Measuring Health Care Need and Coverage on a Probabilistic Scale in Population Surveys

Ajay Tandon* Christopher JL Murray† Bakuti Shengelia‡

March 19, 2004

Abstract

Coverage of health interventions is a critical measure for the assessment of the performance of health systems in achieving objectives such as improving population health and reducing health inequalities. There are many different measures of coverage in the literature. Following Shengelia *et al.* (2004), we define coverage as the probability that an individual receives a health intervention conditional on having a need for it. One implication of this definition is that, in order to measure coverage, one must have a way of assessing health care need at the individual level. This paper proposes a method for doing so by defining health care need on a probabilistic scale estimated from individual responses to symptomatic screening items in a population survey setting. The information in the set of respones to symptomatic items can be used to assess the likelihood that an individual has a disease or condition. This, combined with information on diagnosis and treatment, this can be used to derive a measure of coverage at the individual level and aggegated to the population level.

1 Introduction

Delivery of health interventions to individuals in need is a critical pathway through which health service provision can contribute to social objectives, such as improving population health and reducing health inequalities. To this end, information on coverage of critical health interventions is becoming a corner-stone in the assessment of health service provision function [68],[57],[50],[4],[37],[7]. Coverage is defined as the probability of receiving an intervention conditional on the presence of a health care need. This definition is based on three main premises: (a) the presence of a health care need being a precondition for

*Senior Research Associate, Harvard University Initiative for Global Health, 104 Mt Auburn Street, Cambridge, MA 02138

†Faculty Director, Harvard University Initiative for Global Health, 104 Mt Auburn Street, Cambridge, MA 02138.

‡Scientist, World Health Organization, Geneva, Switzerland.

receiving an intervention; (b) coverage being defined at the individual level; and (c) coverage being an *ex ante* concept, i.e., referring to the anticipation of a certain outcome of the interaction between the individual and the health system when a health care need emerges [57]. The concept of coverage can be extended to that of effective coverage by linking it to health gains from a given intervention. Effective coverage is the ratio of the realized health gain from an intervention to the potential health gain possible with the optimal performance of providers and full adherence to treatment [57]. Aggregation of effective coverage across several critical interventions will provide information on the extent of coverage for the health system as a whole. The focus on health care needs and effective interventions makes the measure of coverage much more concerned with potential health gain attainable in a given health system, as opposed to the total volume of services delivered to the population. Traditional measures of health service provision have not succeeded in maintaining link between the health care needs and the health systems response to those needs [58]. Existing frameworks for measuring health service provision have been focused mainly on the type of services used by population or certain dimensions of access, which limited their applicability for cause-specific decomposition of the gaps between the potential and actual performance of health service provision function. Using coverage as a metric of health service provision helps overcome some of these limitations [57].

In order to determine coverage two key pieces of information are required: who needs the intervention (the denominator) and among those who need the intervention, who received it (the numerator). By describing the relationship between the denominator, the numerator, and the socio-demographic characteristics of the individuals in the focus population, the probability of receiving an intervention conditional on health care need can be estimated [57],[46]. The accurate and precise identification of the denominator is key for the validity of the coverage measure. This is because, by the very definition of coverage, the probability of receiving the intervention is conditioned on the presence of the need for it. Measurement of health care needs typically has been carried out by using such proxies as mortality, socio-economic status and depravation, age, sex, and health service use [25]. For a variety of reasons, these proxies suffer from validity, reliability, and comparability limitations [25]. For the measurement of coverage, health care needs have to be determined at the level of the individual rather than at the population. One common practice so far has been the expression of health care need in terms of prevalence of certain conditions in the population [25]. While this provides a general picture of the magnitude of a health problem, it is not sufficient for the measurement of coverage at the individual level. The numerator of the coverage measure should capture the content of the intervention as precisely as feasible within a given measurement design. For instance in the case of DTP3 vaccination the numerator may include all children who have received three doses of the vaccine regardless of the schedule of administration and the source of verification (card versus mothers' recall). Identification of the numerator through such an approach produced coverage estimates referred to as crude coverage [63],[69]. The other approach is to define the numerator as only valid vaccinations, i.e., vaccinations administered according to the recommended schedule and documented by a health card. In the first case, the numerator does not distinguish between valid – and by implication, effective – vaccinations and those that are invalid. In the second case, the numerator captures only a subset of valid vaccinations and leaves out those valid vaccinations that are not documented [63],[69],[12],[7]. Capturing the numerator

could be more difficult for other more complex curative interventions. The denominator and numerator are necessary but not sufficient for measuring coverage defined as a probability. In order to predict coverage, a set of explanatory variables are required that ideally should capture both health system features and individuals' socio-demographic characteristics. By modelling the relationship between the vector of covariates and the event of receiving the intervention the parameter estimates for each covariate can be determined. These estimates will be used afterwards for predicting coverage for those individuals who were not included in the numerator.

Assessment of coverage typically has been carried out only in relation to a handful of interventions targeted to normative health care needs of certain population groups, where every individual equally needs the intervention in question [67],[40],[45],[66],[43],[44]. Such interventions, more often than not, are of a preventive nature. There is no uncertainty in determining health care need in these cases because the need is based on directly observable objective criteria such as age, sex, exposure to risk factors, etc.. Examples include all pregnant women needing antenatal care, all children requiring full immunization before their first birthday, all children under age five needing to sleep under an insecticide-treated bed net in malaria-endemic areas, etc.. Measurement of coverage for such health care needs entails only estimating the probability of receiving the intervention among certain population sub-group. The only challenge is to predict the probability of receiving the intervention based on *ex post* information. However, when the health care need is not determined by directly observable individual characteristics, the measurement of coverage requires identification of health care need, the condition for receiving the intervention, in the first place. To identify such health care needs with certainty would require sophisticated clinical diagnostic procedures, or at least a qualified medical professional's judgement. This is practical only in a clinical setting. Using only the clinical information obtained from health service statistics as a source of denominator for measuring coverage will introduce a selection bias in the measurement, because the denominator will capture only the fraction of the individuals with a health care need that had access to health services, whilst leaving out those who still have unmet needs. In order to avoid the selection bias health care needs have to be measured on the scale of the entire population. In population surveys, where clinical diagnostic procedures are not feasible, one has to rely on more practical but possibly less accurate tools to identify the presence of health care, at least in a probabilistic sense. It is a serious challenge to define such a probabilistic denominator and relate it to the probability of receiving the intervention. It is important to have a scientifically robust methodology for identifying health care need in the population through surveys without losing information due to imperfection of the tool for determining a health care need.

The purpose of this paper is two-fold. First, to generalize established approaches for evaluating simple diagnostic tests to apply to identification of health care need using multiple survey items. This approach, which we call the probabilistic diagnosis scale (PDS), yields for each individual a probability of diagnosis which can be used for both individual level action or analysis and as an input to measuring coverage or effective coverage. Second, the paper elaborates on methods to combine survey data on diagnosis and treatment with probabilistic diagnosis to yield valid assessments of coverage.

