WORKING PAPER

Determining the Priority Global Health Needs and Quantifying the Health Benefits Resulting From the Introduction of New Diagnostics in the Developing World

FEDERICO GIROSI, MARIA E. RAFAEL, JULIA E. ALEDORT, YEE-WEI LIM, ROB BOER, KAREN RICCI, MOLLY SHEA AND EMMETT B. KEELER

WR-418-HLTH

December 2006

This is supporting material to Girosi F, Olmsted S et al. Developing and interpreting models to improve diagnostics in developing countries. Nature. S1 3-8 (2006).

RAND HEALTH

PREFACE

This working paper contains technical material supporting the article by Girosi et al. "Developing and interpreting models to improve diagnostics in developing countries" *Nature* S1; 3-8 (2006). It is intended to be read in conjunction with that article. This supplement includes additional material referred to in the published article as well as secondary analyses and tables that were not included in the published paper. Although this technical supplement in its current form has not been formally peer-reviewed, an earlier version of this paper, which also contained material that appears in the corresponding *Nature* paper, was reviewed by two outside experts and was revised in response to their comments. The work was funded by the Bill & Melinda Gates Foundation to support the Global Health Diagnostics Forum.

Acknowledgements:

The authors thank Kristin Leuschner (RAND Corporation, California, USA). The authors also thank Gary King (Harvard University) and Richard Hillestad (RAND Corporation, California, USA) for comments on an earlier draft.

Table of Contents

1 Introduction

An essential component of evaluating and improving global health is access to appropriate diagnostic tools. The current diagnostic tests for many diseases do not meet the needs of the developing world[1]. Some tests require technological capabilities and infrastructure that are beyond the resources of developing countries. Others have the required technological capabilities but are too costly to be used.

Developing a rational strategy for investment in diagnostic technologies requires a means to determine the need for, and the health impact of, potential new tools. This paper outlines an approach for modeling the health benefits of new diagnostic tools. The framework was developed by the RAND Corporation in conjunction with the Bill & Melinda Gates Foundation and the partnership they formed in 2004 — known as the Global Health Diagnostics Forum — with domain experts in relevant disease areas, representatives from the diagnostics industry and technology development arena, and experts in the modeling of disease impact and the application of diagnostic technologies. The results of disease-specific interventions and the roles of new diagnostic technologies are reported in several articles in a supplement to *Nature*[1], and are also available in a series of RAND reports[2].

In order to determine the health impact of a new diagnostic test, our approach divides the problem into two tasks[1]: first, we establish the effect that a specific diagnostic tool might have on the reduction of the disease burden; and second, we identify the performance characteristics and user requirements that a diagnostic tool must have to realize that reduction. The first task requires disease-specific modeling of the status quo and the changes that could occur were a new diagnostic to become available in certain settings. The product of this effort is a tool that — given the sensitivity and specificity of a potential new diagnostic, and an estimate of the proportion of people who will have access to it — can predict the health impact of a test using a number of different health outcomes. The second task involves defining the characteristics of diagnostics, such as

[1] http://www.nature.com/diagnostics
[2] http://www.rand.org/health/feature/research/0612_global.html

the type of infrastructure needed to be operational, and estimating the proportion of people that will have access to different types of test. We refer to these characteristics as "user requirements". This task requires us to define representative health-care settings in the developing world, identify their capabilities and estimate access to different levels of care.

The purpose of this report is to describe the methodology we have used, in the fist task, to estimate the attributable benefit associated with the introduction of new diagnostics for the above-mentioned diseases. Our goal is to provide as much details as possible on every aspect of the models we have built, so that any reader could replicate our results just by following what is in this paper. Methodological details on the second task, relating to the relationship between test characteristics, infrastructure requirements and access to care, are described in a separate report[2], while an overview of both tasks is presented in ref. 1.

The plan of the paper is as follows. In Section (2) we provide a qualitative overview of the modeling framework. A detailed formalization of this approach is presented in Section (3). An important set of assumptions on how the new diagnostics is introduced in a country is formalized in Section (4). Section (5) shows how to build on the framework described in the previous section to model more complex scenarios, in which more than one test might be used, or treatment is only partially available, or the test is slow and there is loss-to-follow-up. Section (6) addresses a crucial issue, common to any modeling of diagnostics: the harm associated with treatment and the corresponding trade-off between sensitivity and specificity. In Section (7) we address issues related to sensitivity analysis and estimate of the uncertainty in the model output. Finally, Section (8) concludes with a discussion of some limitations of the approach.

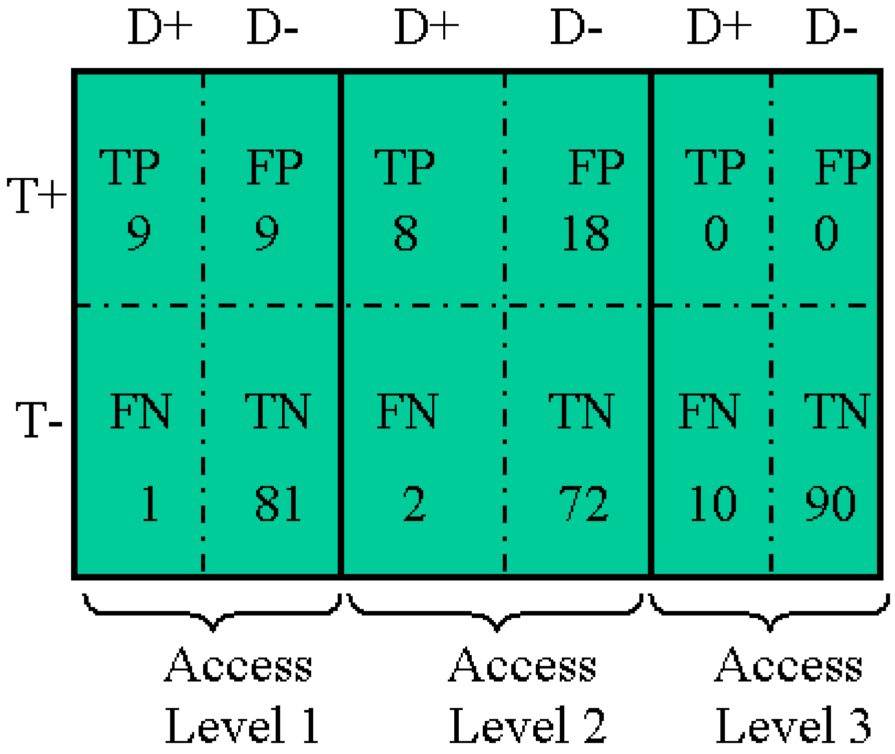
2 Overview of Modeling Framework

We generated models that allow us to compare the status quo with a world in which a new diagnostic is introduced at some point in the continuum of care. Our guiding principle is that in order to estimate the effect of any intervention one must first have a good description of the status quo: once that is available then the intervention is modeled by changing key parameters of the status quo and comparing outcome measures in the status quo with outcome measures in the world where the intervention took place.

2.1 Description of the Status Quo

We take as unit of observation an entire country. In the status quo, one or more tests for the condition of interest may be available. The kind of test individuals get depends on where they seek care. Individuals who seek care at urban hospitals are more likely to get the most sophisticated tests, while people who seek care at village clinics may get very basic tests, if any. Therefore we divide the population in groups, depending on where they access the health care system. We refer to these categories as "access levels", and assume that they can be ranked, from best to worst. For example, we might divide the population in three access levels: people at the first level seek care at a hospital, while people at the second level seek care at a village clinic, and people at the third level either do not seek care or do not receive care because care is not accessible to them. Notice that access in the status quo refers to "actual access", describing where people go when they seek care. This is different from "potential access," which describes where people could go and receive care if they wanted to. We assume that at each access level people are tested with one specific test, if any. This is not restrictive, since it is always possible to re-define access levels in such a way. Therefore, all our modeling begins by dividing the population in groups, which differ by access to care ("access levels"), and assigning one type of test to each access level. Individuals enter one of these levels while experiencing an episode of illness (for example fever and cough), which triggers health care seeking behavior.

The next step consists in modeling what happens to people in different access levels, once they enter the health care system. In all cases it is assumed that a test is administered, when available, and, depending on the outcome of the test, some form of treatment follows. This step naturally divides individuals in four categories, corresponding to the fours possible outcome of a test: true positive, false positive, false negative, true negative (see Figure 1).



	D+	D-	D+	D-	D+	D-
T+	TP 9	FP 9	TP 8	FP 18	TP 0	FP 0
T-	FN 1	TN 81	FN 2	TN 72	FN 10	TN 90
	Access Level 1		Access Level 2		Access Level 3	

Figure 1: A population with three levels of access to care—Here the population is divided in 3 groups, according to the three access levels. Within each access level people can have the disease (D+) or not have the disease (D-), and can test positive for the disease (T+) or negative (T-). Prevalence of the disease is the same across access levels. The test in access level 1 has sensitivity and specificity equal to 90 percent. The test in access level 2 has sensitivity and specificity of 80 percent. In access level 3 there is no care, so the test is 100 percent specific and zero percent sensitive.

The final step consists in assigning a health outcome to individuals, depending on their test outcome. In most cases we are interested in counting the number of people who die of the disease, but we are also interested in the potential negative effects associated with treatment, the so called "harm of treatment" (allergic reaction, utilization of resources which could have been otherwise put to better use, antibiotics resistance, stigma associated with a positive result, to name few). Usually we refer to lives lost because of the disease as "individual lives", and to lives lost because of the harm of treatment as "indirect lives". We use the word "indirect lives" because these lives are lost as the result of an indirect and unintended effect of treatment (for example because of antibiotics resistance). It is also important to note that "indirect lives" is a statistical concept, and does not refer necessarily to the lives lost to particular individuals. For example, if the harm of treatment causes two people to live half the life they would have otherwise lived, we count this as one life lost (here we are focusing on lives, but it would be as easy to use other outcomes, such as DALYs).

