J Healthc Inform Res
https://doi.org/10.1007/s41666-018-0015-z

RESEARCH ARTICLE

TIQS: Targeted Iterative Question Selection for Health Interventions

Keith Feldman[1] · **Spyros Kotoulas**[2] ·
Nitesh V. Chawla[1,3]

Received: 31 August 2017 / Revised: 8 February 2018 / Accepted: 21 February 2018

Abstract While healthcare has traditionally existed within the confines of formal clinical environments, the emergence of population health initiatives has given rise to a new and diverse set of community interventions. As the number of interventions continues to grow, the ability to quickly and accurately identify those most relevant to an individual's specific need has become essential in the care process. However, due to the diverse nature of the interventions, the determination need often requires non-clinical social and behavioral information that must be collected from the individuals themselves. Although survey tools have demonstrated success in the collection of this data, time restrictions and diminishing respondent interest have presented barriers to obtaining up-to-date information on a regular basis. In response, researchers have turned to analytical approaches to optimize surveys and quantify the importance of each question. To date, the majority of these works have approached the task from a univariate standpoint, identifying the next most important question to ask. However, such an approach fails to address the interconnected nature of the health conditions inherently captured by the broader set of survey questions. Utilizing data mining and machine learning methodology, this work demonstrates the value of

✉ Nitesh V. Chawla
nchawla@nd.edu

Keith Feldman
kfeldman@nd.edu

Spyros Kotoulas
Spyros.Kotoulas@ie.ibm.com

[1] Department of Computer Science and Engineering, and iCeNSA, University of Notre Dame, Notre Dame, IN 46656, USA

[2] IBM Research Ireland, IBM Technology Campus, Dublin, Ireland

[3] Wrocław University of Science and Technology, Wrocław, Poland

Published online: 27 March 2018

capturing these relations. We present a novel framework that identifies a variable-length subset of survey questions most relevant in determining the need for a particular health intervention for a given individual. We evaluate the framework using a large national longitudinal dataset centered on aging, demonstrating the ability to identify the questions with the highest impact across a variety of interventions.

Keywords Healthcare informatics · Personalized health · Community health · Data mining

1 Introduction

Marked by the notion that we can no longer reactively treat the sick, society is undergoing a major shift in our approach to healthcare. Over the past decade, clinicians and researchers have demonstrated the value of preemptively addressing and improving the sundry of conditions that impact an individual's health. However, to do so has required a transformation of both the assessment and delivery of care. Whereas health services have traditionally centered on a limited number of formal care entities, such as hospitals and primary care providers, recent methods have adopted a more population health perspective. In turn fostering an explosion of community health interventions that represent a spectrum of care from clinical specialists to social programs [27, 32].

Unfortunately, due to the increasing scale and breath of interventions available, the identification of services aligned with an individual's need presents a non-trivial task. Contributing in large part to this difficulty is the fact that clinical factors available through medical records account for only about 10% of an individual's overall health condition, with the remainder comprised of socioeconomic, environmental, and social factors commonly known as the social determinants of health [8, 26].

Although there exists strong support for the use of survey tools to collect data on these factors, the fluid nature of an individual's health can quickly out-date the collected information [4]. As a result, frequent re-administration of surveys is often required to ensure an accurate assessment of need. Beyond the obvious inconvenience, the repeated administration of surveys has been shown to decrease response rates and reduce data quality, a well-documented experience known as response fatigue [31]. Understanding the need to balance participant experience and data quality with the value of up-to-date information, the work presented in this manuscript is built upon the notion that developing a successful survey framework requires the identification of those questions "which are critical to reduce decisional uncertainty and minimize gaps in the current knowledge base" [42].

Much of the early research into survey tools focused on elements of their construction and administration, including collection mediums, response formats, and cogitative aspects of bias related to how questions are written and presented [10, 22, 25, 28, 29, 36, 37, 39]. However, the need to identify highly informative questions has presented an opportunity to utilize analytical approaches to improve the use of surveys for assessment tasks. Utilizing metrics such as Bayesian variance measures, gain ratio, maximum information criterion, and other estimation methods, current analytical techniques have drawn on the idea of item response theory to quantify the impact of questions related to an outcome [11, 16, 21, 41]. These works have ultimately led to the dynamic optimization of the questions within a survey based on the set of previous answers to construct an optimal test for each examinee, a concept broadly known as adaptive testing [33, 40, 43].

To date, the majority of these works operate in what can be viewed as a sequential or univariate manner, evaluating the importance of each remaining question or category to select the next most informative with respect to some desired outcome. Yet, such an approach fails to capture a vital source of information found in the relations between the questions themselves. It is important to remember that survey questions do not exist in isolation. Rather, the elements on which their questions collect data are often highly interdependent. Thus, it stands to reason that more than simply identifying a ranked list of the most important questions, there is tremendous value in identifying the *set* of most informative questions in relation to the desired outcome.

Healthcare data remains among the most prominent examples of these relations, with numerous works having established how an individual's health conditions are deeply intertwined across all aspects of their social determinants [5, 45]. Recognizing this, community interventions have begun to develop more comprehensive surveys in an effort to identify these relations in the profile of individuals they serve. However, the rising number of factors for which they screen presents a significant barrier to identifying these relations in an efficient manner. Moreover, it is clear the presence of specific relations may differ with respect the outcome being evaluated, resulting in the need to develop and maintain numerous models.

In the face of such complexity, our work aims to demonstrate how data mining and machine learning techniques can be utilized to capture these latent relations, providing the ability to answer the question: *What is the set of most relevant questions that aid in determining an individual's need of a particular health intervention?* To this end, we present TIQS, the Targeted Iterative Question Selection framework (pronounced ticks) designed to identify a personalized subset of survey questions whose combined response maximizes an estimation of need for a user-defined health intervention.

The manuscript begins with a detailed review of each step of the TIQS framework methodology. Once defined, we then evaluate the framework using three real-world interventions of varying size and service area drawn from a national longitudinal aging study in Ireland. For each intervention, we contextualize the observed performance with respect to the set of survey questions selected, the ability of the final question-set to provide the maximal probability estimate of need for an individual, and the ability to generate the question-set quickly. Finally we conclude with a discussion of the frameworks' future development goals and potential use cases.

2 Data

The data for this study were drawn from The Irish LongituDinal Study on Ageing (TILDA) provided by Trinity College in Dublin, Ireland [19]. The TILDA dataset was collected between October 2009 and July 2011, as part of a national study on aging conducted across the Republic of Ireland aimed at understanding the physical, social, and emotional attributes of individuals age 50 and above.

Composed of three parts, TILDA represents a unique and comprehensive data source. Each participant was asked to undergo an extensive data collection exercise, including a structured Computer Aided Personal Interview (CAPI) in their own home with a trained interviewer, an optional physical health assessment, and a "self-completion questionnaire" that includes potentially sensitive questions not asked during the CAPI [7, 44]. For this work, we focus specifically on the survey data collected.

The TILDA survey represents one of the most robust publicly available health surveys, and has been utilized in numerous academic publications. It is a large and representative data from a longitudinal study on aging in Ireland. The questions comprising the questionnaire encompass a breadth of economic (pensions, employment, living standards), health (physical, mental, service needs), and social (contact with friends and kin, social participation) characteristics of an individual's life. Together, these attributes result in a rich real-world dataset that we use to evaluate the inter-related nature of survey questions and of the complex social/physical aspects of an individual's health conditions.