2 Determining Health Care Need

As mentioned in the previous section, the ability to measure health care need at the individual level is a necessary input in the measurement of health system coverage. The measurement of need for some interventions – such as those targeting all members of specific population sub-groups – is straightforward. In other situations, however, the assessment of need would require identification of the presence of a disease, health condition, or risk factor at the individual level. For example, for depression the intervention or treatment should target individuals who are truly depressed, and similarly for other diseases and conditions. We elaborate two different ways of doing this: (a) using clinical diagnostic tests, and (b) using symptomatic screening items in a survey setting. It is not practical to implement (a) in a population setting hence our focus will be more on the latter approach. However, prior to elaborating (b) we outline ways in which (a) can be used for smaller-scale analyses.

2.1 Diagnostic Tests

Ideally, the identification of health care needs in such cases would require a clinical diagnostic test (or tests) on the basis of which the individual would be classified as having a health care need or not. If information on the specificity and sensitivity of the diagnostic test(s) is available, it is common practice to combine this with information on population prevalence using Bayes' theorem and to estimate the probability or likelihood that the individual has a given disease or condition. Mathematically, Bayes' theorem allows us to estimate the probability that an individual i has a disease (D_i^+) conditional on his/her performance on K diagnostic tests $T_{i1}, ..., T_{iK}$:

$$\Pr(D_i^+|T_{i1}, ..., T_{iK}) = \frac{\Pr(D_i^+) \cdot \Pr(T_{i1}^+, ..., T_{iK}^+|D_i^+)}{\Pr(D_i^+) \cdot \Pr(T_{i1}^+, ..., T_{iK}^+|D_i^+) + [1 - \Pr(D_i^+)] \cdot \Pr(T_{i1}^+, ..., T_{iK}^+|D_i^-)}.$$

Assuming independence of tests conditional on disease status, this expression can be rewritten as:[1]

$$\Pr(D_i^+|T_{i1}, ..., T_{iK}) = \frac{\Pr(D_i^+) \cdot \prod_{k=1}^{K} \Pr(T_{ik}^+|D_i^+)}{\Pr(D_i^+) \cdot \prod_{k=1}^{K} \Pr(T_{ik}^+|D_i^+) + [1 - \Pr(D_i^+)] \cdot \prod_{k=1}^{K} \Pr(T_{ik}^+|D_i^-)}, \quad (1)$$

where $\Pr(D_i^+)$ is the population prevalence of the disease (the *prior* in Bayesian parlance). $\Pr(T_{ik}^+|D_i^+)$ is the probability of a positive result on test i conditional on truly having the disease or condition (*sensitivity* of the test in epidemiological parlance). $\Pr(T_{ik}^+|D_i^-)$ is the probability of a positive test result in the absence of having the disease or condition (one minus the *specificity* of the diagnostic test).

[1] Adjustments to the Bayesian approach can be made to allow for non-independence of multiple tests [16].

It is neither feasible nor practical to implement clinical diagnostic tests in a household survey setting. In such settings, symptomatic screening items have been used to assess the presence of a disease or condition. This approach is discussed in the next sub-section.

2.2 Symptomatic Screening Items

In the previous sub-section, we briefly outlined methods that may be used to assess health care need using clinical diagnostic tests. It is more practical, however, to use symptomatic screening items rather than clinical diagnostic tests in a survey setting. What this entails is asking respondents questions on prevalence of symptoms that are characteristic of a given disease or condition. Responses to symptomatic items are then used to identify individuals (and socio-demographic characteristics) that have high likelihood of having the disease. Such symptomatic screening items have a long history of use in diagnosing conditions such as angina, epilepsy, and depression, and have been used quite extensively in large scale population studies [64],[26],[48][49],[29],[2],[65],[39],[5],[3],[38]. Table 1 is an example of a nine-item screening instrument for large-scale epidemiological studies of epilepsy [53]. Symptomatic screening tools typically combine several items none of which have a particularly high sensitivity or specificity. Furthermore, usually not all the symptoms need to be present for a given disease. Therefore, making inference about the presence of a disease with a multiple-item symptomatic screening tool is often more challenging than with diagnostic tests. In terms of analysis of symptomatic screening data, we outline three broad approaches that can be taken. These include: (a) the use of diagnostic algorithms, (b) latent class analysis, and (c) probabilistic diagnostic scale. Each of these is discussed in turn.

Table 1: Short symptomatic screening instrument for epilepsy

1. Have you ever had attacks of shaking of the arms or legs which you could not control?
2. Have you ever had attacks in which you fall and become pale?
3. Have you ever lost consciousness?
4. Have you ever had attacks in which you fall with loss of consciousness?
5. Have you ever had attacks in which you fall and bite your tongue?
6. Have you ever had attacks in which you fall and lose control of your bladder?
7. Have you ever had brief attacks of shaking or trembling in one arm or leg or in the face?
8. Have you ever had attacks in which you lose contact with the surroundings and experience abnormal smells?
9. Have you ever been told that you have or had epilepsy or epileptic fits?

2.2.1 Diagnostic Algorithms

One mechanism for inferring the presence of a disease through multiple-item screening tools is to use special diagnostic algorithms [13],[61],[31],[34],[47],[35],[15]. Typically, algorithms use the set of responses to the multiple items to map individuals into two categories: those having the disease or not. Most algorithms are developed using clinical judgement and are sometimes combined with statistical criteria [60],[21],[32],[19],[11],[1]. If items in a screening instrument are interrelated – i.e., if items are not conditionally independent – then algorithms are often used to determine the combination of the items that ought to be endorsed affirmatively in order to classify the respondent as being disease positive [55],[31]. When the symptomatic items are independent, algorithm often take the form of specifying the minimum number of items that need to be respondent to affirmatively in order to classify a respondent as disease positive. Items in most symptomatic screening instruments are dichotomous or categorical in response. Methods for developing algorithms include setting a minimum number of items (cutoff point), using a linear discriminant function, using logistic regression, and classification tree approach, etc. [6]. Each of these models has its advantages and disadvantages depending on the type of a screening instrument. The outcomes of these models also differ: some are expressed as discrete numbers reflecting the number of items in the instrument, others as outputs of linear or logistic regressions. Some are expressed in terms of probabilities. Regardless of the nature of the outcome variable, a typical practice is to set a decision threshold – a certain outcome value – which is used for mapping individuals into "algorithm positive" and "algorithm negative" groups. The decision threshold can be set using ROC curves. Setting different thresholds moves the point along the ROC curve thereby trading off between sensitivity and specificity.