Quantification of the harm of treatment is a very difficult task, which requires information difficult to collect and analyze. In Section 6 we describe in detail how we bypass the problem of a direct computation of the harm of treatment and how we estimate bounds on it based on more easily collected evidence.

Once the health outcomes have been defined the last step simply consists in aggregating the outcomes over the access levels. This process leads to a description of the status quo that can be validated against observable measures. For example, this procedure leads to counting the number of people who die of the disease in a particular country or region of the world. This figure can then be validated against official death counts available from WHO or similar sources.

2.2 Introduction of the New Diagnostics

Once the status quo has been modeled, it is conceptually easy to simulate the introduction of a new test with given test characteristics. First one has to hypothesize what proportion of the population will have access to the test, and then decide how this proportion is distributed across the different access levels. In conjunction with disease experts we have decided to adopt a "hierarchical access model" for the introduction of the new test. Details about this model can be found in Section 4. The basic assumption is that access levels can be ordered; from best to worst, and that the new diagnostics is available first to individuals with the best access to care, and then to individuals with progressively worse access to care. Suppose for example that 20 percent of the population is in the first (best) access level, and 40 percent of the population is in the second one (next to best). If we assume that the new test is available to less then 20 percent of the population then this proportion falls entirely into the first access level. It is only when access to the new test is over 20 percent that individuals in the second access level gain access to the new test.

For example, if access to the new test is 40 percent, then this proportion covers all the population in the first access level, and half of the population in the second one.

Once it has been determined which groups of people have access to the new test one has to determine whether the new test would actually be used. Clearly a test which is more sensitive and more specific than the status quo test would be adopted, perhaps in conjunction with the status quo test. However, it is possible that one has to consider a test, which is more sensitive than the status quo test, but less specific. Then a criterion must be used to decide which test is better. The criterion will depend on the trade-off between the value placed on sensitivity and specificity, which ultimately will depend on the harm associated with treatment. If there were no harm associated with treatment, of if we were to ignore it, this entire exercise would be useless, since specificity would have no value and treating everybody without testing would be the optimal strategy. Details about how we decide which test is better will be presented in Section 6. In the end, in the world where the new test has been introduced some people will have access to the new test, depending on their access level, and the new test may or may not be adopted, depending on its characteristics. Once the choice between tests has been made, the rest of modeling proceeds as in the status quo: health outcomes are determined at each access level and aggregated, producing counts of individual lives lost to the disease and indirect lives lost to the harm of treatment. Then, these outcomes can be compared to the status quo. In particular, the status quo outcomes can be subtracted from the outcomes obtained with the new diagnostics to produce measure of improvement over the status quo brought by the new test. In general the outcomes are expressed in terms of lives, and therefore the results are expressed in term of individual and indirect lives saved (over the status quo).

3 Methodology

In this section we formalize the approach outlined in the previous section, and provide a generic set of formulas, which can be applied to analyze a variety of scenarios.

In order to make our approach as transparent as possible we start from the simplest setting, that we call "the homogeneous country". This setting describes a country composed of individuals who can differ by the presence of a single disease, but are otherwise identical. We start by describing the status quo, and then proceed to analyze what happens when a new test is introduced.

3.1 The Homogeneous Country: Description of the Status Quo

In the simplest case individuals in a population of interest (for example children age 0 to 5) are screened each year for the presence of a certain disease, and individuals who test positive for the disease are treated. The individual health outcome at the end of the year depends on the presence of the disease and on the results of the test, and the health of the population in defined by the average health outcome. In order to quantify these statements we need to introduce some random variables and their associated probability densities. In order to simplify our notation we will use the same symbol, $P(\cdot)$, for different probability densities: we write $P(x)$ and $P(y)$ instead of $P_x(x)$ and $P_y(y)$ to denote the distinct probability densities of random variables x and y. We define the following:

- $d \in \{D+, D-\}$: this is a random variable which indicates disease status, that is whether an individual has or does not have the disease. $P(D+)$ is the prevalence of the disease in the population of interest.

- $o \in \{T+, T-\}$: this is a random variable which indicates whether the test outcome is positive or negative.

The health outcome is a generic function of health status and test outcome, which we denote by $f(d,o)$. The average health outcome in the population is:

$$\mathrm{E}_{sq}[f] = \sum_{d,o} f(d,o) P(d,o)$$

where the sum $\sum_{d,o}$ runs over the 4 possible combined values of the random variables *d* and *o*. The subscript "*sq*" reminds us that we are describing the status quo. It is useful to rewrite the expression above in terms of conditional expectations:

$$\mathrm{E}_{sq}[f] = \mathrm{E}_d[\mathrm{E}_o[f \mid d]] = \sum_d P(d) \sum_o f(d,o) P(o \mid d) \qquad (1)$$

We call equation (1) a "representation" of the status quo: it captures the fact that the health of the population is the average of the health outcomes of subsets of the population, defined by all the possible combinations of disease status and test outcome. It is useful to make the representation of the status quo clearer by defining the following quantities:

$$f(D+,O+) \equiv TP,\ f(D+,O-) \equiv FN,\ f(D-,O+) \equiv FP,\ f(D-,O-) \equiv TN$$

$$P(T+ \mid D+) \equiv \text{sens},\ P(T- \mid D-) \equiv \text{spec}$$

$$P(D+) \equiv p$$

Using standard diagnostic terminology the quantities *TP*, *FN*, *FP* and *TN* are identified with the health outcomes of true positives, false negatives, false positives and true negatives respectively, while *sens* and *spec* are the sensitivity and specificity of the test, and *p* is the prevalence of the disease in the population being tested.

Using these definitions the representation of the status quo can be rewritten as follows:

$$E_{sq}[f] = p\,sens(TP - FN) + (1-p)spec(TN - FP) + pFN + (1-p)FP \qquad (2)$$

The representation of the status quo (2) is useful because it identifies what the health outcome depends on, and forms the basis for the analysis of interventions. For example, suppose an intervention replaces the current diagnostic test with a perfect test, that is a test that is 100 percent sensitive and 100 percent specific. In this new world health outcome is:

$$\mathrm{E}_{new}[f] = p(TP - FN) + (1-p)(TN - FP) + pFN + (1-p)FP$$

and the improvement in health outcome, relative to the status quo, is immediately computed as:

$$\Delta f \equiv \mathrm{E}_{new}[f] - \mathrm{E}_{sq}[f] = p(1-\mathrm{sens})(TP - FN) + (1-p)(1-\mathrm{spec})(TN - FP)$$

Before moving to more complex examples we note that the representation of the status quo (1) lends itself to an easy graphical depiction, borrowed from the discipline of decision trees, shown in figure (2).

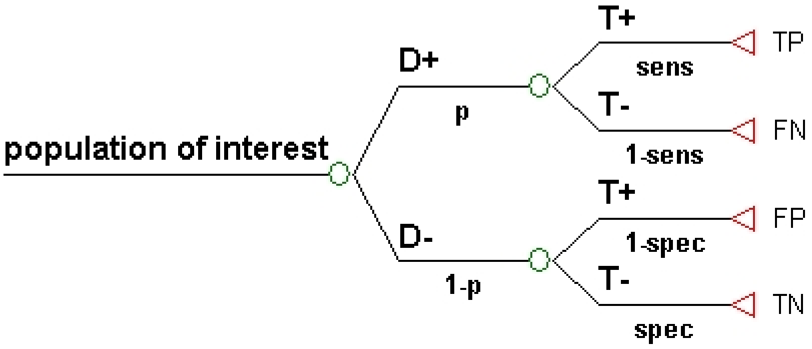


Figure 2: basic decision tree with disease status and test outcome— In this tree the population is first split accordingly to disease status and then according to test outcome. Health outcomes are then attached to each subset of the population so obtained. The health outcome of the population as a whole is an average of the health outcome of its subsets, according to the probabilities indicates under the tree branches.

Such graphical representation will prove to be a useful tool when discussing more complex situations, as we will see shortly.

3.2 The Homogeneous Country: Introducing a New Test

In the world described above every individual is diagnosed with the same test. Here we consider a new world in which a new test is made available to a proportion of the population of interest. Therefore we introduce the following additional random variables:

- $n \in \{N+, N-\}$: this random variable indicates where an individual has or does not have access to the new test. $P(N+)$ is the proportion of the population with access to the new test.
- $t \in \{t_{sq}, t_{new}\}$: these random variables indicate whether an individual is tested with the status quo test or with the new test.

Here we follow the same strategy as before, and write a representation of the status quo in terms of key conditional probabilities:

$$\mathrm{E}[f] = \sum_{d,o,n,t} f(d,o) P(o \mid d,n,t) P(d \mid n,t) P(t \mid n) P(n) \quad (3)$$

The new quantity here, which needs to be modeled further, is $P(t \mid n)$. Where the new test is not available the old status quo test is used with certainty, which implies that $P(t_{sq} \mid n = N-) = 1$. However, we need to model the case in which both test are available, that is $P(t \mid n{=}N+)$. We make the following assumptions:

- There exists a unique way to rank any two tests. In other words, it is always possible to say whether the new test is better than the status quo test. We assume that this is done by assigning a health benefit, $V(t)$, to a test t, and by choosing as best test the test with higher health benefit. In the future we will refer to the health benefit $V(t)$ as the "value of the test". Denoting by $\theta(\cdot)$ the Heaviside step function, the indicator for whether the new test is better than the status quo test can be then written as $\theta(V(t_{new}) - V(t_{sq}))$

- The new test is used if and only if it is better than the status quo test.

