top word to act as a code for this portion of text was identified as “diet” and the second-best code was identified to be “rice.” For Respondent 2’s response regarding how to stay healthy, the top two codes for the passage were machine labeled “exercise” and “active.”

**Discussion**

The results from open coding, a long-established practice in qualitative methods, will be compared to those yielded by both NLP methods. This type of comparison does not appear to have been previously done.
With most questions, one or both NLP methods returned results conceptually similar to the codes created during open coding. The most common open codes created were the parent node of Behavior and its daughter nodes of Behavior–Nutrition and Behavior–Activity. These open codes were created from men’s responses in which they stated health was something that was created or impaired by some

Note.  *Sexually transmitted infection/sexually transmitted disease.  bPunta Gorda town.  cBelize Council for the Visually Impaired.

Conceptually Similar—Questions 1–5, 7–10, 12, and 13

Table 1. Most Common Results for Topic Modeling, Word2Vec, and open coding.

| Interview Questions (Qs)                                      | Most Common Responses for Topic Modeling | Most Common Responses for Word2Vec | Most Common Codes for Open Coding |
|---------------------------------------------------------------|-----------------------------------------|-----------------------------------|----------------------------------|
| Q1: What do you believe health is?                           | Good (.175), exercising (.039), example (.036), healthy (.025), meaning (.024) | Hygiene, flu, pores, sanitation, showers, cleanliness, diabetes, diet | Behavior, Behavior–Nutrition, good, Behavior–Activity |
| Q2: How healthy do you feel? Why?                           | Healthy (.230), good (.121), eat (.071), find (.051), something (.049) | Tablespoons, cholesterol, hypertension, mg, diabetes, flu, aspirin, symptoms | Well, unwell, chronic, age |
| Q3: What is the biggest health problem men have in Toledo?   | Think (.155), problem (.082), alcohol (.067), people (.064), biggest (.061) | Cholesterol, vitamins, marijuana, diabetes, acids, symptoms, checkups, eyesight | Unsure, diabetes, diet, poor self-care |
| Q4: What do you do that keeps you healthy?                   | Healthy (.113), ground (.059), different (.04), tough (.04), meals (.04) | Tortilla, vertebra, tortillas, cilantro, cassava, vitamins, greens, cholesterol | Behavior, Behavior–Nutrition, Behavior–Activity, Sweat |
| Q5: Why do men become sick or unhealthy?                     | Take (.085), exercise (.054), would (.048), body (.043), disease (.026) | Cancers, cholesterol, malnutrition, vitamins, heredity, fertilize, malaria, diabetes | Behavior, Behavior–Nutrition, alcohol, poor self-care |
| Q6: If you have a health problem, where do you go for help?  | Private (.121), health (.084), last (.047), good (.047), community (.041) | Metformin, reflux, hydrated, cholesterol, pill, pills, infections | Public, other provider, private, Hillside |
| Q7: How often do you get help from a hospital, doctor, or clinic? | Went (.145), doctor (.081), hospital (.065), maybe (.041), haven’t (.025) | Flu, fever, inhaler | In last year, over a year, symptomatic, monitoring |
| Q8: How often do you get help from a bush doctor or take bush medicines? | Medicine (.164), use (.138), sometimes (.033), maybe (.027), one (.027) | Coca, vitamins, aloe, antibacterial, cleanser | Yes, Yes-specific, no, Yes-General |
| Q9: Do you ever go for a checkup even when you are feeling healthy? Why or why not? | Time (.055), every (.055), yes (.038), haven’t (.038), would (.038) | Checkup, checkups, immunization, flus | Yes, no, Yes-incorrect, Yes-correct |
| Q10: Can you give me an example of a health problem that you would get help for? | Go (.059), what’s (.059), feel (.040), pain (.021), work (.021) | Reflux, cholesterol, flu, diabetes, strokes, fever, arthritis | Acute, pain, diabetes, chronic |
| Q11: What can you tell me about the health effects of cigarette smoking? Alcohol? Illicit drugs like marijuana? Sexually transmitted illness? | (When referring to smoking) know (.107), bad (.062), think (.062), smoke (.046), tobacco (.046) | Marijuana, syphilis, smoke, nicot ine, kills, condoms, condoms, gonorrhea | Personal, alcohol-wrong, STI-wrong, tobacco-negative |
| Q12: Where would you go for help if you had a problem with your eyes? Your teeth? A lot of worry or unsettled mind? | PGb (.145), teeth (.063), cool (.058), financial (.058), live (.42) | PGb, metformin, blurrriness, cataract, polyclinic, overdose | Teeth-PGb, eyes-BCVI, mind-self-care, other |
| Q13: Of all the health problems men in Toledo may have, which one should we try and fix first? | Sickness (.031), going (.031), hard (.021), pressure (.021), diabetes (.021) | Cholesterol, arteries, gonorrhea, diabetes, hypertension, bacteria, eyesight, lungs | Unsure, diabetes, miscellaneous, lack of knowledge |
form of behavior and were the most common codes for Questions 1, 4, and 5.