Validation studies of multiple-item screening instruments usually entails comparisons of algorithms to a gold standard, as opposed to validating each symptomatic element comprising the screening instrument. This explains the fact that validations studies rarely report sensitivity and specificity separately for symptomatic items, but rather provide validity characteristics of the entire instrument [1],[8],[23],[27],[41]. In this context, the sensitivity of an instrument, for example, would mean the probability of being algorithm-positive given the true presence of a disease. For instance, for the epilepsy items in Table 1, the algorithm suggests classification of subjects as "positives" in the case of any three (or more) affirmative answers out of nine. The authors report a sensitivity of 79.3% for the algorithm and a specificity of 92.9% and a positive predictive value (PPV) – i.e., the probability of having epilepsy given three or more positive responses – of 18.3% [53]. In addition, the authors have kindly provided more detailed information about the sensitivity and specificity of each of the nine items as well. These values are summarized in Table 2. As it can be seen, each item has its own information value. In general, sensitivity of individual items tends to be low, while specificity is reasonably high.

Information content about the presence of the disease determined by an algorithm can vary depending on how the algorithm is developed. In general, the information content is poor when the algorithm is determined based on the minimum number of symptoms present,

Table 2: Validity of individual items of a symptomatic screening question for epilepsy

Question	Sensitivity	Specificity
1	16.7	89.0
2	15.0	88.6
3	16.7	91.0
4	19.2	90.2
5	30.4	94.4
6	34.1	87.8
7	13.1	86.2
8	13.3	86.8
9	40.1	89.3

as is the case in the example with epilepsy. Everyone who meets the threshold is classified as positive, regardless of the combination of responses to different symptomatic items. Each combination will have different sensitivity, specificity and positive predictive value, depending on the symptomatic items comprising the combination. It is quite intuitive to conjecture, however, that an individual who reports seven out of nine symptoms of epilepsy, for example, may be more likely to have the disease than an individual who reports only four of the symptoms. Also, the probability of having epilepsy for two individuals with different combinations of four symptoms will not be the same. However, if the algorithm requires the presence of at least any three symptoms, then all individuals meeting the requirement will be classified as positive without differentiating. This is not optimal since a lot of information is being discarded.

Other methods of developing algorithms – such as the linear discriminant function and logistic regression – do take into account the combination of items and describe the presence of a disease on a continuous scale. Each combination of items renders different values as opposed to mapping responses only to two categories: positive and negative. The differences in the values represent the differences in the likelihood of having a disease depending on the actual responses to the symptomatic questions [6]. Because of the continuous scale of outputs, these methods avoid the loss of information which happens when "minimum number of items present" method is used. An algorithm developed using such methods is consistent with the positive predictive values of the actual responses to symptomatic questions. Conversely, algorithms based on the minimum number of observed symptomatic items are inconsistent because they might classify individuals into test-positive and test-negative groups without taking into account positive predictive values of the actual responses combinations. Despite their merits, algorithms determined through the linear discriminant function and logistic regression have one practical limitation: they require population-representative data on the true presence of a disease obtained through a gold standard test along with responses to the symptomatic items. This can be done in a small validation study, but is not practical to do each time the prevalence of certain diseases needs to be measured in the population.

2.2.2 Latent Class Analysis

Latent class analysis (LCA) is another method that is used in the analysis of symptomatic screening data. LCA models assume that the true disease status of respondents is an unobserved *discrete* latent variable (e.g., the latent variable may be dichotomous if the true disease classification is "has disease" and "no disease"). The observed categorical symptomatic response data are assumed to be related to the underlying discrete latent variable. The goal of LCA analysis is to estimate parameters that characterize the relationship between the latent variable and the observed symptomatic item responses. Usually, LCA analysis is used when gold standard information is unavailable, either because of the nature of the disease or due to other constraints [24]. Suppose we have data on N individuals each of whom has responded to K dichotomous items. As an example, consider two latent classes: $\eta = 0$ (no disease) and $\eta = 1$ (has disease). The likelihood for LCA in this case is given by:

$$L(\mathbf{Y}) = \prod_{i=1}^{N} \sum_{j=0}^{1} \Pr(\eta_i = j) \prod_{k=1}^{K} \Pr(y_{ik} = 1|\eta_i = j)^{y_{ik}} [1 - \Pr(y_{ik} = 1|\eta_i = j)]^{1-y_{ik}},$$

where $\mathbf{Y}$ is the observed data matrix of responses to symptomatic items and $\Pr(y_{ik} = 1|\eta_i = j)$ is the probability that individual i who is in latent class j has a positive response to item k. In LCA, the posterior likelihood of being in a given class can be calculated based on observed response patterns and using estimates of the unconditional likelihood of belonging in a given class, $\Pr(\eta_i = j)$, and the likelihood of a positive and negative responses given latent class membership. The formula is the same as the one given in equation 1 except for the fact that we replace test performance with item responses.

Inferences based on LCA of item response data, however, can be extremely misleading. Unlike the case with diagnostic tests, responses to symptomatic items for most conditions are likely to suffer from differential item functioning (DIF). DIF occurs when there is systematic variation in response probabilities of individuals from different socio-demographic backgrounds. So, for instance, one example of this would be if depressed men and women have significantly different probabilities of responding affirmatively to a symptomatic item related to lack of motivation. There is a voluminous literature on the issue of DIF in health surveys [9],[10],[22],[28],[30],[52],[56],[59],[62]. In the presence of DIF, and in the absence of other exogenous information, it is virtually impossible to detect a difference between a higher probability of having a disease versus a higher likelihood of answering an item affirmatively conditional on having the disease. In other words, there is an identification problem underlying LCA models of symptomatic item responses. As a result, the validity of inferences based on LCA models applied to symptomatic item response data is seriously undermined.

In the next sub-section, we propose a model which we call the "probabilistic diagnosis scale" which measures the presence of a disease in large population surveys on a probabilistic scale and which does not need to be validated using a large-scale gold standard population-representative survey. The model – which is related to LCA and based on Bayes theorem – uses the validity parameters of each symptomatic item obtained from external validation

studies.