- The two statements above imply the following model for the probability that one particular test is used:

$$P(t \mid n = N+) = \delta_{t,t_{new}} \theta\big(V(t_{new}) - V(t_{sq})\big) + \delta_{t,t_{sq}} \theta\big(V(t_{sq}) - V(t_{new})\big) \qquad (4)$$

In order to simplify our notation for the rest of the paper we define $\theta^{new} \equiv \theta\big(V(t_{new}) - V(t_{sq})\big)$ In other words, θ^{new} is an indicator for whether the new test is better than the status quo test. Plugging equation (4) in the representation of the status quo (3), and making explicit the dependence on the availability of the new test, the average health outcome in the world with the new test is then represented by the following expression:

$$\begin{aligned} \mathrm{E}[f] = {} & P(N-)\sum_{d,o} f(d,o)P(o \mid d,t_{sq})P(d \mid N-,t_{sq}) + \\ & +\theta^{new} P(N+)\sum_{d,o,t} f(d,o)P(o \mid d,t^{new},N+)P(d \mid N+,t^{new}) + \qquad (5) \\ & +\big[1-\theta^{new}\big]P(N+)\sum_{d,o,t} f(d,o)P(o \mid d,t_{sq},N+)P(d \mid N+,t_{sq}) \end{aligned}$$

A useful feature of the representation (5) is that it highlights the parameters needed in the model and their dependencies. This in turn prompts an analysis of the simplifying assumptions that will be likely made in modeling more complicated scenarios. For example, in equation (5) we find the conditional probability $P(d \mid n,t)$: this means that in this model we allow the prevalence of the disease to vary with the availability of the new test and with the test actually being used. While it is certainly possible that areas where the new test is being introduce have higher or lower than average prevalence, it is unrealistic, in general, to expect that data will be available to model this dependency. Therefore in most cases one expects to make the assumption $P(d \mid n,t)$. A similar reasoning holds for the conditional probability $P(o \mid d,t,n)$, for which the dependency on n might be dropped. Because assumptions of this type will be made in most parts of the paper we summarize them below:

- *P*(*d* | *n*,*t*)=*P*(*d*). This assumption is usually motivated by data reasons: prevalence data are not very precise, and often only one, average, prevalence is available. Exceptions may happen when prevalence data are available by rural/urban region, and when one model scenarios in which rural and urban areas have different access to the new test.

- *P*(*o* | *d*,*t*,*n*)=*P*(*o* | *d*,*t*). Assumptions of this type are very common: conditional on disease status the outcome of the test does not depend on anything else (in this case it does not depend on whether the new test is available or not). This assumption is equivalent to say that each test is uniquely defined by its test characteristics.

Using the two assumptions above one can rewrite the expected health outcome as:

$$\begin{aligned} \mathrm{E}[f] = {} & P(N-)\sum_{d,o} f(d,o)P(o \mid d,t_{sq})P(d) + \\ & + \theta^{new} P(N+)\sum_{d,o} f(d,o)P(o \mid d,t_{new})P(d) + \\ & + \left[1-\theta^{new}\right]P(N+)\sum_{d,o} f(d,o)P(o \mid d,t_{sq})P(d) \end{aligned} \tag{6}$$

In the language of decision trees equation (6) can be represented as shown in Fig. 3.

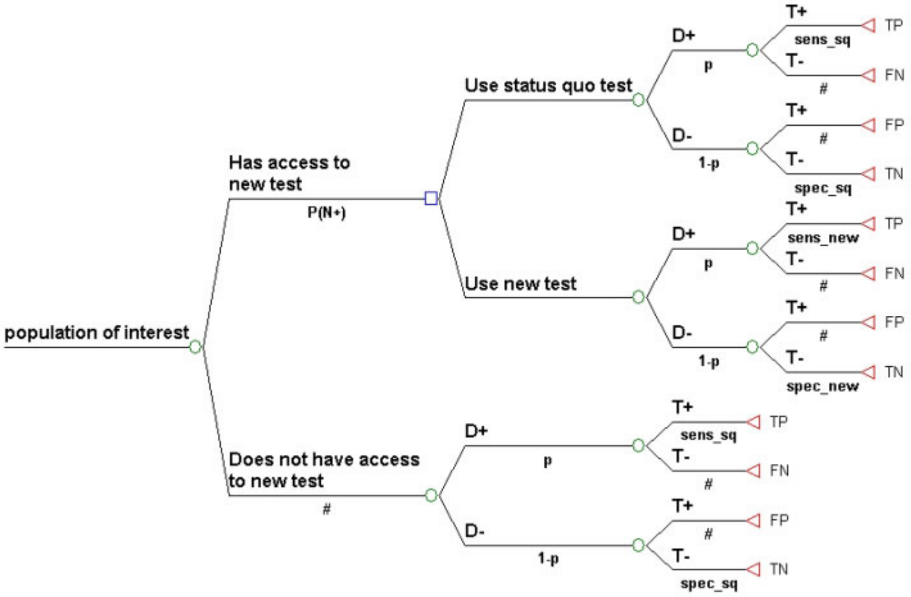


Figure 3—Decision tree for the world with a new test. The population of interest is first split according to whether they have access to the new test. For those who don't, the tree resembles the status quo. For those who have access to the new test there are two options: they can receiver either the new test or the status quo test. Which one they receive depends on which test is better. See text for comments on how to define "better."

There is an important limitation in the graphical representation shown in Fig. 3. The decision tree of Fig. 3 suggests that the decision between the branch "Use status quo test" and "Use new test" is made according to the expected value of health outcome in those branches. In other words, looking at Fig. 3 one would conclude that

$$V(t)= \sum_{d,o} f(d,o)P(d,o \mid t)$$

However, this does not have to be the case, and the health outcome used to make the decision does not have to be the health outcome of which we are computing the expected value. In general we will be interested in computing multiple health outcomes at the same time, and only one of them will be the one used to compute the value of the test $V(t)$.

Before moving to a more complex example we notice that expression (6) is somewhat cumbersome, and does not address the real object of interest. In fact, the quantity we are really interested is the benefit of introducing the new test, relative to the status quo. In other words, our quantity of interest is:

$$\Delta f \equiv E[f] - E_{sq}[f]$$

Noticing that

$$E_{sq}[f] = \sum_{d,o} f(d,o) P(o \mid d, t_{sq}) P(d)$$

and substituting this expression in equation (6) we obtain, after some simple algebra:

$$\mathrm{E}[f] = \mathrm{E}_{sq}[f] + \theta^{new} P(N+) \sum_{d,o} f(d,o) \big[P(o \mid d, t_{new}) - P(o \mid d, t_{sq})\big] P(d)$$

Therefore the improvement over the status quo assume the simple expression:

$$\Delta f = \theta^{new} P(N+) \sum_{d,o} f(d,o) \big[P(o \mid d, t_{new}) - P(o \mid d, t_{sq})\big] P(d)$$

The interpretation of this formula is very simple. The term θ^{new} says that there is improvement only if the new test is preferable to the status quo test. The term $P(N+)$ says that the improvement is proportional to the number of individuals who have access to the new test. The sum $\sum_{d,o}$ is the improvement over the status quo in the case in which the new test replaces the status quo test, which depends on the difference of the test characteristics between the new and the status quo test. We could have written this formula at the very beginning, but the reason for which we went through this more complicated derivation is to highlight the assumptions behind it.

It is worth to make this formula more explicit and express the results in terms of the test characteristics. We use the following notation in the rest of this report:

$$\Delta\text{sens} \equiv \text{sens}_{new} - \text{sens}_{sq}, \Delta\text{spec} \equiv \text{spec}_{new} - \text{spec}_{sq}$$

Using this notation we have:

$$\Delta f = \theta^{new} P(N+)[p\Delta\text{sens}(TP - FN) + (1-p)\Delta\text{spec}(TN - FP)] \qquad (7)$$

3.3 Introducing access to care

The scenario described in the previous section is obviously unrealistic in its homogeneity assumption, and was presented mostly for exposition purposes. Individuals are different and the impact of a new test will affect people differentially according to their characteristics. The challenge here is to find a suitable set of variables that determine what kind of test individuals get and that affect health outcome.

It was agreed with a panel of experts that a key variable is access to care. Individual with an episode of illness enter the health care system at different points, receiving widely different care. Individuals living in urban areas may have access to large urban hospitals, while individuals in rural communities may seek care from a local healer or a pharmacist (or may not seek care at all). When a new test is developed it is very likely that its availability in a country is different at different levels of access to care, and its impact is certainly going to be different. The test characteristics of the new test may be such that the test is not better than what is currently available in, say, a hospital, but might be much better than what is available at a health clinic.

Differentiating among different level of access to care is therefore crucial in the evaluation of the impact of a new test. Formally this is done by adding one variable, that we call "access to care" or "access level", which represents the point of entry in the health care system for individuals who seek care.

Notice that we are explicitly restricting the population of interest to those who seek care. This is an important assumption, since it is possible that health care seeking behavior is modified by the introduction of a new test. Modeling this feature is not difficult in theory, but in practice there is no data to quantify the size of this effect. Qualitatively, it is usually agreed that if a new test is introduced more people may seek care, implying that not taking this factor in account we underestimate the size of the benefit of the new test.