In particular, this work employs a specific subset of survey questions that have been manually extracted by a team at IBM. Each unique question-answer pair in the survey was encoded as a unique entity and will simply be referred to as the "questions" to be optimized though the remainder of the manuscript. In total, 137 questions were encoded to provide a representative set across the economic, health, and social survey sections. The questions were encoded using DBPedia [23], and a small subset of Freebase and ICD10 codes were utilized as a unique identifier to map DBPedia terms to an indexed vocabulary.

In total, 8175 individuals from 6279 households participated in the study. In addition, 329 interviews were also conducted with younger spouses or partners of participants. This resulted in a final dataset of 137 questions collected across 8504 individuals.

3 Framework Details

The TIQS framework can be broken down into three district components, *identifying data*, *generating candidate questions*, and *evaluating question-sets*. Beginning with a discussion of the inputs required at runtime, detailed methodology of each component is provided in the sections to follow. Further, a visual overview of the framework can be found in Fig. 1.

3.1 Runtime Inputs

In an effort to provide the user with a high degree of flexibility, three execution parameters must be provided prior to execution. These include the number of questions to generate (requested question-set size), the individual for which the survey will be generated (target individual), and the intervention on which to assess need (target intervention).

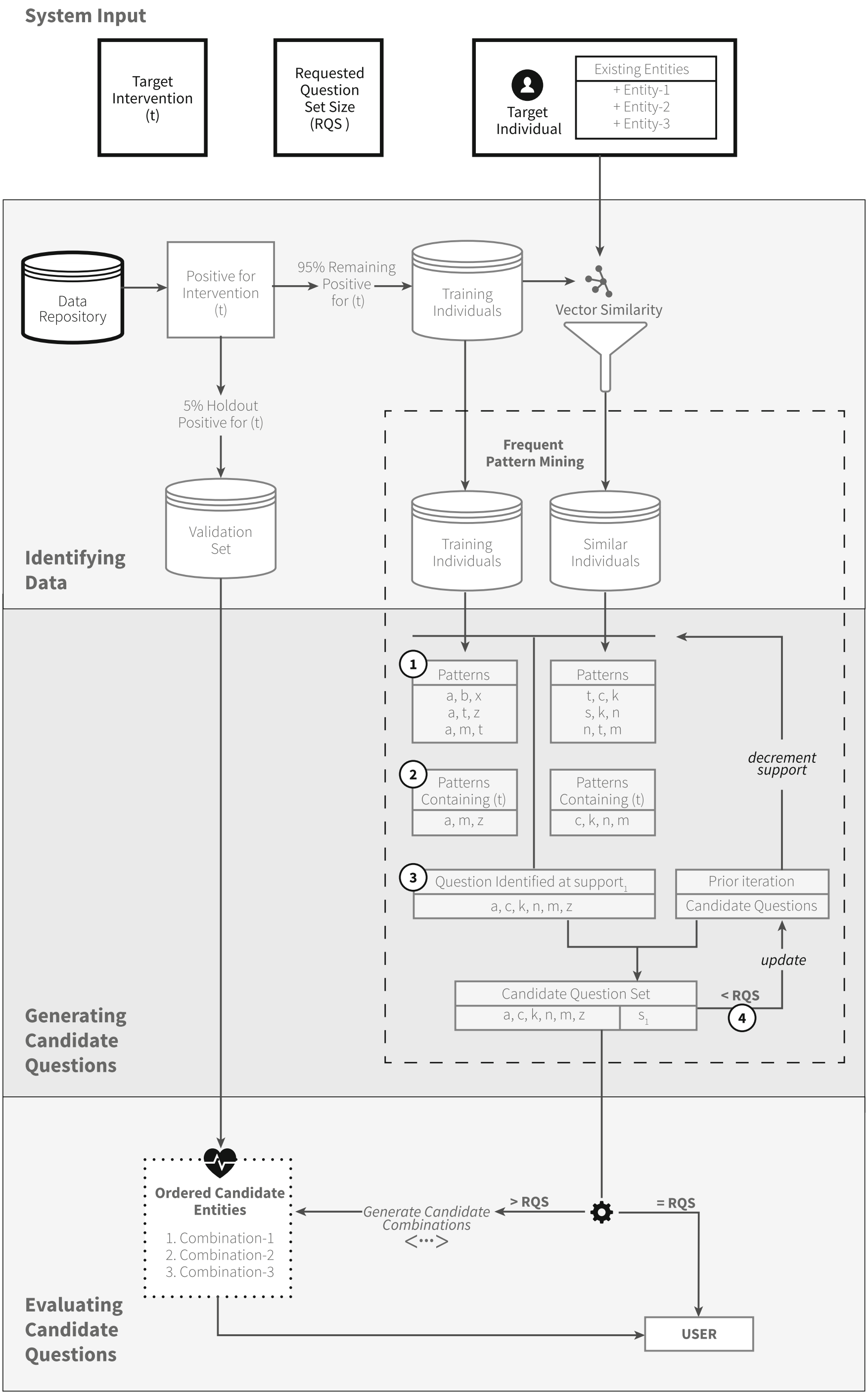


Fig. 1 Visual representation of the TIQS framework

Requested Question-Set Size Beyond minimizing response fatigue, appropriate survey length is particularly important when time is limited. As available time with each individual may vary due to a number of factors, the length of the question-set returned can be determined at runtime. This parameter may range from a single question, to the full length of the survey.

Target Individual A central component of the TIQS framework is the ability to generate personalized questions. This is achieved in large part by utilizing the target individual's historical survey responses. Although, as we will see below, the framework can be utilized even in the case where the target individual is previously unseen, and no existing survey answers are available.

Target Intervention Finally, Fowler reminds us "we are not interested in the answers for their own rake. Rather we are interested in what the answers tell us about something else" [14]. As such, the TIQS framework is designed to not only provide personalized question-sets, but also to target a specific intervention an individual may utilize based on their current health and social conditions.

3.2 Identifying Data

The TIQS framework begins by extracting all individuals whose survey history indicated the target intervention from the data repository. This filtering was done to reduce noise and computational complexity, as those individuals without the target intervention cannot be used to generate viable candidate questions about its use. A detailed examination of the process to generate these questions will be discussed in Section 3.3. Once extracted, 5% of the individuals are randomly selected as a validation set, held out to facilitate the ranking of question-sets returned to the user, while the remaining 95% are labeled as training individuals. The use of the validation set is discussed in full in Section 3.4.

Next, to aid in personalizing the final question-set, we identify a subset of individuals from the training data who are highly similar to the target individual with respect to their prior survey responses. Computing this similarity presents the challenge of comparing the survey history of two individuals. However, due to the breath of topics and specific nature of the questions in the TILDA study, this data is extremely sparse. Individuals in our study ranged from 2 to 61 question responses with a mean of 15.27, median of 14, and a standard deviation of 5.8. Although such sparsity can be expected from any survey data, it presents a notable obstacle in the computation of a metric which sufficiently differentiates individuals. To address this concern, we turn to an established technique from the information retrieval domain and identify similar individuals utilizing a vector space model.