2.2.3 Probabilistic Diagnosis Scale

The idea behind probabilistic diagnostic scale (PDS) is simple. The starting point is the same as that given in equation 1 except for the fact hat that test performance T_{ik} is replaced with item response Q_{ik}:

$$\Pr(D_i^+|Q_{i1},...,Q_{iK}) = \frac{\Pr(D_i^+)\cdot\prod_{k=1}^{K}\Pr(Q_{ik}|D_i^+)}{\Pr(D_i^+)\cdot\prod_{k=1}^{K}\Pr(Q_{ik}|D_i^+) + [1-\Pr(D_i^+)]\cdot\prod_{k=1}^{K}\Pr(Q_{ik}|D_i^-)}. \quad (2)$$

We are assuming for now that the conditional responses for the different items is independent. In LCA, the prevalence as well as the sensitivity and specificity parameters are estimated from item response. However, as mentioned earlier, this approach is susceptible to a non-trivial identification problem: in the absence of additional exogenous information, the estimation procedure cannot discriminate between a higher prevalence rate versus a greater propensity to respond affirmatively given the disease among certain socio-demographic population sub-groups. In order to overcome this problem we propose the PDS approach. In the PDS approach, we estimate some item response parameters using data from a separate study where the likelihood of a positive response is derived from administering the questionnaire to diagnosed-positive respondents. In other words, $\Pr(Q_{ik}|D_i^+)$ is derived from a convenience sample. One advantage of this approach is in reverse-sampling: there is no need to administer the questionnaire in the general population and then also diagnose the respondents using a gold standard in order to estimate the parameters. Instead, a convenience sample of diagnosed respondents is administered the questionnaire. This is an important innovation in terms of practical implementation, especially so for low-prevalence diseases that would require large sample population-based surveys – with a concomitant implementation of gold-standard diagnostic test(s) in the same sample – to get a reasonably accurate estimation of item-response parameters.

The next section describes implementation of this approach for the World Health Survey. In a subsequent section, we provide more details on the practical implementation of the method using data from the World Health Survey.

3 World Health Survey (WHS)

The World Health Survey (WHS) is a nationally-representative survey of adults (older than 18 years of age) in 72 participating countries. Sample sizes in each country vary from 1,000 to 10,000 respondents. Depending on the survey mode, the WHS contains modules on health intervention coverage, health insurance, health expenditure, indicators

of "permanent income" (economic status), health occupations, health state descriptions, health state valuation, risk factors, mortality, health system responsiveness, health system goals, and social capital.[2] As mentioned in the Introduction, the measurement of need is an integral part in the measurement of health system coverage. In the WHS, for major diseases and conditions, need is identified based on responses to symptomatic items such as the ones listed in Table 1. Symptomatic items are included for depression, schizophrenia, angina, and several other conditions. The determination of need requires information on the probabilities of positive item responses with and without the disease or condition. This was done by implementing a separate study – the Diagnosis Item Properties Study – on a convenience sample of respondents with diagnosed diseases and conditions. These respondents were administered the WHS symptomatic screening instrument for coverage.

3.1 Diagnosis Item Properties Study (DIPS)

The Diagnosis Item Properties Study (DIPS) is an auxiliary study to the WHS. Six countries participated in DIPS: Burkina Faso, Czech Republic, Ethiopia, Mexico, Malaysia, and Slovak Republic. In each country site 80 diagnosed respondents were identified for each of the following conditions: arthritis, angina, asthma, depression, schizophrenia, epilepsy, tuberculosis, and cataract.[3] Diagnosis was done based on gold standard tests as reported in Table 3. For each condition, diagnosed respondents were administered the symptomatic screening items corresponding to their condition from the WHS. The idea being that these convenience-sample based DIPS responses could be used in the calculation of an estimate of $\Pr(Q_{ik}|D_i^+)$ which we denote by π_{11}^k. In a second stage, these estimates could then be used in the WHS – where respondents do not have a gold standard diagnosis – to estimate the probability of having a given disease or condition based on the observed pattern of item responses.

In order to implement the Bayesian likelihood, an estimate of $\Pr(Q_{ik}|D_i^-)$ – which we denote by π_{10}^k – is also required. The group of disease negatives will be selected from the respondents of the WHS. As mentioned earlier, the WHS asks respondents the same symptomatic questions as in DIPS. However, in addition, in the WHS each respondent is also asked if he or she has ever been diagnosed with the condition in question by a medical professional. A negative response to the questions about previous diagnosis for all eight conditions in question will be a criterion for considering the respondent as disease negative. Ideally, disease negative status should be identified using the same gold standard diagnostic tests as for the identification of positives. However, selection of negatives through a diagnostic test will be expensive and technically more difficult than its convenient alternative, selection of negatives from the WHS. Given the low prevalence of these eight conditions (most being less than 5%) the chances of having disease positive individuals among randomly-selected WHS respondents who have reported no previous diagnosis is expected to be negligible.

[2] More information on the WHS can be found at http://www.who.int/whs/.

[3] Some countries implemented only a sub-set of DIPS conditions.

Table 3: Gold standard(s) for diagnosing conditions in DIPS

Disease/Condition	**Diagnostic test(s)**
Arthritis	X-ray
Angina	Exercise stress (ECG)/Holter
Asthma	Bronchial hypersensitivity
	Dynamic and static lung volume and capacity tests
	Eosinphil count (>250 to 400 cells/μL)
Depression	Psychiatric examination
Schizophrenia	Psychiatric examination
Epilepsy	Neurological examination
	EEG
	Serum glucose, sodium, magnesium, and calcium
Tuberculosis	Sputum smear exam
	X-ray
Cataract	Opthalmoscope

4 Methods

This section outlines the estimation method in more detail. The starting point is the formula for estimating the probability of having a disease conditional on responses to a series of symptomatic questions using Bayes' theorem applied to the WHS:

$$\Pr(D_i^+|Q_{i1},...,Q_{iK}) = \frac{\Pr(D_i^+)\cdot\prod_{k=1}^{K}\pi_{11}^k}{\Pr(D_i^+)\cdot\prod_{k=1}^{K}\pi_{11}^k + [1-\Pr(D_i^+)]\cdot\prod_{k=1}^{K}\pi_{10}^k}, \tag{3}$$

which is basically equation (2) except for the fact that $\Pr(Q_{ik}|D_i^+)$ and $\Pr(Q_{ik}|D_i^-)$ are replaced by their corresponding estimates from the DIPS study: i.e., by π_{11}^k and π_{10}^k, respectively. All that remains to compute the posterior probability of having the disease or condition is some knowledge of $\Pr(D_i^+)$, i.e., of the prior. Estimates of the prior can come from several sources. One strategy could be to simply assume that no prior information is available, i.e., this would correspond to the situation with uninformative priors. In other words, the assumption would be that the prevalence of the disease or condition could take any value between 0% and 100%, each being equally likely. This situation is depicted in Figure 1. In this case, posterior estimates would be derived solely from the information content in the set of individual responses to the symptomatic items. A second option would be to base the prior on estimates derived from other sources such as epidemiological studies of prevalence. A third option would be to derive priors from the WHS itself using information on π_{11}^k and π_{10}^k from DIPS. This is because – on average – the aggregate proportion of positive responses to any given symptomatic response item k – $\Pr(Q_k^+|D^+)$ – can be related to π_{11}^k and π_{10}^k as follows:

$$\Pr(Q_k^+|D^+) = \pi_{11}^k \cdot D^+ + \pi_{10}^k \cdot (1 - D^+). \tag{4}$$

Here, D^+ is any given value of the population prevalence of the disease or condition. Note that $\Pr(Q_k^+|D^+)$ is now at the aggregate level, i.e., it is the proportion of positive responses for an item k across all N individuals in the sample. With a sample size N and any given value of prevalence D^+, $\Pr(Q_k^+|D^+)$ can be derived from the normal distribution approximation: it is the likelihood of observing the observed proportion of positive responses assuming the mean is $N \cdot [\pi_{11}^k \cdot D^+ + \pi_{10}^k \cdot (1 - D^+)]$ and the standard deviation is $N \cdot [\pi_{11}^k \cdot D^+ + \pi_{10}^k \cdot (1 - D^+)] \cdot [1 - \pi_{11}^k \cdot D^+ - \pi_{10}^k \cdot (1 - D^+)]$. Now, we can assume that the prior for the prevalence is uninformative and use the equation (4) as our likelihood. The posterior from the *aggregate-level* data can then be used as a prior for the *individual-level* analysis in a subsequent step.