Both NLP methods returned responses that were conceptually similar to the two most common open codes. In Question 1, exercising and diet were common responses and have a similar meaning to the Behavior nodes. In Question 4, the words “ground” (which in a Belizean context refers to homegrown food) and “meals” were yielded by Topic Modeling, while Word2Vec returned a wider variety of food terms. Both are consistent with the code Behavior–Nutrition. In Question 5, causes for ill health, were again related to behavior during open coding, while Topic Modeling and Word2Vec listed exercise, and malnutrition and vitamins, respectively.

In Question 2, responses for how healthy respondents felt were also conceptually similar. Open coding returned fairly predictable results that respondents felt well or unwell. The results of healthy and good for Topic Modeling are consistent with the result of well for open coding, while the results for Word2Vec included a number of disease states that are likewise consistent with feelings of unwellness returned by open coding.

Men’s health reasons for seeking care were evaluated in Questions 3, 9, 10, and 13. In all cases, the most common results listed diabetes or an acute problem (pain, stroke, fever). This may be due to the results from both Topic Modeling and open coding indicating men felt mostly healthy, well, or good and so did not seek treatment unless there was an acute or chronic problem. In Question 9, the term “haven’t” in Topic Modeling supports the Open Codes of No and Yes—Incorrect in regard to checkups, which indicates that men either do not come for checkups or they defined the term checkup incorrectly.

While there was not agreement among all the methods in Questions 7 or 8, there was consistency between open coding and Topic Modeling. Question 7 asked about seeking help, and Question 8 queried respondents about their use of bush medicines. Both open coding and Topic Modeling methods indicated infrequent use of health resources in Question 7 (maybe 4.1%, haven’t 2.5%) and “over a year,” respectively. The NLP methods in Question 7 had difficulty interpolating the time frame indicated in the responses (see below for more about this). Nevertheless, both Word2Vec and open coding provided responses consistent with only coming to care when the respondent perceived a problem with the terms flu, fever, symptomatic.

In Question 12, which inquired into the availability of local health resources, all three methods showed a similar awareness that dental and vision care was available in the town of Punta Gorda where a public polyclinic is located. Neither NLP method indicated a clear awareness of where respondents would go for mental health treatment, while open coding returned “Mind-Self Care,” which indicated that mental health issues were addressed by the respondent trying to calm his emotions or mood by himself.

Not Conceptually Similar—Questions 6 and 11

Question 6 also inquired about men’s understanding of the availability of local health resources. In Question 6, there was a different result produced by each method. Open coding reflected a preference for public resources, while Topic Modeling instead suggested private resources (12.1%) were more common than public ones (4.1%). Word2Vec did not address the question clearly.

The responses for Question 11 were not entirely consistent among the three methods. This question asked about the risks of alcohol, tobacco, sexually transmitted infection, and illicit drugs. The code “Personal” created during open coding showed that if men had no personal experience with any of these items, they were either unable to name any of its risks, or if they did discuss a risk, it was frequently factually incorrect—for example, the major risk of alcohol was described as a headache rather than liver damage. Neither NLP method returned results consistent with open coding, although both contained responses suggesting negative consequences of these behaviors.