In order to model this scenario we need to introduce the following:

- $a\varepsilon\{a_1, a_2, ..., a_{\mathrm{A}}\}$. This is the level of access to care (from best to worst). For example we could have A=3 and $a_1 = hospital, a_2 = healthclinic, a_3 = nocare$..

- $t\varepsilon\{t_1, t_2, ..., t_{\mathrm{A}}, tnew\}$. Here t_a is the status quo test available at the level a of access to care. We assume that the new test being introduced is the same across levels of access to care.

The strategy to derive a representation of the status quo and the corresponding improvement over the status quo is the same as outlined in the previous section: one writes the expected value of the health outcome as a series of conditional expectations. The algebra is straightforward, so we only report the final formula for the improvement over the status quo and then highlight the assumptions we have made to get to this formula. The net health benefit is:

$$\Delta f = \sum_a \theta^a P(a, N+)\left[p\Delta\mathrm{sens}_a(TP - FN) + (1-p)\Delta\mathrm{spec}_a(TN - FP)\right] \qquad (8)$$

where we have defined:

$$\theta_a \equiv \theta\big(V(t_{new}) - V(t_a)\big)$$

We have also defined:

$$\Delta\mathrm{sens}_a \equiv \mathrm{sens}_{new} - \mathrm{sens}_a,\ \Delta\mathrm{spec}_a \equiv \mathrm{spec}_{new} - \mathrm{spec}_a$$

where $sens_a$ and $spec_a$ are the status quo sensitivity and specificity at level a of access to care.

The interpretation of equation (8) is simple, and it says that the net health benefit is the sum of several contributions, one for each level of access to care. The factor θ^a says that there is a contribution from level a only if at that level the new test is better than the status quo test t_a. The factor $P(a,N+)$ is the proportion of the population whose access to care is *a and* that has access to the new test. This is a key quantity, which needs to be modeled and that will be discussed in Section 4. Finally, the remaining factor measures the improvement in health outcome conditional on having access to the new test and being at level *a* of access to care.

It is important to highlight the assumptions we have made in the derivation of equation (8):

- $P(d \mid n,a,t)=P(d)\equiv p$. There is only one prevalence, independent of access level. This assumption is dictated by data restrictions. It is likely that prevalence differs by access level, but it is difficult to have data about it. If data are available to take this source of variation in account one could write $P(d \mid n,a,t)=P(d \mid a)\equiv p_a$ and equation (8) should be modified as follows:

$$\Delta f=\sum_a \theta^a P(a,N+)\left[p_a \Delta \mathrm{sens}_{\mathrm{a}}(TP-FN)+(1-p_a)\Delta \mathrm{spec}_{\mathrm{a}}(TN-FP)\right]$$

- $P(o \mid d,n,a,t)=P(o \mid d,t)$: conditional on disease status the outcome of the test does not depend on access parameters.

In order for equation (8) to be useful in practice one needs to have a model for the joint probability of access to care and access to the new test, which is discussed in the next section.

4 Access to the New Diagnostics: Hierarchical Access Model

In the model described above a key quantity is the joint probability of access to care and access to the new diagnostics, $P(a,N+)$. There is common agreement that these two variables are correlated: whenever a new test is introduced into a country the probability that an individuals has access to the new test depends on the level of access to care. Therefore we exclude a priori the model $P(a,N+)=P(a)P(N+)$ as unrealistic. A more realistic approach is to assume that access levels can be ordered, from best to worst (a_1 being the best), and that the new diagnostics is available first to individuals with the best access to care, and then to individuals with progressively worse access to care. For example we could have

$a \in \{a_1 = hospital, a_2 = healthclinic, a_3 = nocare\}$.

For access level a_1 this is formalized by the following:

$$P(a_1, N+) \equiv \min[P(a_1), P(N+)]$$

which implies the following model for the conditional probability:

$$P(a_1 \mid N+) = \frac{\min[P(a_1), P(N+)]}{P(N+)} = \min\left[\frac{P(a_1)}{P(N+)}, 1\right]$$

For example, if $P(N+)$=0.4 and $P(a_1) = 0.2$,, then $P(a_1 \mid N+) = 0.5$ because the set of people in access category 1 is a subset of the set of people having access to the new test. On the other side, if $P(N+)$=0.1 and $P(a_1)$=0.2 , then $P(a_1 \mid N+) = 1$,, because the set of people with access to the new test is a subset of the set of people in access category 1. When $P(N+) > P(a_1)$ there is a non-zero "residual" of access to the new test, equal to $P(N+) - P(a_1)$, which can be "allocated" to the next best access category (level 2). In general, the residual can be written as:

$$\varepsilon_1 \equiv \max[0, P(N+) - P(a_1)]$$

The joint probability of being at access level a_2 and having access to the test is obtained by an expression similar to the one for level a_1, only with the residual replacing $P(n)$:

$$P(a_2, N+) \equiv \min[P(a_2), \varepsilon_1]$$

Therefore the probability of being in access level 2 conditional on having access to the new test is:

$$P(a_2 \mid N+) = \frac{P(a_2, N+)}{P(N+)} = \frac{\min[P(a_2), \max[0_2 P(N+) - P(a_1)]]}{} \quad (9)$$

$$= \min\left[\frac{P(a_2)}{P(N+)}\right], \max\left[0{,}1 \frac{Pa_1}{P(N+)}\right] \qquad (10)$$

The last expression makes clear that $P(a_2)$ is not a monotonic function of $P(n)$: it first increases with $P(n)$ and then decreases. The conditional probabilities for the other access categories can be written in similar way: the generic formula for k>1 is:

$$P(a_k \mid N+) = \frac{\min\left[P(a_k), \max\left[0, P(N+) - \sum_{j=1}^{k-1} P(a_1)\right]\right]}{P(N+)}$$

We summarize we rewriting the explicit formulas for 4 levels of access:

$$P(a_1 \mid n) = \frac{\min[P(a_1), P(n)]}{P(n)} \qquad (11)$$

$$P(a_2 \mid n) = \frac{\min[P(a_2), \max[0, P(n) - Pa_1]]}{P(n)} \qquad (12)$$

$$P(a_3 \mid n) = \frac{\min[P(a_3), \max[0, P(n) - P(a_1) - P(a_2)]]}{P(n)} \qquad (13)$$

$$P(a_4 \mid n) = \frac{\min[P(a_4), \max[0, P(n) - P(a_1) - P(a_2) - P(a_3)]]}{P(n)} \qquad (14)$$

In the attached interactive spreadsheet (`hierarchical_access.xls`) we implemented the formulas above in the case of three levels of access (hospital, health clinic, no care). The spreadsheet allows to vary the level of access to the new test and shows how the people who will have access to it are distributed across the access levels.

5 Modeling More Complex Scenarios

The models we have discussed so far attach an health outcome to each test outcome. However, we have not been specific about how these outcomes are computed and what they depend on. In the simplest case we are interested in counting how many people survive an episode of the disease. Therefore, denoting by m_{tp} and m_{fn} the case fatalities for true positives and false negatives (treated and untreated disease cases), assuming a unit population, we could set

$$TP = m_{tp},\ FP = 0,\ FN = m_{fn},\ TN = 0$$

Such a choice would lead to count the number of people who die of the disease of interest. However, in many cases this description is too simplistic: it assumes that a test is performed, the patient is treated and she/he either will survive or die. In many cases other events take place after the test is performed, which may affect outcomes. In the following we list some of the most common, and show how to include them in the model.

5.1 Tests with Loss to Follow-up

In many cases the test which is first performed on the patient is slow, and the results are not immediately available (sputum test for TB is an example, some RPR tests for syphilis are another). This implies that the patient will have to return one or few more days later to pick up the test results and be treated, if necessary. In developing world countries the patient return rate is often low, and must be included in the analysis. The effect of such loss to follow-up is clear: it lowers the overall sensitivity of the test, while making specificity higher. In fact, individuals with positive test results who do not return experience the same health outcome of individuals who return and have false negative results. Similarly, individuals with false positive results who do not return experience the same health outcome of individuals who return and are true negatives.

A probability tree representing this process is shown in Fig. 4, where individuals have a probability of returning equal to p_{return}.