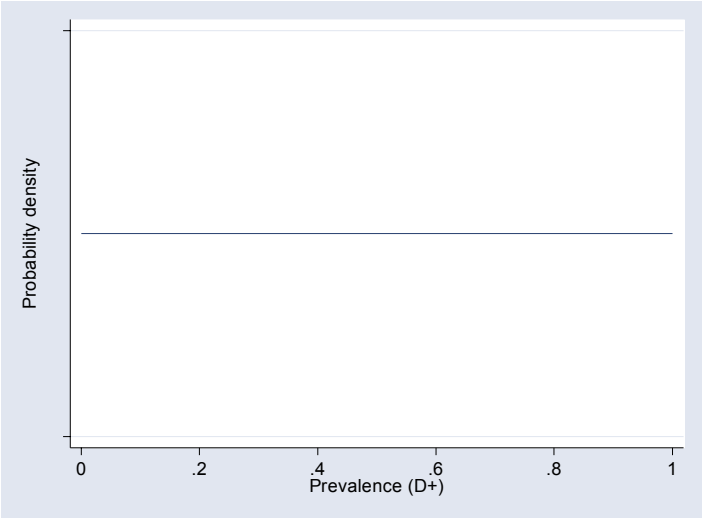

Figure 1: Uninformative prevalence priors

5 Results

Data from the WHS and DIPS studies have just started to come in, and we report some preliminary results from those data in this paper. The first sub-section summarizes the results from the DIPS data for the estimation of π_{11}'s. Subsequently, we discuss the estimation of π_{10}'s from WHS data.

5.1 π_{11} Estimation using DIPS

We begin by considering the case of depression. There were seven symptomatic screening items for depression as shown in Table 4. Figure 2 plots the proportion of diagnosed

depressed respondents that answered affirmatively to each of the seven symptomatic items in three countries: Czech Republic, Malaysia, and Mexico. As can be seen, for the first three items, the proportion of positive responses ranges between 60% and 80%. There is more variance across the countries for the remaining four items. This is something that needs to be investigated further as more data become available. There are several reasons this may be occurring. One possibility is due to varying severities of depression in the DIPS sample across the three countries. The other possibility could be due to cultural (or other) DIF-related factors: there may be differing propensities of an affirmative response conditional on disease status. A similar pattern of outcomes is observed for the case of asthma (Table 5 and Figure 3).

Table 4: Symptomatic screening items for depression

1. Have you had a period lasting several days when you felt sad, empty, or depressed?
2. Have you had a period lasting several days when you lost interest in most things you usually enjoy such as hobbies, personal relationships, or work?
3. Have you had a period lasting several days when you have been feeling your energy level decreased or that you were tired all the time?
4. Was this period for more than two weeks?
5. Was this period most of the day, nearly every day?
6. During this period, did you lose your appetite?
7. During this period, did you notice any slowing down in your thinking?

Table 5: Symptomatic screening items for asthma

1. Have you experienced attacks of wheezing or whistling breathing?
2. Have you experienced attacks of wheezing that came on after you stopped exercising or doing some other physical activity?
3. Have you experienced a feeling of tightness in your chest?
4. Have you woken up with a feeling of tightness in your chest in the morning or any other time?
5. Have you had an attack of shortness of breath that came on without obvious cause when you were not exercising or any other time?

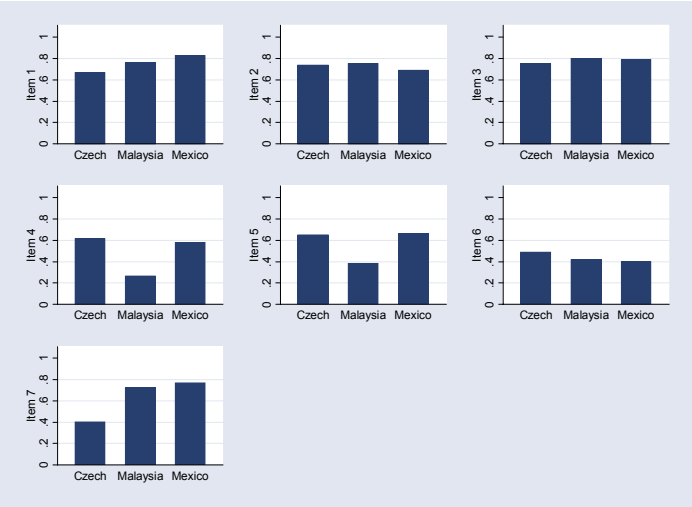

Figure 2: Preliminary estimates using DIPS: Depression

5.2 π_{10} Estimation using WHS

As mentioned earlier, π_{10} was estimated from the WHS. Disease negatives were assumed to be respondents who had a negative response to all eight questions about previous diagnosis. Although this method of identifying negatives may not be perfect, it is unlikely to have a major contaminating effect given the low prevalences of all the conditions in question. Figures 4, 5, and 6 plot the distribution of responses disaggregated by sex, age, and education for 12 WHS countries for which data were available. There are a couple of interesting things to note. First, there are big cross-country differences in the proportion of affirmative responses across countries for all three items. The proportion of affirmative responses is very low in China, Ethiopia, Sri Lanka, and Vietnam. On the other hand, there proportion is fairly high as in Bangladesh and the Czech Republic. These differences are almost certainly due to DIF. Differences in education, awareness regarding depression, income, and cultural propensities for self-reporting health problems are likely reasons for such large observed differences in responses. Evidence of DIF brings into question the use of standardized algorithms for diagnosing conditions such as depression as the results are not comparable across countries. The problem exists even within countries: there is evidence of wide-spread differences in the proportions of affirmative responses even within countries as evidenced by the distribution of responses in a given country by age, sex, and education.

