Can computers measure the chronic disease burden using survey questionnaires?

The Symptomatic Diagnosis Study

Spencer L. James

A thesis submitted in partial fulfillment of the requirements for the degree of
Master of Public Health

University of Washington
2012

Committee:
Bernardo Hernandez, Chair
Emmanuela Gakidou
Abraham Flaxman
Christopher J.L. Murray

Program Authorized to Offer Degree:
Public Health – Global Health

University of Washington

Abstract

Can computers measure the chronic disease burden using survey questionnaires?
The Symptomatic Diagnosis Study

Spencer L. James

Chair of the Supervisory Committee:

Professor Bernardo Hernandez

Department of Global Health

Background
Improved information collection systems are critical for more accurately estimating the burden of different chronic conditions around the world. Current estimates are limited by the low amount of medical resources in the developing world and the lack of biometry tests for chronic conditions such as depression or arthritis. Yet, chronic conditions form a substantial part of the global disease burden. Computer-based diagnosis and estimation based on self-reported signs and symptoms ("Symptomatic Diagnosis", or SD) may be a promising method for collecting higher quality information on the chronic disease burden.

Methods
As part of the Population Health Metrics Research Consortium study, we collected nearly 1,400 questionnaires in Mexico from individuals who suffered from chronic conditions that had been diagnosed with gold standard diagnostic criteria, and individuals who did not suffer from any of

the 10 target conditions. We implemented four techniques adopted for verbal autopsy cause-of-death calculation: the Tariff Method, Simplified Symptom Pattern, Random Forest, and King-Lu Direct Estimation. We analyzed the comparative performance between these methods, and compared their performance to current epidemiological measurement techniques for select conditions.

Results and discussion

The top-performing analytical methods are capable of achieving 68% concordance with true diagnosis, and 0.826 accuracy in their ability to calculate fractions of different causes. SD is also capable of matching or outperforming the performance of current estimation techniques for conditions estimated by questionnaire-based methods.

Conclusion

Symptomatic Diagnosis is a viable method for producing more detailed estimates of the burden of chronic conditions in areas with low health information infrastructure. This technology can provide myriad benefits to the field of epidemiology, such as higher resolution prevalence data, more flexible data collection, and potentially individual diagnosis for certain conditions.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I wish to express sincere thanks to my thesis committee members Dr. Bernardo Hernandez, Dr. Emmanuela Gakidou, Dr. Abraham Flaxman, and Dr. Christopher J.L. Murray for connecting me with this project and for providing their guidance, insight, and creativity in my work on this study. I am deeply grateful to Dr. Rafael Lozano for the additional support in my research. I also wish to thank the Population Health Metrics Research Consortium Team in Mexico: Minerva Romero, Sara Gomez, and Dolores Ramirez, and Dr. Osvaldo González La Rivere, Dra. Araceli Martínez González, Dr. Miguel Ángel Martínez Guzmán, Dr. Argemiro José Genes Narr, Dr. Antonio Manrique Martin, Dr. Adrian Ramírez Alvear, Dr. Benjamín Méndez Pinto, Dr. Enrique Garduño Salvador, Dr. Rogelio Pérez Padilla, Dra. Cecilia García Sancho, Dr. Mauricio Moreno Portillo, Dr. Carlos Tena, Dra. Lucía Yáñez, Dra. Ma. Elena Medina-Mora, Dr. Lino Palacios, and Dr. Eduardo Barragán Padilla. I also thank the Secretary of Health of the Federal District in Mexico City, Dr. Armando Ahued, and the coordinator of high specialty hospitals of the Ministry of Health, Dr. Bernardo Bidart, for their help in accessing medical records needed for this study. I also very much appreciate the help and support of Summer Ohno in guiding the organization of this project. Finally, I am ever grateful to my family for their love and support.

DEDICATION

To my cohort of Post-Bachelor Fellows at the Institute for Health Metrics and Evaluation, Stephanie Ahn, Ruru Wang, Leslie Mallinger, Laura Dwyer-Lindgen, Kathryn Andrews, Lisa Rosenfeld, Megan Costa, Joseph Hoisington, and Ray Zhang: thank you for your wonderful friendship, continual inspiration, and myriad memories over the past three years.

Background

Chronic conditions form a substantial part of the global burden of disease, yet there is a severe lack of high quality methods for collecting information on their prevalence in many areas of the world. Recent studies have shown that a condition such as chronic obstructive pulmonary disease (COPD) contributed approximately 70 million disability-adjusted life years (DALYs) to the global burden of disease in 2010, while unipolar depressive disorders formed an estimated 60 million DALYs. Contrary to popular belief, chronic conditions are as much a problem in the developing world as they are in higher income areas. Cardiovascular and circulatory conditions, for example, are the 3rd leading cause of DALYs in both Mexico and France. Depression, as a second example, has a higher rate of YLDs in Mali than in Canada based on recent Global Burden of Disease research. Mexico and much of Latin America in general have seen a relative increase in their chronic disease burden in the past 20 years, with conditions such as heart disease, arthritis, and vision loss steadily increasing in terms of DALYs [1].

Despite the substantial importance of chronic conditions around the world, it continues to be difficult to collect high quality information on their prevalence, particularly in areas that lack consistent or accessible health care. In part, this is due to the inherent challenge in accurately diagnosing these conditions. While information on some infectious diseases such as HIV, malaria, and TB can collected through various types of biometric blood or sputum tests, such an equivalent does not exist for most chronic conditions. The diagnostic criteria for a condition such as COPD, for example, require medical knowledge and resources that are simply not available in many parts of the world. Moreover, even if they do exist, medical examinations as

part of a household survey to estimate prevalence of different chronic conditions would be expensive and would divert already-scarce medical personnel and resources from other areas.

The paradox that areas with high burden of chronic disease can also pose the most challenges in collecting prevalence information motivates the analytic question of interest in this study: is it possible to use self-reported signs and symptoms to diagnose chronic conditions using data-driven computational models? This concept is appealing in its flexibility. Prevalence estimates on different chronic conditions could be collected in low-resource settings if a survey instrument/questionnaire was combined with a computational algorithm to virtually "diagnose" different chronic conditions without requiring any clinical expertise. This idea has been explored, developed, and tested extensively in the area of verbal autopsy (VA) research. In VA analysis, interviewers ask family members questions about the signs and symptoms leading up to a death that occurred in the household, and computer models can be used to classify the estimated cause of death. These techniques are described in more extensive detail elsewhere [2–9], though a fundamental finding has been that computer algorithms are capable of outperforming physician-certified verbal autopsy, a result that bodes well for the idea that computers can also be used for diagnosing chronic conditions in the living.

In this study, we investigate four techniques for diagnosing chronic conditions and/or predicting the fraction of a population with a given condition using self-reported data collected in Mexico as part of the Population Health Metrics Research Consortium (PHMRC). The PHMRC Mexico Study is an offshoot of the Gates Grand Challenge 13 PHMRC Project, an international collaborative focused on developing better ways to measure health. This study strives to

develop better instruments and methods for measuring population health, particularly in resource-poor settings. The information collected will be used to improve strategies for population health measurement and produce instruments that are science-based, standardized, and widely-applicable across different resource-poor settings. It will additionally enable policymakers and researchers to help address persistent inequities in health outcomes in both the developed and the developing world. We term the technique explored in this study "symptomatic diagnosis" (SD), which is intended to refer to the general practice of using computation algorithms to diagnose health conditions based on responses to a questionnaire.

Methods

Overview of study design

The symptomatic diagnosis study consisted of two main components: data collection and model validation. These components are described in more detail below. A flowchart showing the data collection component of the symptomatic diagnosis study plan is provided in Figure 1. Another flowchart describing the model validation component is provided in Figure 2.

Questionnaire

The SD study developed a questionnaire that focused on 10 priority conditions: Angina Pectoris, Rheumatoid Arthritis, Cataracts, Asthma, COPD, Symptomatic Cirrhosis, Vision Loss, Hearing Loss, Depression, and Osteoarthritis. These focus causes were chosen since they contribute considerably to the burden of disease in Mexico, and because current methods for collecting prevalence data on these conditions is expensive and time-consuming. This questionnaire also

collected socio-demographic information. This questionnaire was adapted from the World Health Survey [10] and the PHMRC Household Survey [11]. Information on the signs and symptoms of the respondent is collected, but the questionnaire also asks questions that relate to the respondent's experience, if any, with health care providers. These questions ask about whether the respondent has ever been diagnosed with different conditions, and whether certain medical procedures or protocols have occurred. These questions are tremendously useful for diagnostics from a modeling standpoint, but they also imply that the respondent has had access to the health care system. Since one of the main purposes of the SD study is to develop methods that can estimate prevalence of chronic conditions in low-resource areas without health care, it is important to validate each potential method both with and without the use of health care experience ("HCE") information. This analysis is described in more detail in the Results section. The list of items in the questionnaire and an HCE indicator is provided in Table 1.

Data

Building the dataset used to develop and test symptomatic diagnosis (SD) methods involved three main components: identifying cases for the 10 conditions of interest, identifying controls who did not suffer from any of the chronic conditions, and then implementing the SD questionnaire at the household of each case and each control.

Cases

A team of trained coders located a total of approximately 1,200 cases (120 of each of the morbid conditions under study) in 11 public hospitals in the Mexico City area (Hospital de

Especialidades Belisario Domínguez, Hospital General Dr. Enrique Cabrera, Hospital General Balbuena, Hospital General Gregorio Salas, Hospital Juárez de México, Hospital General Dr. José G Parres, Hospital General Iztapalapa, Hospital General La Villa, Instituto Nacional de Enfermedades Respiratorias, Hospital General Dr. Manuel Gea González, Clínica de Detección y Diagnóstico Automatizado (CLIDDA, ISSSTE)) and 120 cases from the collaborating psychiatric hospital Instituto Nacional de Psiquiatría (for cases of depression).

For each condition, a case was defined to be a patient that a physician had diagnosed with the condition and meets a specific set of gold standard criteria. A gold standard diagnosis refers to diagnosis of a specific disease with the highest level of accuracy possible. This involves checking that the diagnosis is based on positive results from a laboratory test or appropriate cabinet and/or the recording and documentation of appropriate symptoms of the disease were observed during the development of clinical records. To be acceptable, the symptoms of the disease should be observed or documented in a medical record by a physician. The gold standard criteria for each condition are provided below.

Angina pectoris

Gold standard cases for angina pectoris have chest pain when doing physical exertion or when they feel strong emotions. The pain is relieved by rest and must be determined by one of the first three tests and a chest radiograph:

- Test 1: Resting electrocardiogram: QRS-segment deviation, Q waves and ST segment changes and T waves, or

- Test 2: Electrocardiogram with exercise or stress test: QRS segment deviation, Q waves, and ST segment changes and T waves, or
- Test 3: Resting electrocardiogram during chest pain: QRS segment deviation, Q waves and ST segment changes and T waves
- Chest x-ray in patients with signs or symptoms of congestive heart failure, valve disease, heart disease, pericardial disease, or aortic dissection/aneurysm.

Cataract

Gold standard cataract cases require the opacity of the lens to be confirmed by an slit lamp examination. It can also include cataract with retinopathy.

Symptomatic Cirrhosis

The gold standard cases were required to have four of the following results of liver function tests:

- Anemia (detected on a complete blood count)
- Abnormalities of coagulation
- Elevated liver enzymes
- Elevated bilirubin
- Low serum albumin
- Enlarged liver (seen on an abdominal radiograph)
- ALT (alanine aminotransferase) > 2.1

Test	**Men**	**Women**
ALT (IU/L)	10 – 40	8 – 35

- AST (aspartate aminotransferase) > 2

Test	**Men**	**Women**	**Children**
AST (IU/L)	20 – 40	15 – 30	Newborn: 25 – 75 Baby: 15 – 60

- ALP (alkaline phosphatase)

Test	Men	Women
ALP (IU/L)	50 – 120	50 – 120

- Prothrombin

Test	Men	Women
Prothrombin time (seconds)	> 30	> 30

- Albumin

Test	Men	Women	Children
Albumin (g/dL)	1-35 years: 3.5-4.8	1-35 years: 3.5-4.8	0-1 years: 2.9 – 5.5
Albumin: globulin > 1	Reduced in > 40 years	Reduced in > 40 years	

- GGT(gamma-glutamyl transpepsidase) > 2

Test	Men	Women
GGT (IU/L)	2 – 30	1 – 24

Decrease or loss of hearing

Gold standard cases must meet one of the first two test-based requirements, and then undergo a hearing test:

1. The inability to hear a whisper, normal speech, and the ticking of a clock, or
2. The inability to hear a tuning fork through air and hear the pitch of the bone.

In a hearing test with detailed audiometry, patient must not be able to hear tones from 250 Hz - 8,000 Hz at 25 dB or lower.

Osteoarthritis

Gold standard cases for osteoarthritis must have all of the following elements:

- Pain in the large joints (particularly knees),
- Swelling and limited movement,

- Radiography showing loss of joint cartilage, narrowing of joint space between adjacent bones, the formation of bone spurs, and decreased joint space and spicules.

Rheumatoid arthritis

Gold standard cases for rheumatoid arthritis must have four of the following:

- Stiffness in the morning,
- Arthritis in three or more joint areas,
- Arthritis in the joints of the hands,
- Symmetrical arthritis,
- Rheumatoid nodules,
- Blood serum rheumatoid factor,
- Radiographic changes typical of arthritis.

Asthma

The gold standard cases for asthma must have spirometry showing reduced forced expiratory volume in 1 second (FEV1), reduced ratio of FEV1 to forced vital capacity (FVC), and reduced peak expiratory flow (PEF).

Chronic obstructive pulmonary disease (COPD)

Gold standard cases for COPD required all of the following:

- Spirometry: FEV1/FVC < 0.7
- Spirometry: FEV1 ≥ 80% predicted

Decrease or loss of visual acuity

Gold standard cases must have a visual acuity test with one of the following results:

1. Mild visual disability (farsightedness):
   Visual acuity in the eye to see better than 6/10 to 6/18 (20/32 to 20/63 inclusive)
2. Moderate visual disability (farsightedness):
   Visual acuity in the eye to see better than 6/24 to 6/48 (20/80 to 20/160 inclusive)
3. Severe visual disability (farsightedness):
   Visual acuity in the eye to see better than 6/60 to 3/60 (20/200 to 20/400 inclusive)

4. Visual impairment deep (farsightedness):
   Visual acuity in the better eye to see 2/60 (= 20/500 to 20/1000 inclusive)
5. Near to blindness (farsightedness):
   Visual acuity in the better eye to see 1/60 or worse (= worse than 20/1000 inclusive)
6. Near visual disabilities:
   Near binocular visual acuity worse than 6/10 (= 20/32) and distant visual acuity in the eye to see better than 6/7.5 (20/25) or better

Depression

The study protocol for depression cases in the SD study followed a different format than the other conditions. Upon a patient's arrival to the Instituto Nacional de Psiquiatría, a third-year psychiatry resident applied the Mini-International Neuropsychiatric Interview (MINI) questionnaire and diagnosed cases as depressed using the Diagnostic and Statistical Manual of Mental Disorders, 4th edition (DSM-IV) criteria [12]. Gold standard cases were required to meet the following criteria:

- Criterion A: The presence of at least five of the following symptoms for at least two weeks. One of the symptoms must be the first (sad mood or anhedonia):
  - Sad mood, dysphoric or irritable most of the day, almost every day, as reported by the individual,
  - Anhedonia or diminished ability to enjoy or show interest and/or pleasure in usual activities, most of the day, almost every day,
  - Decrease or increase in weight or appetite almost every day,
  - Insomnia or hypersomnia, almost every day,
  - Psychomotor agitation or retardation almost every day (big enough to be observed by others, not just feelings of agitation or retardation),
  - Asthenia, almost every day,
  - Recurrent feelings of worthlessness or guilt, almost every day (not only self-blaming for the fact of being sick),
  - Decreased ability or intellectual ability, almost every day (either a subjective attribution or an observation by another person)
  - Recurrent thoughts of death or suicidal (not only fear of death), recurrent suicidal thoughts without a specific plan or suicide attempt.
- Criterion B: There are no signs for mixed affective disorders (manic and depressive), schizoaffective disorder, or schizophrenia disorders.

- Criterion C: The symptoms have a negative impact on the social, occupational, or other vital areas of the patient.
- Criterion D: The symptoms are not explained by the consumption of toxic substances or drugs, or by another illness.
- Criterion E: The symptoms are not explained by a reaction of grief at the loss of an important person in the patient's life, the symptoms remain for more than 2 months or indicate functional inability, morbid inutility concerns, suicidal ideation, psychotic symptoms or psychomotor retardation.

Interviews

We included only cases living in Mexico City, and which had an address that was identifiable through the hospital records. Once the cases were identified, an interviewer from the same team that administered the SD questionnaire to the controls visited each household to administer a SD questionnaire to the cases. The informed consent letter obtained prior to the interview is provided in Annex File 1. The project was approved by the institutional review board of the University of Washington and by the research, ethics, and biosafety committees of the National Institute of Public Health and participant institutions.

Controls

We located a population of controls from the records of the Automated Detection and Diagnosis Clinic (CLIDDA) in Mexico City. CLIDDA performs a battery of diagnostic tests on people who are affiliated with the ISSSTE insurance scheme (ISSSTE is the insurance scheme for government employees in Mexico.). We defined a control to be someone who attended the CLIDDA in the last 6 months prior to the data collection, diagnosed not to have any history of the morbid conditions being studied, in the sex and same age range as the individuals in the sample of cases, living in the urban area of Mexico City and whose address was locatable from the CLIDDA records. The person must not have any obvious other disease. We identified a

sample of 240 controls. Once the controls were identified by trained coders from the CLIDDA database, appointments were made, and one visited each household to administer an SD questionnaire to individuals from the control group.

Processing

The SD dataset was processed into a format usable by statistical models using the same protocol as described in the PHMRC VA study [9]. Specifically, the duration or continuous survey items are converted to a dichotomous "long duration" item using a median absolute deviation (MAD) estimator, where the item is considered to be endorsed if it is greater than the long duration cutoff. The cutoffs for each continuous item are provided in Table 2. Categorical items are expanded into being separate dichotomous items for each level or category. For the purpose of clarity, the term "feature" will be used to refer to the dichotomized (endorsed versus not endorsed) items or information used by the model/estimation process, while the term "cause" will be used to refer to "condition" or "illness" or to healthy controls.

Natural language processing

The symptomatic diagnosis dataset is composed almost entirely of structured (ie dichotomous, continuous, or categorical) questionnaire items, but free response and text transcription items also provide a data-rich unstructured component of the instrument. However, since the responses to these items are not classified in a dichotomous, categorical, or continuous format, we implemented further techniques to capture the "free response" information. There were two survey items that held free response information. Instrument question SD6.4 asked the interviewer to transcribe text found on any drug containers in the household. Example

responses to this item read, “composed chlorphenamine” and “metoprolol”. The second free response item essentially asked the interviewer to write down any other pertinent information about the interview that he/she felt was useful. An example of this response was, “Mrs. (name) says that she has some problems to hear from her left ear and the light bothers her a little and she wears glasses with lenses that don’t have a strong prescription.”

Ongoing research in text mining and natural language processing has examined their value in data classification and machine learning techniques [13–17]. In this study, we were interested in identifying text signals that held some diagnostic value (for example, the word “alcohol” is more useful for diagnostics than “suffers”, because “suffers” is generic for many diseases but “alcohol” likely is more associated select diseases), and then in “tokenizing” the free text into data features that could be used by computational algorithms. Tokenization refers to the process of 1) identifying that a free response item in a given interview includes some target word such as “albuterol”, and 2) marking or “endorsing” a dichotomous item that was created for that word. For the text feature “alcohol”, an interview would have a 1 if that word appears in the free response section and a 0 if it did not.

One challenge with implementing text mining or natural language processing (NLP) techniques is that some words or expressions are essentially synonymous for data classification purposes. For example, “alcohol”, “alcoholism”, and “alcoholic” can hold different meanings in the context of a medical interview. However, it seems reasonable to assume that the signal-to-noise ratio will be improved if we treat the root of the word (in this case, “alcohol”) as the actual text feature instead of the entire word itself. Thus, any word with “alcohol” contained

within the word would be replaced with “alcohol”. This process is known as stemming. Similarly, misspellings, mistranslations, or variations in medical terminology may disperse useful signals across too many individual text features to be singularly useful. For example, the abbreviations “PNC-E” (post-necrotic cirrhosis secondary to alcohol) and “MNLC” (macronodular liver cirrhosis) refer to different medical characteristics, but the signal-to-noise ratio is again assumed to be better if the diagnostic model treats these terms equivalently. To this end, we utilized the dictionary developed for verbal autopsy analysis that mapped roughly-synonymous words to a single text feature. This dictionary is presented in Annex File 2. Note that the dictionary incorporates a number of terms that are not likely to be a part of an SD interview, though their inclusion should not adversely affect the dictionary’s value.