In the context of information retrieval, vector space models formalize the relation between documents and the terms that comprise them. By projecting the terms for each document into a set feature space, basic vector operations can be applied to determine how well a document aligns with a specific query, or to another document projected into the same space [2]. Our decision to compute the similarity between individuals utilizing these models was twofold. Firstly, the implementation of the model requires only a document-term matrix that can be easily constructed from the set of survey responses. In our case, each individual represents a document, and the question responses from their survey represent the terms. Secondly, vector space models are commonly utilized on large sparse datasets and have demonstrated repeated success in the healthcare domain, deriving similarity between diagnosis histories and constructing meaningful vector mappings of clinical elements for use in statistical modeling and deep learning applications [6, 9, 30].

Employing the vector space model to compute a similarity with the target individual begins by feeding the survey histories of the training individuals through an inverse document frequency (IDF) transformation [35]. Through this transformation, we first obtain a standardized feature vector for each individual that indicate the presence or absence of every survey response, while also providing higher feature values for infrequent responses with respect to their occurrence across the training individuals. Next, the target individual is transformed into the same IDF space, dropping any responses not seen in the training data. From here, a cosine similarity is computed between the IDF vectors of the target individual and each of the training individuals. Cosine similarity was chosen as it accounts for both the matching items and their magnitude. This results in a higher ranking not only for those individuals with the largest percentage of overlapping answers, but also for those who indicated similar infrequent answers. Finally, those individuals with the top 20% of similarity scores are labeled as the "similar individuals" and are carried forward to the *generating candidate questions* step.

Unfortunately, the survey history on which this similarity is calculated is not always available. Particularly in the healthcare domain, cases where the individual has relocated, enters a new medical system, or has simply never taken the particular survey are in fact quite common. This lack of a survey history presents an issue, as the vector model constructed would be empty, in turn making the identification of similar individuals impossible. This represents a well-established phenomenon known as the cold start problem [34]. To account for this scenario, the complete set of training individuals is also carried forward to the *generating candidate question* step. Although inevitably nosier, carrying these individuals forward will allow the framework to generate question-sets even in the case where similar patients cannot be identified. Further, as an additional benefit, the complete training data may capture population-level associations that may not be well defined when looking at only the subset of similar individuals.

3.3 Generating Candidate Questions

Once the sets of training and similar individual's has been identified, the second step of the TIQS framework moves to identify the patterns of questions connected with the occurrence of the target intervention. However, it is important to note that, due to the large number of possible questions, the number of distinct combinations can quickly become intractable to evaluate. As such, we turn to a set of techniques focused on the efficient identification of item occurrence patterns, namely the use of frequent itemset generation.

Ranging from the analysis of purchasing patterns for online shoppers to similarity searches of complex structured data, frequent itemsets form the basis of numerous works that have shown success in identifying relationships within large transactional databases [17]. These techniques have also been utilized for a wide variety of applications in the healthcare domain, uncovering underlying relationships among symptoms, health conditions, diseases, and drugs [38, 47]. The work presented in this manuscript will demonstrate how such methodology can be employed to systematically identify a set of high-value "candidate questions" from which a final question-set will be created.

Formally, frequent itemsets are defined as the set of questions (Q) for which the number of individuals containing Q is equal or above a predefined support value [24]. From this definition, we note that, although the mining process is mathematically straightforward, the extraction of meaningful itemsets is highly dependent on the selection of an appropriate support parameter. While oftentimes this value is set experimentally, using techniques such as cross validation, the tuning process assumes the parameter is estimated on data representing a single underlying population. However, in the scenario addressed by this work, the target intervention is dynamic, potentially requiring a significant tuning effort for each target intervention.

To account for this variability, frequent itemsets are mined in an iterative fashion, by sweeping the parameter space with decrementing support values from 1 to 0.1, at a step size of 0.1. Beyond identifying the questions themselves, the use of support allows us to associate a measure of "strength" to a questions occurrence with the target intervention.

The process of generating candidate questions can be broken down into four primary steps, each discussed in detail below. For the reader's convenience, the numeric label associated with each step corresponds to the framework numbers found within the *generating candidate questions* section in Fig. 1.

1. Utilizing an individual's survey question responses as a "basket", frequent itemsets are mined separately on both the training and similar individuals (if applicable) using the open source pattern mining library SPMF [13].
2. The itemsets that contain the target intervention are then extracted, and the questions comprising each (excluding the target intervention) are flattened into a single set.
3. If data from both the training and similar individuals is present, the set of questions generated by each are aggregated and labeled as *candidates*. These candidate questions are then associated with the support value of the iteration. However, if a question has been labeled as a candidate at a previous iteration, it is ignored and remains associated with the prior (higher) support value.
4. If the total number of unique candidate questions generated is less than the requested question-set size, the support value is decremented, and the process is repeated. Otherwise, the mining process is terminated, and the framework continues to the next step. As a caveat, in the case of requested question-set sizes under 5, a minimum of 5 candidate questions must be generated prior to proceeding. Such a requirement was found to reduce noise and ensure a more robust ranking when returning small question-sets to the user.

It should be noted that although the problem of identifying the occurrence of a target intervention lends itself well to the rule generation problem, we chose to utilize frequent itemset as rule mining presents two sub-optimal qualities. First, it introduces an additional parameter (confidence); the automated tuning of which would require another iterative weight sweep as is done with the support parameter [1]. Second, from an implementation standpoint, the computational complexity of rule mining is significantly higher than itemset mining alone. Minimizing such complexity is especially important as health-based surveys may consist of a large number of questions and are often expected to be completed in a limited time window.

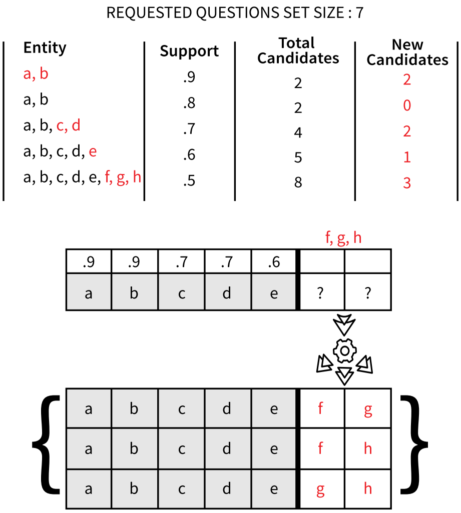
3.4 Evaluating Candidate Questions

Ultimately, the terminating condition of the candidate mining step can result in two distinct scenarios. First, the number of candidate questions is exactly the size of the required question-set. In this case, the question-set is returned to the user and the process is complete. However, as the mining process undergoes multiple iterations, and the support value is continually lowered, it is highly likely the terminating iteration will generate a set of candidate questions that exceeds the requested question-set size. In this second scenario, the TIQS framework undergoes an additional expansion process to generate and rank a set possible question-sets in order to provide the user with a single set of the requested size. This process is described in detail in the following two sections.