6 Discussion

In this paper, we have proposed one method for estimating the need for health care using responses to symptomatic items in population surveys. The identification of respondents

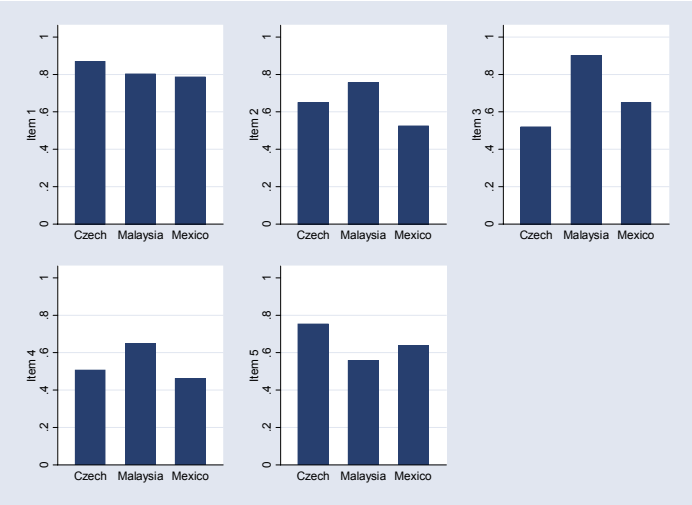

Figure 3: Preliminary estimates using DIPS: Asthma

who truly have a health care need is important for the measurement of health system coverage, i.e., for the measurement of the probability of receiving a health intervention given that the individual has a need for it. The method is based on the application of Bayes theorem to estimate the likelihood of having a disease or condition based on a given pattern of responses. The innovation in this method is that we derive the likelihood of a positive response conditional on having the disease from a separate study of diagnosed respondents who are also administered the symptomatic items. This allows for the incorporation of exogenous gold-standard based information into the estimation without the implementation of a gold-standard test in the entire population sample as only a convenience sample can be used for estimation of the parameters. Some preliminary results were tabulated from the WHS and DIPS studies. As more data become available, the method will allow for the DIF-free estimation of prevalence of diseases as well of the coverage of health systems, the latter being an important indicator and monitoring tool to measure the performance of the health system.

References

[1] Anderson C, S Laubscher, and R Burns (1997), "Validation of the Short-Form 36 (SF-36) Health Survey Questionnaire Among Stroke Patients," *Stroke*, 27(10):1812-1816.

[2] Anderson DW, FA Bryan, BS Harris, JT Lessler, and JP Gagnon (1985), "A Survey Approach for Finding Cases of Epilepsy," *Public Health Reports*, 100(4):386-393.

[3] Arolt V. (1994), "Psychotherapy of Depression in Psychiatrist's Private Practices: Results of an Epidemiological Survey," *Psychotherapie Psychosomatik Medizinische Psychologie*, 44(6):177-183.

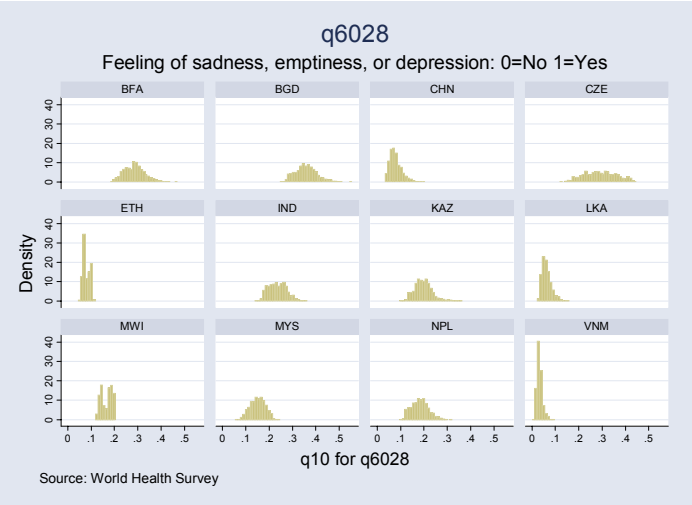

Figure 4: Distribution of responses for disease negatives: first depression item

[4] Atting IA and IN Egwu (1991), "Indicators of Accessibility to Primary Health Care Coverage in Rural Odukpani, Nigeria," *Asia-Pacific Journal of Public Health*, 5(3):211-216.

[5] Blazer DG, RC Kessler, KA McGonagle, and MS Swartz (1994), "The Prevalence and Distribution of Major Depression in a National Community Sample: The National Comorbidity Survey. *American Journal of Psychiatry*, 151(7):979-986.

[6] Bloch DA, LE Moses LE, and BA Michel (1990), "Statistical Approaches to Classification. Methods for Developing Classification and other Criteria Rules," *Arthritis and Rheumatism*, 33(8):1137-1144.

[7] Bos E and A Batson (2000), "Using Immunization Coverage Rates for Monitoring Health Sector Performance: Measurement and Interpretation Issues," *World Bank, Health, Nutrition and Population,* Washington, DC 2000.

[8] Bridges KW, and DP Goldberg (1986), "The Validation of the GHQ-28 and the Use of the MMSE in Neurological In-Patients," *British Journal of Psychiatry*, 148:548-553.

[9] Chen, L, and CJL Murray (1992), "Understanding Morbidity Change," *Population and Development Review*, 18:481-504.

[10] Cole, SR (1999), "Assessment of Differential Item Functioning in the Perceived Stress Scale-10," *Journal of Epidemiology and Community Health*, 53(5):319-320.

[11] Cook DG, AG Shaper, and PW MacFarlane PW (1989), "Using the WHO (Rose) Angina Questionnaire in Cardiovascular Epidemiology," *International Journal of Epidemiology*, 18(3):607-613.

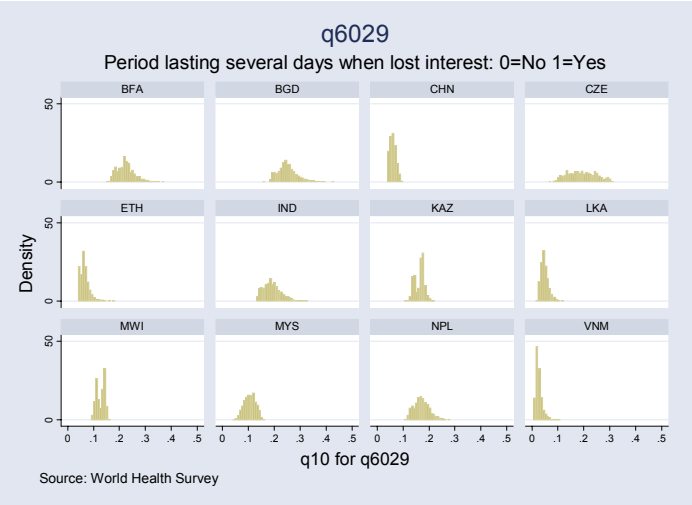

Figure 5: Distribution of responses for disease negatives: second depression item

[12] Cutts FT, RJ Waldman, and HMD Zoffman (2001), "Surveillance for the Expanded Programme on Immunization," *Bulletin of the World Health Organization*, 71(5):633-639.