We utilized the TM package in R [18], which allows the user to use a built in stemming function and to set a threshold for the number of times a text feature must appear in the data in order for it to be tokenized. We used a threshold of 4 for our analyses, and additionally implemented the removeWords (to remove stop words such as “the”), removePunctuation, stripWhitespace, removeNumbers, and stemDocument functions.

Train-test environment

The goal of SD is to predict chronic conditions based on the questionnaire responses. A critical component of developing and validating data classification models is constructing an appropriate validation environment. A given model must be “trained” on a randomly-selected portion of the dataset and then “tested” on an uncontaminated separate portion of the dataset. This separation ensures that the predictive validity is calculated in an out-of-sample

environment, and that the models do not overfit the data and produce deceptively high performance.

The train-test environment used in previous studies of this nature split the entire dataset into 75% train data and 25% test data, where the components are sampled by the outcome variable (in this case, the chronic condition condition). This train-test split is repeated 500 times to conduct 500 simulations and to estimate uncertainty around the predictive validity estimates.

The 25%-75% test-train split results in test and train sets with roughly equivalent cause compositions (e.g. if 10% of the subjects in the test split have COPD, then roughly 10% of the subjects in the train split will also have COPD). Previous research in verbal autopsy has shown that 1) predictive validity is artificially enhanced when test and train composition are similar [3] and 2) the estimated performance of a method is largely a function of the cause composition of the test dataset [19]. Murray et al. further show that an effective method for conducting sensitivity "stress" tests is to deliberately vary the composition of the test data by resampling with replacement based on an uninformative Dirichlet distribution. The Dirichlet distribution is a continuous probability distribution of a multinomial outcome variable $x_1, \ldots, x_K$, where $x_i \in [0,1]$ and $\sum x_i = 1$. That is, each cause fraction is randomly varied between 0 and 1 but still sum to 100%. This results in a test data composition that is different than the train data composition and also varied across the 500 splits of test data.

Models

Four data-based models for verbal autopsy classification were tested and validated as part of the Population Health Metrics Research Consortium (PHMRC) study. Due to the high level of

analogy between verbal autopsy and symptomatic diagnosis, we were able to adapt each of these models for use in the SD analysis. Three of the models – Tariff, Simplified Symptom Pattern (SSP), and Random Forest (RF) – are capable of diagnosing individual subjects and estimating cause fractions, while the King-Lu (KL) algorithm can only estimate cause fractions. In the PHMRC verbal autopsy study, each model was shown to have different strengths and weaknesses. These models are described in more detail below.

Tariff

Tariff is a simple additive algorithm that uses the calculation of a "tariff" for each cause-feature combination followed by a summation and ranking function to predict the most likely causes for each subject in the test dataset. The tariff for a given cause-feature combination quantifies how uniquely and strongly predictive a given data feature is for a given cause. The tariff for cause *i* and feature *j* is calculated as:

$$Tariff_{ij} = \frac{x_{ij} - Median(x_j)}{Interquartile\ Range\ x_j}$$

where $tariff_{ij}$ is the tariff for cause i, feature j, x_{ij} is the fraction of subjects with cause i for which there is a positive response for item j, median(x_j) is the median fraction with a positive response for feature j across all causes, interquartile range x_j is the interquartile range of positive response rates for feature j averaged across causes.

For each subject in the SD dataset, we compute summed tariff scores for each cause:

$$Tariff\ Score_{ki} = \sum_{j=1}^{w} Tariff_{ij} x_{jk}$$

The tariff scores for each cause are ranked across all subjects, and the top-ranked cause for each subject is assigned as the diagnosis for that subject. The tariffs for each feature-cause combination are provided in Annex File 3. Blank values indicate that the tariff for that feature-cause was not significantly different than 0.

Simplified Symptom Pattern

SSP uses Bayes Theorem to calculate a posterior probability of each cause for each subject in the test data after conditional probabilities of features given cause are calculated in the training data:

$$P(D_i = j|S_i) = \frac{P(S_i|D_i=j)P(D_i=j)}{\sum_{j'=1}^{J} P(S_i|D_i=j')P(D_i=j')}$$

Where S_i is the response pattern on a set of *k* features in the subject's responses (not simply one feature), and where $P(D_i = j|S_i)$ is the probability of individual i having cause j, conditional on the observed vector of feature responses, S_i . The various options and specifications for SSP are described in more detail elsewhere [8], and for implementation in the SD environment we opted to use the same specifications as Murray et al. in their verbal autopsy analysis.

The output from SSP implementation is a posterior probability of each cause for each subject. The diagnosis is made based on the highest posterior for each subject.

Random Forest

Machine learning (ML) is a discipline in computer science that seeks to use computer-automated methods to make predictions based on examples. Random Forest (RF) is a machine learning classifier that uses large number of decision trees to classify data in a test dataset after the trees are built with the training data set. Random Forest has consistently proven itself as one of the strongest performing machine learning classifiers, and in verbal autopsy analysis has been demonstrated to outperform Tariff, SSP, and physician-certified verbal autopsy [RF citation]. The RF approach developed by Flaxman et al. [2] for verbal autopsy analysis also implemented a pairwise coupling procedure, where each cause is essentially scored against each other cause as opposed to against all other causes. For example, a given decision tree will contain features that result in a subject being voted as either COPD or cirrhosis, as opposed to being voted as either COPD or not COPD. We adopted the VA specifications from Flaxman et al. for our implementation of RF on symptomatic diagnosis data. This process is described in more detail elsewhere [2].

King-Lu

The King-Lu (KL) algorithm is a statistical model described in more detail elsewhere [3, 20] that directly estimates cause-specific mortality fractions for verbal autopsy using “hospital” data to train the algorithm and “community” data as the test equivalent. Due to the similarity of verbal autopsy and symptomatic diagnosis, the assumptions and model structure of the KL approach should be applicable to SD data. Furthermore, it was observed in research by Vahdatpour et al. that KL performed more strongly when the cause list was shorter; for example, in the PHMRC

study, KL achieved higher performance in the neonate VA module which had 11 causes than in the child module which had 21 causes. The main weakness of the KL method is that it does not make diagnoses for each subject in the dataset, but rather estimates the fraction due to each cause directly. In our implementation of the KL algorithm, we used the parameters of 12 feature clusters and 400 iterations per split, as suggested by the documentation.

Performance metrics

We assessed and compared the capability of the different SD models using two verbal autopsy performance metrics described by Murray et al. [19]. Symptomatic diagnosis is capable of predicting whether or not an individual suffers from different chronic conditions and estimating the fraction of individuals in a population who suffer from a given condition, and consequently the performance of each method should be quantified in both of these domains.

Chance-corrected concordance

As described by Murray et al. [19], chance-corrected concordance is a kappa-like measure of a method's ability to correctly diagnose a condition in an individual. For example, if a population has 100 people who truly have COPD, and a model correctly diagnoses 70 of them as having COPD, then it achieves concordance with the true cause 70% of the time. However, because random assignment of n different causes would be correct 1/N times, this metric must also be adjusted for random chance.

The formal calculation of chance-corrected concordance for cause j (CCCj) is:

$$CCC_j = \frac{\left(\frac{TP_j}{TP_j + FN_j}\right) - \left(\frac{1}{N}\right)}{1 - \left(\frac{1}{N}\right)}$$

where TP is true positives, FN is false negatives and N is the number of causes (11 for symptomatic diagnosis).

CSMF accuracy

Murray et al. also critique the use of absolute error or relative error in verbal autopsy as a stand-alone metric for a VA method's ability to estimate cause-specific mortality fractions (CSMF), since the measured performance of a method using these metrics will be largely dependent on the cause composition of the test dataset. They propose an alternative metric, CSMF accuracy, which is capable of generalizing the performance of different methods in their CSMF estimation regardless of the number of causes. CSMF accuracy, which is an aggregate measure across all causes k, is formally defined as:

$$CSMFAccuracy = 1 - \frac{\sum_{j=1}^{k} \left| CSMF_j^t - CSMF_j^{\Pr ed} \right|}{2(1 - Min(CSMF_j^t))}$$

Where the superscript for CSMF refers to true ("t") or predicted ("pred") cause fractions. The denominator reflects the maximum possible CSMF error in the given test split:

$$CSMF\ Maximum\ Error = 2\left(1 - Minimum(CSMF_j^{true})\right)$$

Hence, CSMF accuracy can be described as 1 minus the sum of absolute errors divided by the maximum error. A CSMF accuracy of 1 would indicate perfect cause fraction predictions, while 0 would indicate the worst possible model. We note that we use the term "CSMF accuracy" for evaluating symptomatic diagnosis methods despite the SD cause fractions measuring prevalence, not mortality. Despite the weaknesses of using absolute error as a performance metric, we also report these results since it allows for performance evaluation for specific causes.

Analysis details

Stata, R, and Python were used for all analysis and data management. All data and code are available from the authors upon request.

Results

Symptomatic diagnosis interviews

The number of SD interviews conducted for each of the 10 conditions and for controls are provided in Table 3. This table also provides the age and sex distribution by condition, and shows that the project gathered approximately the target number of interviews or more for each condition, and that the target number of controls was also reached.

Methods validation

We assessed the performance of each method in terms of the two metrics described above, chance-corrected concordance (CCC) and CSMF accuracy, and in terms of cause fraction

absolute error, which allows for inspection of each method's performance for a given cause. Each level of validation was conducted across 500 splits of data as described in the Methods section. We tested each method under two conditions: with all data features and with all data features excluding HCE information. The rationale for this testing environment is that it is important to analyze how SD methods perform in areas where the respondents are unlikely to have access to health care or to have been clinically diagnosed with one of the target conditions.

Individual diagnosis

After a cause assignment is made using one of the methods described above, it is marked as correct if the estimated cause matches the true cause. Chance-corrected concordance for each cause is calculated using the methods described above, and then the median CCC across 500 splits is calculated for each cause, and then also the average across the 11 causes. As in VA analysis, there tends to be considerable variation in the performance of different methods in the individual diagnosis predictive validity. For example, depression has high CCC in each of the four methods, regardless of whether HCE information is used, while vision loss and osteoarthritis tend to experience considerably lower performance. Some causes, such as rheumatoid arthritis, are more greatly affected than others by the inclusion of HCE information. This is likely caused by the nature of the signs and symptoms associated with different causes. Diagnosing depression or angina pectoris from questionnaire responses is likely an easier exercise for a computational model because the signs and symptoms that the model use are not entirely different than the questions a doctor or psychiatrist would ask to diagnose these

conditions. This question is explored in more detail in subsequent sections. Table 4 provides the mean chance-corrected concordance across causes for each method across 500 splits, with and without HCE. The KL method is not shown since it does not make individual-level cause assignments. Table 5 provides the median chance-corrected concordance for each cause across 500 splits, with and without HCE. Figure 3 shows the median chance-corrected concordance for each cause in the SD dataset across 500 splits, with and without HCE. Overall, we found that the Simplified Symptom Pattern method achieved the highest mean CCC, though as discussed above we observed a cause-specific performance hierarchy that varied from method to method.

Population cause fraction estimation

We used each method to calculate the estimated and true cause fractions for each test split of data. These true and estimated cause fractions were used to calculate absolute errors and CSMF accuracy across 500 splits. We found that the Tariff method with HCE achieved the highest CSMF accuracy (though the point estimate for CSMF accuracy for Tariff with HCE was within the 95% confidence interval for SSP with HCE, but not the other way around), while KL with no HCE resulted in the lowest CSMF accuracy, though similarly the point estimate for RF without HCE was within this confidence interval. Table 6 provides the median CSMF accuracy for each method across 500 splits, with and without HCE information.

To inspect the performance of each method for each specific cause, we calculated the absolute error (absolute value of the true minus estimated CSMF) for each split, for each cause, and then determined the median absolute error across 500 splits. For cause-by-cause inspection, Table 7

shows the median CSMF absolute error for each cause for each method. The table is displayed as a heat map, where darker red colors indicate greater error and darker green colors indicate lower error. This table shows that in general, SSP and Tariff produce the most accurate cause fractions, though the RF method appears to better estimate the cause fractions for depression among the three methods capable of also making individual cause assignment. The figure also shows how there is at least one high-performance method for each cause. This issue is discussed in more detail in the discussion. Figure 4 shows the same information expressed as a bar chart.

We also analyzed whether SD methods systematically over- or underestimate the prevalence fractions. Using the true and estimated prevalence fractions from 500 splits, we conducted linear regressions where the estimated prevalence fraction was a function of true. This analysis quantifies any systematic bias in a given method for a given cause, and also suggests a way to correct prevalence fractions using this bias. The results (coefficient, intercept, RMSE, and R-squared value) from this analysis are provided in Table 8. An illustration of this analysis for angina pectoris from the Tariff Method with HCE is provided in Figure 5. This figure and associated coefficient and intercept illustrate how a computational method can estimate a cause fraction when there are actually 0 cases, but how the Tariff Method for this cause tends to generally slightly underestimate the prevalence of angina pectoris, except for at very low true prevalence fractions. In contrast, the equivalent scatterplot in Figure 6 for hearing loss shows more overestimation when the true prevalence fraction is 0 but a general systematic underestimation for larger prevalence fractions.

Cross-classification and cause aggregation

We found that most methods achieve high CCC for the causes of angina pectoris, depression, and cirrhosis. However, vision loss and osteoarthritis experienced lower performance. With vision loss, we suspected that there may be some cross-classification with cataracts due to the similar clinical presentation of these conditions. We investigated this question using cross-classification or "confusion" matrices. A confusion matrix shows the frequency of different cause assignments for each true cause in a given test split. An example of a confusion matrix for a single split (prior to undergoing the Dirichlet-based resampling) is shown in Table 9. This confusion matrix shows how 8 out of 24 true vision loss cases were correctly classified as vision loss but 8 were misclassified as cataracts. Out of 27 true cataracts cases, 10 were correctly classified as cataracts but 5 were misclassified as vision loss.

This investigation of cause assignments indicated that there could be considerable cross-classification but also that the features most strongly associated with vision loss or cataracts had a much weaker association than other feature-cause combinations. For example, the highest tariff feature for vision loss had a tariff of 15.5, while the highest tariff for cirrhosis (which had high performance) had a tariff of 60. With osteoarthritis, we also observed cross-classification error where osteoarthritis was being classified as rheumatoid arthritis and vice versa. This suggests the possibility of increase performance by combining similar cause. That is, instead of differentiating between vision loss and cataracts, a prediction of vision loss OR cataracts would become a prediction for the combined category of "vision loss or cataracts." We measured the effect of this aggregation on performance using the Tariff Method and found

that overall chance-corrected concordance increased by approximately 3% in absolute terms and that CSMF accuracy increased by .032 in absolute terms. The chance-corrected concordance for the 9-cause aggregate list and the mean chance-corrected concordance with and without HCE across causes are provided in Table 10. The cause fraction absolute errors and the CSMF accuracy with and without HCE are provided in Table 11.

At the cause-specific level with the inclusion of HCE, the combined arthritis category achieved a chance-corrected concordance of 0.703 (compare to .358 for osteoarthritis and 0.664 for rheumatoid arthritis), while the combined vision loss or cataracts category had a chance-corrected concordance of 0.574 (compare to 0.400 for cataracts and 0.391 for vision loss). In terms of prevalence fraction estimation, the combined arthritis category had an absolute error of 0.029 (compare to 0.029 for osteoarthritis and 0.017 for rheumatoid arthritis), and the combined vision loss or cataracts category had an absolute error of 0.036 (compare to 0.027 for cataracts and 0.038 for vision loss). While the aggregation conferred a clear advantage in chance-corrected concordance, its effect on prevalence fraction estimation was not as dramatic. Future users of SD may have different analytical or research tasks at hand. It may be important in some arenas to differentiate between rheumatoid arthritis and osteoarthritis, while in other applications the user may be more interested in measuring overall vision problems and would want to aggregate vision loss and cataracts. As a result, we opted to report all results at the more granular 11-cause level instead of for the aggregated cause list, though the option to estimate for 9 causes is retained in the available analysis code in case future users

would prefer to retain higher chance-corrected concordance at the expense of higher resolution prediction.

Comparison to other information systems

In order to investigate the prospect of computational methods being a viable option for estimating prevalence in areas with low chronic condition prevalence information, we reviewed the approaches that other studies have used to collect such information. Each of the 10 conditions (Angina Pectoris, Rheumatoid Arthritis, Cataracts, Asthma, COPD, Symptomatic Cirrhosis, Vision Loss, Hearing Loss, Depression, and Osteoarthritis) has been estimated in the Global Burden of Disease study. We reviewed the data that was used to inform each of these estimates and investigated the sensitivity of the diagnostic or data collection techniques actually used in the field.

Asthma

Asthma prevalence and incidence estimates for epidemiological modeling can be measured via self-reported signs, symptoms, or reported diagnosis in population-representative surveys such as the World Health Survey, or by physician “current diagnosis” or “ever diagnosis” estimates in literature studies. The latter approach, which relies on physician diagnosis, is not viable in many areas of the world due to low medical infrastructure. Self-reported data such as in the World Health Survey, the instrument can determine the prevalence of “ever diagnosed with asthma” and the prevalence of asthma symptoms, namely “current wheeze.” This approach is similar to that of SD. However, SD’s diagnosis of asthma is based on empirical, data-driven models that are capable of capturing other important clues in asthma diagnosis. In order to estimate the

accuracy with which a survey instrument like the WHS could estimate asthma, we conducted a simple simulation. We used the criteria specified by To et al. [21] to estimate the number of asthma cases in the SD dataset, where we also knew with the underlying gold standard diagnosis whether the subject actually had clinical asthma.

To et al. estimated three fractions of asthma: doctor diagnosed asthma, based on the question "Have you ever been diagnosed with asthma?", clinical asthma, based on a positive response to either the previous diagnosis item or "Have you been taking any medications or treatment for asthma during the last 2 weeks?", and symptoms of asthma, based on a positive response to any, all, or two of the two questions above and a third question, "During the last 12 months have you experienced attacks of wheezing or whistling breath?". To et al. also limited this analysis to adults aged 18 to 45 to eliminate cross-classification with COPD, and report that these are similar questions to items used by the ISAAC and ECRHS surveys [22–24]. We replicated this analysis with the SD dataset, and calculated the cause fraction absolute error with each asthma classification. Since it seems plausible that some areas have no access to health care (and therefore the previous doctor diagnosis or asthma medications items would never have a positive response), we also looked at the scenario where only wheezing/whistling was used. The results for this simulation are shown in Table 12. The results show that this approach can accurately estimate cause fractions, except for when wheezing/whistling is used, and that the cause fraction is misestimated from 2 to 9% in absolute terms, with the higher error occurring when the whistling/wheezing information is also used. In comparison, the Tariff Method achieves a median absolute error of only 1.4% when HCE information is used, and 2.5%

when HCE information is excluded, and can consequently be considered a superior approach to current practices.

COPD

COPD prevalence data is similar to asthma but is more involved since it requires a forced expiratory volume in 1 second (FEV1) to forced vital capacity (FVC) measurement, or pre- and post-bronchodilator spirometry. Consequently, the only approach to estimating prevalence of COPD with household surveys is to include a medical examination sufficient to measure FEV1/FVC ratio, which requires a bronchodilator, a spirometer, and the knowledge to use them correctly. Prevalence estimates of COPD typically use these measurements conducted in the field or in health facilities [25–35], though SD is a more flexible approach in areas which do not have the resources to measure FEV1/FVC comprehensively across a population or sample.

Angina pectoris

Angina pectoris is “chest pain or discomfort that often occurs with activity or stress” [36] and is measured in epidemiological surveys using the Rose questionnaire for angina [37]. The Rose questionnaire asks the questions listed in Figure 7. Since angina pectoris is by definition a collection of signs and symptoms which may be associated with different heart conditions, such a questionnaire could hypothetically have a sensitivity of 100% assuming respondents answer honestly. That is, positive endorsements of the Rose questionnaire items are the diagnostic criteria angina pectoris. The SD instrument essentially conducts the Rose questionnaire within its questionnaire, and consequently the instrument could be used as is, where any respondent positively endorsing the Rose items would be classified as having angina pectoris. However, it is

also possible that these items would be endorsed when a different underlying diagnosis had been found. For example, a COPD patient could report symptoms for angina pectoris.