3.4.1 Generating Question-Sets

Although smaller than set of all survey questions, depending on the requested question-set size, and number of candidate questions generated at the final iteration, it is clear the space of possible combinations remains intractable and cannot be evaluated directly. However, to approximate the best question-set while minimizing the number of potential candidate combinations, we postulate that only those candidates associated with the lowest recorded support value need to be evaluated for inclusion, i.e., those generated at the iteration for which the candidate set exceeded the requested question-set size. The procedure for generating the list of possible question-sets for evaluation can be found bevlow, and a visual example of the process canbe found in Fig. 2.

First, all questions associated with support values higher than the final iteration are added directly to a list deemed the baseline question-set. The number of these questions in this list must be less than the requested question-set size, or the framework would have terminated at a prior iteration. The difference between the size of the baseline question-set and the requested question-set size denotes the number of "free questions" (*FQ*). Taking the set of new candidates generated at the final iteration (*C*), we compute permutations of size FQ. Each of the permutations are then appended to a copy of the baseline question-set, resulting in $\binom{FQ}{C}$ possible question-sets for evaluation. In this way, entities mined at higher support values are always included, while only those at lower support values are evaluated combinatorially for inclusion.



REQUESTED QUESTIONS SET SIZE : 7

Entity	Support	Total Candidates	New Candidates
a, b	.9	2	2
a, b	.8	2	0
a, b, c, d	.7	4	2
a, b, c, d, e	.6	5	1
a, b, c, d, e, f, g, h	.5	8	3

f, g, h

.9	.9	.7	.7	.6		
a	b	c	d	e	?	?

a	b	c	d	e	f	g
a	b	c	d	e	f	h
a	b	c	d	e	g	h

Fig. 2 Expansion example. Here FQ = 2, representing the two remaining questions for a question-set size of 7. The candidate set = 3, representing the three questions generated during the final iteration during which the question-set size was exceeded

3.4.2 Ranking Question-Sets

In order to obtain a ranking between the possible question-sets, the candidate questions comprising each question-set are used in a classification paradigm. To create the classification task, all individuals, both with and without the target intervention, are extracted from the original data repository, excluding those contained in the validation set. For each, a 137-element feature vector was constructed, encoding every question-answer pair as a unique feature. Each feature was populated with a binary value (0/1) representing if the respective question was marked within their survey history. Finally, each vector was assigned a binary class label indicating if the individual had indicated the target intervention in their survey history.

Utilizing these instances, a classifier was trained to predict need of the target intervention based on survey question responses. For the results presented within this work, we utilize the Java implementation of the AdaBoost Algorithm provided by the Weka package [15, 46]. Utilizing a set of weak classifiers, AdaBoost provides an optimal linear combination in order to create a stronger classifier, a particularly valuable quality given the sparse nature of survey data [12]. It should be noted that, within the developed framework, the learning task has been abstracted in such a way that any classifier may be substituted, provided it can provide probability estimates for each class.

Next, to compare and rank the possible question-sets, we derive a score from the validation set. For each possible question-set, the features corresponding to the candidate questions were artificially marked as present (value of 1) across each individual's feature vector. The classifier was then run to predict the class-wise probabilities for each validation instance. As all individuals in the validation set must possess the target the intervention, each question-set was associated with a score representing the mean positive-class probability. Validation instances were then "reset" to their initial question states, and the process was repeated for the next question-set. Upon completion, the highest scoring question-set was selected and returned to the user, with ties broken randomly.

For comparison, we will discuss a set of alternative candidate selection methods within the evaluation section. It should be noted that as target interventions can represent highly specialized services, it would not be uncommon for the classification to result in highly imbalanced classes. To account for this, oversampling was utilized to balance the training data classes prior to classification. Although modern sampling methods may improve the classifier performance, our aim was not to optimize the prediction task, but rather to allow for a standardized measure between possible question-sets, to identify which maximizes the classifier performance.

4 Evaluation

The evaluation of the TIQS framework was focused on three aspects of its execution. First, we looked to the frequent pattern mining utilized in the generation of candidate questions selected from the pool of survey questions. Second, we evaluated the learning process utilized to rank the possible question-sets resulting from the iterative component of question generation. Finally, we also ensured that the framework could provide question-sets within a reasonable timespan.

However, we would like to note two important caveats around missing data. First, it is important to note that that healthcare suffers from an intrinsic problem where the lack of a diagnosis does not necessarily indicate a negative class (disease/condition absent). It is very much possible the condition or disease was never formally recorded in the data. Second, with regard to community interventions, the individual may have failed to associate their participation in an activity with survey questions discussing a service, utilized a service after the survey was administered, or chose not to indicate their involvement in a service due to social or privacy concerns.

With this understanding, all evaluations in this work were treated as one-class problems, evaluating the framework's performance against those whose survey responses indicate the target intervention. Further, in an effort to approximate the circumstances in which the framework would be used, leave-one-out (LOO) testing was performed to simulate the occurrence of identifying intervention need for a specific individual. To achieve such an evaluation, the current target (test) individual in the LOO paradigm was stripped of the target intervention and the framework run on their existing survey responses. Given the knowledge the individual has already utilized the target intervention, we can accurately evaluate the frameworks ability to identify the set of questions, which, if present, would best indicate provide the maximal estimation of need with respect to the alternative possible question-sets.

4.1 Frequent Pattern Mining

Fundamentally, the strength of the proposed TIQS framework lies in its capacity to generate a set of meaningful candidate questions from which a final question-set can be constructed and returned to the user. As such, our first evaluation assessed the ability of a question-set to identify need of a target intervention. The evaluation centered on comparing the final question-set selected by TIQS with question-sets derived using two alternative methods of generating candidate questions, as well as a baseline where no additional questions were selected. An overview of the evaluation methodology and alternative candidate selection methods can found in the respective sections below.

4.1.1 Evaluation Methodology

For each alternative selection method, a question-set was created in the same manner as with the TIQS methodology. The AdaBoost classifier was run on the individuals' survey history, augmented with the question-set of the respective method, and the performance was quantified by the probability of the target intervention. Resulting from the one-class nature of our evaluation, we know individuals evaluated have utilized the target intervention. As such, a higher predicted probability of intervention need is desirable for this evaluation.

It should be noted, there exists the possibility that the performance of a selection method may vary as a function of the number of questions generated. To account for this, each of the candidate generation methods is evaluated over requested question-set sizes ranging from 1 to an upper bound of 14, which represents 10% of the total question bank.

4.1.2 Alternative Candidate Selection Methods

Our initial comparison looked to the *random* method, which, as the name suggests, populated a question-set by generating a random sampling of candidate questions from across the survey. As the survey from which the questions are drawn is designed to capture a measure of an individual's health condition, there exists the possibility that any additional information may increase the ability to estimate the probability of need for the target intervention. The comparison to random should aid in accounting for this inherent bias, when comparing to questions selected by the proposed framework.

Next, we evaluate a *ranked* entity selection method utilizing Pearson's correlation coefficient. Utilizing the complete database, again excluding the validation set, a univariate correlation coefficient was computed between each survey question and the specified target intervention. Beginning with the highest correlation, entities are then added to the target individual response set in descending order, again with ties broken randomly. The need for such an evaluation lies in the belief that, although correlation to the target intervention may constitute a component of identifying relevant questions, a number of additional factors are likely to influence an individual's particular use. It is possible that there exists a higher-order interaction between questions on the survey, and this comparison allows us to differentiate between a capturing these relations and methodology that simply identifies the set of highly correlated questions.