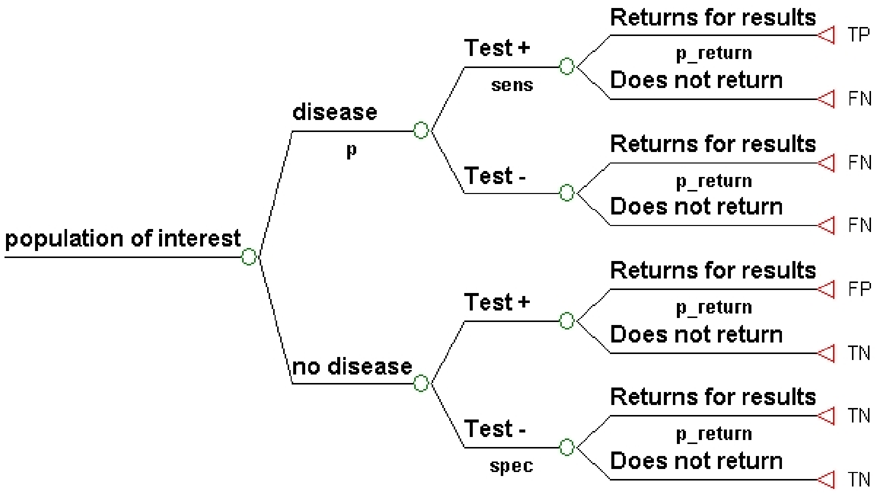


Figure 4—Tree for test with loss to follow up

Notice that, in this model, for people who test negative it does not matter whether people return or not. In practice, however, a negative result may have different outcomes depending on whether the person returns or not. For example, a negative result may trigger further testing, and this will have to be modeled too (see next section). It is easy to show that the expected outcome associated with the tree of Fig. 4 is as follows:

$$\begin{aligned}\mathrm{E}_{sq}[f] &= p\,p_{return}\mathrm{sens}(TP - FN) + \\ &+ (1-p)(1 - p_{return}(1-\mathrm{spec}))(TN - FP) + p\,FN + (1-p)FP\end{aligned} \tag{15}$$

Comparing this equation with equation (2) (the case with p_{return}=1) we see that having a rate of return which is less than one is equivalent to have a rate of return equal to one, but lower sensitivity and higher specificity. In particular, the sensitivity and specificity should be adjusted as follows:

$$\begin{aligned}\mathrm{sens}' &\equiv p_{return}\mathrm{sens} \\ \mathrm{spec}' &\equiv 1 - p_{return}(1-\mathrm{spec})\end{aligned} \tag{16}$$

For example, a “slow” test which is 80 percent sensitive and specific but has a return rate of 50 percent is equivalent to a “fast” test, with return rate 100 percent, which is only 40 percent sensitive but 90 percent specific. In terms of decision trees, this result says that we can replace the tree of Fig. 4 with the tree of Fig. 5

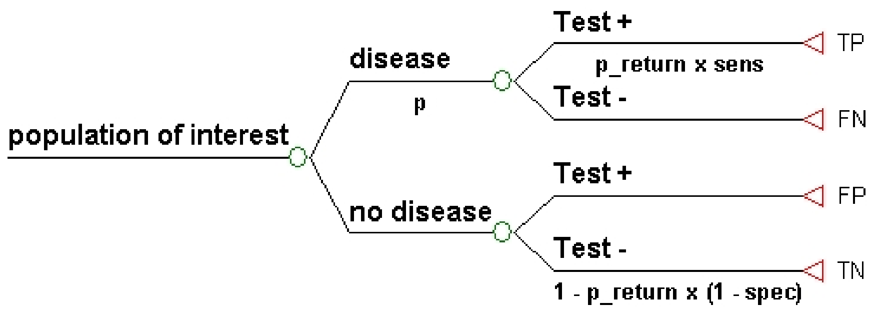


Figure 5—An alternative tree with loss to follow up. This tree has no explicit loss to follow up, but it is equivalent to the tree with loss-to-follow up of Fig. 4.

The fact that one can replace nominal sensitivity and specificity of the test with "effective" sensitivity and specificity in order to model more complex scenario is a common theme, as we will see in the following section. It is also a very useful fact, since it allows to write complex models with relatively simple trees.

5.2 Ineffective or Unavailable Treatment

A common problem, in addition to the slowness of the test, is that a positive test result is not a guarantee that the patient receives treatment. Usually this happens because treatment supply may not be fully available, or because treatment is simply not effective on some patients. For example, a stage 3 AIDS patient may test positive for a low CD4 count and be recommended treatment with anti-retrovirals (ART), but ART may only be available to a small fraction of the population. Similarly, a child may by hypoxic, test positive for severe pneumonia and referred to a hospital for oxygen therapy, but an hospital may not be within reach, and even if it is it may not have a supply of oxygen.

Conceptually these cases are equivalent to a test with loss-to-follow-up: some individuals with the disease who test positive will become false negatives, and some individuals without the disease and test positive will become true negative. Therefore, as in the previous section, we can model this scenario by simply replacing the nominal test characteristics with those specified in equation (16) above and replacing p_{return} with the probability of receiving effective treatment.

5.3 Follow-Up Tests

In some cases the outcome of a test may trigger a further round of testing. This may happen, for example, because a test with the desired characteristics is not available, or because a test is expensive and needs to be rationed. In two typical cases a second test is performed only if the outcome of the first test is positive, or negative.

In Fig. 6 we show the tree corresponding to the case in which the second test is performed only to confirm a positive test result. This situation may arise, for example, when screening for active syphilis in pregnant women, when a rapid test is used to confirm a positive results obtained with a standard test.

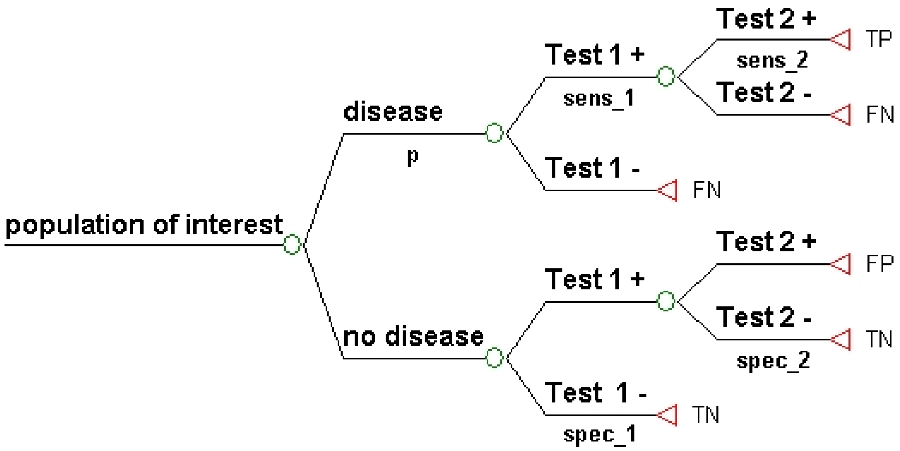


Figure 6—A confirmatory test for a positive result. Some true positives from the first test are "transformed" into false negatives, while some false positives from the first test are "transformed" into true negatives. The net effect is lower sensitivity and higher specificity.

As in the previous section, we can think of the consecutive tests as one test with certain test characteristics. For the positive confirmatory scenario of Fig. 6 the test characteristics of the combined test are as follows:

$$\begin{aligned}\text{sens}' &\equiv \text{sens}_1\text{sens}_2\\ \text{spec}' &\equiv 1-(1-\text{spec}_1)(1-\text{spec}_2)\end{aligned}$$

The effect of the second test is clear: because it catches the positive results from the first test it will reduce the overall sensitivity and increase the overall specificity.

The opposite scenario, in which the second test is used only if the result of the first test is negative, has the exact opposite effect: the sensitivity of the overall test is increased, but the specificity is reduced. The formulas for the test characteristics of the overall test are as follows:

$$\begin{aligned}\text{sens}' &\equiv 1 - (1 - \text{sens}_1)(1-\text{sens}_2)\\ \text{spec}' &\equiv \text{spec}_1\text{spec}_2\end{aligned}$$

6 Harm of Treatment and the Trade-off between Sensitivity and Specificity

When a new diagnostic is introduced it is crucial to be able to say whether the new test is better or worse than the status quo test. The answer depends on the health outcomes of interest associated with the test. Outcomes of interest might be the number of people who survive because appropriately treated, the number of people who die because of adverse reaction to treatment, or the cost associated with unnecessary treatments. Some of these outcomes depend on the sensitivity of the test, and some on the specificity, and therefore different tests are associated with different vectors of health outcomes. Unfortunately there is no obvious way to compare these vectors, since health outcomes will be in general not comparable (unless we are in the lucky and unrealistic situation in which one test dominates the other along each health outcome).

The typical situation we face in this paper is one in which there are negative externalities (cost to society) associated with treatment. Therefore the number of treatments administered is a key outcome, together with mortality. While mortality is controlled by sensitivity, the number of treatments is controlled by both sensitivity and specificity. Therefore tests with different characteristics will lead to different combinations of mortality and number of treatments.

Consider, for example, test A, which leads to the use of 500,000 treatments and saves 100,000 children, and test B, which leads to the use of only 300,000 treatments but saves only 80,000 lives. It is not obvious a priori which of the two tests is preferable: test A saves 20,000 more lives than test B, but does so at the price of an additional 200,000 treatments. If the negative externalities associated with treatment are sufficiently large, test B might be preferable to test A, even if it saves the lives of fewer children in the short term. In order to tell which test is better we need to compare apples and oranges, that is a reduction in mortality with an increase in number of treatments. This is a difficult task, for which a solution, however, must be found, otherwise it is impossible to make choices among options, nor it is possible to compare the new test with a status quo test.

In order to compare apples and oranges we need to say how many apples in a orange: in this context this means that we need to convert the number of treatments in a mortality figure. Therefore we need to say that for each individual treated the negative externalities associated with treatment lead to the loss of a certain number of lives (or life years, which can then be converted into lives), possibly at some point in the future. Since these are not necessarily individual lives, we refer to them as "indirect lives". The number of lives lost for each treatment (usually a very small fraction) is called "harm of treatment" and is denoted in this paper by *C*.

Let us consider, for the sake of example, the case of a disease which requires antibiotics treatment. The harm of treatment would include at the very least the following:

- Each time we treat a patient with antibiotics we increase the chance of development of antibiotics resistance. Antibiotics resistance implies that at some point in the future some people may die because infected with a resistant strain of bacteria for which treatment may not be available. In addition, once resistance has built up, a new line of drug has to be administered, leading to additional cost, using up resources which would have been otherwise used to save lives.
- Each time we treat a patient with antibiotics there is a small possibility of an adverse drug reaction, leading to loss of lives.
- Each time we treat a patient with antibiotics we utilize scarce resources: we need to pay for the cost of treatment itself and to provide the labor necessary for administering the treatment. These resources would have been otherwise used to treat other patients, which will now go untreated, leading to loss of life.