[13] Dawson-Saunders B, and RG Trapp (1994), *Basic and Clinical Biostatistics: Second Edition*, Appleton & Lange, Norwalk, Connecticut.

[14] De Vet HCW, T van der Weijden, JWM Muris, J Heyrman, and F Buntinx (2001), "Systematic Reviews of Diagnostic Research: Considerations about Assessment and Incorporation of Methodological Quality," *European Journal of Epidemiology*, 17:301-306.

[15] Deal LW, and VL Holt VL (1998), "Young Maternal Age and Depressive Symptoms: Results from the 1988 National Maternal and Infant Health Survey," *American Journal of Public Health*, 88(2):266-270.

[16] Dendukuri, N, and L Joseph (2001), "Bayesian Approaches to Modeling the Conditional Dependence Between Multiple Diagnostic Tests," *Biometrics*, 57(1):158-167.

[17] Editorial (2000), "Role of Bronchial Responsiveness Testing in Asthma Prevalence," *Thorax*, 52:352-354.

[18] Erikssen J, K Forfang, and O Storstein (1997), "Angina Pectoris in Presumably Healthy Middle-Aged Men: Validation of Two Questionnaire Methods in Making the Diagnosis of Angina Pectoris," *European Journal of Cardiology*, 6(4):285-298.

[19] Failde I, I Ramos (2000), "Validity and Reliability of the SF-36 Health Survey Questionnaire in Patients with Coronary Artery Disease," *Journal of Clinical Epidemiology*, 53(4):359-365.

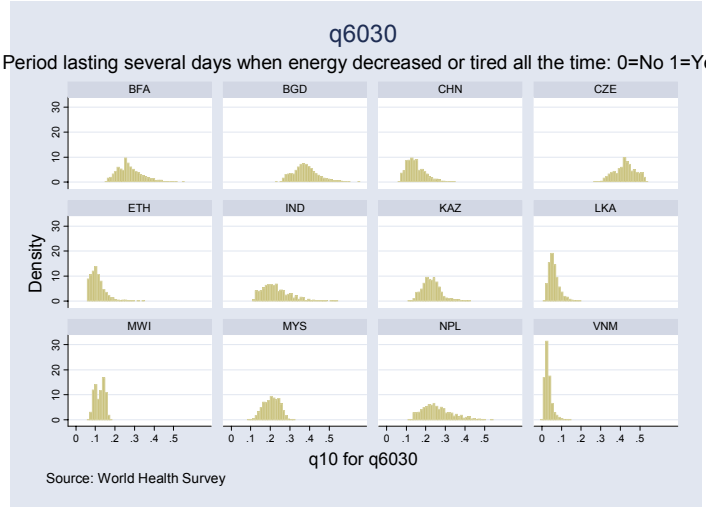

Figure 6: Distribution of responses for disease negatives: third depression item

[20] Farr BM (2001), "Diagnostic Tests for Healthcare Epidemiology," *Current Opinion in Infectious Diseases*, 14(4):443-447.

[21] Fink P, J Jensen, I Borgquist, and JI Brevik JI (1995), "Psychiatric Morbidity in Primary Public Health Care: A Nordic Multicentre Investigation," *Acta Psychiatrica Scandinavica*, 92:409-418.

[22] Fleishman, JA, and WF Lawrence (2003), "Demographic Variation in SF-12 Scores: True Differences or Differential Item Functioning?" *Medical Care*, 41(7 Suppl):III75-III86.

[23] Gandek B, JE Ware JE, NK Aaronson NK *et al.* (1998), "Cross-Validation of item selection and scoring for the SF-12 Health Survey in nine countries: results from the IQOLA Project. International Quality of Life Assessment. *Journal of Clinical Epidemiology*, 51(11):1171-1178.

[24] Garrett ES, WW Eaton, and S Zeger (2002), "Methods for Evaluating the Performance of Diagnostic Tests in the Absence of a Gold Standard: A Latent Class Model Approach," *Statistics in Medicine,* 21(9):1289-1307.

[25] Gibson A, S Asthana, P Brigham, G Moon, and J Dicker (2002), "Geographies of Need and the New NHS: Methodological Issues in the Definition and Measurement of the Health Needs of Local Populations," *Health and Place*, 8:47-60.

[26] Glader EL, and B Stegmayr (1999), "Declining Prevalence of Angina Pectoris in Middle-Aged Men and Women: A Population-Based Study within the Northern Sweden MONICA Project. Multinational Monitoring of Trends and Cardiovascular Disease," *Journal of Internal Medicine*, 246(3):285-291.

[27] Goldberg DP, K Rickels, R Downing, and Hesbacher P (1976), "A Comparison of Two Psychiatric Screening Tests," *British Journal of Psychiatry*, 129:61-67.

[28] Iwata, N, and S Buka (2002), "Race/Ethnicity and Depressive Symptoms: A Cross-Cultural/Ethnic Comparison Among University Students in East Asia, North and South America," *Social Science and Medicine*, 55(12):2243-2252.

[29] Jacoby A, GA Baker, N Steen, and D Buck (1999), "The SF-36 as a Health Status Measure for Epilepsy: A Psychometric Assessment," *Quality of Life Research*, 8(4):351-364.

[30] Jones, RN (2003), "Racial Bias in the Assessment of Cognitive Functioning of Older Adults," *Aging and Mental Health*, 7(2):83-102.

[31] Katz BP, DA Freud, RS Heck *et al.* (1996), "Demographic Variation in the Rate of Knee Replacement: A Multi-Year Analysis," *Health Services Research*, 31(2):126-140.

[32] Keller SD, JE Ware, PM Bentler *et al.* (1998), "Use of Structural Equation Modeling to Test the Construct Validity of the SF-36 Health Survey in Ten Countries: Results from the IQOLA Project," *Journal of Clinical Epidemiology*, 51(11):1179-1188.

[33] King, G (1998), *Unifying Political Methodology: The Likelihood Theory of Statistical Inference*, Ann Arbor: University of Michigan Press.

[34] Lehtinen V, M Joukamaa, K Lahtela *et al.* (1990), "Prevalence of Mental Disorders Among Adults in Finland: Basic Results from the Mini Finland Health Survey," *Acta Psychiatrica Scandinavica*, 81(5):418-425.

[35] Lemkau JP, B Mann B, D Little D, P Whitecar, P Hershberger, and JA Schumm (2000), "A Questionnaire Survey of Family Practice Physicians' Perceptions of Bereavement Care," *Archives of Family Medicine*, 9(9):822-829.

[36] Leonard, T, and JSJ Hsu (1999), *Bayesian Methods: An Analysis for Statisticians and Interdisciplinary Researchers*, Cambridge, UK: Cambridge University Press.

[37] Lever P (1998), "Management by Targets: Is Coverage an Adequate Measure for Health Care?" *Health Policy*,14(3):356-359.