In order to assess the prevalence fraction estimation capability of using the Rose questionnaire alone, we conducted the same analysis as we did for asthma. Specifically, we calculated the cause fraction error based on the Rose questionnaire response pattern required for define angina as used in Ugurlu et al. [38]: "Definite angina was defined as having pain or discomfort in the chest when walking uphill or hurrying and fulfilling all of the following criteria: (i) situated in the sternum or the left anterior chest with or without left arm, (ii) caused the subject to stop or slow down, (iii) went away when the subject stopped or slowed down and (iv) was relieved within 10 min." The SD instrument did not have an item related to the 4th requirement ("was relieved within 10 min"), but using the other requirements we calculated the true and estimated cause fractions for the same 500 test datasets as the other computation methods. The results show that the angina pectoris prevalence fractions are misestimated by approximately 8%, whereas the Tariff Method with HCE can estimate the prevalence fraction within 2% and the SSP method without HCE can estimate the prevalence fraction within 2.6%. Hence, SD is a viable and better alternative to using the Rose questionnaire. The results for this analysis are also shown in Table 12.

Arthritis

Epidemiological measurement of arthritis in the developed world can use hospital data, but arthritis (both rheumatoid arthritis and osteoarthritis) is not commonly measured at population levels in areas of the developing world. This dearth of data in the developing world is a

particularly important issue due to aging populations [39]. Some efforts, notably the Community Oriented Program for Control of Rheumatic Disease (COPCORD) [40], use a screening questionnaire followed by examination by a rheumatologist [41–46], though this approach is fairly resource-intensive particularly in a developing country due to its personnel requirement.

Cirrhosis

The epidemiology of cirrhosis has historically been difficult to measure because diagnosis or prevalence estimates require hospital admission data. Cirrhosis has two clinical phases: subclinical, which is asymptomatic and requires a biopsy for identification, and decompensated, which causes manifestations such as bleeding, encephalopathy, and jaundice. In epidemiological investigations, it is assumed that in developed countries, people with signs and symptoms of cirrhosis will visit the hospital. Hence, hospital admissions data can be used to estimate the incidence of cirrhosis in developed countries with access to health care [47, 48]. A model with a covariate for health care access can then be used to estimate the prevalence of cirrhosis in areas of the world with nonexistent hospital data or low health system access. SD may be a plausible alternative method for estimating cirrhosis. It achieved high performance with the different models in terms of both chance-corrected concordance and prevalence fraction absolute error. It also allows for greater flexibility since the signs and symptoms of the actual population of interest can be collected instead of modeling the cirrhosis epidemiology in a data-sparse country as a function of hospital data in other countries.

Vision loss

Vision loss is typically characterized in terms of sequela thresholds: blindness (best-corrected distance vision less than 3/60), low vision (distance vision impairment worse than 6/18 and better than or equal to 3/60), and near vision impairment, characterized by the inability to see or read items at close proximity. Prevalence data on vision loss requires a visual acuity test using a Snellen or tumbling-E chart conducted in the sample population or collection of clinical records with vision acuity test results. This approach is currently used for population level prevalence estimates [49–57], and whether or not SD should be used as an alternative to this approach when such resources is an important future topic of discussion. As discussed in the cause aggregation section, if users are interested in accurate information but do not need to differentiate between cataracts and general vision loss, then aggregation to 9 causes may increase the viability of SD methods in population vision loss measurement.

Cataracts

Population estimates of cataracts prevalence can use one of three approaches. The first method is to measure visual acuity deficit and then attribute a fraction of its prevalence to cataracts. The second is to measure opacification by itself, and the third is to measure the combination of visual acuity deficit and the opacification of the eye's lens [58]. Classification of opacification requires an examination using slit lamp images conducted by an ophthalmologist, and then the use of the Lens Opacity Classification System (LOCS) [18]. An alternative measurement approach is the Wisconsin Cataract Grading System [59] which uses retro-illumination photographic grading. These measurement approaches have been used for a

number of population-level prevalence estimates [58, 60–67], though more globally-comprehensive studies may be limited by the resource requirements of measuring cataract prevalence. Hence, while symptomatic diagnosis is not as accurate at diagnosing cataracts as examination using the required instruments and expertise of an ophthalmologist, it may still provide a meaningful tool for collecting more comprehensive population estimates of cataract prevalence.

Hearing loss

Similar to vision loss, the prevalence of hearing loss is typically estimated at different thresholds that are stepwise inclusive so that each higher threshold includes the lower thresholds. Measuring hearing loss requires audiometric equipment and ideally a silent room. Some studies have conducted cross-sectional studies of hearing loss prevalence by using audiometric measurement in the field or in communities [68–70], while others have used health care records or existing clinical data [71]. The requirement of using audiometric equipment to measure hearing loss prevalence is arguably undesirable since it imposes technological and resource requirements for estimating hearing loss disease burdens, though the lower performance of SD methods on hearing loss may not be a reasonable tradeoff.

Depression

Depression is the only mental health condition included in this study. Estimates of depression prevalence typically implement questionnaires that have been validated to have high sensitivity in depression diagnosis. The most common is the WHO's Composite International Diagnostic Interview (CIDI) Short Form for Major Depression (SFMD) [72], which requires a positive

response (if screening questions are positively endorsed) to 5 of the 9 major depression symptoms: depressed mood, diminished interest, weight/appetite change, insomnia/hypersomnia, agitation/retardation, fatigue/lack of energy, worthlessness/guilt, diminished ability to think, thoughts of death. The CIDI-SFMD questionnaire has been used extensively in past population estimates of depression prevalence. The SD questionnaire was designed to include these questions.

We conducted a simulation of how depression prevalence would be estimated using the responses to these questions, and measured the true versus estimated prevalence fractions based on this use of the CIDI-SFMD questions within our data/instrument and the DSM-IV criteria specified elsewhere [12]. Specifically, we "screened" for depression cases with positive endorsements of either "having a period lasting several days when you felt sad, empty, or depressed" or "having a period lasting several days when you lost interest in most things" that lasted more than 2 weeks and lasted most of the day, nearly every day. Of those respondents that passed the "screening", we classified them as having depression if they positively endorsed 4 or more of the 8 remaining questions. Based on CIDI-SFMD DSM-IV criteria, we reclassified them as not having depression if the symptoms 1) did not cause significant discomfort, or 2) could be caused completely by loss of a loved one, or 3) could be explained by current drugs or medication, or 4) could be explained by a concurrent illness.

After this depression classification, we calculated the prevalence fraction of depression for the same 500 test datasets as our other analyses. We found a median absolute error of 5.9%. In contrast, the top-performing SD method for estimating depression prevalence fractions was

King-Lu, which had an absolute error of 1.6%. Consequently, SD seems also to be a viable alternative to the CIDI questionnaire. The results from this analysis are also provided in Table 12.

Discussion

Overview

The Population Health Metrics Research Consortium Symptomatic Diagnosis study presents a novel source of data and an innovative application of verbal autopsy research to computational estimation of chronic disease burden. The study identified cases of 10 chronic conditions that had been diagnosed with gold standard criteria and then conducted a questionnaire with over 100 patients with each condition. The questionnaire was designed with the intent of virtual diagnosis using data-driven methods. Such a study and dataset provides a validated testing ground for the different methods explored in this study: Tariff, Simplified Symptom Pattern, Random Forest, and the King-Lu Direct Estimation technique. By validating these methods on the PHMRC dataset, we sought to demonstrate the capability of SD methods in future epidemiological research concerning the burden of chronic diseases, which is a critical global health challenge. We further simulated the application of these methods in the field by testing the performance with the inclusion/exclusion of health care experience information, which allowed us to determine the viability of using SD methods in areas with no health care, and by testing performance in samples of test data with random cause compositions. The results of the study can thus be considered robust due to 1) the gold-standard validation of the questionnaire

responses, 2) calculating predictive validity in various test data compositions, and 3) conducting comparisons of the performance of various computer-based methods.

Speculation on findings

In general, we observed that symptomatic diagnosis is a promising approach to collecting prevalence data on the chronic conditions outlined in this study. The SD methods we tested are capable of estimating prevalence fractions within 3% for each condition. Additionally, 5 of the 11 conditions (asthma, depression, rheumatoid arthritis, angina pectoris, and cirrhosis) have a chance-corrected concordance of over 80% for at least one method. Estimation of prevalence fractions within 3% for these chronic conditions would allow for higher resolution epidemiological information in areas with sparse data, and for the conditions with high performance in terms of chance-corrected concordance, this technology could be used to diagnose in the field without requiring extensive medical expertise or other tools or resources.

In our literature-based comparison to current methods, we observed that SD was capable of matching or outperforming the questionnaire-based methods such as the Rose questionnaire and the CIDI depression questions in terms of prevalence fraction estimation. The SD questionnaire essentially includes the Rose questionnaire, CIDI depression questions, and the WHS asthma questions. It may be surprising that the data-driven SD methods can achieve higher performance than these current approaches, though the result in some ways parallels the finding in verbal autopsy research that computational methods can outperform physician-certified verbal autopsy. This performance hierarchy can be partly explained by the idea that certain conditions may have signs and symptoms that are not part of a textbook clinical

presentation of the condition. The Tariff Method and Random Forest can utilize over 700 data features, allowing them to identify subtle signals and to differentiate between different causes. The Simplified Symptom Pattern and King-Lu algorithms capture statistical interdependent relationships in data features that are essentially only observable with computational power. In contrast, diagnosis of these conditions in the current approaches does not incorporate the consideration that other conditions could cause certain symptoms. For example, in the WHS approach, a case could present with asthma symptoms when the underlying condition is actually COPD. Applied systematically to an entire dataset, this could lead to significant over exaggeration of the asthma disease burden. However, if a feature with a high tariff for COPD was positively endorsed, then the Tariff Method would likely decide the condition was COPD instead of asthma, while in the WHS approach the case could have been classified as asthma. Furthermore, the inclusion of free text is a powerful source of unstructured data that cannot be harnessed using the other questionnaire techniques.

Current estimation of COPD, both types of arthritis, vision loss, hearing loss, cirrhosis, and cataracts can be undertaken with a high level of accuracy in a clinical setting, but it requires specialized equipment and/or training such as a spirometer or the medical diagnostic knowledge. Certainly it is most desirable to collect the most accurate information possible, but access to these tools and resource is simply not possible in all areas of the world. As discussed in the introduction, this is an unfortunate paradox since the areas lacking these resources are also likely the areas that have the worst health. SD in this regard is an extremely valuable alternative to collecting more refined information in a resource-poor setting. A household

survey can be conducted virtually anywhere in the world. Currently Demographic and Health Surveys and World Health Surveys cover areas of the world such as Sudan, Code d'Ivoire, and Congo Democratic Republic, for example. If access to the tools and expertise to diagnose these conditions in these areas is unavailable, then SD-based epidemiology could be a practical alternative. The use of SD as an epidemiological smoke signal in low-resource or inaccessible areas to identify and focus attention on the chronic disease burden could also help ease the aforementioned paradox. Furthermore, training field workers to conduct an SD survey is significantly less expensive than making available the resources to diagnose all of the conditions outlined in this study. Cumulatively, this flexibility makes SD a compelling alternative strategy for estimating these conditions.

One of the unique results from this study is found in the comparison of different SD methods. Random Forest arguably provides the most sophisticated and powerful machinery for this data classification task, yet it does not yield consistently superior performance compared to other methods. Tariff, in contrast, is a simple, additive algorithm but produces remarkably high performance for many of the causes analyzed in this study. Simplified Symptom Pattern achieves superior performance in certain comparisons in this study, and KL is overwhelmingly the most accurate method for estimating prevalence fractions of depression, though it cannot predict individual cases. Essentially, no one method is superior for every situation. Some users may value the parsimony of Tariff, while others may opt for the strong performance of KL in estimating depression prevalence. This finding also hints that one of the most powerful approaches may be to use an ensemble of different methods. The concept of ensemble

approaches has been demonstrated to confer a significant advantage in research areas such as the Netflix Challenge [73] and in epidemiological modeling research such as the causes of death ensemble model CODEm [74]. Future research in SD may yield further advances in predictive validity using some variant of an ensemble model.

Limitations

Our study had some inherent limitations. One of the main limitations and questions in verbal autopsy research is that it seems plausible that the questionnaire responses in a "community" death where an interview is conducted will be systematically different than the responses from the "hospital" deaths on which computational algorithms are changed. For example, the family members of a person who died in a hospital may be from higher socioeconomic status than the family responding to a verbal autopsy for someone who died outside of a hospital. If the response patterns are sufficiently different, then the computational methods could perform differently than expected. However, this limitation, which applies also to symptomatic diagnosis, is essentially a normative question. It is not possible to develop data-driven models unless this limitation is accepted, and as previous research in verbal autopsy has shown, data-driven models outperform expert-based models by a substantial margin. We attempted to simulate cases where the respondent did not have access to health care by conducting analyses in which we withheld "health care experience" features, though this does not avoid the limitation entirely.

Since prevalence data are sparse in many areas of the world, it is important to consider the potential implementation of the SD methods outlined in this study in countries besides Mexico.

The 10 chronic conditions considered in this study are also highly prevalent in areas of Africa and Asia, and this consideration raises the question of there exist systematic cultural variations in questionnaire response data. For example, the symptoms of depression may be more commonly acknowledged in one country over another, and these variations could affect the performance of SD models. This limitation can be addressed by further collection of validated SD questionnaire responses in other countries. In fact, additional validated SD questionnaire response data would strengthen the performance of the existing models. Furthermore, the computational SD methods can readily be re-trained on any further validation data that is collected.

Another limitation of the study derives from the inclusion of “healthy” controls. The controls are referred to as “healthy” controls, though the use of that term is somewhat subjective. The controls do not suffer from the target conditions, though they can be ill with other conditions. The inclusion is important, however, because it allows for a model to predict that a person does not suffer from a given condition despite possibly presenting some of the signs and symptoms associated with that condition. For example, 50% of the controls report a non-productive cough, which is not a dramatically lower endorsement rate than asthma with 58% of cases reporting a non-productive cough. The inclusion of controls highlights one of the important differences between SD and VA: every person who dies has an underlying cause of death, but not every living person has an underlying illness. We observed that the performance of each method was also somewhat a function of the composition of conditions in the “test” population, which was our rationale for imposing the Dirichlet-based prevalence fraction

variation. Since the fraction of health controls tended to have the lowest performance in terms of prevalence fraction error, it seems likely that SD methods would be more accurate in areas with higher chronic disease burden and fewer healthy people. The healthy controls also have different characteristics from the rest of the study participants. Specifically, based on Table 3, the controls tend to be slightly younger and are more frequently female than the other conditions.

The relatively poor performance of hearing loss in SD models is a surprising result since it seems reasonable that hearing loss patients would express somewhat salient signs and symptoms compared with other conditions. We investigated the cause for the low performance by analyzing the endorsement rates of all data features by the true cause. This analysis showed that one of the obstacles in diagnosing hearing loss based on this questionnaire derives from a lack of items with a high endorsement rate for hearing loss but not for other causes. For example, one item asks, “Have you ever had your hearing checked by a provider?” While the endorsement rates hearing loss are high (91.2%) for this item, cases with the other conditions also endorse this item at a fairly high frequency. For example, 82.2% of the controls and 71.8% of the asthma cases positively endorsed this item. Similarly, for the question “Do you have deafness or trouble hearing in one or both ears without the help of a hearing aid?”, 39.5% of hearing loss cases positively endorsed the item, and 36.1% of cataracts positively endorsed this item. In contrast, causes that had better performance tend to have an item that has a high endorsement rate for that cause but not for others. For example, the question “Have you had a period lasting several days when you lost interest in most things that you usually enjoy such as

pastimes, relationships, or work?" has a 90% endorsement rate among depression cases, but no higher than 30.3% endorsement for other causes. Future work in the field of computational diagnosis should explore which other types of questionnaire items are more specific to hearing loss compared to other chronic conditions.

Future implementation

The ultimate goal of the symptomatic diagnosis study is to develop a tool that can be used effectively in the field to collect higher quality prevalence data on the 10 chronic conditions listed in the study. This implementation requires two steps. The first step is making the questionnaire available in a standard format for any researchers to use. Current work in verbal autopsy is moving towards using tablet devices that can use questionnaire software such as ODK Collect to facilitate data collection. The second required step is developing a user-friendly software package that readily conducts the methods described in this study so that researchers do not need fluency in Stata, R, and Python to implement a method. Reducing these barriers will facilitate more rapid use of the methods outlined in this study to improve the collection of health information for chronic conditions.

List of abbreviations used

KL: King-Lu algorithm for verbal autopsy

ML: Machine Learning

PHMRC: Population Health Metrics Research Consortium

RF: Random Forest

SD: Symptomatic Diagnosis

SSP: Simplified Symptom Pattern

VA: Verbal Autopsy

References

1. Global Burden of Diseases and Injuries for 21 regions, 1990-2010: A systematic analysis. *In preparation.* .

2. Flaxman AD, Vahdatpour A, Green S, James SL, Murray CJ: Random forests for verbal autopsy analysis: multisite validation study using clinical diagnostic gold standards. *Population Health Metrics* 2011, 9:29.

3. Flaxman AD, Vahdatpour A, James SL, Birnbaum JK, Murray CJ: Direct estimation of cause-specific mortality fractions from verbal autopsies: multisite validation study using clinical diagnostic gold standards. *Population Health Metrics* 2011, 9:35.

4. Hernández B, Ramírez-Villalobos D, Romero M, Gómez S, Atkinson C, Lozano R: Assessing quality of medical death certification: Concordance between gold standard diagnosis and underlying cause of death in selected Mexican hospitals. *Popul Health Metr* 2011, 9:38.

5. James SL, Flaxman AD, Murray CJ: Performance of the Tariff Method: validation of a simple additive algorithm for analysis of verbal autopsies. *Population Health Metrics* 2011, 9:31.

6. Lozano R, Freeman MK, James SL, Campbell B, Lopez AD, Flaxman AD, Murray CJ: Performance of InterVA for assigning causes of death to verbal autopsies: multisite validation study using clinical diagnostic gold standards. *Population Health Metrics* 2011, 9:50.

7. Lozano R, Lopez AD, Atkinson C, Naghavi M, Flaxman AD, Murray CJ: Performance of physician-certified verbal autopsies: multisite validation study using clinical diagnostic gold standards. *Population Health Metrics* 2011, 9:32.

8. Murray CJ, James SL, Birnbaum JK, Freeman MK, Lozano R, Lopez AD: Simplified Symptom Pattern Method for verbal autopsy analysis: multisite validation study using clinical diagnostic gold standards. *Population Health Metrics* 2011, 9:30.

9. Murray CJ, Lopez AD, Black R, Ahuja R, Ali SM, Baqui A, Dandona L, Dantzer E, Das V, Dhingra U, Dutta A, Fawzi W, Flaxman AD, Gómez S, Hernández B, Joshi R, Kalter H, Kumar A, Kumar V, Lozano R, Lucero M, Mehta S, Neal B, Ohno SL, Prasad R, Praveen D, Premji Z, Ramírez-Villalobos D, Remolador H, Riley I, Romero M, Said M, Sanvictores D, Sazawal S, Tallo V: Population Health Metrics Research Consortium gold standard verbal autopsy validation study: design, implementation, and development of analysis datasets. *Popul Health Metr* 2011, 9:27.

10. WHO | World Health Survey Instruments and Related Documents [http://www.who.int/healthinfo/survey/instruments/en/index.html].

11. Population Health Metrics Research Consortium | Institute for Health Metrics and Evaluation [http://www.healthmetricsandevaluation.org/research/project/population-health-metrics-research-consortium].

12. DSM-IV-TR Index.

13. Cohen AM, Hersh WR: A Survey of Current Work in Biomedical Text Mining. *Brief Bioinform* 2005, 6:57–71.

14. Dale R, Moisl H, Somers H: Handbook of Natural Language Processing. *Computational Linguistics* 2001, 27:602–603.

15. Kao A, Poteet SR: *Natural Language Processing And Text Mining*. Springer; 2007.

16. Kim J-D, Ohta T, Tateisi Y, Tsujii J: GENIA corpus--a semantically annotated corpus for bio-textmining. *Bioinformatics* 2003, 19:i180–i182.

17. Rajman M, BESANÇON R, Besancon R: Text Mining: Natural Language techniques and Text Mining applications. In *In Proceedings of the 7 th IFIP Working Conference on Database Semantics (DS-7). Chapam*. Hall; 1997:7–10.

18. tm - Text Mining Package [http://tm.r-forge.r-project.org/].

19. Murray CJ, Lozano R, Flaxman AD, Vahdatpour A, Lopez AD: Robust metrics for assessing the performance of different verbal autopsy cause assignment methods in validation studies. *Popul Health Metr* 2011, 9:28.

20. Verbal Autopsy Methods with Multiple Causes of Death. *Statistical Science* 2008, 23:78–91.

21. To T, Stanojevic S, Moores G, Gershon AS, Bateman ED, Cruz AA, Boulet L-P: Global asthma prevalence in adults: findings from the cross-sectional world health survey. *BMC Public Health* 2012, 12:204.