In most cases, a detailed computation of the harm of treatment is not directly available. However, this does not mean that we do not know anything about the harm of treatment. In fact, whenever a new test is developed a range of options is considered, with different values of sensitivity and specificity. These options are usually summarized in a ROC curve and some criterion is used to choose among the options. If we had access to this criterion, we could use it to rank any pair of tests. Unfortunately this is usually not the case, and it is possible that the criterion used is experts' consensus, which is not easily captured. However, the choice made over the options reveals some information about the preference over the different tests. This is best explained with the following example.

Example

In ref. 3 the authors consider the task of finding good clinical indicators for diagnosis of acute bacterial meningitis in children at a rural Kenya district hospital. The authors compared several combinations of clinical indicators (such as bulging fontanel, cyanosis, etc.) and produced a table which listed for each combination the corresponding sensitivity and specificity. They then decided that the combination with 79 percent sensitivity and 80 percent specificity was the best option. In particular they wrote "additional improvements in sensitivity (e.g., up to 97 percent) are possible but only at the expense of an unacceptable loss of specificity". To be precise, they decided that a test that is 97 percent sensitive but only 44 percent specific is inferior to their best option. In other words, gaining an additional 18 percentage points in sensitivity was not worth losing 36 percentage points in specificity. Similarly, they deemed a test which was 71 percent sensitive but 88 percent specific was suboptimal. In other words, gaining an additional 8 percentage points in specificity is not worth losing 8 percentage points in sensitivity.

The data cited in the paper are actually sufficient to "back out" an approximate description of the criterion used to make the choice, since they basically provide a description of the trade-off between sensitivity and specificity. The details of how this can be done will be given in the next section. Here we want to outline our general strategy for devising a way to compare tests in absence of detailed information about the harm of treatment:

- Observe choices made over tests with different characteristics;
- Analyze the observed choices and back out, approximately, the criterion used to make the choice, or, equivalently, the harm of treatment;

- Used the criterion so constructed to compare any two tests.

On the surface, this strategy seems impractical because ROC curves, in our setting of interest, are not easily found. However, we point out that whenever a test is considered by the medical community fit to be used, a choice is being made: it is considered better to use this test than treating no one (a test which is 0 percent sensitive and 100 percent specific), and it is also considered better to use this test than treating everyone (a test which is 100 percent sensitive and 0 percent specific), since these choices are always available.

For example, if we are willing to use a test that is highly sensitive but has very poor specificity, we are implicitly saying that the harm associated with treatment is not very large, since we are willing to tolerate a large number of false positives. Conversely, if the clinical community is reluctant to use a test unless it has at least a specificity of, say, 95 percent, this suggests that the harm of treatment may be high (these considerations depend clearly on the prevalence of the disease being tested).

As we will see in details in the next section, this line of reasoning implies that by merely observing the characteristics of a test currently used we can "back out" an estimate, possibly subject to large uncertainty, of the harm of treatment. The appealing feature of this idea, which has flavor of the revealed-preference approach of neo-classical economics[4], is that it can be used in any situation and it will always produce an answer with a lower and an upper bound. Its disadvantage is that it is not interpretable: it summarizes the collective decision of the medical community about whether a test should or should not be used, but it does not give us any insight on the factors which influenced that decisions. It is often possible to corroborate the findings of this method with ad hoc calculations or by consulting with an experts' panel, making the results more credible and less uncertain. In all cases sensitivity analysis can be used to study how the estimate of the harm of treatment affects the results.

6.1 Comparing Two Tests

We consider the standard binary test, which is graphically represented by the tree in Fig. 2. We use the following notation:

- p: prevalence of the condition to be tested;
- sens: sensitivity of the test;
- spec: specificity of the test;
- TP,FN,FP,TN: net health benefit of a true positive, false negative, false positive, true negative[3].

It is crucial to note that the net health benefits *TP*,*FN*,*FP* and *TN* include both the health benefits to the individuals and the harm of treatment. We are not specifying at this point how the harm of treatment is computed, and how it is converted into an outcome which is comparable with the outcome at individual level. This step will be discussed in the next section. Here we are simply assuming that the net benefit corresponding to a test outcome can be indeed summarized by a single number. The point of this section is to demonstrate what is the minimal amount of information needed in order to compare two tests, and how to estimate it from available data.

Using the notation introduced above the health benefit associated with the test is:

$$V = p[\text{sens}\ TP + (1-\text{sens})\ FN] + (1-p)[(1-\text{spec})\ FP + \text{spec}\ TN] \qquad (17)$$

This expression can be rewritten as:

$$V = p\ \text{sens}\ (TP-FN) + (1-p)\text{spec}\ (TN-FP) + pFN + (1-p)FP \qquad (18)$$

Since we use the value V above only as mean to compare two tests, it is always possible to redefine it with an appropriate linear transformation whose parameters do not depend on the test characteristics. Therefore we redefine V by subtracting the term $pFN+(1-p)FP$ and by dividing by $(1-p)(TN-FP)$, obtaining:

[3] We refer to true positive, false negative, false positive, true negative as “test categories.”

$$V = \frac{p}{1-p}\frac{TP-FN}{TN-FP}\text{sens}+\text{spec} \qquad (19)$$

Using equation (19) we see that if we have two tests, the first will be preferred to the second if and only if:

$$\text{spec}_1 - \text{spec}_2 > \frac{p}{1-p}\frac{TP-FN}{TN-FP}(\text{sens}_2 - \text{sens}_1) \qquad (20)$$

The conclusion we draw from equation (20) is that in order to compare two tests we do not need to know all the four quantities *TP*, *FN*, *TN* and *FP*, but only a particular combination of them, that is the ratio $\frac{TP\text{-}FN}{TN\text{-}FP}$. This factor has a very simple interpretation which we know present.

Let us denote by *N* the number of individuals in the population of interest, and let us denote by N_c the number of individuals in test category *c*, so that the following holds:

$$\begin{aligned} N_{tp} &= Np\text{sens} \\ N_{\text{fn}} &= Np(1-\text{sens}) \\ N_{fp} &= N(1-p)(1-\text{spec}) \\ N_{tn} &= N(1-p)\text{spec} \end{aligned}$$

The we can easily see that, aside from a constant which does not depend on test characteristics, the average health outcome is:

$$V = \frac{TP-FN}{TN-FP}N_{tp} - N_{fp}$$

Therefore, if we are considering a new test which leads to one additional true positive and ΔN_{fp} additional false positives, this test will be preferred as long as the additional false positives remain bounded as follows:

$$\Delta N_{fp} < \frac{TP-FN}{TN-FP} \equiv \gamma \qquad (21)$$

In other words, the ratio $\frac{TP\text{-}FN}{TN\text{-}FP}$ is equal to the maximum number of false positives we are willing to accept in order to gain one additional true positive. We define this ratio as the number of "equivalent false positives", and denote it by γ in the rest of the paper. This number is the only quantity that is necessary to specify in order to compare two tests, and, when combined with prevalence information, it uniquely determines the trade-off between sensitivity and specificity. In fact, knowledge of this numbers allows answering the following question: *how many percentage points in specificity are we willing to give up in order to gain one percentage point in sensitivity?* Using equation (19) it is easy to show that the answer is simply:

$$\Delta\text{spec} = \frac{p}{1-p}\gamma \times 0.01 \qquad (22)$$

For example, if the number of equivalent false positives were γ=100 and the prevalence of the condition being tests were p=5%, we would be willing to lose 5.2 percentage points in specificity in order to gain 1 percentage point in sensitivity.

Similarly, if we are given a measure of the trade-off between sensitivity and specificity we can easily compute the number of equivalent false positives. As a simple exercise, we go back to the example on bacterial meningitis presented in the previous section. In that example we reported that the authors of the study decided that gaining an additional 18 percentage points in sensitivity was not worth losing 36 percentage points in specificity. This equivalent to say that gaining an additional 1 percentage points in sensitivity is not worth losing 2 percentage points in specificity, implying that $\Delta spec \leq 0.02$ in equation (22). Since the prevalence of bacterial meningitis at that site was about 2 percent, we find that the number of equivalent of false positives γ for that case was less than 98.

6.2 Bounds on the Number of Equivalent False Positives

In practical situations we may not be able to compute the net health benefits associated with the test outcomes. In particular, the computation of *TP* and *FP* is usually problematic because requires to quantify the harm of treatment. Therefore the number of equivalent false positives γ is often not known. However, information about γ might be gained if there is a commonly accepted status quo test. In fact, if a body of experts agreed on a certain test, it must have been certainly considered several other alternatives and ruled that the current test is preferable. Notice that this process might have not used an explicit value of γ, but this is irrelevant: the main point is that there must be a range of values of γ which is consistent with the observed choice.

More formally, let us denote by sens and spec the sensitivity and specificity of the status quo test, and let us denote by sens_i and spec_i, i=1,…,k, the sensitivity and specificity of k alternative tests which have been ruled inferior to the status quo. From equation (20) we know that the following must hold:

$$\text{spec} - \text{spec}_i > \frac{p}{1-p}\gamma(\text{sens}_i - \text{sens}) i = 1,\ldots,k$$

We notice now that each of the expressions above provides a bound on γ, provided we know the sensitivity and specificity of the k alternative tests. If we had an ROC curve for the status quo test, then we would have a large number of bounds, which could identify γ with great accuracy. This is unfortunately not a common case, and most of the times we do not have much knowledge about alternative tests to the status quo.