Finally, we evaluate the *baseline* case where no additional questions are selected. Such a baseline is important, as there exists a scenario in which any additional questions may introduce noise into an individual's survey data, negatively influencing the overall ability to estimate the probability of the target intervention as a result.

To ensure consistency, the number of questions added for each selection method is held constant, with the exception of the baseline, for which no questions were added. For the ranked addition method, if the validation individual previously indicated the selected question in their survey history, the next highest is tested until a new question is selected. Similarly, the random method continues to draw questions without replacement until one not already present in the individual's survey history is found. The process is repeated until the set of new questions reaches the size of the requested question-set.

4.2 Optimal Set Ranking

While the question-set provided to the user represents the ultimate goal of the TIQS framework, the evaluation does not assess the ranking of possible question-sets resulting from the candidate expansion process. As all question-sets are drawn from the same set of candidate questions, it remains possible that any question-set generated will provide a similar benefit or that another ranking would have proved more effective.

To address the first prospect that any combination of candidate questions provide a similar benefit, we evaluate the results of the ranking as a distance across the estimated probability for each question-set. For each target individual, the list of possible question-sets was ranked as before utilizing the external validation set. The estimated probability of the highest ranked question-set was then compared to the mean estimated probability across all remaining candidate question-sets. Values near zero indicate the estimated probability target intervention need was similar across all candidate question-sets, while large values indicate a beneficial ordering. As before, to account for variability within the requested question-set size, we evaluate this metric across requested question-set sizes from 1–14.

To evaluate the second prospect, in which an alternative ranking would have provided a more effective question-set to the user, we look to the framework methodology. As rankings are obtained through use of the validation set, we re-frame the question to investigate the scenario where the maximal ranking of the question-sets from the validation set does not generalize to that of the target individual. Thus, we perform an evaluation in which to determine if the final question-set returned in fact represented the optimal probability estimation of intervention need with respect for the particular test patient.

To achieve this, we again perform the standard ranking for all possible questions sets. However, after evaluation on the validation set, the complete list of possible candidate question-sets is returned with the corresponding rankings for each. The target individual is marked "correct" if the question-set ranked first (highest estimated probability resulting from the validation set) provided greater than or equal to the maximal estimated probability among the alternative question-sets; otherwise, it was marked incorrect. The results are reported as the average percentage of individuals for whom the highest ranked question-set would have provided the optimal estimate of need for question-set sizes 1–14.

4.3 Framework Efficiency

While our prior two evaluations focused on the question-sets generated by the framework, due to the time-sensitive nature of many healthcare applications, it is also important to ensure an individual's final question-set be generated in a timely manner. Thus, for the final analysis, we investigated the computational efficiency of the TIQS framework.

In particular, we report the total runtime of the framework's three major components: first, the identification of similar individuals through the vector space models; next, the iterative pattern mining for candidate questions; and finally, the possible question-set ranking. Although the time to identify similar individuals is clearly linear with the size of the training individual data, we include all three elements in the overall timing to ensure a robust evaluation. As the iterative pattern mining is highly dependent on the requested question-set size, the runtime is again computed across a question-set size 1–14.

5 Results

We present the results for each evaluation in the sections to follow. As the TIQS framework contains an iterative component, we evaluated the performance across target interventions of both varying size and area of service. In particular, we began with a large clinically focused intervention, the use of an optician, recorded by 515 individuals. We then considered two smaller social services: (1) a specialized program known as respite care that offers temporary institutional care of an individual, providing relief for their usual caregivers (33 individuals), and (2) the Meals on Wheels program, which was utilized by 46 individuals.

5.1 Frequent Pattern Mining

Figure 3 reports the performance results for both the TIQS framework and alternative candidate question selection methods across the set of 14 requested question-set sizes evaluated. It is important to note that, generally, any additional question information (increasing the requested question-set size) tends to provide a level improvement in performance across all of the selection methods. However, even at the largest requested question-set sizes, the TIQS framework displayed an improvement of almost 10% over the next best (ranked) selection method.

Going further, we tested the significance of this improvement. The set of probability estimates across each of the requested question-set sizes were extracted as repeated measurements for the baseline, random, and ranked methods and compared to the estimate values exacted for the TIQS framework. However, as there is no guarantee that these estimates adhere to a normal distribution, we cannot employ the

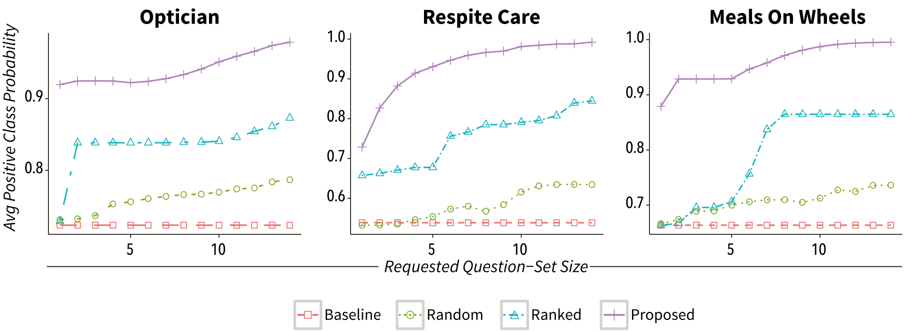


Fig. 3 Average probability of intervention need with each candidate selection method

standard paired *t* test. Rather we utilized the non-parametric Wilcoxon Sum-rank test. All three comparisons were found to be statistically significant at $p < 0.05$ for each of the three interventions tested.

5.2 Ranking Evaluation

In our second set of evaluations, we work to demonstrate if the ranking process of possible question-sets provides utility in obtaining an estimation of intervention use. The results of the initial ranking evaluation in which we compare the performance of the highest ranked question-set to the average performance of the remaining possible question-sets can be found in Fig. 4. Overall, we find an impressive performance of the ranking framework, with average distance across the evaluated interventions of 6.97%, at 6.64, 5.80, and 8.46% for the meals on wheels, optician, and respite care interventions, respectively.

Moving on, we next evaluated the performance of all question-sets with respect to the target individual. Figure 5 displays the percentage of target individuals for each intervention where the highest ranking question-set provided equal to or greater than the highest probability estimate across the total set of possible question-sets generated during the expansion step. As we can see, the ranking was quite effective, providing the optimal estimation probability for over 80% of the target individuals, across more than 90% of the evaluated requested question-set sizes.