22. Variations in the prevalence of respiratory symptoms, self-reported asthma attacks, and use of asthma medication in the European Community Respiratory Health Survey (ECRHS). *Eur. Respir. J.* 1996, 9:687–695.

23. Worldwide variation in prevalence of symptoms of asthma, allergic rhinoconjunctivitis, and atopic eczema: ISAAC. The International Study of Asthma and Allergies in Childhood (ISAAC) Steering Committee. *Lancet* 1998, 351:1225–1232.

24. Worldwide variations in the prevalence of asthma symptoms: the International Study of Asthma and Allergies in Childhood (ISAAC). *Eur. Respir. J.* 1998, 12:315–335.

25. Buist AS, Vollmer WM, Sullivan SD, Weiss KB, Lee TA, Menezes AMB, Crapo RO, Jensen RL, Burney PGJ: The Burden of Obstructive Lung Disease Initiative (BOLD): Rationale and Design. *COPD: Journal of Chronic Obstructive Pulmonary Disease* 2005, 2:277–283.

26. Caballero A, Torres-Duque CA, Jaramillo C, Bolívar F, Sanabria F, Osorio P, Orduz C, Guevara DP, Maldonado D: Prevalence of COPD in five Colombian cities situated at low, medium, and high altitude (PREPOCOL study). *Chest* 2008, 133:343–349.

27. Fukuchi Y, Nishimura M, Ichinose M, Adachi M, Nagai A, Kuriyama T, Takahashi K, Nishimura K, Ishioka S, Aizawa H, Zaher C: COPD in Japan: the Nippon COPD Epidemiology study. *Respirology* 2004, 9:458–465.

28. Jiang R, Luo D, Huang C, Li W: [Study on the prevalence rate and risk factors of chronic obstructive pulmonary disease in rural community population in Hubei province]. *Zhonghua Liu Xing Bing Xue Za Zhi* 2007, 28:976–979.

29. Jyrki-Tapani K, Sovijärvi A, Lundbäck B: Chronic obstructive pulmonary disease in Finland: prevalence and risk factors. *COPD* 2005, 2:331–339.

30. Liu S, Zhou Y, Wang X, Wang D, Lu J, Zheng J, Zhong N, Ran P: Biomass fuels are the probable risk factor for chronic obstructive pulmonary disease in rural South China. *Thorax* 2007, 62:889–897.

31. Menezes A, Macedo SC, Gigante DP, da Costa JD, Olinto MT, Fiss E, Chatkin M, Hallal PC, Victora CG: Prevalence and risk factors for chronic obstructive pulmonary disease according to symptoms and spirometry. *COPD* 2004, 1:173–179.

32. Menezes AMB, Perez-Padilla R, Jardim JRB, Muiño A, Lopez MV, Valdivia G, Montes de Oca M, Talamo C, Hallal PC, Victora CG: Chronic obstructive pulmonary disease in five Latin American cities (the PLATINO study): a prevalence study. *Lancet* 2005, 366:1875–1881.

33. Shirtcliffe P, Weatherall M, Marsh S, Travers J, Hansell A, McNaughton A, Aldington S, Muellerova H, Beasley R: COPD prevalence in a random population survey: a matter of definition. *Eur. Respir. J.* 2007, 30:232–239.

34. Yao W, Zhu H, Shen N, Han X, Liang Y, Zhang L, Sun Y, Hao Z, Zhao M: [Epidemiological data of chronic obstructive pulmonary disease in Yanqing County in Beijing]. *Beijing Da Xue Xue Bao* 2005, 37:121–125.

35. The Burden of Obstructive Lung Disease Initiative (BOLD): Rationale and Design, COPD: Journal of Chronic Obstructive Pulmonary Disease, Informa Healthcare [http://informahealthcare.com/doi/abs/10.1081/COPD-57610].

36. Stable angina - PubMed Health [http://www.ncbi.nlm.nih.gov/pubmedhealth/PMH0001247/].

37. ROSE GA: The diagnosis of ischaemic heart pain and intermittent claudication in field surveys. *Bull. World Health Organ.* 1962, 27:645–658.

38. Ugurlu S, Seyahi E, Yazici H: Prevalence of Angina, Myocardial Infarction and Intermittent Claudication Assessed by Rose Questionnaire Among Patients with Behcet's Syndrome. *Rheumatology* 2008, 47:472–475.

39. Miranda VS, Decarvalho VB, Machado LA, Dias JM: Prevalence of chronic musculoskeletal disorders in elderly Brazilians: a systematic review of the literature. *BMC musculoskeletal disorders* 2012, 13:82.

40. COPCORD Website [http://www.copcord.org/].

41. Chaaya M, Slim ZN, Habib RR, Arayssi T, Dana R, Hamdan O, Assi M, Issa Z, Uthman I: High burden of rheumatic diseases in Lebanon: a COPCORD study. *International Journal of Rheumatic Diseases* 2012, 15:136–143.

42. Dans LF, Tankeh-Torres S, Amante CM, Penserga EG: The prevalence of rheumatic diseases in a Filipino urban population: a WHO-ILAR COPCORD Study. World Health Organization. International League of Associations for Rheumatology. Community Oriented Programme for the Control of the Rheumatic Diseases. *J. Rheumatol.* 1997, 24:1814–1819.

43. Haq SA, Darmawan J, Islam MN, Uddin MZ, Das BB, Rahman F, Chowdhury MAJ, Alam MN, Mahmud TAK, Chowdhury MR, Tahir M: Prevalence of rheumatic diseases and associated outcomes in rural and urban communities in Bangladesh: a COPCORD study. *J. Rheumatol.* 2005, 32:348–353.

44. Joshi VL, Chopra A: Is there an urban-rural divide? Population surveys of rheumatic musculoskeletal disorders in the Pune region of India using the COPCORD Bhigwan model. *J. Rheumatol.* 2009, 36:614–622.

45. Minh Hoa TT, Darmawan J, Chen SL, Van Hung N, Thi Nhi C, Ngoc An T, Damarwan J, Shun Le C: Prevalence of the rheumatic diseases in urban Vietnam: a WHO-ILAR COPCORD study. *J. Rheumatol.* 2003, 30:2252–2256.

46. Obregón-Ponce A, Iraheta I, García-Ferrer H, Mejia B, García-Kutzbach A: Prevalence of Musculoskeletal Diseases in Guatemala, Central America: The COPCORD Study of 2 Populations. *J Clin Rheumatol* 2012, 18:170–174.

47. Haukeland JW, Lorgen I, Schreiner LT, Frigstad S-O, Brandsaeter B, Bjøro K, Bang C, Raknerud N, Konopski Z: Incidence rates and causes of cirrhosis in a Norwegian population. *Scand. J. Gastroenterol.* 2007, 42:1501–1508.

48. Sørensen HT, Thulstrup AM, Mellemkjar L, Jepsen P, Christensen E, Olsen JH, Vilstrup H: Long-term survival and cause-specific mortality in patients with cirrhosis of the liver: a nationwide cohort study in Denmark. *J Clin Epidemiol* 2003, 56:88–93.

49. Bourne R, Dineen B, Jadoon Z, Lee PS, Khan A, Johnson GJ, Foster A, Khan D: The Pakistan national blindness and visual impairment survey--research design, eye examination methodology and results of the pilot study. *Ophthalmic Epidemiol* 2005, 12:321–333.

50. Ezepue UF: Magnitude and causes of blindness and low vision in Anambra State of Nigeria (results of 1992 point prevalence survey). *Public Health* 1997, 111:305–309.

51. Jadoon MZ, Dineen B, Bourne RRA, Shah SP, Khan MA, Johnson GJ, Gilbert CE, Khan MD: Prevalence of blindness and visual impairment in Pakistan: the Pakistan National Blindness and Visual Impairment Survey. *Invest. Ophthalmol. Vis. Sci.* 2006, 47:4749–4755.

52. Kyari F, Gudlavalleti MVS, Sivsubramaniam S, Gilbert CE, Abdull MM, Entekume G, Foster A: Prevalence of blindness and visual impairment in Nigeria: the National Blindness and Visual Impairment Study. *Invest. Ophthalmol. Vis. Sci.* 2009, 50:2033–2039.

53. Limburg H, Barria von-Bischhoffshausen F, Gomez P, Silva JC, Foster A: Review of recent surveys on blindness and visual impairment in Latin America. *Br J Ophthalmol* 2008, 92:315–319.

54. Maberley DAL, Hollands H, Chuo J, Tam G, Konkal J, Roesch M, Veselinovic A, Witzigmann M, Bassett K: The prevalence of low vision and blindness in Canada. *Eye (Lond)* 2006, 20:341–346.

55. Mathenge W, Bastawrous A, Foster A, Kuper H: The Nakuru Posterior Segment Eye Disease Study: Methods and Prevalence of Blindness and Visual Impairment in Nakuru, Kenya. *Ophthalmology* 2012.

56. Richard AI: Causes of blindness and low vision in Bayelsa State, Nigeria: a clinic based study. *Nig Q J Hosp Med* 2010, 20:125–128.

57. Xu L, Wang Y, Li Y, Wang Y, Cui T, Li J, Jonas JB: Causes of blindness and visual impairment in urban and rural areas in Beijing: the Beijing Eye Study. *Ophthalmology* 2006, 113:1134.e1–11.

58. Acosta R, Hoffmeister L, Román R, Comas M, Castilla M, Castells X: [Systematic review of population-based studies of the prevalence of cataracts]. *Arch Soc Esp Oftalmol* 2006, 81:509–516.

59. Tan ACS, Wang JJ, Lamoureux EL, Wong W, Mitchell P, Li J, Tan AG, Wong TY: Cataract prevalence varies substantially with assessment systems: comparison of clinical and photographic grading in a population-based study. *Ophthalmic Epidemiol* 2011, 18:164–170.

60. Bao Y, Cao X, Li X, Chen J, Hu J, Zhu T: [Prevalence of age-related cataract among adults aged 50 and above in four rural areas in western China]. *Zhonghua Yi Xue Za Zhi* 2008, 88:1697–1702.

61. Carlos GA, Schellini SA, Espíndola RF de, Lana FP, Rodrigues ACL, Padovani CR: Cataract prevalence in Central-West region of São Paulo State, Brazil. *Arq Bras Oftalmol* 2009, 72:375–379.

62. Hashemi H, Hatef E, Fotouhi A, Feizzadeh A, Mohammad K: The prevalence of lens opacities in Tehran: the Tehran Eye Study. *Ophthalmic Epidemiol* 2009, 16:187–192.

63. Jacob S, Boveda S, Bar O, Brézin A, Maccia C, Laurier D, Bernier M-O: Interventional cardiologists and risk of radiation-induced cataract: Results of a French multicenter observational study. *International journal of cardiology* 2012.

64. Mrena S, Kivelä T, Kurttio P, Auvinen A: Lens opacities among physicians occupationally exposed to ionizing radiation--a pilot study in Finland. *Scand J Work Environ Health* 2011, 37:237–243.

65. Raman R, Pal SS, Adams JSK, Rani PK, Vaitheeswaran K, Sharma T: Prevalence and risk factors for cataract in diabetes: Sankara Nethralaya Diabetic Retinopathy Epidemiology and Molecular Genetics Study, report no. 17. *Invest. Ophthalmol. Vis. Sci.* 2010, 51:6253–6261.

66. Ravindran RD, Vashist P, Gupta SK, Young IS, Maraini G, Camparini M, Jayanthi R, John N, Fitzpatrick KE, Chakravarthy U, Ravilla TD, Fletcher AE: Inverse association of vitamin C with cataract in older people in India. *Ophthalmology* 2011, 118:1958–1965.e2.

67. Vashist P, Talwar B, Gogoi M, Maraini G, Camparini M, Ravindran RD, Murthy GV, Fitzpatrick KE, John N, Chakravarthy U, Ravilla TD, Fletcher AE: Prevalence of cataract in an older population in India: the India study of age-related eye disease. *Ophthalmology* 2011, 118:272–278.e1–2.

68. Westerberg BD, Skowronski DM, Stewart IF, Stewart L, Bernauer M, Mudarikwa L: Prevalence of hearing loss in primary school children in Zimbabwe. *Int. J. Pediatr. Otorhinolaryngol.* 2005, 69:517–525.

69. Taha AA, Pratt SR, Farahat TM, Abdel-Rasoul GM, Albtanony MA, Elrashiedy A-LE, Alwakeel HR, Zein A: Prevalence and risk factors of hearing impairment among primary-school children in Shebin El-kom District, Egypt. *Am J Audiol* 2010, 19:46–60.

70. Bastos I, Mallya J, Ingvarsson L, Reimer A, Andréasson L: Middle ear disease and hearing impairment in northern Tanzania. A prevalence study of schoolchildren in the Moshi and Monduli districts. *Int. J. Pediatr. Otorhinolaryngol.* 1995, 32:1–12.

71. Saunders JE, Vaz S, Greinwald JH, Lai J, Morin L, Mojica K: Prevalence and etiology of hearing loss in rural Nicaraguan children. *Laryngoscope* 2007, 117:387–398.

72. The World Mental Health Composite International Diagnostic Interview [http://www.hcp.med.harvard.edu/wmhcidi/index.php].

73. The Netflix Prize [http://www.netflixprize.com/].

74. Foreman KJ, Lozano R, Lopez AD, Murray CJ: Modeling causes of death: an integrated approach using CODEm. *Population Health Metrics* 2012, 10:1.

Illustrations and figures

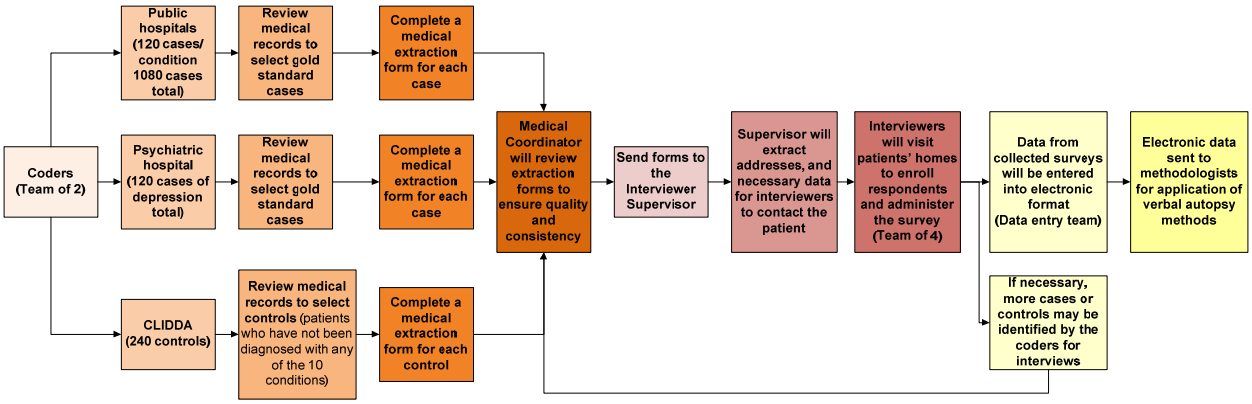

Figure 1: Study design and data collection process for the symptomatic diagnosis project.