Divergences in Responses

Our use of conceptual similarity does not necessarily mean that there is total agreement among the three methods when there is conceptual similarity—it means that the authors judge at least two of the methods to be in agreement. Thus, our investigation of differences in results returned by the three methods will
include cases where conceptually similar responses have differences (e.g., Questions 7 and 8) as well as questions in which there was not conceptual similarity (Questions 6 and 11). In particular, the differences among the responses to Questions 6, 7, 8, and 11 are worth exploring further. Reasons for these differences may be a function of the nature of the question, the nature of the interviewee’s response, or both. For example, in Question 6, the fact that Topic Modeling suggested a preference for private over public health care, while open coding suggested the reverse, might be due to the fact that Topic Modeling aggregated responses that open coding interpreted as “other” or “Hillside” into “private” (and the NLP methods might not have known how to classify Hillside, which is a local clinic near Punta Gorda—if so, this suggests that NLP methods might need to incorporate knowledge of the locale to be more effective).

Vague responses seemed to cause an issue with Questions 7 and 8, which were intended to be straightforward. Respondents were expected to provide a factual response that could be located somewhere in time (e.g., last month, about a year ago) indicating when they last came to medical attention. During open coding, however, it became apparent that few responses actually provided a fixed point in time when care was sought. Frequently, respondents to Question 7 provided lengthier answers reflecting on the nature of some illness or symptom they had suffered, and the time frame for care-seeking, while implied, was buried in a longer response that did not, strictly speaking, provide a factual answer to the question. Consequently, during open coding, creating the codes of less than a year or more than a year had to be interpolated from the transcripts. Neither NLP method proved able to interpolate a result from a somewhat vague transcript that did not provide a clear answer. There were similar issues with Question 8.

Similarly, asking for a factual risk of liver disease, cancer, and so on in Question 11 proved difficult when (1) any factual risk respondents provided was incorrect or (2) respondents provided an indirect answer indicating their lack of personal experience with the subject. As with Question 7, identifying a simple fact proved difficult for the respondents, and as noted above, neither NLP method was able to generate the responses that converged to those provided by open coding. Again, as with Question 7, the human coder seemed better suited to interpret difficult responses.

In contrast to the difficulties NLP encountered in Questions 6, 7, 8, and 11 when encountering responses that were vague or nonfactual, or required local knowledge, both NLP methods were able to clearly identify data reflecting respondents’ opinions that health is a consequence of various behaviors when questions sought an opinion (e.g., Question 1—What do you believe health is?). In these cases, NLP methods were not required to identify a single fact otherwise hidden in a complex and perhaps murky response, and both methods returned a clear result.

A summary of how NLP and open coding provided conceptually similar or divergent results is shown in Table 3.

### Limitations

Several limitations to the study should be considered. Use of only a single analyst for open coding may have introduced bias into the results. We attempted to adjust for this by having four experts who reviewed the Open Codes independently and confirmed they appeared to be empirically reasonable.

When comparing results from different methods or analysts, typically, agreement between analysts would be determined by a κ score or similar statistic. In this study, this was not possible since the process of creating codes, or generating topics, or scores in Word2Vec was heterogeneous. Consequently, the conceptual similarity of the results still required qualitative judgment by the authors, which was achieved by discussions among all authors until agreement could be reached.

We only compared two NLP methods to open coding. Different analysts using other NLP methods may have different results.

We only studied the role of NLP for only one qualitative method and cannot comment on the potential value of NLP for other qualitative methods.

We compared the NLP methods to open coding, which is only the first step in a qualitative analysis. Although the results across all three methods returned conceptually similar codes and concepts, neither NLP method was able to critically evaluate its results and synthesize them into actionable insights, which would be the next step in a thematic analysis.

### Conclusions

The current study’s suggestion that the form of a question affects NLP’s ability to negotiate complex and imprecise responses notwithstanding, it does appear that NLP nevertheless may be a useful adjunct to traditional qualitative research methods.

Our results suggest that both Topic Modeling and Word2Vec quite frequently return results that are conceptually similar to those created in the production of Open Codes. This happened in 9 of the 13 questions (Questions 1–5, 9, 10, 12, and 13) and to a lesser degree in Questions 7 and 8, where Topic Modeling was consistent with open coding.

### Table 3. Conceptual Similarity or Divergence Across Methods.

| Question | Conceptual Similarity? |
|----------|------------------------|
| 1–5      | Yes—Overall agreement among all three methods |
| 6        | No—Little agreement among the methods |
| 7 and 8  | Yes—Overall agreement between Topic Modeling and open coding; limited agreement between Word2Vec and open coding |
| 9 and 10 | Yes—Overall agreement among all three methods |
| 11       | No—Limited agreement among all three methods |
| 12 and 13| Yes—Overall agreement among all three methods |
Researchers may find NLP helpful to support their analysis in two ways:

1. Performing NLP after traditional qualitative methods have been completed permits researchers to evaluate the probable accuracy of the codes created. In this study, the frequent mentions about food in the NLP results confirms that the codes related to nutrition and other codes used in the open coding are likely to be correct. NLP may thus serve as a way to corroborate the value of the codebook.