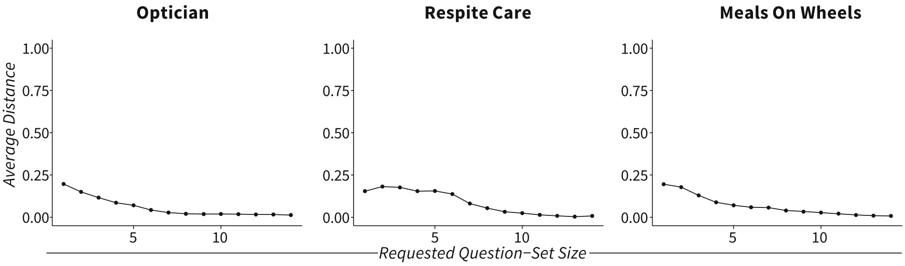


Fig. 4 Individual's average distance between optimal and lower ranked question-sets

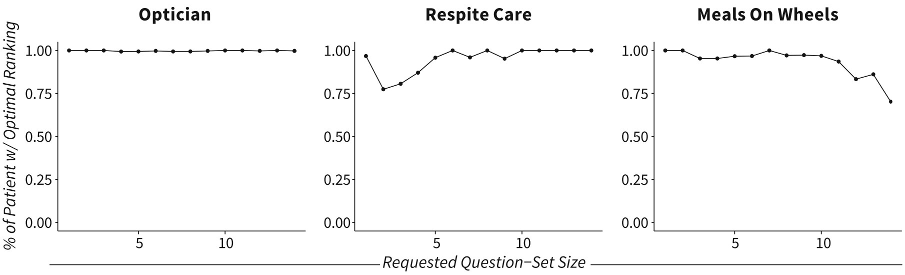


Fig. 5 Percentage of individuals for which the optimal question-set was ranked highest

5.3 Framework Efficiency

The results of the timing evaluation were calculated as the average runtime needed to generate the final question-set size for each individuals under the leave-one-out paradigm across the range of respective question-set sizes. We find the results highly encouraging, with mean runtimes of 10.10 (± 0.70), 8.84 (± 0.16), and 12.08 (± 12.30) seconds across the 14 question-set sizes of the optician, respite care, and meals on wheels interventions, respectively.

Further, the standard deviations indicate a fairly stable runtime across increasing question-set sizes, with the exception of the meals on wheels intervention. However, the large standard deviation for the meals on wheels intervention stems from a single individual, a comprehensive evaluation of whom can be found in Section 6.3 of the manuscript. Removing this individual results in a mean runtime and standard deviation in line with the other interventions at 9.07 (± 1.09) seconds.

6 Discussion

Today, population and community health initiatives have emerged on a global scale, taking aim at high-level goals such as reducing disparities in care, improving the care experience, and promoting overall health across all stages of life, all while reducing cost [3, 18]. While a lofty goal, the TIQS framework fits directly into this mission, aiding individuals from all walks of life by providing a tool to efficiently evaluate their need for a specific health interventions. Such an approach may in turn reduce costs both of survey administration time and potentially in the cost of care by allowing individuals to receive more appropriate and timely preemptive clinical and social services.

As such, we centered our discussion around three questions through which to assess the performance of each component of the TIQS framework. First, with respect to the frequent pattern mining process by which candidate questions are generated, we ask: *Do the candidate questions provide the most information with regard to the set of possible survey questions?* Turning next to the ranking of the question-sets, resulting from our iterative designation of the support parameter, we ask: *Are we providing an optimal ordering of the possible candidate question-sets to the user?*

Finally, looking to the intended use cases of the framework, we ask: *Can TIQS generate a question-set in an efficient manner?* An analysis of the evaluation results with respect to these questions can be found in the respective sections below.

6.1 Do the Candidate Questions Provide the Most Information?

Our initial evaluation aimed to establish the validity of the candidate questions, demonstrating that questions selected by the TIQS framework provided significantly higher probability estimates for the need of a specified target intervention than those selected randomly or through correlation alone.

6.1.1 Alternative Question Selection Methods

Baseline As the prediction of intervention use for a single individual represents a binary prediction task, it is unsurprising that, regardless of intervention, the performance of adding no additional questions resides at approximately 50%.

Random Looking to Fig. 3, we find that, while the estimated probability resulting from randomly selecting candidate questions is elevated in relation to the baseline, such improvement is marginal when compared to other more sophisticated selection methods.

Further, it should be noted, that in a few instances at smaller question-set sizes, the random selection method was found to provide a lower average probability estimates than the baseline. As the baseline method adds no additional information, such a case highlights the fact that, if picked without regard to the target intervention, survey questions may only serve to add noise to the identification task.

Ranked Finally, with respect to the alternate selection methods, we find that ranked candidate selection often provides the highest probability estimates. However, the performance remains well below that of the TIQS framework. This observation leads back to the notion that there exists higher-order interactions between questions when ascertaining the probability of need for a specific target intervention. Although such interactions may be captured by candidates generated by pattern mining, they would be ignored when adding candidate entities based on univariate correlation to the target alone.

A closer inspection of the ranked selection performance reveals two distinct cases: the first in which the ranked performance remains fairly stable and spikes over a short increase in question-set sizes and the second where performance gradually increases. This variability can likely be explained with the understanding that specific highly correlated features may provide valuable information with regard to the target intervention. However, not all correlations will result in the increased ability to identify use, and simply selecting a large number of questions based on their correlation alone may introduce a significant amount of noise into the system, stagnating overall performance.

6.1.2 TIQS Framework

Looking finally to the questions selected by the proposed framework, we note that appropriate selection of survey questions can provide significantly more information, producing classifier probability estimates averaging well above 90% across a range of question-set sizes. This is a notable result, as the second highest selection method (ranked selection) provides estimates of almost 10% less, averaging around 80%.

Taking a deeper look at the performance of the TIQS framework, it is worth noting that we find performance benefits even for a question set of size 1. For the random selection method, such a boost in performance is logical as not all questions drawn from the survey are relevant to the target intervention. However, for the ranked selection method, this performance difference can be explained with the understanding that the question identified based on Pearsons correlation is the same for all individuals. Yet it may not be the most useful for all individual based on their history. Rather, another question not as highly correlated in a univariate manner to the target intervention may provide a higher increase in predicted probability for the target individual, reinforcing the importance of selecting the *right* questions.

6.2 Is the Ordering of the Possible Question-Sets Effective?

Having established the value of the final question-set generated by TIQS when compared to alternative selection methods, our second question looked to quantify the framework's ability to rank the possible question-sets generated from the set of candidate questions. To do so, we look and compare the highest ranked question-set returned to the user to the complete list of possible question-sets generated during candidate expansion.

6.2.1 Distance

Looking first to the average distance for each of the evaluated interventions, we note that the highest ranked question-set provides a notable improvement in probability estimates than the alternative sets. However, it is worth noting that, as the requested question-set size is increased, the average distance metric tends to decrease. Although it would seem the ideal case represents that in which this number would remain at the highest initial point, elements of the framework design must be considered in our interpretation of this metric.

As the requested question-set size is increased, often an increasing number of iterations are required to mine frequent itemsets containing a sufficient number of candidate questions. With each iteration, the support parameter is lowered, resulting in an increasing number of number of candidate questions. As the number of candidates identified during the final iteration can become quite, the number of possible question-sets to be evaluated can also grow quite large.

Following from the assumption that the questions generated at the highest support values are the most valuable, the increased number of possible question-sets will create the scenario where many alternative question-sets contain the questions contributing most to the increase in probability estimates. This in turn presents the case where the average performance across the highest ranked question-set and the alternative question-sets appears more similar. However, even a 1% increase across thousands of possible question-sets at the largest requested question-set sizes represents a clear need to perform such a ranking after the expansion of candidate questions.