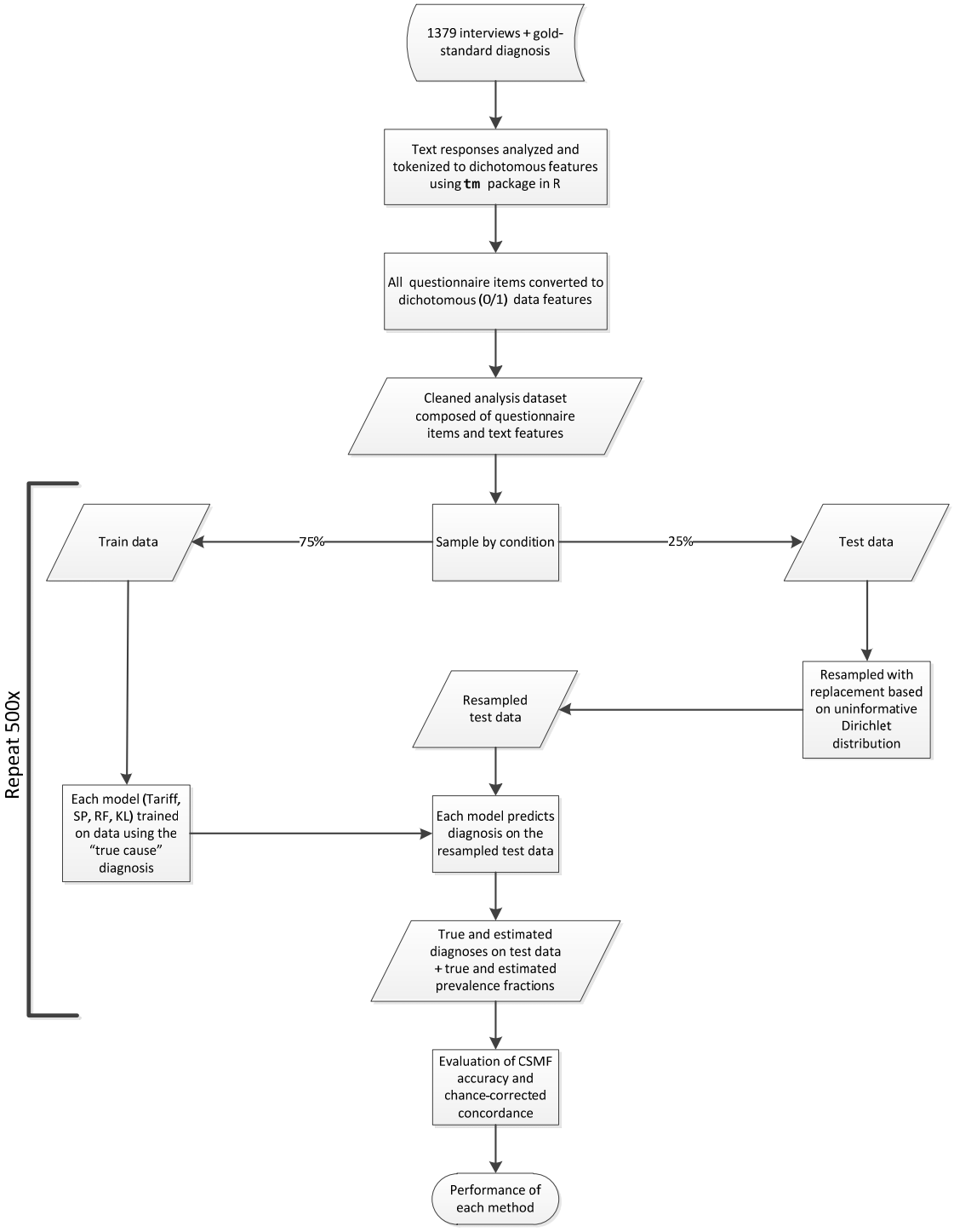

Figure 2: Model validation process for SD methods

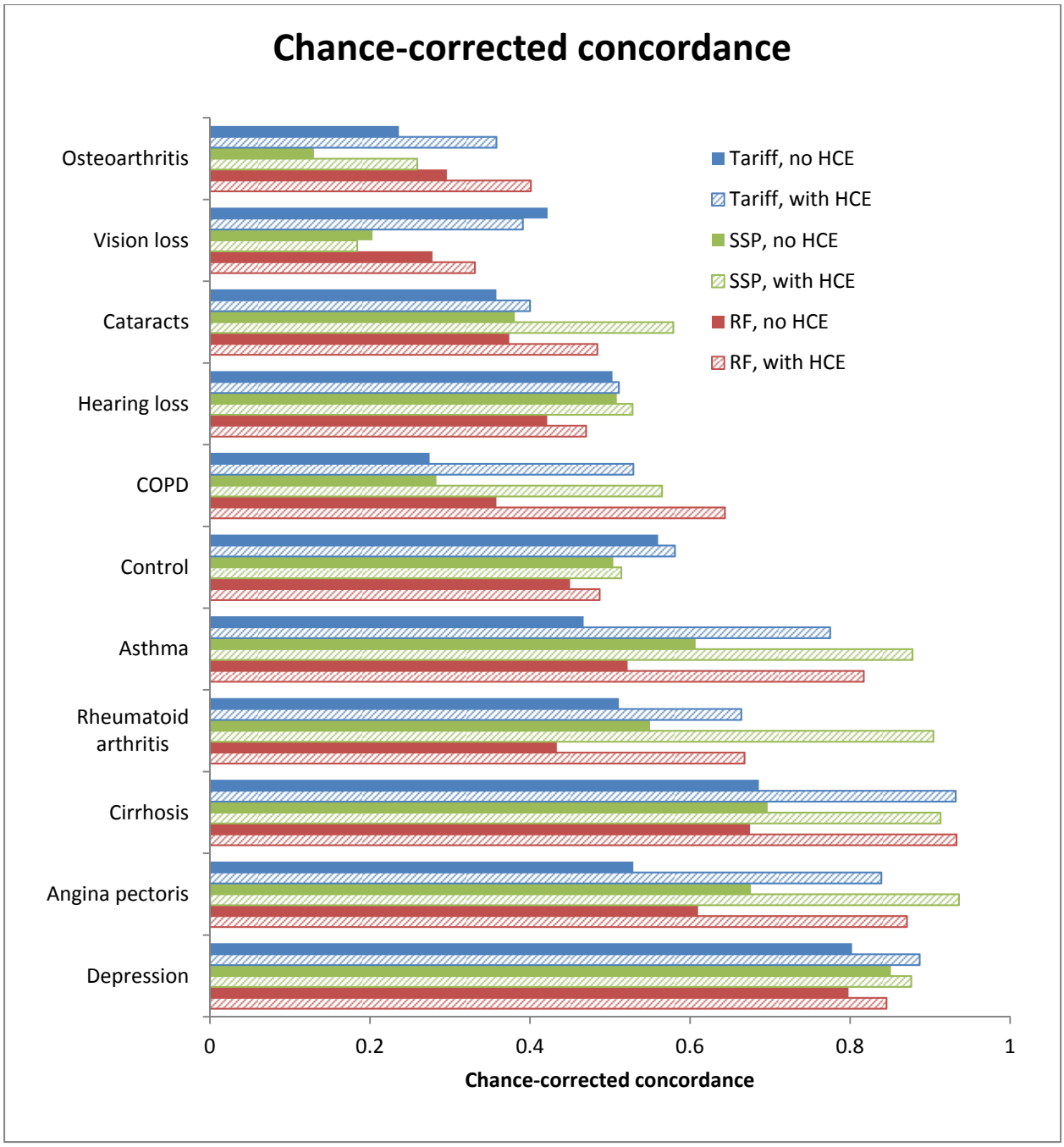

Figure 3: Cause-specific chance-corrected concordance with and without HCE

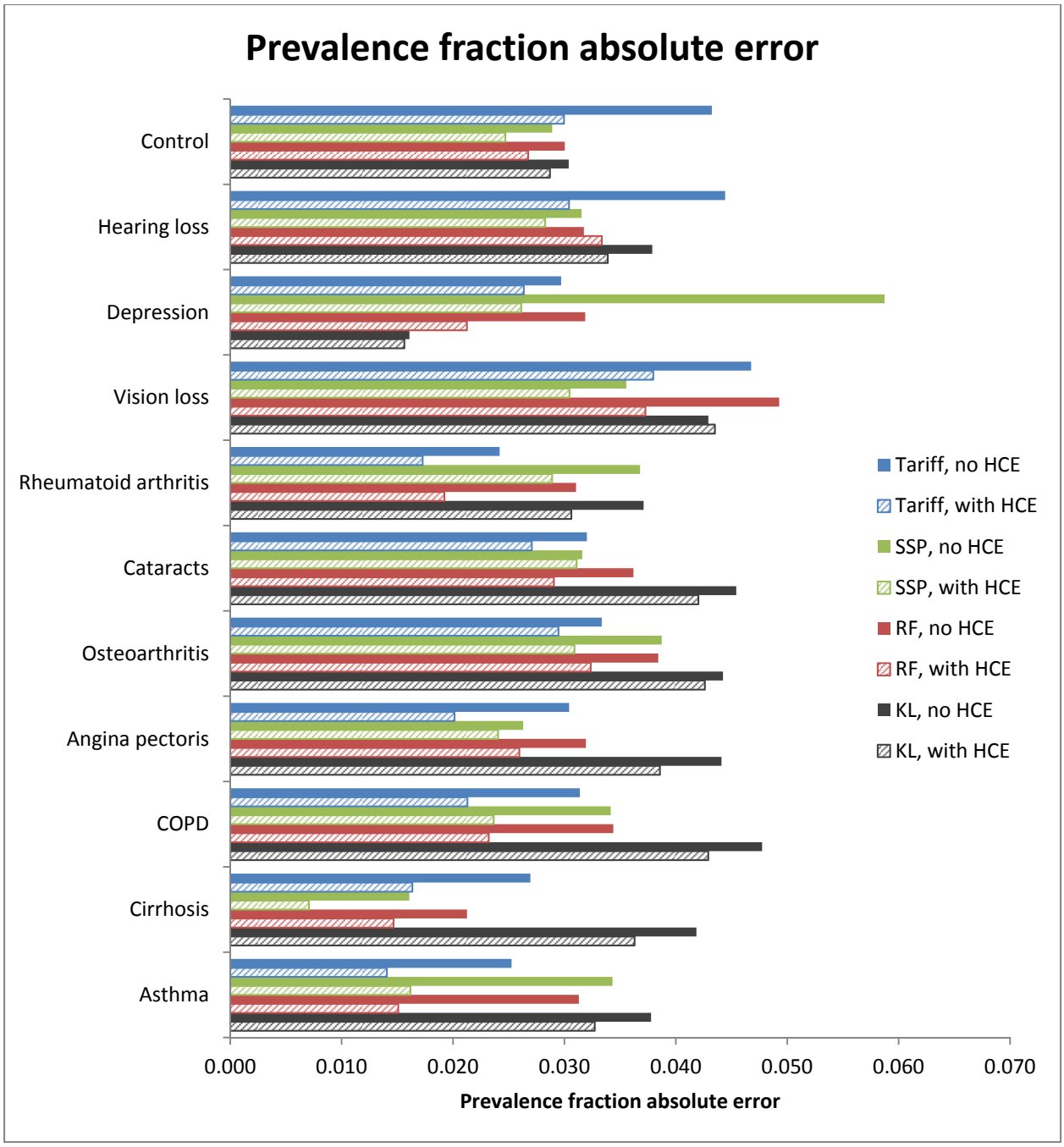

Figure 4: Cause-specific prevalence fraction error

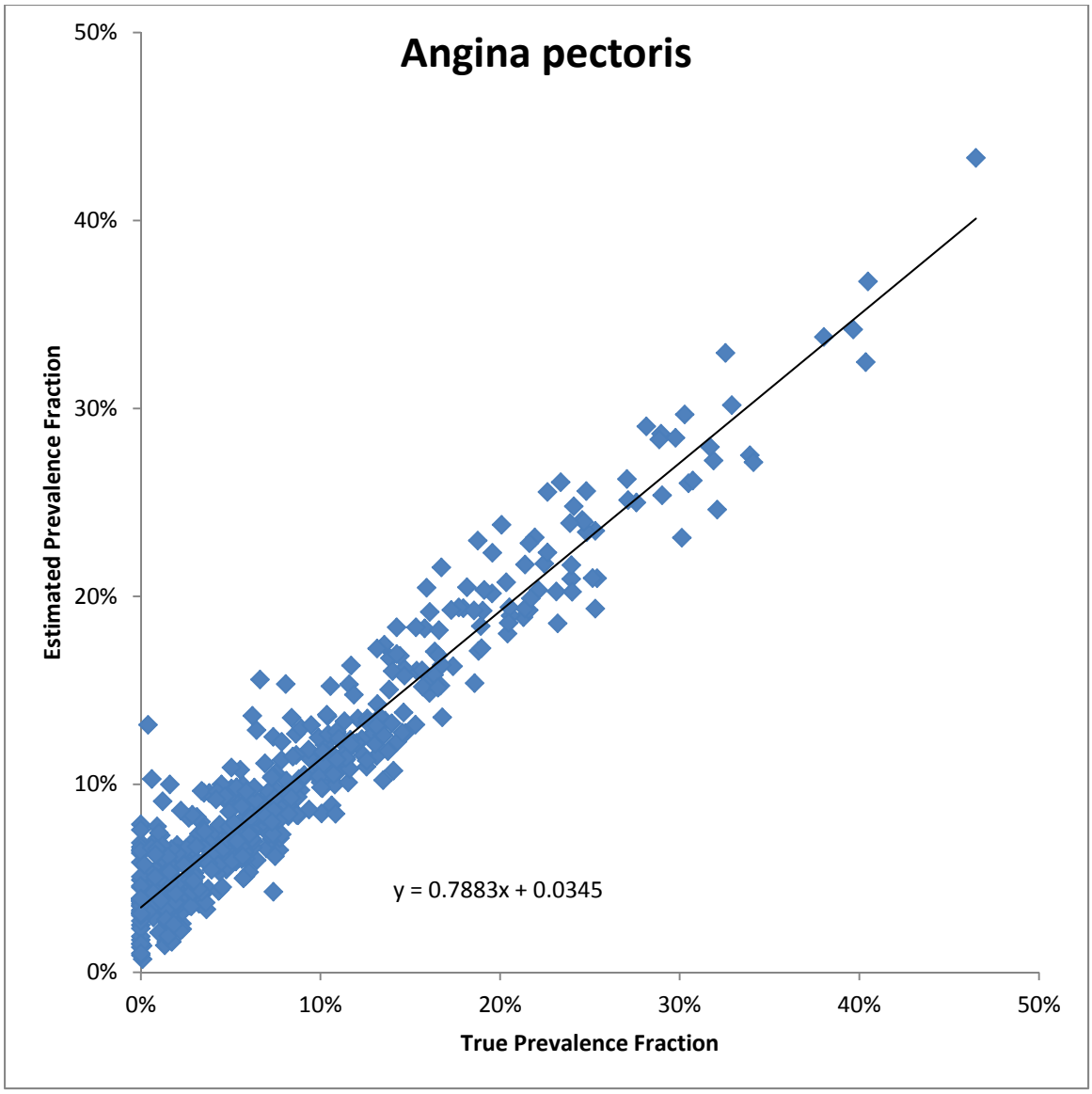

Figure 5: True and estimated prevalence fractions using the Tariff Method with HCE for 500 splits for angina pectoris

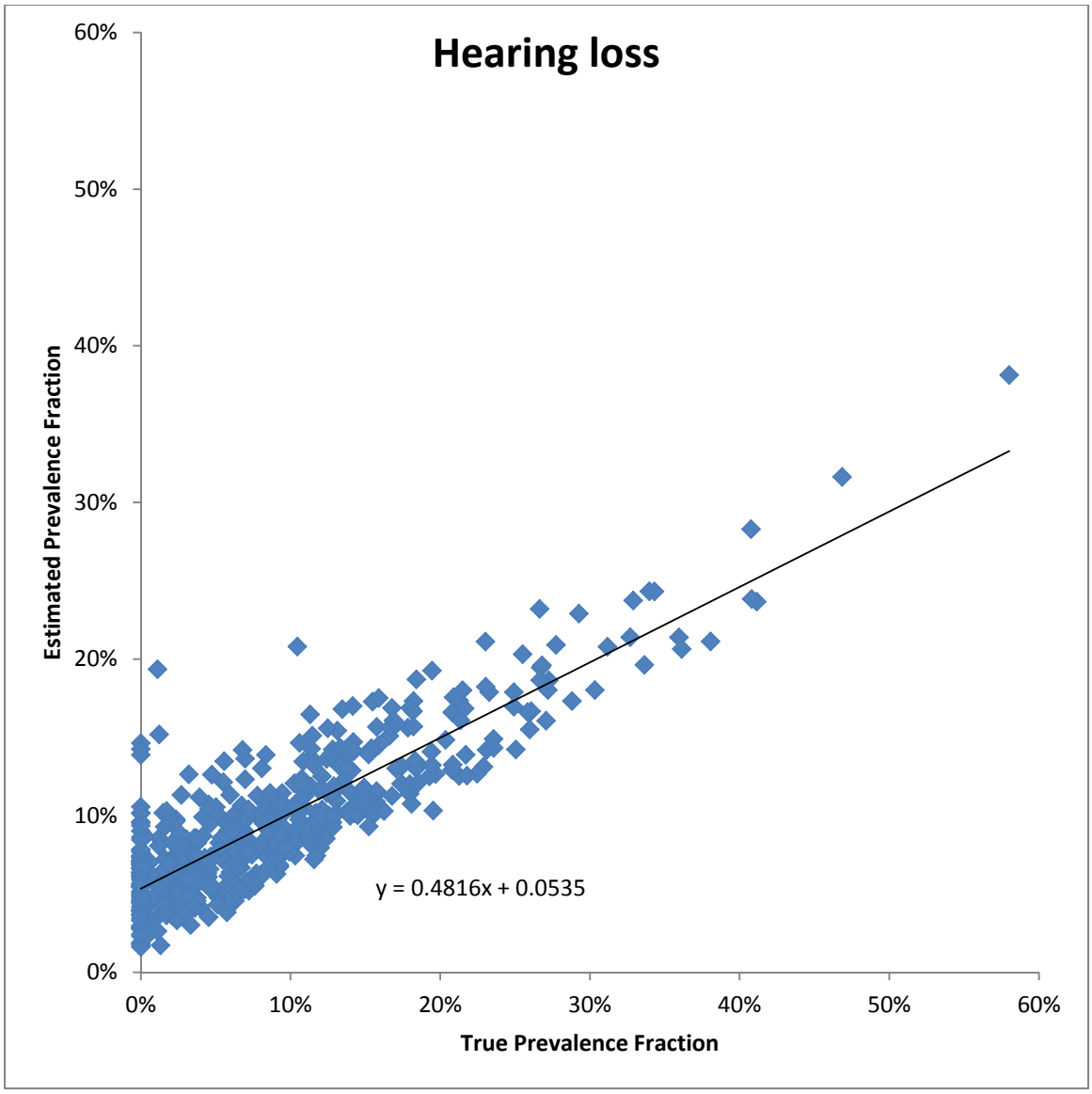

Figure 6: True and estimated prevalence fractions using the Tariff Method with HCE for 500 splits for hearing loss

During the last 12 months, have you experienced any of the following:

<table>
<tr><td>Q6012</td><td>Pain or discomfort in your chest when you walk uphill or hurry?</td><td colspan="2">Yes</td><td>No</td><td>Never walks uphill or hurries</td><td></td></tr>
<tr><td>Q6013</td><td>Pain or discomfort in your chest when you walk at an ordinary pace on level ground?</td><td colspan="3">Yes</td><td>No</td><td>If Q6012 and Q6013 No: Go to Q6017</td></tr>
<tr><td rowspan="3">Q6014</td><td rowspan="3">What do you do if you get the pain or discomfort when you are walking?</td><td colspan="5">1. Stop or slow down</td><td rowspan="5"></td></tr>
<tr><td colspan="5">2. Carry on after taking a pain relieving medicine that dissolves in your mouth</td></tr>
<tr><td colspan="5">3. Carry on</td></tr>
<tr><td>Q6015</td><td>If you stand still, what happens to the pain or discomfort?</td><td colspan="3">Relieved</td><td colspan="2">Not relieved</td></tr>
<tr><td>Q6016</td><td>Will you show me where you usually experience the pain or discomfort? RECORD ALL AREAS OF BODY MENTIONED OR SHOWED</td><td>Upper or middle chest</td><td>Lower chest</td><td>Left arm</td><td>Other</td></tr>
</table>

Figure 7: Rose questionnaire for angina pectoris

Tables and captions

Table 1: Symptomatic diagnosis questionnaire items

Item	Question *(HCE items are italicized)*
sd4_01	*Have you ever been told by a health provider that you have chronic bronchitis, emphysema, or chronic obstructive pulmonary disease (COPD)?*
sd4_02a	*How long ago, in months or years, were you told by a health provider that you have chronic bronchitis, emphysema, or chronic obstructive pulmonary disease (COPD)? [specify]*
sd4_02b	*How long ago, in months or years, were you told by a health provider that you have chronic bronchitis, emphysema, or chronic obstructive pulmonary disease (COPD)?*
sd4_03	*Are you currently taking medication for chronic bronchitis, emphysema, or chronic obstructive pulmonary disease (COPD)?*
sd4_04	*Have you ever been told by a health provider that you have heart failure?*
sd4_05a	*How long ago, in months or years, were you told by a health provider that you have heart failure?*
sd4_52b	*How long ago, in months or years, were you told by a health provider that you have heart failure?*
sd4_06	*Are you currently taking medication for heart failure?*
sd4_07	*Have you ever been told by a health provider that you have cirrhosis?*
sd4_8a	*How long ago, in months or years, were you told by a health provider that you have cirrhosis?*
sd4_8b	*How long ago, in months or years, were you told by a health provider that you have cirrhosis?*
sd4_09	*Are you currently taking medication for cirrhosis?*
sd4_10	*Have you ever been told by a health provider that you have liver failure?*
sd4_11a	*How long ago, in months or years, were you told by a health provider that you have liver failure?*
sd4_11b	*How long ago, in months or years, were you told by a health provider that you have liver failure?*
sd4_12	*Are you currently taking medication for liver failure?*
sd4_13	*Have you ever been told by a health provider that you have angina?*
sd4_14a	*How long ago, in months or years, were you told by a health provider that you have angina?*
sd4_14b	*How long ago, in months or years, were you told by a health provider that you have angina?*
sd4_15	*Are you currently taking medication for angina?*
sd4_16	*Have you ever been told by a health provider that you have angina?*
sd4_17	*What type of arthritis did they say you had?*
sd4_18a	*How long ago, in months or years, were you told by a health provider that you have arthritis?*
sd4_18b	*How long ago, in months or years, were you told by a health provider that you have arthritis?*
sd4_19	*Are you currently taking medication for arthritis?*
sd4_20	*Have you ever been told by a health provider that you have asthma?*
sd4_21a	*How long ago, in months or years, were you told by a health provider that you have asthma?*

sd4_21b	*How long ago, in months or years, were you told by a health provider that you have asthma?*
sd4_22	*Are you currently taking medication for asthma?*
sd4_23	*Have you ever been told by a health provider that you have depression?*
sd4_24a	*How long ago, in months or years, were you told by a health provider that you have depression?*
sd4_24b	*How long ago, in months or years, were you told by a health provider that you have depression?*
sd4_25	*Are you currently taking medication for depression?*
sd4_26	*Have you ever been told by a health provider, including an optician, that you have a cataract in one or both of your eyes (that is, an opacity in the lens of the eye)?*
sd4_27	*Have you ever had eye surgery to remove your cataract(s)?*
sd4_28	*Have you ever had your eyes checked by a health provider, including an optician?*
sd4_29a	*How long ago, in months or years, was your vision checked by a health provider?*
sd4_29b	*How long ago, in months or years, was your vision checked by a health provider?*
sd4_30	Do you wear glasses or contact lenses?
sd4_31	Wearing your glasses or contact lenses, how much difficulty do you have in seeing and recognizing a person you know across the road (i.e. from a distance of about 20 meters)?
sd4_32	Wearing your glasses or contact lenses, how much difficulty do you have in seeing and recognizing an object at arm's length or in reading?
sd4_33	If you are NOT wearing glasses or contact lenses, how much difficulty do you have in seeing and recognizing a person you know across the road (i.e. from a distance of about 20 meters)?
sd4_34	If you are NOT wearing your glasses or contact lenses, how much difficulty do you have in seeing and recognizing an object at arm's length or in reading?
sd4_35	Does respondent use glasses/contacts or not (for skip pattern)
sd4_36	How much difficulty do you have in seeing and recognizing a person you know across the road (i.e. from a distance of about 20 meters)?
sd4_37	How much difficulty do you have in seeing and recognizing an object at arm's length or in reading?
sd4_38	*Have you ever had your hearing checked by a health provider?*
sd4_39a	*How long ago, in months or years, was your hearing last checked by a health provider?*
sd4_39b	*How long ago, in months or years, was your hearing last checked by a health provider?*
sd4_40	Do you have deafness or trouble hearing in one or both ears without the help of a hearing aid?
sd4_41	*Are you currently wearing a hearing aid?*
sd4_42	*Do you have trouble hearing in one or both ears even with a hearing aid?*
sd5_1	Have you had productive cough for at least two weeks in a year especially in the cold seasons?
sd5_2	Have you had shortness of breath during the time you had a productive cough?
sd5_3	Have you had non-productive cough?
sd5_4	Have you had attacks of shortness of breath?
sd5_5	Have you had shortness of breath that gets worse when you lie down, like during sleep?
sd5_6	Have you had wheezing?
sd5_7	Have you had chest pain?