[38] Madianos MG, D Gefou-Madianou, and CN Stefanis (1994), "Symptoms of Depression, Suicidal Behaviour and Use of Substances in Greece: A Nationwide General Population Survey," *Acta Psychiatrica Scandinavica*, 89(3):159-166.

[39] Maier W, M Gansicke, R Gater, M Rezaki, B Tiemens, and RF Urzua (1999), "Gender Differences in the Prevalence of Depression: A Survey in Primary Care,". *Journal of Affective Disorders*, 53(3):241-252.

[40] Malison MD (1987), "Estimating Health Service Utlization, Immunization Coverage, and Childhood Mortality: A New Approach in Uganda. *Bulletin of the World Health Organization*, 65(3):325-330.

[41] Mari JJ, and P Williams (1985), "A Comparision of the Validity of Two Psychiatric Screening Questionnaires (GHQ12 and SRQ20) in Brazil, Using Relative Operating Characteristic Analysis," *Psychological Medicine*, 651-659.

[42] Miller TD, VL Roger, JJ MIlavetz *et al.* (2001), "Assessment of the Exercise Electrocardiogram in Women versus Men using Tomographic Myocardial Perfusion Imaging as the Reference Standard," *American Journal of Cardiology*, 87(7):868-873.

[43] Mobarak AB (1980), "Study on Coverage, Effectiveness and Efficiency of Rural Health Delivery Service in Egypt," Ministry of Health of Egypt, Cairo.

[44] Monteith RS, CW Warren, E Stanziola, UR Lopez, and MW Oberle (1987), "Use of Maternal and Child Health Services and Immunization Coverage in Panama and Guatemala," *Bulletin of the Pan American Health Organization*, 21(1):1-15.

[45] Montoya-Aguilar C, and MA Marin-Lira (1986), "International Equity in Coverage of Primary Health Care Examples from Developing Countries," *World Health Statistics Quarterly*, 39(4):336-344.

[46] Murray CJL, B Shengelia, N Gupta, S Moussavi, and M Thieren (2002), "Routinely Reported Immunization Coverage: How Accurate is the Evidence for Performance Assessment."

[47] Newman SC, RC Bland RC, and HT Orn (1998), "The Prevalence of Mental Disorders in the Elderly in Edmonton: A Community Survey using GMS-AGECAT," *Canadian Journal of Psychiatry*, 43(9):910-914.

[48] Nicholson A, IR White, P Macfarlane, E Brunner, and M Marmot (1998), "Rose Questionnaire Angina in Younger Men and Women: Gender Differences in the Relationship to Cardiovascular Risk Factors and Other Reported Symptoms," *Journal of Clinical Epidemiology*, 52(4):337-346.

[49] Nicoletti A, A Reggio, A Bartoloni *et al.* (1999), "Prevalence of Epilepsy in Rural Bolivia: A Door-to-Door Survey," *Neurology*, 53(9):2064-2069.

[50] Paganini JM (1998), "Health Service Coverage in Latin America and the Caribbean," *Pan American Journal of Public Health*, 4(5):305-310.

[51] Peat JK, BG Toelle, GB Marks, and CM Mellis (2001), "Continuing the Debate about Measuring Asthma in Population Studies," *Thorax*, 56:406-411.

[52] Peterson, MA, M Groenvol, JB Bjorner et al. (2003), "Use of Differential Item Functioning to Assess Equivalence of Translations of a Questionnaire," *Quality of Life Research*, 12(4):373-385.

[53] Placencia M, JSWAS Sander, SD Shorvon, RH Ellison, and SM Cascante (1992), "Validation of a Screening Questionnaire for the Detection of Epileptic Seizures in Epidemiological Studies," *Brain*, 115:783-794.

[54] Roberts RE, PM Lewinsohn, and JR Seeley (1995), "Symptoms of DSM-III-R Major Depression in Adolescence: Evidence from an Epidemiological Survey," *Journal of the American Academy of Child & Adolescent Psychiatry*, 34(12):1608-1617.

[55] Roe GA, H Blackburn, RF Gillum, and RJ Prineas (1982), "Cardiovascular Survey Methods," WHO Monograph Series No 56. Geneva.

[56] Sen, A (2002), "Health: Perception versus Observation," *British Medical Journal*, 324:860-861.

[57] Shengelia B, CJL Murray, and O Adams (2002), "Beyond Access and Utilization: Defining and Measuring Health-System Coverage," Geneva: World Health Organization.

[58] Shengelia, B, CJL Murray, and A Tandon (2004), "Measuring Health System Coverage," Cambridge: Harvard Global Health Initiative.

[59] Smith, LL, and SP Reise (1998), "Gender Differences on Negative Affectivity: An IRT Study of Differential Item Functioning on the Multidimensional Personality Questionnaire Stress Reaction Scale," *Journal of Personality and Social Psychology*, 75(5):1350-1362.

[60] Surtees PG, NWJ Wainwright, WR Gilks, TS Brugha, H Meltzer, and R Jenkins (1997), "Diagnostic Boundaries, Reasoning and Depressive Disorder II: Application of a Probabilistic Model to the OPCS General Population Survey of Psychiatric Morbidity in Great Britain," *Psychological Medicine*, 27(4):847-860.

[61] Tarnopolsky A, DJ Hand, EK McLean, H Roberts, and RD Wiggins (1979), "Validity and Uses of Screening Questionnaire (GHQ) in the Community," *British Journal of Psychiatry*, 134:508-515.

[62] Tennant, A, M Penta, L Tesio et al. (2004), "Assessing and Adjusting for Cross-Cultural Validity of Impairment and Activity Limitation Scales through Differential Item Functioning within the Framework of the Rasch Model: The PRO-ESOR Project," *Medical Care*, 42(1 Suppl):I37-I48.

[63] Tonglet R, M Soron'gane, M Lembo, MM Wa, M Dramaix, and P Hennart (1993), "Evaluation of Immunization Coverage at Local Level," *World Health Forum*, 14(3):275-281.

[64] Udol K, and N Mahanonda (2000), "Comparison of the Thai Version of the Rose Questionnaire for Angina Pectoris with the Exercise Treadmill Test," *Journal of the Medical Association of Thailand*, 83(5):514-522.

[65] Upton MW, M Evans M, DP Goldberg, and DJ Sharp (1999), "Evaluation of ICD-10 PHC Mental Health Guidelines in Detecting and Managing Depression within Primary Care," *British Journal of Psychiatry*, 175:476-482.

[66] *WHO National assessments of health care coverage and of its effectiveness and efficiency.*1983.

[67] WHO *Coverage of maternity care*: a listing of available information.1997.

[68] WHO/AMRO Critical Issues in Health System Performance Assessment. *Regional Consultation on Health System Performance Assessment*. WHO/AMRO, Washington D.C.2001.

[69] WHO/EIP *Immunization Services Assessment Guidelines*.2000.