6.2.2 Optimal Ranking

Although the increase in estimated probabilities may appear low in absolute terms, the benefits of ranking represent a critical aspect in the overall effectiveness of the framework. To quantifiably establish this premise, we moved to evaluate how each of the identified question-sets would perform if selected for the target individual, removing uncertainty of the rankings on the validation set alone.

Looking to the results, we find that the ranking proves highly effective, with well over 90% of the individuals receiving optimal rankings across the three interventions, and requested question-set size from 1–14 questions. Such a result is incredibly important, as it helps to affirm the assumption that utilization of an external validation set can provide a highly transferable ranking of estimation of probability estimation for the target intervention with respect to an unseen target individual. Again reinforcing the value of the framework was proposed in this work, over simply training a generalized model for future use.

We do note a dip at the largest question-set sizes for the meals on wheels intervention. It is important to remember that generation of large number of candidate questions for interventions with such a low number of participants is extremely difficult, with question-sets receiving noise from the smaller sets in pattern mining. However, we attempt to account for this noise through our iterative generation of candidates. A result indicating that, even at the lowest point, 75% of individuals receiving optimal rankings is, in any case, highly promising.

6.3 Can TIQS Generate a Question-Set in an Efficient Manner?

With execution times averaging well under 60 seconds per individual, it is clear the TIQS framework is capable of rapidly generating a question-set to aid in the determination of an individual's need for a specific community intervention. Moreover, looking to the ability of the framework to generate variable-length question-sets, we find the performance remains stable, with less than 5 seconds differentiating the runtimes of the largest and smallest requested question-set size for each intervention. Such a result highlights the value of our iterative framework, demonstrating the utility in evaluating only those questions generated at the lowest support levels, rather than the set of all of the possible questions.

As it stands, we believe these results present a compelling use case for our framework. One in which the framework could be utilized in an ad-hoc manner to determine need for community intervention for a specific individual should it be required. However, in our evaluation we encountered one irregularity. While stable across question-set sizes 1–13 ($\mu = 8.79$, $\sigma = 0.093$), the largest question-set value in the meals on wheels evaluation jumps to an average runtime of 54.8 seconds. A closer inspection reveals this increase is the direct result of a single individual, whose computation time exceeded 20 minutes.

In an effort to understand this performance, we examined the survey history of the individual in question. Finding the responses centered on the use of three highly distinct interventions. In addition to meals on wheels, the individual highlighted the use of a dentist, with corresponding responses indicating tooth loss and the use of dentures, as well as use of optician, with responses indicating, hyperopia, cataracts, eye disease, and the use of corrective lenses.

With only 46 individuals indicating need for the meals on wheels service, the number of similar individuals was bounded to only 9 (20%). Thus, it is possible this data became highly diverse, including those individuals with similar response patterns for each of the three interventions. In turn requiring a significantly higher number of iterations to generate candidate questions targeted to the meals on wheels intervention.

Importantly, it does not appear the increase in runtime was tied to the number of questions answered. As the individual being evaluated had the same or fewer total responses than six of the others in the meals on wheels population, each of whom had runtimes of under 30 seconds. Additionally, as the individual also utilized the Optician intervention included in our evaluations, we found relevant to look into their performance at a requested question-set of size 14. Finding it required only a fraction of the time, at 27 seconds, suggesting that the anomaly in execution time for the meals on wheels intervention results as a product of the small population size, and large number of requested questions.

Although we find it extremely promising that across the 8316 total timing runs in our evaluation (594 individuals, across 14 different question-set sizes), we incur such a performance hit only once, we do plan to explore additional optimizations, in particular, assessing the performance benefits of computing question-sets without generating the intermediate candidates of each frequent itemset though efficient pattern mining structures such as FP-trees.

7 Future Work

Although, from an academic perspective, we have demonstrated the promise of utilizing personalized question-sets to identify the need of a targeted intervention, we are currently investigating two extensions to expand the practical impact of our framework.

First, we address the scenario in which the question-set derived by TIQS becomes incomplete during the administration of the survey. Such a scenario may result due to a lack of available time, or in the case where an individual does not present all of the social, behavioral, and physical factors included in the question-set. As the strength of TIQS lies in the ability to address the combinational aspect between survey questions, the lack of a specific question may decrease the probability estimates for the target intervention. To account for this, we are extending the evaluation of candidate question-sets to assess not only at the performance of each possible question-set, but also of the combinations of candidate entities that comprise them. Unfortunately, to do so would require a factorial number of evaluations for each question-set. To address this, we are currently investigating solutions utilizing parallel computing, and optimization techniques such as dynamic programming.

The second extension looks to address the temporal aspect of the data collected. As noted prior, one of the motivating factors for this work was the awareness that the dynamic nature of an individual's health often required the re-administration of these surveys. However, currently, the TIQS framework does not consider the time at which the previous surveys were administered. As such, we may be ignoring potentially valuable information around the progression of an individual's health state. Currently we are augmenting the framework to account for the time at which the risk factors were identified, allowing recent responses to contribute to the final question-set with a higher weight than those recorded in the distant past.

8 Conclusion

The landscape of healthcare is shifting, and the TIQS framework represents an example of how emerging informatics techniques can personalize and advance population-screening tools developed by the healthcare community. Drawing on an individual's survey history, the work presented in this manuscript has demonstrated how TIQS can efficiently generate a personalized set of survey to accurately estimate the probability of need for a designated community intervention. We have evaluated the frameworks' effectiveness on three community interventions of differing sizes and types, with similar results across the range of question-set sizes.

Although our work focused on applications to the healthcare domain, as no domain knowledge required to obtain the question-set, the TIQS approach has potential applications to a number of fields. Like many pattern mining techniques, the ability to identify relations from a set of likely interconnected data has the direct benefits to various user- and item-based analyses. However, TIQS takes this one-step further, providing the ability to identify items, which, if associated with a user, indicate a likely association with a user-defined item of interest. Additionally, the TIQS methodology highlights how iterative models can be utilized efficiently to improve generalizability, reducing the need to tune individual models when evaluating multiple target item.

Furthermore, the TIQS framework provides two additional benefits for general use: first, the ability to request question-sets of variable length provides the flexibility to utilize the framework in time-sensitive applications where a quick coarser result may be needed and second, in the ability to generate question-sets a priori. Allowing optimized surveys to be offered through a number of medium, including verbally, or by asynchronous medium such as email, or post. Such flexibility is critical when considering the use across a diverse set of subject populations. This is particularly true for those who may have cognitive or physical limitations, making their familiarity, comfort, and ability to interact with traditional computerized methods limited or even prohibitive.

Finally, reflecting on the many ideas which inspired this work, we look to Kilbourne et al., who posit that a "primary reason for the research-to-practice gap is the lack of a framework for implementing effective interventions in community-based organizations that maintains fidelity while maximizing transferability when the interventions are adopted across different settings" [20]. It is our hope that our work represents a step in addressing that gap, providing an efficient method for care works to estimate an individual's probability of need for targeted community interventions.

Funding Information This work was done as part of an internship at IBM Research Ireland. This work was also supported in part by the National Science Foundation (NSF) Grant IIS-1447795 at the University of Notre Dame.