sd5_8	Have you had swelling around your ankle?
sd5_9	Have you had a period lasting several days when you felt sad, empty or depressed?
sd5_10	Did this period last more than 2 weeks?
sd5_11	Did this period last most of the day?
sd5_12	Was this period nearly every day?
sd5_13	During this period, did your appetite increase or decrease?
sd5_14	During this period, did you lose or gain weight without it being your intention?
sd5_15	During this period did you notice any slowing down in your thinking?
sd5_16	During this period, did you have insomnia or sleep excessively most of the time?
sd5_17	During this period, did you feel tired and without energy all of the time?
sd5_18	During this period, did you feel guilty or useless?
sd5_19	During this period, did you have trouble concentrating or making decisions?
sd5_20	During this period, did you want to hurt yourself or be dead, or did you think of how to kill yourself or commit suicide?
sd5_21	Have you had a period lasting several days when you lost interest in most things that you usually enjoy such as pastimes, relationships, or work?
sd5_22	Did this period last more than 2 weeks?
sd5_23	Did this period last most of the day?
sd5_24	Was this period nearly every day?
sd5_25	During this period, did your appetite increase or decrease?
sd5_26	During this period, did you lose or gain weight without it being your intention?
sd5_27	During this period did you notice any slowing down in your thinking?
sd5_28	During this period, did you have insomnia or sleep excessively most of the time?
sd5_29	During this period, did you feel guilty or useless?
sd5_30	During this period, did you have trouble concentrating or making decisions?
sd5_31	During this period, did you want to hurt yourself or be dead, or did you think of how to kill yourself or commit suicide?
sd5_32	During this period, did you feel tired and without energy all of the time?
sd5_33	During this period, did those depressive symptoms cause signifiant discomfort or make it difficult to work or socialize, or affect your life in general in any other way?
sd5_34	During this period, were those symptoms caused completely by the loss of a loved one?
sd5_35	During this period, were those symptoms similar to those that someone in similar circumstances would experience?
sd5_36	During this period, do you remember having taken any medicine or drug right before or associated with the start of those depressive symptoms?
sd5_37	During this period, do you remember having suffered from or acquired an illness just before or associated with the beginning of those depressive symptoms?
sd5_38	Do you know if any family member such as your daughter, son, mother, father, grandfather or grandmother suffered from or were treated for depression at any point?
sd5_39	Do you have difficulty following a conversation in a noisy environment?
sd5_40	Are you able to hear out of both of your ears?
sd5_41	Are you able to hear when you are using a phone?
sd5_42	Do you have ringing in the ears?

sd5_43	Have you ever experienced back pain (including disc problems) during the last 30 days?
sd5_44a	How many days did you have this back pain for during the last 30 days? [specify]
sd5_44b	How many days did you have this back pain for during the last 30 days?
sd5_45	Have you experience joint inflammation in a symmetrical pattern (both sides of the joint affected rather than just one side)?
	During the last 12 months, have you experienced pain, aching, stiffness or swelling in or around the joint which was not related to an injury and lasted for more than a month?
sd5_46a	…A (neck)
sd5_46b	…B (right shoulder)
sd5_46c	…C (left shoulder)
sd5_46d	…D (right elbow)
sd5_46e	…E (left elbow)
sd5_46f	…F (right hand)
sd5_46g	…G (left hand)
sd5_46h	…H (right hip)
sd5_46i	…I (left hip)
sd5_46j	…J (right knee)
sd5_46k	…K (left knee)
sd5_46l	…L (right ankle)
sd5_46m	…M (left ankle)
sd5_46n	…N (left foot)
sd5_46o	…O (right foot)
sd5_46p	…P (wrist)
sd5_46q	…Q (thumb)
sd5_46r	…R (pinkie finger, lower joint)
sd5_46s	…S (ring finger, lower joint)
sd5_46t	…T (middle finger, lower joint)
sd5_46u	…U (index finger, lower joint)
sd5_46v	…V (pinkie finger, upper joint)
sd5_46w	…W (index finger, upper joint)
sd5_46x	…X (ring finger, upper joint)
sd5_46y	…Y (middle finger, upper joint)
sd5_47	Stiffness in the joint in the morning after getting up from bed, or after a long rest of the joint without movement?
sd5_48	How long does this stiffness last?
sd5_49	Does this stiffness go away after exercise or movement in the joint?
sd5_50	Have you experienced attacks of wheezing or whistling breathing?
sd5_51	Attack of wheezing that came on after you stopped exercising or some other physical activity?
sd5_52	A feeling of tightness in your chest?
sd5_53	Waking up with a feeling of tightness in your chest in the morning or any other time?
sd5_54	An attack of shortness of breath that came on without obvious cause when you were not exercising or doing some physical activity?

sd5_55	Pain or discomfort in your chest when you walk uphill or hurry?
sd5_56	Pain or discomfort in your chest when you walk at an ordinary pace on level ground?
sd5_57	Chest discomfort or pain for skip pattern
sd5_58	What do you do if you get the pain or discomfort when you are walking?
sd5_59	If you stand still, what happens to the pain or discomfort? Is it…
	Will you show me where you usually experience the pain or discomfort?
sd5_60a	…A: right shoulder
sd5_60b	…B: right side chest
sd5_60c	…C: neck area
sd5_60d	…D: upper middle chest
sd5_60e	…E: lower middle chest
sd5_60f	…F: left side chest
sd5_60g	…G: left shoulder
sd5_60h	…H: abdomen
sd5_61	In the past 12 months have you experienced cloudy or blurry vision?
sd5_62	Vision problems with light, such as seeing glare from bright lights, or seeing halos around lights?
sd5_63	Have you noticed, or has anyone told you that you have whiteness in your eye?
sd5_64	Have you passed dark urine during the past two weeks (dark yellow or plain tea color)?
sd5_65	Did you have icterus (yellow tinge in your body, especially the conjunctiva, palms and skin) during the past two weeks?
sd5_66	Was your skin itchy during the past two weeks?
sd5_67	Did you have malena (dark brown or black stools) during the past two weeks?
sd5_68	Have you ever vomited blood (haematemesis)?
sd5_69	Have you noticed abdominal enlargement during the past two weeks?
sd5_70	Did you notice swelling around your ankles during the past two weeks?
sd5_71	*Have you ever had hepatitis in your life or has a diagnosis of hepatitis ever been made by a health provider?*
sd6_1	*Do you have any medicines in the house that a health provider has prescribed for you or given to you?*
sd6_2	*May I see what medicines you personally have been using in the last 2 weeks?*
sd6_3	*Name of prescription from bottle/label:*
sd6_4a	*How frequently do you use this medicine?*
sd6_4b	*How frequently do you use this medicine?*
sd6_5	*Name of prescription from bottle/label:*
sd6_6a	*How frequently do you use this medicine?*
sd6_6b	*How frequently do you use this medicine?*
sd6_7	*Name of prescription from bottle/label:*
sd6_8a	*How frequently do you use this medicine?*
sd6_8b	*How frequently do you use this medicine?*
sd6_9	*Name of prescription from bottle/label:*
sd6_10a	*How frequently do you use this medicine?*

sd6_10b	*How frequently do you use this medicine?*
sd6_11	*Name of prescription from bottle/label:*
sd6_12a	*How frequently do you use this medicine?*
sd6_12b	*How frequently do you use this medicine?*
sd7_1	*Is there anything the research team should be aware of about this respondent that may have affected the quality of data?*

Table 2: Cutoff for each continuous/duration questionnaire item

Item	Question	Cutoff
sd4_02b	How long ago, in months or years, were you told by a health provider that you have chronic bronchitis, emphysema, or chronic obstructive pulmonary disease (COPD)?	9.25 years
sd4_05a	How long ago, in months or years, were you told by a health provider that you have heart failure?	1 year
sd4_11b	How long ago, in months or years, were you told by a health provider that you have liver failure?	5.25 years
sd4_14b	How long ago, in months or years, were you told by a health provider that you have angina?	4.7 years
sd4_18b	How long ago, in months or years, were you told by a health provider that you have arthritis?	9.75 years
sd4_21b	How long ago, in months or years, were you told by a health provider that you have asthma?	17.75 years
sd4_24b	How long ago, in months or years, were you told by a health provider that you have depression?	5.4 years
sd4_29b	How long ago, in months or years, was your vision checked by a health provider?	3 years
sd4_39b	How long ago, in months or years, was your hearing last checked by a health provider?	3.4 years

Table 3: Characteristics of the study participants for each condition

Condition	Number of interviews	Mean age	Standard deviation age	% Male	% Female
Angina pectoris	107	62.7	11.9	69%	31%
Asthma	117	42.7	12.8	26%	74%
COPD	108	69.1	11.0	43%	57%
Cataracts	108	68.1	13.4	38%	62%
Cirrhosis	104	51.4	11.3	81%	19%
Control	198	40.3	6.0	14%	86%
Depression	100	39.4	15.2	30%	70%
Hearing loss	205	47.3	9.1	29%	71%
Osteoarthritis	107	62.1	11.5	20%	80%
Rheumatoid arthritis	119	52.1	12.3	8%	92%
Vision loss	106	54.9	16.8	39%	61%

Table 4: Mean chance-corrected concordance across causes

Method	HCE	Chance-corrected concordance
Tariff	no HCE	0.534 (0.532, 0.539)
	with HCE	0.661 (0.656, 0.665)
SSP	no HCE	0.539 (0.535, 0.543)
	with HCE	0.679 (0.674, 0.683)
Random Forest	no HCE	0.524 (0.521, 0.527)
	with HCE	0.669 (0.665, 0.673)

Table 5: Cause-specific chance-corrected concordance

	Tariff		SSP		RF	
Chance-corrected concordance	*no HCE*	*with HCE*	*no HCE*	*with HCE*	*no HCE*	*with HCE*
Depression	0.803	0.887	0.851	0.8765	0.798	0.846
Angina pectoris	0.529	0.839	0.676	0.936	0.61	0.871
Cirrhosis	0.686	0.932	0.697	0.913	0.675	0.933
Rheumatoid arthritis	0.511	0.664	0.550	0.904	0.434	0.668
Asthma	0.467	0.775	0.607	0.878	0.522	0.817
Control	0.56	0.581	0.504	0.514	0.45	0.487
COPD	0.275	0.529	0.283	0.565	0.358	0.644
Hearing loss	0.503	0.511	0.509	0.528	0.422	0.470
Cataracts	0.358	0.400	0.381	0.579	0.374	0.484
Vision loss	0.422	0.391	0.203	0.184	0.278	0.331
Osteoarthritis	0.236	0.358	0.130	0.259	0.296	0.401

Table 6: Median CSMF accuracy with and without HCE information

Method	HCE	CSMF accuracy
Tariff	no HCE	0.772 (0.765, 0.779)
	with HCE	0.826 (0.818, 0.834)
SSP	no HCE	0.764 (0.753, 0.771)
	with HCE	0.817 (0.809, 0.828)
Random Forest	no HCE	0.775 (0.767, 0.781)
	with HCE	0.822 (0.814, 0.829)
King-Lu	no HCE	0.736 (0.729, 0.745)
	with HCE	0.765 (0.758, 0.772)

Table 7: Cause-specific prevalence fraction error

Prevalence fraction absolute error	Tariff		SSP		RF		KL	
	no HCE	*with HCE*	*no HCE*	*with HCE*	*no HCE*	*with HCE*	*no HCE*	*with HCE*
Asthma	0.025	0.014	0.034	0.016	0.031	0.015	0.038	0.033
Cirrhosis	0.027	0.016	0.016	0.007	0.021	0.015	0.042	0.036
COPD	0.031	0.021	0.034	0.024	0.034	0.023	0.048	0.043
Angina pectoris	0.030	0.020	0.026	0.024	0.032	0.026	0.044	0.039
Osteoarthritis	0.033	0.029	0.039	0.031	0.038	0.032	0.044	0.043
Cataracts	0.032	0.027	0.032	0.031	0.036	0.029	0.045	0.042
Rheumatoid arthritis	0.024	0.017	0.037	0.029	0.031	0.019	0.037	0.031
Vision loss	0.047	0.038	0.036	0.030	0.049	0.037	0.043	0.043
Depression	0.030	0.026	0.059	0.026	0.032	0.021	0.016	0.016
Hearing loss	0.044	0.030	0.032	0.028	0.032	0.033	0.038	0.034
Control	0.043	0.030	0.029	0.025	0.030	0.027	0.030	0.029

Table 8: Results from true versus estimated prevalence fraction linear regressions

Model	HCE	Condition	Coefficient	Intercept	RMSE	R^2
Tariff	no HCE	Angina pectoris	0.49	0.03	0.02	0.67
		Osteoarthritis	0.24	0.03	0.02	0.36
		Rheumatoid arthritis	0.51	0.02	0.02	0.73
		Asthma	0.45	0.03	0.02	0.64
		Cataracts	0.34	0.03	0.01	0.62
		COPD	0.34	0.02	0.01	0.64
		Cirrhosis	0.64	0.04	0.02	0.75
		Depression	0.74	0.06	0.02	0.69
		Hearing loss	0.30	0.14	0.03	0.15
		Vision loss	0.35	0.06	0.02	0.36
		Control	-0.23	0.45	0.04	0.00
	with HCE	Angina pectoris	0.80	0.02	0.01	0.89
		Osteoarthritis	0.35	0.03	0.02	0.49
		Rheumatoid arthritis	0.62	0.02	0.02	0.80
		Asthma	0.73	0.02	0.01	0.87
		Cataracts	0.41	0.02	0.01	0.78
		COPD	0.52	0.02	0.01	0.78
		Cirrhosis	0.87	0.03	0.02	0.86
		Depression	0.80	0.05	0.02	0.72
		Hearing loss	0.34	0.13	0.03	0.20
		Vision loss	0.35	0.04	0.02	0.36
		Control	-0.34	0.50	0.04	0.00
SSP	no HCE	Angina pectoris	0.62	0.03	0.02	0.75
		Osteoarthritis	0.12	0.03	0.02	0.11
		Rheumatoid arthritis	0.49	0.05	0.02	0.50
		Asthma	0.56	0.04	0.02	0.59
		Cataracts	0.38	0.03	0.02	0.59
		COPD	0.28	0.02	0.01	0.50
		Cirrhosis	0.68	0.02	0.01	0.83
		Depression	0.72	0.07	0.02	0.68
		Hearing loss	0.38	0.15	0.04	0.21
		Vision loss	0.22	0.04	0.02	0.26
		Control	-0.90	0.75	0.04	0.01
	with HCE	Angina pectoris	0.86	0.02	0.01	0.91
		Osteoarthritis	0.24	0.03	0.02	0.34
		Rheumatoid arthritis	0.79	0.04	0.02	0.74
		Asthma	0.83	0.02	0.01	0.90
		Cataracts	0.53	0.03	0.02	0.66
		COPD	0.53	0.03	0.02	0.69
		Cirrhosis	0.85	0.01	0.01	0.95

		Depression	0.80	0.05	0.02	0.74
		Hearing loss	0.36	0.16	0.04	0.18
		Vision loss	0.20	0.03	0.02	0.24
		Control	-0.99	0.78	0.04	0.01
	no HCE	Angina pectoris	0.53	0.04	0.02	0.69
		Osteoarthritis	0.23	0.04	0.02	0.24
		Rheumatoid arthritis	0.37	0.04	0.02	0.46
		Asthma	0.47	0.04	0.02	0.63
		Cataracts	0.33	0.04	0.02	0.44
		COPD	0.38	0.03	0.02	0.55
		Cirrhosis	0.61	0.03	0.02	0.77
		Depression	0.68	0.06	0.02	0.72
		Hearing loss	0.23	0.14	0.04	0.09
		Vision loss	0.19	0.08	0.03	0.09
RF		Control	-0.42	0.47	0.04	0.00
	with HCE	Angina pectoris	0.80	0.04	0.02	0.83
		Osteoarthritis	0.37	0.04	0.02	0.44
		Rheumatoid arthritis	0.60	0.02	0.01	0.83
		Asthma	0.72	0.02	0.01	0.88
		Cataracts	0.42	0.02	0.01	0.71
		COPD	0.59	0.02	0.01	0.80
		Cirrhosis	0.83	0.03	0.02	0.85
		Depression	0.75	0.05	0.02	0.72
		Hearing loss	0.16	0.17	0.04	0.04
		Vision loss	0.28	0.06	0.02	0.22
		Control	-0.30	0.42	0.04	0.00
	no HCE	Angina pectoris	0.20	0.04	0.02	0.24
		Osteoarthritis	0.13	0.04	0.02	0.08
		Rheumatoid arthritis	0.33	0.03	0.02	0.35
		Asthma	0.26	0.05	0.02	0.30
		Cataracts	0.09	0.05	0.02	0.06
		COPD	0.16	0.04	0.02	0.15
		Cirrhosis	0.15	0.05	0.02	0.19
		Depression	0.66	0.01	0.02	0.73
KL		Hearing loss	0.24	0.15	0.04	0.09
		Vision loss	0.17	0.06	0.02	0.13
		Control	0.40	0.16	0.07	0.00
	with HCE	Angina pectoris	0.27	0.03	0.01	0.42
		Osteoarthritis	0.15	0.04	0.02	0.13
		Rheumatoid arthritis	0.43	0.03	0.02	0.52
		Asthma	0.32	0.04	0.02	0.47
		Cataracts	0.14	0.05	0.02	0.15

	COPD	0.22	0.04	0.02	0.26
	Cirrhosis	0.25	0.05	0.01	0.45
	Depression	0.66	0.01	0.02	0.73
	Hearing loss	0.28	0.14	0.04	0.10
	Vision loss	0.15	0.05	0.02	0.14
	Control	0.59	0.08	0.07	0.00

Table 9: Confusion matrix for true and estimated cause classifications

Predicted cause:	Angina pectoris	Osteoarthritis	Rheumatoid arthritis	Asthma	Cataracts	COPD	Cirrhosis	Depression	Hearing loss	Vision loss	Control	**Total**
True cause:												
Angina pectoris	24	1	0	0	0	0	0	0	1	0	1	**27**
Osteoarthritis	2	8	6	0	4	0	0	1	2	0	1	**24**
Rheumatoid arthritis	1	4	19	0	1	0	1	2	0	0	0	**28**
Asthma	0	1	0	15	0	1	0	0	1	0	0	**18**
Cataracts	1	1	0	0	10	4	1	2	1	5	2	**27**
COPD	1	0	0	1	1	2	0	0	0	0	0	**5**
Cirrhosis	0	0	0	0	0	0	20	0	0	2	1	**23**
Depression	0	0	0	0	0	0	0	11	0	1	0	**12**
Hearing loss	0	0	0	0	0	0	0	5	5	0	2	**12**
Vision loss	3	0	0	0	8	0	1	1	3	8	0	**24**
Control	2	0	0	3	1	0	0	3	21	0	20	**50**
Total	**34**	**15**	**25**	**19**	**25**	**7**	**23**	**25**	**34**	**16**	**27**	**250**

Table 10: Chance-corrected concordance for 9-cause aggregation using the Tariff Method

	Tariff	
Chance-corrected concordance	*no HCE*	*with HCE*
Cirrhosis	0.824	0.906
Depression	0.875	0.887
Angina pectoris	0.801	0.841
Asthma	0.689	0.764
Arthritis	0.634	0.703
Control	0.565	0.578
Vision loss or cataracts	0.570	0.574
COPD	0.416	0.495
Hearing loss	0.471	0.479
Mean CCC	0.650	0.692

Table 11: Prevalence fraction absolute error and CSMF accuracy for 9-cause Tariff Method aggregation

	Tariff	
CSMF absolute error	*no HCE*	*with HCE*
Cirrhosis	0.013	0.013
Asthma	0.017	0.014
COPD	0.025	0.024
Angina pectoris	0.030	0.027
Arthritis	0.035	0.029
Depression	0.032	0.029
Control	0.036	0.031
Hearing loss	0.035	0.032
Vision loss or cataracts	0.038	0.036
CSMF Accuracy	0.842	0.858

Table 12: Performance comparison of SD methods to literature-based approaches

Condition	Test	Absolute Error	Top SD Method
Asthma	World Health Survey (WHS): Doctor (MD) diagnosis (Dx)	0.023 (0.020, 0.025)	**Tariff with HCE:** 0.014 (0.012, 0.016)
	WHS MD Dx OR asthma medications (Rx)	0.023 (0.020, 0.025)	
	WHS MD Dx OR asthma Rx OR wheezing/whistling attacks	0.092 (0.087, 0.095)	
	Wheezing/whistling attacks	0.060 (0.055, 0.066)	**Tariff with no HCE:** 0.025 (0.023, 0.029)
Angina pectoris	Rose questionnaire	0.082 (0.073, 0.088)	**Tariff with HCE:** 0.020 (0.018, 0.022)
Depression	CIDI questionnaire	0.059 (0.054, 0.064)	**KL with HCE:** 0.016 (0.015, 0.017)

Annex Files

Annex File 1: Informed consent letter obtained prior to the interview.

Informed Consent Form

"Estimating prevalence of chronic conditions through population based surveys"

October 2, 2009

The National Institute of Public Health is conducting a research study in collaboration with some Mexico City hospitals, regarding the prevalence of certain diseases in Mexico City. We are working to better understand the number of ill people suffering from a particular chronic disease or condition in the city. Based on medical records from the hospital, we are aware that you recently had a medical exam. It is important to clarify that the team I work with, including myself, does not have any information on any diagnosis that may have been given to you. We ask that you permit us to read the following information to explain the study.