2. Alternatively, researchers can perform NLP prior to open coding and use the NLP results to guide creation of the codes. Thus, to maximize the utility of NLP results, researchers may wish to pretest their proposed interview questions against NLP methods.

In both cases, NLP has the potential to guide coding in ways that can make the process less arduous and time-consuming. Looking toward the future, it is certainly possible that as issues such as those discussed above with Questions 6, 7, 8, and 11 arise, NLP methodologies will improve and the algorithms will be better able to interpolate vague, complex, and unexpected responses or to identify entities particular to the locale.

At this point, however, it does not appear that the two NLP methods we evaluated would be able to proceed beyond the first step of open coding. Although the results across all three methods returned conceptually similar codes and concepts, neither NLP method was able to critically evaluate its results and identify themes that would normally be the final product of a thematic analysis. Thus, while NLP appears to be a useful adjunct to qualitative methods, it does not yet appear to be an adequate alternative to human analysis.

In other fields, we are seeing a machine–human symbiosis. A combination of computer analytics and human insight produces optimal results,\(^1\) so perhaps the sweet spot in qualitative analysis is a combination of machine learning and human analysis, where we take advantage of NLP to speed analysis and cross-check open coding, yet continue to rely on humans to critically evaluate results.

**Acknowledgments**
We gratefully acknowledge the assistance of Matthew Nicasio, MPH, for assistance with study logistics, local expertise, and providing cultural guidance. We also would like to acknowledge the assistance of Ann Colbert, MD; Jessica Gosch, PharmD; Anna Hanson, RN; and Carley Kirsch, DPT, for reviewing the results of the open coding and thematic analysis.

**Declaration of Conflicting Interests**
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**Funding**
The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This project was funded in part by the John R. Ellis Research Endowment, the Jan Harris Research Endowment and a Nelson Pressing Global Issues Grant, all from Drake University.

**ORCID iD**
John Rovers doi: https://orcid.org/0000-0002-7992-1587

**Note**
1. For example, business decision-making (Jarrahi, 2018), sports analytics (“Blending Old and New: How the Florida Panthers Try to Predict Future Performance at the NHL Entry Draft;” “How Analytics is Shaping the 2018 World Series and Baseball’s Future,” 2018), and election forecasting, where The Cook Report’s blend of quantitative and qualitative methods has led to an enviable success in projecting elections (Silver, 2012) and where the website FiveThirtyEight.com incorporates expert predictions into its computer models (“How FiveThirtyEight’s House, Senate and Governor Models Work,” 2018).

**References**
Alshenqeeti, H. (2014). Interviewing as a data collection method: A critical review. *English Linguistics Research, 3*, 39–45. doi:10.5430/elr.v3n1p39
Baldwin, K. B. (2008). Evaluating healthcare quality using natural language processing. *Journal for Healthcare Quality, 30*, 24–29. doi:10.1111/j.1945-1474.2008.tb01151.x
Bekhuis, T., Kreinaecke, M., Spallek, H., Song, M., & O’Donnell, J. A. (2011). Using natural language processing to enable in-depth analysis of clinical messages posted to an Internet mailing list: A feasibility study. *Journal of Medical Internet Research, 13*, e98. doi:10.2196/jmir.1799
Blei, D. M. (2012). Topic modeling and digital humanities. Retrieved from http://journalofdigitalhumanities.org/2-1/topic-modeling-and-digital-humanities-by-david-m-blei/
Blei, D. M., Ng, A., & Jordan, M. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Blending old and new: How the Florida Panthers try to predict future performance at the NHL entry draft. MIT Sloan Analytics Conference. Retrieved from http://www.sloansportsconference.com/content/blendinng-old-new-florida-panthers-try-predict-future-performance-nhl-entry-draft/
Chen, J., Druhl, E., Ramesh, B., Houston, T., Brandt, C., Zulman, D. M., . . . Yu, H. (2018). A natural language processing system that links medical terms in electronic health record notes to lay definitions: System development using physician reviews. *Journal of Medical Internet Research, 20*, e26. doi:10.2196/jmir.8669
Creswell, J. W. (2007). *Qualitative inquiry and research design* (2nd ed.). Thousand Oaks, CA: Sage.
Crowston, K., Allen, E. E., & Heckman, R. (2012). Using natural language processing technology for qualitative data analysis. *International Journal of Social Research Methodology, 15*, 523–543. doi:10.1080/13645579.2011.625764
Crowston, K., Liu, X., & Allen, E. E. (2010). Machine learning and rule-based automated coding of qualitative data. *Proceedings of the American Society for Information Science and Technology, 47*, 1–2. doi:10.1002/meet.14504701328