Compliance with Ethical Standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Agrawal R, Srikant R et al (1994) Fast algorithms for mining association rules. In: Proceedings of the 20th international conference on very large data bases, VLDB, vol 1215, pp 487–499
2. Berry MW, Drmac Z, Jessup ER (1999) Matrices, vector spaces, and information retrieval. SIAM Rev 41(2):335–362
3. Berwick DM, Nolan TW, Whittington J (2008) The triple aim: care, health, and cost. Health Aff 27(3):759–769
4. Billings JR, Cowley S (1995) Approaches to community needs assessment: a literature review. J Adv Nurs 22(4):721–730
5. Braveman P (2011) Accumulating knowledge on the social determinants of health and infectious disease. Public Health Rep 126(3_suppl):28–30
6. Choi E, Schuetz A, Stewart WF, Sun J (2016) Medical concept representation learning from electronic health records and its application on heart failure prediction. arXiv:1602.03686
7. Cronin H, O'regan C, Finucane C, Kearney P, Kenny R (2013) Health and aging: development of the irish longitudinal study on ageing health assessment. J Am Geriatr Soc 61(s2):S269–S278
8. Dahlgren G, Whitehead M (1991) Policies and strategies to promote social equity in health. Institute for Future Studies, Stockholm
9. Davis DA, Chawla NV, Christakis NA, Barabási AL (2010) Time to care: a collaborative engine for practical disease prediction. Data Min Knowl Disc 20(3):388–415
10. de Leeuw ED (1992) Data quality in mail, telephone and face to face surveys. ERIC
11. Edelen MO, Reeve BB (2007) Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. Qual Life Res 16(1):5
12. Emanet N, Öz HR, Bayram N, Delen D (2014) A comparative analysis of machine learning methods for classification type decision problems in healthcare. Decision Analytics 1(1):6
13. Fournier-Viger P, Lin JCW, Gomariz A, Gueniche T, Soltani A, Deng Z, Lam HT (2016) The SPMF open-source data mining library version 2. In: Joint European conference on machine learning and knowledge discovery in databases. Springer, pp 36–40
14. Fowler FJ (1995) Improving survey questions: design and evaluation, vol 38. Sage, Thousand Oaks
15. Freund Y, Schapire RE (1995) A desicion-theoretic generalization of on-line learning and an application to boosting. In: European conference on computational learning theory. Springer, pp 23–37
16. Hambleton RK, Swaminathan H, Rogers HJ (1991) Fundamentals of item response theory, vol 2. Sage, Thousand Oaks
17. Han J, Cheng H, Xin D, Yan X (2007) Frequent pattern mining: current status and future directions. Data Min Knowl Disc 15(1):55–86
18. Ireland H (2013) A framework for improved health and wellbeing 2013–2025. Department of Health
19. Kenny R (2014) The Irish longitudinal study on ageing (tilda), 2009–2011. icpsr34315-v1. Ann Arbor, MI: Interuniversity Consortium for Political and Social Research [distributor], pp 07–16

20. Kilbourne AM, Neumann MS, Pincus HA, Bauer MS, Stall R (2007) Implementing evidence-based interventions in health care: application of the replicating effective programs framework. Implement Sci 2(1):42
21. Kingsbury GG, Zara AR (1989) Procedures for selecting items for computerized adaptive tests. Appl Meas Educ 2(4):359–375
22. Krosnick JA, Presser S (2010) Question and questionnaire design. Handbook of Survey Research 2(3):263–314
23. Lehmann J, Isele R, Jakob M, Jentzsch A, Kontokostas D, Mendes P, Hellmann S, Morsey M, van Kleef P, Auer S, Bizer C (2015) DBpedia—a large-scale, multilingual knowledge base extracted from Wikipedia. Semantic Web Journal 6(2):167–195
24. Leskovec J, Rajaraman A, Ullman JD (2014) Mining of massive datasets. Cambridge University Press, Cambridge
25. McFarland SG (1981) Effects of question order on survey responses. Public Opin Q 45(2):208–215
26. McGovern L, Miller G, Hughes-Cromwick P (2014) Health policy brief: the relative contribution of multiple determinants to health outcomes, health affairs, August 21
27. Merzel C, D'Afflitti J (2003) Reconsidering community-based health promotion: promise, performance, and potential. Am J Public Health 93(4):557–574
28. Opdenakker R (2006) Advantages and disadvantages of four interview techniques in qualitative research. In: Forum qualitative sozialforschung/forum: qualitative social research, vol 7
29. Pasek J, Krosnick JA (2010) Optimizing survey questionnaire design in political science: insights from psychology. In: Oxford handbook of american elections and political behavior, pp 27–50
30. Pham T, Tran T, Phung D, Venkatesh S (2017) Predicting healthcare trajectories from medical records: a deep learning approach. J Biomed Inform 69:218–229
31. Rolstad S, Adler J, Rydén A (2011) Response burden and questionnaire length: is shorter better? A review and meta-analysis. Value Health 14(8):1101–1108
32. Roussos ST, Fawcett SB (2000) A review of collaborative partnerships as a strategy for improving community health. Annu Rev Public Health 21(1):369–402
33. Sands WA, Waters BK, McBride JR (1997) Computerized adaptive testing: from inquiry to operation. American Psychological Association, Washington
34. Schein AI, Popescul A, Ungar LH, Pennock DM (2002) Methods and metrics for cold-start recommendations. In: Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval. ACM, pp 253–260
35. Sparck Jones K (1972) A statistical interpretation of term specificity and its application in retrieval. J Doc 28(1):11–21
36. Sudman S, Bradburn NM, Schwarz N (1996) Thinking about answers: the application of cognitive processes to survey methodology. Jossey-Bass, San Francisco
37. Tanur JM (1992) Questions about questions: inquiries into the cognitive bases of surveys. Russell Sage, New York
38. Tomar D, Agarwal S (2013) A survey on data mining approaches for healthcare. International Journal of Bio-Science and Bio-Technology 5(5):241–266
39. Tourangeau R, Rips LJ, Rasinski K (2000) The psychology of survey response. Cambridge University Press, Cambridge
40. Van der Linden WJ, Glas CA et al (2000) Computerized adaptive testing: theory and practice. Springer, Berlin
41. Veerkamp WJ, Berger MP (1997) Some new item selection criteria for adaptive testing. J Educ Behav Stat 22(2):203–226
42. Velentgas P, Dreyer NA, Nourjah P, Smith SR, Torchia MM et al (2013) Developing a protocol for observational comparative effectiveness research: a user's guide. GPO, Washington
43. Wainer H, Dorans NJ, Flaugher R, Green BF, Mislevy RJ (2000) Computerized adaptive testing: a primer. Routledge, Evanston
44. Whelan BJ, Savva GM (2013) Design and methodology of the irish longitudinal study on ageing. J Am Geriatr Soc 61(s2):S265–S268
45. Wilkinson RG, Marmot M (2003) Social determinants of health: the solid facts. World Health Organization
46. Witten IH, Frank E, Hall MA, Pal CJ (2016) Data mining: practical machine learning tools and techniques. Morgan Kaufmann, San Mateo
47. Yoo I, Alafaireet P, Marinov M, Pena-Hernandez K, Gopidi R, Chang JF, Hua L (2012) Data mining in healthcare and biomedicine: a survey of the literature. J Med Syst 36(4):2431–2448