Study Objective:
This study strives to develop better instruments and methods for measuring population health, particularly in resource-poor settings. The objective of this study is to accurately measure how common certain chronic disease are in a given population, without diagnostic testing in hospitals or taking biological samples at the time of the survey.

Type of Participation:
If you agree to participate, we will ask you a series of questions related to your health in the last 12 months as well as some socio-demographic questions.

Risk of your participation:
The questionnaire contains questions regarding symptoms you may have experienced in the past 12 months. It is possible that answering personal questions about your health may make you feel uncomfortable. You have the right to decline to answer any questions that make you feel uncomfortable.

Benefits of your participation:
You will not receive any direct benefit, economic or other compensation for your participation in this study, but the information that you provide will allow us to develop this questionnaire to understand the prevalence of chronic conditions in other settings where this information is not available through other sources. Understanding the major chronic conditions in a community will help policy makers and public health officials better allocate resources to provide health interventions for the most pressing health concerns. If you feel the need for emotional support because of emotional tension provoked by the interview, we will provide information for where you can receive free emotional support.

Cost of your participation:
Participation in this study will not cost you anything.

Confidentiality of the information you provide:
Privacy is of great importance to us. Information that you provide us will be kept strictly confidential. Results of the information you provide us may be published in medical literature, but your name will not be revealed. The information collected during the study will be transferred to an electronic database in which your name will be substituted for a record number or identification number.

Voluntary Participation / Withdrawal:
Your participation in the study is completely voluntary. You have the right to suspend the interview in any moment that you wish, or decline to answer questions that make you uncomfortable. You may also decide to withdraw your participation at any time. Your decision to participate, or not, will in no way affect the way you are treated for health services.

Thank you very much for your attention. If you have any doubts or would like clarification with regards to the study, I will leave a card with contact information for the researcher responsible for the study as well as the ethics commission.

Your signature here indicates your willingness to participate in the study.

Participant Name

Participant Signature Date

Annex File 2: Dictionary developed for verbal autopsy analysis

Identified:	Replaced with:
BAPA	back pain
BaP	
LBP	
LBPS	
ALBP	
CHLBP	
CLBP	
LUMBAGO	
Low Back Pain	
Low Back Pain with Sciatica	
Acute Low Back Pain	
Chronic Low Back Pain	
IBP	inflammatory back pain
BLNE	blindness
CRB	congenital retinal blindness
XN	night blindness
ABLEPSY	blindness
CONY	convulsion
COVUJ	
COVUJS	
conv	
GTCC	
FC	febrile convulsions
FCs	
C	cyanosis
CYA	
cy	
CN	
BID	brought in dead
FILAR	elephantiasis
GAGR	gangrene
MORMAL	
MORTIFICATION	
LL	elephantiasis
LLp	
LLs	
RLL	

ILBP	inflammatory low back pain
SID	sudden inexplicable death
DIC	disseminate intravascular coagulopathy
IOM	maternal death
AFFP	
EP	
HRP	
IPs	
IUP	
IUPD	
IUPX	
PBCS	
PDCS	
PPCS	
PR	
Pg	
UWP	
grav I	
grav II	
preg	
APH	maternal death due to hemorrhage
PPH	
PPH/DIC	
MH	
HALO	
LPPH	
PPH	
Placenta	
Placenta accreta	
Placenta praevia	
placenta previa	
PPT	maternal death due to sepsis
SEPS	
MILK LEG	
post mtp with sepsis	
PE	hypertensive disorder(maternal)
PPIH	
PRCLA	
TOXEMIA OF PREGNANCY	

PE	
PPIH	
PRCLA	
LPT	
PPIH	
Proteinuric Pregnancy-Induced Hypertension	
SHSP	
SPIH	
Superimposed Pregnancy Induced Hypertension	
TOX	
eclampsia	
OL	obstructed labor
MD&IDA	maternal death due to anemia
MAP	
PAM	other defined causes of death as a consequence of pregnancy
IOM	
AFILP	
AFLP	
AHPP	
APLF	
APO	
ECPR	
ICP	
PRAD	
PT	
PUL	
TUPG	
UEP	
VHP	
VIG	
UMD	unspecified maternal death
IFID	infectious diseases
AAIO	
AID	
APGN	
HID	
ID	
INDSE	
INF	
INFOUS	

TIS	
amyloidosis	
ARE	aids
HIV	
AIDS	
ISS	
iris/irts	
HIV/AIDS	
STD/HIV	
VIH	
HIVR	
KS	
CM	
CRME	
PJP	
PCP	
PCP	
AIDS	
IFID	
AIDS-KS	
AIDS-NHLS	
HIV	
AIDS-SI	
HIV+	
HIVE	
HIVP	
SIDS	
AIDS with TB	aids with tb
ADD	diarrheal diseases
DEH20	
ADI	
ADS	
BWD	
D	
D&D	
WDS	
Enteritis	
Colitis	
Acute Diarrheal Illness	
Diarrhea and Dehydration	

Bacillary White Diarrhea	
Watery Diarrhea Syndrome	
ENTERITIS	
Dehydration	
SUMMER COMPLAINT	
Cholera	
dysentery	
gastroenteritis	
ENTCOL	
NE	
NEC	
NNE	
NNEC	
PNE	
V&D	
G/E	
AING	
EID	
electrolyt	
acute colitia	
Diarrhea	
Dysentery	dysentery
ABP	pneumonia
ABR	
AFPP	
AIP	
AP	
BAPN	
CAP	
CAPN	
PERIPNEUMONIA	
LUNG FEVER	
WINTER FEVER	
ABPN	
CALP	
Flu	
LAGRIPPE	
GRIPPE/GRIP	
CATARRHAL	
GNBP	

H.flu	
HEIN	
HI	
HAP	
IPCP	
IPF	
IPN	
NIP	
NIPF	
klp	
KLPN	
KP	
KPCS	
KPN	
L/Pneumomia	
LAGI	
laryn	
laryn	
APHONIA	
CROUP	
MP	
NBP	
NOPN	
Nosocomial Pneumonia	
NPCP	
OPP	
OPPV	
PAP	
Ovine Progressive Pneumonia Virus	
PN	
PNA	
PNE	
PNM	
Pn	
Pnm	
pneu	
pneum	
PPB	
PPV	
RTI	

SCAP	
TWAR	
PIDS	
Pulmonary Infectious Diseases	
AIB	
IB	
IBV	
Infectious Bovine Rhinotracheitis Virus	
Infectious Bronchitis Coronavirus	
Infectious Bronchitis Virus	
PID	
AMT	tb
DTB	
ATB	
GDTB	
ITB	
EPTB	
FGTB	
GUTB	
UPTB	
TB	
PTB	
APT	
APTB	
PT	
T	
CONSUMPTION	
LONG SICKNESS	
LUNG SICKNESS	
Active Pulmonary Tuberculosis	
PHTHISIS	
KING'S EVIL	
POTT'S DISEASE	
WHITE SWELLING	
SCROFULA	
TB	
TB	
TB	
TB	
TB	

TBM	
TM	
tuberculoma	
AFM	malaria
BLACK FEVER	
BLACKWATER FEVER	
CM	
CMD	
MALA	
MALAR	
CONGESTIVE CHILLS	
CONGESTIVE FEVER	
REMITTING FEVER	
AGUE	
NCM	
PMCS	
malaria	
PIE	meningitis/encephalitis
ABM	meningitis
BAME	
ACM	
AVM	
CGM	
GNBM	
GNM	
MEGI	
men	
HS	
FMS	
ESC	
VHSV	
BLOOD POISONING	
BLOODY FLUX	
ADE	encephalitis
AES	
DROPSY OF THE BRAIN	
AHL	
E	
AHLE	
JE	

JBE	
ME	
PIE	
Encephalitis	
Other Infectious	other infectious
Neonatal tetanus	neonatal tetanus
AME	measles
AMED	
Measles	
WC	pertussis
MWC	
KRUCHHUSTEN	
TUSSIS CONVULSIVA	
CHIN COUGH	
Pertussis	
DEN	hemorrhagic fever (dengue)
DF	
DHF	
BREAKBONE	
PID	pelvic inflammatory disease
Other Childhood Infectious Diseases	other childhood infectious diseases
DIP	other infectious diseases
MALIGNANT SORE THROAT	
MEMBRANOUS CROUP	
PUTRID FEVER	
BLADDER IN THROAT	
SORE THROAT DISTEMPER	
APE	
CMVE	
CMV Encephalitis	
HZE	
DIC	
NP	
BLACK PLAGUE OR DEATH	
bubonic plague	
RAB	
HYDROPHOBIA	
CANINE MADNESS	
SF	
SCFE	

Scarlatina	
SP	
black small pox	
BLACK POX	
VARIOLA	
CS	
NS	
SY	
SYPH	
VD-S	
VDS	
L	
LAT	
Pestilence	
Lues	
BAD BLOOD	
FRENCH POX	
GREAT POX	
LUES DISEASE	
YF	
JYF	
YFMD	
AMERICAN PLAGUE	
BRONZE JOHN	
DOCK FEVER	
STRANGER'S FEVER	
YELLOWJACKET	
yellow fever	
GS	
CO	
cancrum oris	
gangrenous stomatitis	
HERPES	
Zoster	
Shingles	
Other Infectious Diseases	
AIM	
HIM	
IH	
IHH	

IM	
IMN	
SID	
Neonatal	neonatal conditions
SID	
SIDS	
FA	birth asphyxia
PA	
Birth asphyxia	
Pneumonia (Serious Infection)	pneumonia (serious infection)
neomonia	
Pneumonia and Diarrhea	pneumonia and diarrhea
Sepsis with Local Bacterial Infection	sepsis with local bacterial infection
IBNS	
NBIS	
AOP	preterm delivery (<33 weeks ga) without respiratory distress syndrome
NIA	
Preterm	
ARDS	respiratory distress syndrome (<33 wks ga)
RDS	
ARF	
Respiratory distress syndrome (<33 wks GA)	
Respiratory distress syndrome (33-36 wks GA)	
Sepsis (Serious Infection)	sepsis (serious infection)
sepsi	
sepsis	
Stillbirth	stillbirth
Preterm Delivery (without RDS) and Birth Asphyxia	preterm delivery (without rds) and birth asphyxia
Preterm Delivery (with or without RDS) and Sepsis	preterm delivery (with or without rds) and sepsis
Preterm Delivery (without RDS) and Sepsis and Birth Asphyxia	preterm delivery (without rds) and sepsis and birth asphyxia
Sepsis/Septicemia (without local bacterial infection)	sepsis/septicemia (without local bacterial infection)
Sepsis/Septicemia (with local bacterial infection)	sepsis/septicemia (with local bacterial infection)
BFC	other defined causes of child deaths
PARAXYSM	
BFNC	
BNFC	

NEC	
DDR	
VDDR	
VDR	
Other Defined Causes of Child Deaths	
hypoternia	
ASB	congenital malformation
SB	
SBA	
SBHC	
SBO	
HPH	
HYCEP	
IH	
malformation	
AHNC	malignant neoplasms
CA	
CAN	
SCIRRHUS	
cancerous tumors	
sol	
MLC	
Malignant Neoplasms	
metastasi	
carcinoma	
mets	
Mouth/Orapharynx Cancer	mouth/orapharynx cancer
Malignant neoplasm of digestive organs	malignant neoplasm of digestive organs
Esophageal Cancer	esophageal cancer
AGC	stomach cancer
Stomach Cancer	
ACC	colorectal cancer
ACRC	
CC	
Colorectal Cancer	
HCC	liver cancer
Liver Cancer	
Female genital cancer	female genital cancer
Cervical Cancer	cervical cancer

cavanoma of the cervix	
cevical	
ADOVCA	ovarian cancer
AEOC	
Advanced Epithelial Ovarian Cancer	
Ovarian Cancer	
Uterine Cancer	uterine cancer
ABC	breast cancer
BC	
BCA	
BRCR	
BrCa	
Lung Cancer	lung cancer
Prostate Cancer	prostate cancer
Leukemia/Lymphomas	leukemia/lymphomas
AML	
leukemia	
CLL	
Leukemia	
NHL	lymphomas
Lymphomas	
lymphonia	
BCC	other defined cancers
BCSC	
BLCA	
SCC	
Other Defined Cancers	
Cardiovascular Diseases	cardiovascular diseases
IHD	ihd
CAD	
infarct	
ischem	
cad	
cad	
IHD	ihd : acute myocardial infarction
CHD	
AMI	
MI	
ISHD	
MCI	

MYIN	
coronary heart disease	
Acute myocardial infarction	
myocardial infarction syndrome	
Acute Myocardial Infarction	
ACHF	ihd : congestive heart failure
ADHF	
AHF	
Acute Congestive Heart Failure	
CCF	
CHFX	
EDEMA OF LUNGS	
Dropsy of lungs	
CCHF	
CHF	
CLHF	
ESHF	
HF	
IHF	
LVH	
Left Ventricular Hypertrophy due to HTN	
RVF	
Heart Failure	
cardiomegali	
Inflammatory Heart Disease	inflammatory heart disease
ARHD	
DCM	cardiomyopathy
DCMP	
CCM	
EMF	
Congestive Cardio Myopathy	
Endo Myocardial Fibrosis	
Cardiomyopathy	
IE	endocarditis
ABE	
AIE	
BE	
ENCAR	
NBTE	
Acute Bacterial Endocarditis	

Acute Infective Endocarditis	
Bacterial Endocarditis	
Endocarditis	
Nonbacterial Thrombotic Endocarditis	
Endocarditis	
AIE	
IE	
Pericarditis	pericarditis
HCVD	stroke
APOPLEXY	
SOFTENING OF BRAIN	
HIE	
SAH	
SDH	
tia	
Stroke	
cerebr	
cerebrovascular	
cva	
subarachnoid	
cva	
AHPP	hypertension
BP	
HT	
HTN	
AHTD	
HTA	
blood pressure	
blood Hypertension	
Arterial Hypertension Disease	
Hypertension Artérielle (French)	
Hypertension	
MR	other specified cardiovascular diseases
RHD	
CRHD	
DRHD	
RF/RHD	
Active Rheumatic Heart Disease	
Chronic Rheumatic Heart Disease	
Decompensated Rheumatic Heart Disease	

Rheumatic Fever And Rheumatic Heart Disease	
DVT	
THROMBOSIS	
Other Specified Cardiovascular Diseases	
Respiratory Diseases	respiratory diseases
COPD	copd
Asthma	asthma
CFRD	diabetes
CFRDM	
DB	
DDM	
DIA	
DIAB	
DIMEL	
DM	
DKA	
DMKA	
Insulin-Dependent Diabetes Mellitus	
Diabetes Mellitus	
Diabetes Mellitus	
diabetes ketoacidosis	
Diabetes	
Diabetes with Coma	diabetes with coma
Diabetes with Renal Failure	diabetes with renal failure
nefropatia diabetica	diabetes with renal failure
Diabetes with Skin Infection/Sepsis	diabetes with skin infection/sepsis
Digestive Diseases	digestive diseases
AC	cirrhosis
ALC	
ALCIR	
Alc	
PNC-E	
Alcoholic Liver Cirrhosis	
Alcoholic Cirrhosis	
Postnecrotic cirrhosis-ethanol	
APBC	
CAHC	
CIC	
CILI	
CIR	

CIRR	
CL	
LC	
DLC	
FNC	
FATTY LIVER	
MNLC	
NALC	
NALD	
PC	
Cirrhosis	
liver	
cirrosis	
ABC	other specified digestive diseases
BC	
BICO	
Biliary Colic	
GACO	
an abdominal pain and cramping	
CRAMP COLIC	
I.O	
DIOS	
Distal Intestinal Obstruction Syndrome	
PUD	
Other Specified Digestive Diseases	
g.neurological conditions	neurological conditions
Dementia	dementia
ABFEC	epilepsy
BCE	
Benign Childhood Epilepsy	
EPI	
EPMR	
GE	
GM	
epil	
Epilepsy with Mental Retardation	
Generalized Epilepsy	
Grand Mal (Epilepsy)	
FALLING SICKNESS	
EEL THING	

IGE	
LOE	
PE	
PM	
PTEP	
SGE	
Epilepsy	
ARF	renal failure
ATN	
ANRF	
APORF	
IARF	
SARF	
Acute Non-Inflammatory Renal Failure	
Acute Postischemic Renal Failure	
Ischemic Acute Renal Failure	
Severe Acute Renal Failure	
CRF	
CHRF	
CRFX	
CTRF	
ESRF	
SCRF	
Chronic Terminal Renal Failure	
End Stage Renal Failure	
Severe Chronic Renal Failure	
BRIGHT'S DISEASE	
NEPRITIS	
NEPHROSIS	
EDEMA	
RF	
RIDM	
RNFA	
TRF	
esk	
ESRD	
Terminal Renal Failure	
End stage kidney	
End stage renal disease	
ARF	

RF	
nephropathi	
SCA	other non-communicable diseases
SCD	
SSA	
HbSS	
SS disease	
haemoglobin S	
Sickle Cell Disease	
drepanocytosis	
AA	
AAA	
ACD	
AHA	
APAN	
SAAA	
Acquired aplastic anemia	
Anemia of Chronic Disease	
Acquired hemolytic anemia-Acute hemolytic anemia	
Aplastic Anemia	
Severe Acquired Aplastic Anemia	
ANM	
AOCD	
BLA	
Anemia of Chronic Disease	
GREEN FEVER/SICKNESS	
Blood Loss Anemia	
HA	
HHA	
Hereditary Hemolytic Anemia	
IDA	
SIDA	
Severe Iron-Deficiency Anemia	
CHLOROSIS	
MA	
MAP	
Megaloblastic Anemia Of Pregnancy	
HHS	
DI	

PIG	
MF	
DMF	
Diphasic Milk Fever	
PUKING FEVER	
SLOES	
GBS	
BPH	
Non-communicable Diseases	
CNID	
metabolic encephalopathy	
ASPH	group 3 injuries
cyanotic (lack of oxygen)	
PHS	
ABIS	
AOI	
BBMI	
BCIS	
BCVI	
BICI	
CHI	
CHIS	
CSCI	
DOI	
H&S	
HFNI	
IAIS	
ICH	
ICP	
MSTI	
OHI	
SCIS	
SKFX	
SOIS	
STI	
STIS	
Soft Tissue Injuries	
UHS	
PTM	
injuri	

head	
head injury	
unintentional injuries	unintentional injuries
Bite of Venomous Animal	bite of venomous animal
bite	
bitten	
Drowning	drowning
Falls	falls
Fires	fires
burns	
burnt	
Poisonings	poisonings
MVTI	road traffic
PEDI	
RTA	
RTI	
RTIS	
Road traffic injuries	
Road Traffic Injuries	
Road Traffic	
Other Injuries	other injuries
Intentional injuries	intentional injuries
Homicide	homicide
Suicide	suicide

Annex File 3: Tariffs for each feature-cause combination

Values are rounded to the nearest 1 for clarity. Blank values indicate that the tariff for that feature-cause combination was not significantly different than 0. 0 values indicate that the tariff was less than 0.5 and was rounded to 0.