However, we know the following: we always have the alternative of using a random test. A random test is a test whose positive outcome is determined by flipping a biased coin, that is a test whose outcome is positive with probability z, $0\leq z\leq 1$. Such a test has sensitivity equal to z and specificity equal to 1-z. Particular cases of random tests are tests which always produce a positive result and always produce a negative result. It seems reasonable to assume that the status quo is better than any random test, otherwise it would have not been introduced. Therefore the following must be true:

$$s\text{pec} - 1 + z > \frac{p}{1-p}\gamma(z - \text{sens}) \forall z \in [0,1]$$

The inequality above can be rewritten as:

$$\begin{aligned} &\gamma < \frac{1-p}{p}\frac{\text{spec}-1+z}{z-\text{sens}} \qquad && z > \text{sens} \\ &\gamma > \frac{1\text{-p}}{\text{p}}\frac{1-\text{spec}-z}{\text{sens}-z} \qquad && z < \text{sens} \end{aligned} \qquad (23)$$

Since the inequalities above hold for any value of *z* in the appropriate range, we can choose the values of *z* which make the upper and lower bounds the highest possible, which turn out to be 1 and 0, for the first and second inequality of equation (23) respectively. In other words, *it is sufficient to assume that the status quo is always better than tests whose results are either always positive (100 percent sensitivity and 0 percent specificity) or always negative (100 percent specificity and 0 percent sensitivity).* The bounds for γ can now be rewritten as:

$$\frac{1-p}{p}\frac{1-\text{spec}}{\text{sens}} < \gamma < \frac{1-p}{p}\frac{\text{spec}}{1-\text{sens}} \qquad (24)$$

Just to give the reader an idea of the kind of numbers involved, a test which is 80 percent sensitive and 70 percent specific, with a prevalence of 5 percent leads to an lower an upper bound for γ of 7.1 and 66.5 respectively.

Notice that the bound becomes tighter for increasing levels of prevalence. In fact, the width of the bound is given by

$$\Delta\gamma = \frac{1-p}{p}\frac{\text{sens}+\text{spec}-1}{\text{sens}(1-\text{sens})}$$

In addition, the bound gets tighter as the sensitivity and specificity of the status quo test approach those of a random test (that is as their sum gets closer to one), and as the sensitivity gets closer to the value of 0.5.

It is useful to notice that the bound 24 takes a particularly simple form when stated in terms of the number of individuals falling in the different test categories; In fact, it is easy to see that this bound can be rewritten as:

$$\frac{N_{fp}}{N_{tp}} < \gamma < \frac{N_{tn}}{N_{fn}} \qquad (25)$$

In the next section we put this result in the perspective of a model with a detailed description of health outcomes.

6.3 Harm of Treatment

In the previous section we have shown how to bound the number of equivalent false positive given an observed status quo test. The only assumption made in that section was that it is possible to define a net health benefit for each of the test outcomes. In this section we go into more details, and show how the notion of harm of treatment can be quantified and how it relates to the number of equivalent false positives.

We consider the case of a test that is used to detect cases which needs treatment for a specific disease. The goal is therefore to avoid deaths from that specific disease, and the natural outcome to consider is disease specific mortality (this framework can easily adapted to other cases). If mortality were the only outcome of interest, the health outcomes would be as follows:

$$\begin{aligned} TP_{ind} &= m_{tp} \\ FN_{ind} &= m_{fn} \\ FP_{ind} &= 0 \\ TN_{ind} &= 0 \end{aligned}$$

Here the subscript *ind* reminds us that this outcome refers to individuals, as opposed as society at large. With this choice the outcome of the test would be simply the number of people dying from the disease of interest, which, assuming a unit population, would be computed as:

$$V = p\,\mathrm{sens}\,m_{tp} + p(1-\mathrm{sens})m_{fn}$$

However, if this were the case we would not need to develop a new diagnostics: because there is no harm of treatment specificity does not appear in the equation above, and treating everybody is optimal.

In practice we do care about false positives, and there are negative externalities associated with treatment. The only way to combine some measure of harm of treatment with our outcome of interest, people dying of a specific disease, is to express it in terms of number of lives lost. These should be considered as "public health" lives, or "statistical lives", and not as individual lives. Therefore, the natural way to introduce harm of treatment in this setting is to assume that each time an individual is treated, some fraction C of a life is lost at some point in the future. Formally this is equivalent to associate to true positives and false positives a mortality rate equal to C, defining the following health outcome:

$$\begin{aligned} TP_{indirect} &= C \\ FN_{indirect} &= 0 \\ FP_{indirect} &= C \\ TN_{indirect} &= 0 \end{aligned}$$

Here the subscript *indirect* reminds us that these deaths in general cannot be attributed to specific individuals[4], but rather are statistical deaths at some point in the future. The harm of treatment may have a component which relates to the individual being treated (for example an allergic reaction to treatment). However, usually this is the least interesting part of the harm of treatment, and therefore we keep referring to the deaths related to harm of treatment as future deaths.

It is plausible that one wants to assign different weights to future deaths than to individual deaths, for example using a discounting factor. This would be done by simply substituting C in the equation above with wC, for some weight w. However, since C is usually unknown and has to be estimated anyway, it seems preferable to absorb whatever weight one wants to use into the definition of C

Having defined the harm of treatment in terms of deaths, it is now meaningful to sum the two health outcomes defined above. Since it is common to define an health outcome in such a way that more is better, we also switch from a mortality outcome to survival, by

[4] Except when the treatment causes harm to the person receiving the treatment, as in the case of allergic reactions.

changing the sign of the previous definitions and subtracting them from 1. Therefore we finally define our net health benefit of the test as:

$$TP \equiv 1 - \left(TP_{ind} + TP_{future}\right) = 1 - m_{tp} - C \qquad (26)$$

$$FN \equiv 1 - \left(FN_{ind} + TP_{future}\right) = 1 - m_{fn} \qquad (27)$$

$$FP \equiv 1 - \left(FP_{ind} + FP_{future}\right) = 1 - C \qquad (28)$$

$$TN \equiv 1 - \left(TN_{ind} + TN_{future}\right) = 1 \qquad (29)$$

When we use this outcome to measure the improvement brought by the new test over the status quo test, we refer to the improvement as the number of "adjusted lives" saved, which can be obviously decomposed into an "individual lives" and "indirect lives" component.

The only task left is now to estimate the harm of treatment *C*. This can rarely be done directly, although sometimes the component related to harm to the individual receiving treatment is known. This is the case, for example, if the treatment following the test includes penicillin. In this case a component of the harm of treatment is the case fatality for anaphylactic reaction to penicillin.

In general, the harm of treatment can be bound using the methods described in the previous section, by bounding the number of equivalent false positives γ and then using the relationship between γ and the health benefit of the test, expressed in equation (21). Substituting equation (29) in equation (21) we obtain the following expression:

$$\gamma \equiv \frac{TP - FN}{TN - FP} = \frac{m_{fn} - m_{tp} - C}{C} \qquad (30)$$

Solving for *C* we then obtain:

$$C = \frac{m_{fn} - m_{tp}}{1 + \gamma} \tag{31}$$

Now we can use the bounds on γ (equation 2) to derive the following bounds on *C*

$$\frac{p(1-\text{sens})(m_{fn} - m_{tp})}{p(1-\text{sens}) + (1-p)\text{spec}} \leq C \leq \frac{p\text{sens}(m_{fn} - m_{tp})}{p\text{sens} + (1-p)(1-\text{spec})} \tag{32}$$

Somewhat surprisingly the bounds above can be rewritten in a very simple form in terms of conditional probabilities associated with the test:

$$P(D+ \mid T-) \leq \frac{C}{m_{fn} - m_{tp}} \leq P(D+ \mid T+)$$

and using the usual definitions of positive predictive value (PPV) and negative predictive value (NPV) we finally obtain:

$$1 - \text{NPV} \leq \frac{C}{m_{fn} - m_{tp}} \leq \text{PPV}$$

This expression make clear that the method described here is more informative, that is it produces tighter bounds, when the status quo test has poor characteristics that is it has low positive and negative predictive values.

Once the bounds have been estimated it remains the problem of choosing a specific value within these bounds. In absence of other information the middle point is usually a good choice, but alternatives are available. It is for example possible to present the bounds on *C* (and/or γ) to a panel of experts and have them express a preference. In our experience experts tend to prefer estimates that are closer to the lower bound. This is probably because experts feel that the status quo is better than the treating everybody, but not a whole lot better. On the contrary, experts usually tend to agree that not treating anyone because of fear of harm of treatment is bad policy. These preferences imply much more confidence in the lower bound than in the upper bound.

6.4 Limitations of Method for Bounding the Harm of Treatment

The method outlined in the previous section provides a crude way of quantifying the harm associated with treatment. It allows to quantify the trade-off between sensitivity and specificity, or between true positives and false positives, and in so doing it allows to compare any two tests for the same disease. The method does not rely on a direct estimate of the harm of treatment, but rather it bounds it by making the assumption that the test(s) currently used are better than any random test. In particular, it assumes that it is better to use the status quo test than to treat everybody or to treat no one. In a sense, this method attempts to "read the mind" of the medical community when, as a whole, it has accepted the usage of a certain test.

The bounds on the harm of treatment are extremely simple to compute: they depend on the prevalence of the condition being tested, the sensitivity and specificity of the test currently used, and the difference in mortality between false negatives (untreated individuals with disease) and true positives (treated individuals with disease). Within the uncertainty introduced by the above mentioned parameters, these bounds are quite solid: asserting that the harm of treatment is outside these bounds is equivalent to assert that the medical community has made a huge mistake, and is either treating some individuals when nobody should be treated, or not treating some individuals when everybody should be treated.