Guest, G. (2012). In G. Guest, K. M. MacQueen, & E. E. Namey (Eds.), *Applied thematic analysis* (pp 49–78). Los Angeles, CA: Sage.

How analytics is shaping the 2018 World Series and baseball’s future. (2018). Retrieved from https://www.sporttechie.com/analytics-shaping-2018-world-series-baseball-future-mlb/

How FiveThirtyEight’s House, Senate and Governor models work. (2018). Retrieved from https://fivethirtyeight.com/methodology/how-fivethirtyeights-house-and-senate-models-work/

Jarrahi, M. H. (2018). Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. *Business Horizons, 61*, 577–586. doi:10.1016/j.bushor.2018.03.007

Kiefer, A. (2019). A defense of pure connectionism. *CUNY Academic Works*. Retrieved from https://academicworks.cuny.edu/gc_eds/3036

Lyra, M. (2017). Evaluating topic models. Retrieved from https://pydata.org/berlin2017/schedule/presentation/54/

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. Retrieved from https://arxiv.org/abs/1301.3781

Miles, M. B. (1994). In M. B. Miles & A. Huberman (Eds.), *Qualitative data analysis: An expanded sourcebook* (2nd ed.) (pp. 1–12). Thousand Oaks, CA: Sage.

Nath, C., Albaghdadi, M. S., & Jonnalagadda, S. R. (2016). A natural language processing tool for large-scale data extraction from echocardiography reports. *PLoS One, 11*, e0153749. doi:10.1371/journal.pone.0153749

Rahman, M. S. (2017). The advantages and disadvantages of using qualitative and quantitative approaches and methods in language “testing and assessment” research: A literature review. *Journal of Education and Learning, 6*, 102–112.

Ranard, B. L., Werner, R. M., Antanavicius, T., Schwartz, H. A., Smith, R. J., Meisel, Z. F., . . . Merchant, R. M. (2016). Yelp reviews of hospital care can supplement and inform traditional surveys of the patient experience of care. *Health Affairs, 35*, 697–705. doi:10.1377/hlthaff.2015.1030

Rehurek, R., & Sojka, P. (2010, May 22). *Software framework for topic modelling with large corpora*. Paper presented at the New Challenges for NLP Frameworks Programme, Valletta, Malta.

Reiter, E., & Dale, R. (Eds.). (2000). *Building natural language generation systems*. Cambridge, England: Cambridge University Press.

Saldana, J. (2016). *The coding manual for qualitative researchers* (3rd ed.). Los Angeles, CA: Sage.

Schakel, A. M. J., & Wilson, B. J. (2015). Measuring word significance using distributed representations of words. Retrieved from https://www.openaire.eu/search/publication?articleId=od__________l8:e19dc843dbb002394227fd5ce19ef827f6

Silver, N. (2012). *The signal and the noise: Why most predictions fail—But some don’t*. New York, NY: Penguin Press.

Smith, J. A. (2008). *Qualitative psychology: A practical guide to research methods* (2nd ed.). Thousand Oaks, CA: Sage.

Smith, J. A., & Osborn, M. (2015). Interpretative phenomenological analysis as a useful methodology for research on the lived experience of pain. *British Journal of Pain, 9*, 41–42. doi:10.1177/2049463714541642

Stanford University. (2008). Dropping common terms: Stop words. Retrieved from https://nlp.stanford.edu/IR-book/html/htmledition/dropping-common-terms-stop-words-1.html

Uzuner, Ö., Goldstein, I., Luo, Y., & Kohane, I. (2008). Identifying patient smoking status from medical discharge records. *Journal of the American Medical Informatics Association, 15*, 14–24. doi:10.1197/jamia.M2408