Question	Angina pectoris	Osteoarthritis	Rheumatoid arthritis	Asthma	Cataracts	COPD	Cirrhosis	Depression	Hearing loss	Vision loss	Control
Have you ever been told by a health provider that you have chronic bronchitis, emphysema, or chronic obstructive pulmonary disease (COPD)?				2	1	7					0
How long ago, in months or years, were you told by a health provider that you have chronic bronchitis, emphysema, or chronic obstructive pulmonary disease (COPD)?	1			2		5					
Are you currently taking medication for chronic bronchitis, emphysema, or chronic obstructive pulmonary disease (COPD)?				2	1	8					0
Have you ever been told by a health provider that you have heart failure?	6		1			2			1		1
How long ago, in months or years, were you told by a health provider that you have heart failure?	9					5					
Are you currently taking medication for heart failure?	9		1			2			1		1

Have you ever been told by a health provider that you have cirrhosis?						40				
Are you currently taking medication for cirrhosis?										
Have you ever been told by a health provider that you have liver failure?						36				
How long ago, in months or years, were you told by a health provider that you have liver failure?						29				
Are you currently taking medication for liver failure?						46				
Have you ever been told by a health provider that you have angina?	17									
How long ago, in months or years, were you told by a health provider that you have angina?	34			1	1					
Are you currently taking medication for angina?	40				1				1	
Have you ever been told by a health provider that you have angina?		8	12				1	1		
What type of arthritis did they say you had?		2		0			0		0	
What type of arthritis did they say you had?		6	20					1		1
What type of arthritis did they say you had?		9	1							
How long ago, in months or years, were you told by a health provider that you have arthritis?		6	13					1		
Are you currently taking medication for arthritis?		11	19				1	1		
Have you ever been told by a health provider that you have asthma?		0		10		3	0	1		0

How long ago, in months or years, were you told by a health provider that you have asthma?	1		8		2	1			
Are you currently taking medication for asthma?			22		6				
Have you ever been told by a health provider that you have depression?		1				1	9		0
How long ago, in months or years, were you told by a health provider that you have depression?						2	8		
Are you currently taking medication for depression?						1	11		1
Have you ever been told by a health provider, including an optician, that you have a cataract in one or both of your eyes (that is, an opacity in the lens of the eye)?				3	0	0	1	1	1	1
Have you ever had eye surgery to remove your cataract(s)?				3	1		0	0	2	0
Have you ever had your eyes checked by a health provider, including an optician?			1	0	1	1		0	1	0
How long ago, in months or years, was your vision checked by a health provider?							1	1	1	1
Do you wear glasses or contact lenses?								1	1	1
Wearing your glasses or contact lenses, how much difficulty do you have in seeing and recognizing a person you know across the road (i.e. from a distance of about 20 meters)?				1			1	1		0
Wearing your glasses or contact lenses, how much difficulty do you have in seeing and recognizing a person you know across the road (i.e. from a distance of about 20 meters)?										1

Wearing your glasses or contact lenses, how much difficulty do you have in seeing and recognizing a person you know across the road (i.e. from a distance of about 20 meters)?			1	
Wearing your glasses or contact lenses, how much difficulty do you have in seeing and recognizing a person you know across the road (i.e. from a distance of about 20 meters)?	2		2	
Wearing your glasses or contact lenses, how much difficulty do you have in seeing and recognizing an object at arm's length or in reading?	1		2	1
Wearing your glasses or contact lenses, how much difficulty do you have in seeing and recognizing an object at arm's length or in reading?				1
Wearing your glasses or contact lenses, how much difficulty do you have in seeing and recognizing an object at arm's length or in reading?	2		2	
Wearing your glasses or contact lenses, how much difficulty do you have in seeing and recognizing an object at arm's length or in reading?	1			
If you are NOT wearing glasses or contact lenses, how much difficulty do you have in seeing and recognizing a person you know across the road (i.e. from a distance of about 20 meters)?			2	

If you are NOT wearing glasses or contact lenses, how much difficulty do you have in seeing and recognizing a person you know across the road (i.e. from a distance of about 20 meters)?	2			1		1
If you are NOT wearing glasses or contact lenses, how much difficulty do you have in seeing and recognizing a person you know across the road (i.e. from a distance of about 20 meters)?						1
If you are NOT wearing glasses or contact lenses, how much difficulty do you have in seeing and recognizing a person you know across the road (i.e. from a distance of about 20 meters)?		2	1		2	1
If you are NOT wearing your glasses or contact lenses, how much difficulty do you have in seeing and recognizing an object at arm's length or in reading?				1		
If you are NOT wearing your glasses or contact lenses, how much difficulty do you have in seeing and recognizing an object at arm's length or in reading?				2		1
If you are NOT wearing your glasses or contact lenses, how much difficulty do you have in seeing and recognizing an object at arm's length or in reading?						1
If you are NOT wearing your glasses or contact lenses, how much difficulty do you have in seeing and recognizing an object at arm's length or in reading?				1	1	1

Does respondent use glasses/contacts or not (for skip pattern)										
How much difficulty do you have in seeing and recognizing a person you know across the road (i.e. from a distance of about 20 meters)?			1	1					1	3
How much difficulty do you have in seeing and recognizing a person you know across the road (i.e. from a distance of about 20 meters)?	1									
How much difficulty do you have in seeing and recognizing a person you know across the road (i.e. from a distance of about 20 meters)?		1							1	0
How much difficulty do you have in seeing and recognizing a person you know across the road (i.e. from a distance of about 20 meters)?				4	1			0		1
How much difficulty do you have in seeing and recognizing an object at arm's length or in reading?			1	1	0	1	1		0	2
How much difficulty do you have in seeing and recognizing an object at arm's length or in reading?	1	1			1			0		
How much difficulty do you have in seeing and recognizing an object at arm's length or in reading?					1		1	1		1
How much difficulty do you have in seeing and recognizing an object at arm's length or in reading?				2	1					1
Have you ever had your hearing checked by a health provider?			1			1		2		1

How long ago, in months or years, was your hearing last checked by a health provider?			1			1			1			1
Do you have deafness or trouble hearing in one or both ears without the help of a hearing aid?	1		0	1					1			1
Are you currently wearing a hearing aid?												
Do you have trouble hearing in one or both ears even with a hearing aid?												
How long ago, in months or years, were you told by a health provider that you have cirrhosis?					1	26				1		
Have you had productive cough for at least two weeks in a year especially in the cold seasons?	1		1		2				0			1
Did this period last more than 2 weeks?							6		1			2
Did this period last most of the day?							7		2			2
Was this period nearly every day?							8		2			2
During this period, did your appetite increase or decrease?							5		1			1
During this period, did you lose or gain weight without it being your intention?							7					1
During this period did you notice any slowing down in your thinking?							4		1			2
During this period, did you have insomnia or sleep excessively most of the time?		1					4		1			1
During this period, did you feel tired and without energy all of the time?		1					6		1			2
During this period, did you feel guilty or useless?							4		1			1

During this period, did you have trouble concentrating or making decisions?						7		1
Have you had shortness of breath during the time you had a productive cough?			6		6			1
During this period, did you want to hurt yourself o be dead, o did you think of how to kill yourself or commit suicide?		1				7	1	1
Have you had a period lasting several days when you lost interest in most things that you usually enjoy such as pastimes, relationships, or work?		1				5	1	1
Did this period last more than 2 weeks?		1		1		5	1	1
Did this period last most of the day?		1				5	1	1
Was this period nearly every day?	1	1				5		0
During this period, did your appetite increase or decrease?				1		5	1	1
During this period, did you lose or gain weight without it being your intention?				1		6	1	1
During this period did you notice any slowing down in your thinking?						5	1	1
During this period, did you have insomnia or sleep excessively most of the time?						5	1	1
During this period, did you feel guilty or useless?						6	1	1
Have you had non-productive cough?					1		1	
During this period, did you have trouble concentrating or making decisions?				1		6	1	1

During this period, did you want to hurt yourself o be dead, o did you think of how to kill yourself or commit suicide?			1		6	1	1
During this period, did you feel tired and without energy all of the time?					5	1	1
During this period, did those depressive symptoms cause signifiant discomfort or make it difficult to work or socialize, or affect your life in general in any other way?		1			6	0	1
During this period, were those symptoms caused completely by the loss of a loved one?	1				3		1
During this period, were those symptoms similar to those that someone in similar circumstances would experience?	2				4		
During this period, do you remember having taken any medicine or drug right before or associated with the start of those depressive symptoms?		1	1	1	2		
During this period, do you remember having suffered from or acquired an illness just before or associated with the beginning of those depressive symptoms?			1			1	1
Do you know if any family member such as your daughter, son, mother, father, grandfather or grandmother suffered from or were treated for depression at any point?					13		
Do you have difficulty following a conversation in a noisy environment?					2		2

Have you had attacks of shortness of breath?	1		0	2		2			1
Are you able to hear out of both of your ears?							1		1
Are you able to hear when you are using a phone?	2	2							1
Do you have ringing in the ears?		1				1			1
Have you ever experienced back pain (including disc problems) during the last 30 days?		1	2			1			
How many days did you have this back pain for during the last 30 days?		1	1	0		1			1
Have you experience joint inflammation in a symmetrical pattern (both sides of the joint affected rather than just one side)?		6	10						1
…A (neck)		1	2		0	1			0
…B (right shoulder)		1	1	0	0	1	0		1
…C (left shoulder)		1	1	0		1	0	0	0
…D (right elbow)		3	7			1			0
…E (left elbow)		2	6			1			
…F (right hand)		3	7					1	
…G (left hand)		3	7				0		
…H (right hip)		3	4						0
…I (left hip)		2	4						0
...J (right knee)		5	4	1					1
…K (left knee)		4	4	1					1
…L (right ankle)		2	6						1
…M (left ankle)		3	7						1
…N (left foot)		1	3		1	1			1
…O (right foot)		1	3		1	0			
…P (wrist)		3	5			1	1		1
…Q (thumb)		4	6				1		1
…R (pinkie finger, lower joint)		2	4			0	1		0
…S (ring finger, lower joint)		3	4			1	1		1
…T (middle finger, lower joint)		2	3			1	1		1
…U (index finger, lower joint)		2	4			1	0		1
…V (pinkie finger, upper joint)		2	4				0		1
…W (index finger, upper joint)		3	5			1			1

…X (ring finger, upper joint)		2	4				1	0		1
…Y (middle finger, upper joint)		2	4				1	0		1
Stiffness in the joint in the morning after getting up from bed, or after a long rest of the joint without movement?		3	4					1		1
How long does this stiffness last?		1	1		1			1		1
Does this stiffness go away after exercise or movement in the joint?		2	2					1		1
Have you had shortness of breath that gets worse when you lie down, like during sleep?	1			2		2		0		0
Have you experienced attacks of wheezing or whistling breathing?				6		4		1		1
Attack of wheezing that came on after you stopped exercising or some other physical activity?	1		0	2		2				0
A feeling of tightness in your chest?	1			1		1	1	0	0	0
Waking up with a feeling of tightness in your chest in the morning or any other time?	1			1		1		0	0	0
An attack of shortness of breath that came on without obvious cause when you were not exercising or doing some physical activity?	1			1		1				0
Pain or discomfort in your chest when you walk uphill or hurry?	2			1		1				0
Pain or discomfort in your chest when you walk at an ordinary pace on level ground?	2			1		1		0	0	0
Chest discomfort or pain for skip pattern	2			1		1				0
What do you do if you get the pain or discomfort when you are walking?	4					2				1

What do you do if you get the pain or discomfort when you are walking?	2			4		1						
What do you do if you get the pain or discomfort when you are walking?												
If you stand still, what happens to the pain or discomfort? Is it…	2			1		1						1
If you stand still, what happens to the pain or discomfort? Is it…	3			2								
Have you had wheezing?				5		3			1			2
…A: right shoulder	4											
…B: right side chest	2					1					1	
…C: neck area	1			1		1						0
…D: upper middle chest	2			1		1						0
…E: lower middle chest												
…F: left side chest	3		1									1
…G: left shoulder	6											
…H: abdomen												
In the past 12 months have you experienced cloudy or blurry vision?					1			1	1			2
Vision problems with light, such as seeing glare from bright lights, or seeing halos around lights?					2							1
Have you noticed, or has anyone told you that you have whiteness in your eye?					3				1	1		1
Have you passed dark urine during the past two weeks (dark yellow or plain tea color)?						1	4		1			1
Did you have icterus (yellow tinge in your body, especially the conjunctiva, palms and skin) during the past two weeks?		1					7					
Was your skin itchy during the past two weeks?							3		1			1
Did you have malena (dark brown or black stools) during the past two weeks?				1			5					

Have you ever vomited blood (haematemesis)?							14				
Have you noticed abdominal enlargement during the past two weeks?				1			4				1
Have you had chest pain?	1			1		1			0		0
Did you notice swelling around your ankles during the past two weeks?		1	1			0	1	0	1		1
Have you ever had hepatitis in your life or has a diagnosis of hepatitis ever been made by a health provider?							3	1			
Have you had swelling around your ankle?			1					1	1		1
Have you had a period lasting several days when you felt sad, empty or depressed?			1					5	1		1
Free Text Words:											
abdomen											
abil		1		1			1	1			1
abl	0										
absentmind											
accept	0						1	1	2		1
acetylsalicyl	2				1			0			0
acid			4	1				1			1
activ											
address			1	1		1		0			0
adepsiqu		1		1				4			
adjust											
affect											
affirm			1						2	1	
afraid							2		1		4
age											
ago				2				1	1		1
agre	0										1
ailment					7						
air											
alcohol											
aldacton											
allergi											
allopath				1		1					1

allopurinol												
altern												
ambroxol												
amlodipin					1	1						
amoxicillin												
ampicillin												
angina	41											
ankl										1		
answer	1							1	2	1	3	
anxious								1	1		1	
anymor					2							
appar												
appear							3					
appli				3								
appoint									3		4	
approach												
approxim												
aralen												
arava												
arm												
arriv												
arthriti		4	12					1	1		1	
aspirin	4				1	1						
assist												
assur												
asthma				15		3		1			0	
atorvastatin												
attack	6											
attend												
attent									11		7	
auditori												
autom												
avenu												
azulfidin												
bad												
beclomethason												
bed												
begin	1									1		
behav				1					3		7	
bezafibr	1									1		
bit									2		2	
block				2								

blood			1	1			1			
bodi					4					
bone										
breath				6						
brief		2		1						2
bromid										
bronchial			1	2		2				
bronchiti				2		2		1		1
budesonid										
bus										
buy										
buzz								2		
calcium		1	1							
call								3		2
calm										2
caltrat		10	11							
camino										
cancer										
capac									1	
captopril		1			1		0		1	0
carbamazepin										
cardiac	3									
care				1					1	
carri										
cartigen										
cartilag		6								
cataract					3	0	1	1	2	1
caus				2						
celebrex										
chang				4						
check					1			1		1
chest	1									
child										
chloroquin										
chronic				3		10		1		1
ciprofloxacin										
cirrhosi							57			
citalopram										
clidda										
cliida										
clinic					1			20		19
clonazepam										

clopidogrel												
close												
cold			4		3							
combiv				22		22						
comment						1	1		1			
commit			2						3		1	
complet												
complex							2			2		
complic												
condit			2		2	1						
confid												
connect												
consent									2		2	
consult												
contact												
continu									2		3	
control												
convinc											2	
cook												
cooper									1		1	
copd				3		11						
corneal												
correct		2		1	1							
cough											2	
cpod						10						
cri												
crisi				4								
current				4	4				1			
daili	2		2								1	
danger												
dark												
date			2		6				1			
daughter					5							
day						1			1		1	
death					3							
decemb												
deform		2	4							1		
deni												
depress			1					2	1		1	
depression								13				
despit									1			
detect	1								20		18	

develop		1		4		1					
devic											
dexamethason			1	3		2					
diabet					5	1		1		3	1
diagnos											
diagnosi									4	1	3
diazepam											
diclofenac		3	4	1							1
dicloxacillin											
die											
difficult			2						1		1
dinitr	4	1			1						
dipropion				2		2				1	
disappear											
discomfort											
diseas				1		2		1	1		1
diskus											
distanc		1						1	2		
distract											
doctor								1			
door											
doubt											
drawback											
drink							8				
drop			1		1					2	
due					1				1		
ear		1				0	0	0	2		
eat											
effort											
emphysema						4					
enalapril	2				1	1	0		0		0
enriqu											
enter							1		8		6
entir		2			8				1		
epoc											
especi											
espironolacton							7				
exercis											3
expens	1		1	3							
explain					1			1	1		2
explan											
explicit			1		4	1					

extrem											
eye					7			1		6	
eyesight					10					8	
failur	2					1	2		0		0
famili					3						
fast											
father											
februari											
feel				2				1			
feet											
fell					6						
felodipin											
final											
fine			1								
finger			4								1
finish									9		6
five											
fluoxetin	1							22		1	
folic			8				2			1	
follow											
form											
fort		3			1			1			
found											
frequent											
furosemid						2	5			1	
gallbladd	1	1			3						
gastriti		2									
genteal	1		2							2	
get											
girl											
glass					2			1	2		
glaucoma		2			1					2	
glibenclamid	1				2			0		1	0
glucosamin		6									
goe							1				
gone											
govern			1		1		2				
half		3									
hand			4								
happen											
hard								1			1
head											

health										2
healthi								1	2	7
hear		1		1		0	0	1		
heard								1	2	3
heart	5			1	2			0		0
help				4						
hepat						15				1
hernia										
hip		1	3	1						
home			2				1	0		
hospit			0			1	1	1		
hour										
hous			1				1	1		1
hurt							1			
husband										
hypertens	1			1						
ibuprofen			7							
ill			1						1	
immedi						2		1		3
improv					1			2	1	
includ										
inconveni										
indomethacin		2	2	1						
inform										
inject										
injuri										
insid										
insist										
institut								1		2
insulin				1		1			1	
interrupt								1		3
interview		0		1			0	2	0	2
interviewe			1	1			1	1	1	
intraocular				3						
ipratropium			2		11	1				
isosorbid	7									
issst										
itch								2		
jan			3	5				1		
job	1		2		1					
joint		1	2							
juarez										

ketorolac											
kidney											
kind				1					2		4
knee		4	2								
lack											
lactulax											
ladi	0	1	1		1		0	0			
last			3						1		2
late											
leav											
left								1	2		
leg											
len					4						
letter								1	2		4
levothyroxin											
light											1
lipitor	16										
liquid		1					7			1	
listen											2
littl								1			2
live			1								
liver							12				
locat				1							
look								1			
lopez											
loratadin											
losartan	5										
lose									1	2	
loss									2		2
lost					2					2	
lot											
lung											
main											
major											
make											
market											
mateo											
meaning											
medic				1	0				1	0	1
medicin								1			
meloxicam											
mental		2			1					1	

mention							1	1			
metfomin											
metformin					3			0	0	1	1
methotrex		1	46							1	
meticorten			2	2	1						
metoclopramid				3							
metoprolol	6			0			0				0
metroprolol	4				3						
micardi											
miflonid											
mind											1
minibus											
minut									1		
misgiv											
mishap											2
miss								2			3
moment											26
money			3								
month											2
moreov					3				1		
morn											
mother											
move											
murmur											
name								1			1
naproxen											
nation											
neighborhood		3	1							1	
nevertheless					3						
nice			1		1		0	1			1
nifedipin						3					
night						10				4	
nimesulid											
notic									3		
numb			4								
number											
nurs											
object					2				2		1
observ					4						
obstruct				9		30					
octob											
omeprazol	1						6				1

one										
oper		1		2		0	1		1	0
osteoarthr		7	1							
osteoporosi										
oxygen										
paid										
pain		5	9							
paracetamol			2	1						
paralysi										
paroxetin										
particip	1						1	2		2
pass										2
past										
pectori	36									
penicillin				1					1	2
pentoxifyllin					2					
peopl										
perform				2					1	1
period								1		
person										1
phone										2
piec			3	4						1
pill			3							
piroxicam										
plaquenil		1	27	1						
plavix										
polit										7
pravastatin	7								3	
prednison			3	2	1					
prescrib										
prescript										
pressur									2	
pressure	3	1							1	
propanolol										
proper										
propranolol						21	3			
prosthesi		2	2	1						
protect	5			1	1	0	0			
public										
pulmonari			7		26					
question							1	1	1	1
questionnair										3

quick										2				4
ranitidin														
ray														
read										3				1
realiz														2
reason														2
receiv														
recent														
regist														
relat														
relev								1		3				1
rememb					2				1					
repeat														
respiratori				3		5								
respond										1				
rest														
restless														
result														3
retain				1		2	8							
rheumatoid		5	20							1				1
rivotril					3									
rush														2
sad					3									
salbutamol				9	1	4								
schedul	1		4			1								
section														
secur										2				3
sent														
seretid				18		16								
serious														
servic										2				4
sever														
shot				3		1					1			
shoulder														
sick					1				1					
sight											3			
sign	1	1							1					1
signific														
sister														
six														2
skin														
sleep		1				3								1

smoke										
smooth										4
social		1						3		5
sociodemograph										
socioeconom								2		
sodium										
sometim										
son				1						
soon										
specif										
spend										
spiriva			1	1	10					
spironolacton					5	26				
spite		3		4						1
spray			12		8			1		
stand										
start		2	2							
state										2
station										
stay										
stiff		4								
stomach										
stop									2	
street			2							
strong			2							
studi					1		1	2		1
subway			3	1	3					
sudden			4							
suffer		1						3		3
suffoc					5					
sulfat										
support		3		4				1		
surgeri				2					1	0
suspend										
swell										
swollen										
symptom							4			
syndrom										
tablet		1								
take									1	
talk				4						
telephon										

teofilin				3	1		1					
teolong												
term												
test												
theophyllin												
therapi			3					2		1		
thyroid												
time		1				2			1			
tire												
told								2				
transplant												
transport												
treat												
treatment		1	1	1							0	
tri											1	
twice	2											
type												
uncomfort												
undergo					3							
undergon												
understand									1		1	
understood												
underw										4		
unit												
urin												
use						4			2			
useless												
varicos												
vartalon												
vein												
verapamil		2				2		1				
vision								3				
visit							1	1	1		1	
visual				1				1	1		1	
vital												
vitamin												
wait								1				
walk	0		2					0				
weak												
wear									2			
weather												
week			4									

weight								
wife				2				
will				2		1		6
woman						1	2	
worker			1			3		5
worn		7						
worri		1				1		3
write								
wrong								
yellow								