Clearly the bounds provided here can be greatly improved if the status quo test is known to be better than some other test (other than the random tests). In addition, the estimate of the harm of treatment can sometimes be validated against other, ad hoc, calculations. For example, if the cost-effectiveness of administering treatment is known, it may be possible to estimate how many lives will be lost by not investing in more cost-effective interventions.

Finally it should be pointed out that, even if the bounds on the harm of treatment can be loose, it is often the case that they are the best we have. It is important to recognize that often, without an estimate for the harm of treatment, an analysis of the benefit of introducing a new diagnostics would be very hard to perform, especially in those cases in which over-treatment is a great concern. Without knowing how much sensitivity we are willing to give up in order to gain specificity the only options we can consider are tests that have both better sensitivity and specificity than the status quo, severely restricting the scope of the analysis.

6.5 Alternative Calculations for Harm of Treatment

In many cases it is possible to obtain an alternative estimate for the harm of treatment, based on an opportunity cost consideration. Let us assume that the cost of administering the treatment is C_{treat} (this includes both capital and labor cost). Since resources are limited, if C_{treat} is spent on treating an individual, that amount of money cannot be spent on another intervention that could also save lives. Let us assume that there exists an intervention that can save one life at the cost of C_{int}, and let $\eta = C_{\text{int}} / C_{treat}$. Then, for every η treatments administered we have spent an amount equal to C_{int}, and therefore missed the opportunity of save a life. In other words, for every η treatments administered an indirect life is lost. Since the harm of treatment refers to the lives lost to one single treatment, this implies that the harm of treatment is at least $1/\eta$. We say "at least" because this calculation only takes in account the opportunity cost of treatment, while there might be additional sources of harm.

This type of calculation is useful because it help to set the order of magnitude of the harm of treatment. For instance, assume that the cost of treatment is 50 US cents (a reasonable value for antibiotics). Then for every 1,000 treatments administered, US$500 is spent. If there is at least one intervention that can save the life of one child at a cost of US$500, then for every 1,000 treatments administered, we miss the opportunity to save one child, and the calculated harm of treatment is C = 0.001.

7 Sensitivity Analysis

Once a model for the status quo and for the world with the new test has been built, the model output is a vector of health outcomes expressed as a function of the model parameters. Some of the parameters of the model refer to the status quo, and are known with uncertainty. Others refer to the world with the new diagnostics, and therefore are "free", or "design" parameters. By this we mean that we are interested in exploring how the outcomes of interest vary as a function of these parameters, which are not estimated, because refer to a hypothetical world. A typical table of parameters will look as shown in Table (1). It reports the estimated values of the status quo parameters, together with their range. For the design parameters the range has a different meaning, and simply restricts the variation of these parameters to regions of interest (for example we are unlikely to be interested in sensitivities below 50 percent).

Table (1) refers to a status quo world in which there are three levels of access to care, the test is slow (so there is return rate which is less than 100 percent) and effective treatment is not universally available. The world with the new diagnostics is modeled similarly to the status quo, but it is allowed to have different key parameters. For example, the return rate in the world with the new test may be 100 percent because the test is fast, and access to treatment could be different because treatment might be introduced along with the new test.

To each table of parameters we can assign a vector of health outcomes. It is clearly of great interest to know how sensitive is the outcome of interest to variation in the design parameters. For example, we would like to know whether an increase in access to the test of 5 percentage points is better or worse than an increase in the sensitivity of the new test by a similar amount. However, it is also of interest to know how the outcome varies as a function of the status quo parameters: since they are known with different degrees of uncertainty it is useful to know how much the uncertainty on one parameter will affect the outcome.

	Parameter	Value	Lower Bound	Upper bound
Status Quo				
	Prevalence of disease			
	Probability of access to level 1			
	Probability of access to level 2			
	Sensitivity of Status Quo test			
	Specificity of Status Quo test			
	Rate of return for test results			
	Probability of receiving effective treatment			
	Case fatality for true positives			
	Case fatality for false negatives			
	Harm of treatment			
New Test				
	Probability of access to the new test			
	Sensitivity of new test			
	Specificity of new test			
	Rate of return for results of new test			
	Probability of receiving effective treatment			

Table 1. A Typical Table of Model Parameters

Variations in outcome associated with changes in specific parameters can usually reported as elasticities. Therefore to each row of the parameter table we can assign an additional column, reporting the elasticity of the outcome of interest with respect to the corresponding parameter. The resulting set of elasticities, often plotted as a "tornado diagram", is useful to assess which parameters have the greatest impact on the model output.

A limitation of this analysis is that the results vary depending on the base case values of the design parameters. For example, if the sensitivity and specificity of the new test are set to 95 percent the elasticity of outcome[5] with respect to the sensitivity of the status quo test may be large, but if we set sensitivity and specificity of the new test to 60 percent it

[5] Outcome is always measured as improvement over the status quo.

may be zero, because the new test may not improve over the status quo. Therefore we recommend that one chooses carefully a scenario of particular interest, and computes elasticities around that particular base case, without trying to generalize the results to other scenarios.

Elasticities and tornado diagrams, however, only tell part of the story. We are also interested, in general, in knowing how much uncertainty there is on the benefit of the introduction of a new test with certain characteristics. Clearly we can only estimate the uncertainty coming from the model parameters, and have no general way to estimate the model uncertainty. The uncertainty originating from uncertainty in the status quo model parameters is estimated using a multi-way Monte Carlo simulation. Each parameter is allowed to vary in a range, estimated from the literature of from experts' opinion. Then we draw at random parameters within their ranges and compute the corresponding net health benefit. Repeating this procedure a large number of times we obtain a distribution of net health benefits, which we can then summarize. If the distribution is approximately normal we can simply summarize it with the mean and the standard deviation, while if it is skewed reporting quantile is a better option.

In our simulations we use triangular probability densities to sample from the allowed range of a parameter, setting the mean of the probability distribution around the actual value used in the parameter table. However, it is often the case that there is no triangular density with the specified mean. This happens when the mean is close to one of the ends of the range. When this is the case we switch to a monotonic Beta distribution. Remember that the Beta distribution in the interval [0,1] has density proportional to $x^{a}(1-x)^{\beta}$. When α=0 or β=0 the density is monotonically decreasing or increasing, and includes, as special case, the limit of the triangular distribution where the peak of the distribution is at one end of the range. Unlike the triangular distribution the monotonic Beta distribution can have any mean, and fixing the mean uniquely determines the α or β parameter of the density.

8 Limitations

In this final section we comment on some of the limitation of the approach described here.

- **Static model:** the models we have discussed are static, and compare two worlds, with and without the new test, over a fixed period of time, usually one year. No feedback or any dynamic component has been taken in account: next year the net health benefit will be the same as this year (except for trivial changes, such as population growth, which we take as exogenous). In cases where disease transmission plays an important role this may be an oversimplification. Since the introduction of the new test leads to more individuals being cured, one could expect that in a dynamic framework the introduction of the test would lead, in the longer term, to a reduction in the prevalence of the disease. However, the new test may bring some behavioral changes, especially if the test is introduced along with treatment, and it could be very difficult to estimate the combined effect of these factors. In principle one could merge the static approach of this model with the dynamic approach of compartmental models. In practice, however, it is not clear that the data to support such an effort are available. This does not mean that these models are totally silent about the longer term effects of the introduction of the new test. One possibility, for example, is to make use of other available models, or ad hoc models, to estimate "multipliers". In other words, it may be possible to estimate that for each additional case of the disease cured a number of infections will be avoided down the road. While this is a crude way to address the transmission of infectious disease, it may be still acceptable as long as one is aware of the potentially large uncertainty on the results. For example, it may turn out that even under the most conservative scenario, there would be a huge benefit from introducing a new test, in terms of the number of avoided future disease cases.

- **Single Disease Analysis:** within the framework presented in this paper one considers diseases one at a time: there is no interactions among diseases. This is often not the case. For example, it may be more clinically meaningful to analyze malaria and ALRI at the same time, or to take in account the interaction between HIV and TB. It is in principle possible to expand on this framework to analyze multiple disease at the same time. Again, though, the data needed in order to perform such an analysis may not be available. It has been our experience that even relatively simple single disease scenarios pose huge problems in terms of data requirements. Therefore it is possible that the type of analysis we considered in this paper is all we can do, conditional on available data.

- **Provider Behavior** In most cases we have not attempted to model the behavior of providers. In some cases providers may not act according to the result of the test, and may take in account other factors in order to make a decision. In addition, it is possible that providers' behavior may change when a new diagnostics is introduced.

- **Binary classification**. In most cases people are classified as either having the disease or not having it. In practice there are other dimensions to take in account. One is severity of the disease, which may influence health care seeking behavior as well as the test characteristics of clinical diagnosis, when used, and case fatalities. In the type of models we have discussed so far we consider the average patient, which receives an average test and experiences average case fatalities. This approach does not take in account the correlation among these quantities (for example more severe cases of bacterial pneumonia are easier to diagnose, so that the sensitivity of the clinical diagnosis is higher). One solution is to explicitly add additional variables representing the severity of the case. Another, which we have used in some models, is to make ad hoc choices in the parameters to reflect severity considerations. For example, it is reasonable to assume that only fairly sever cases of pneumonia are brought to a health provider. This can be modeled by using a sensitivity which is higher than average.

References

1. Girosi, F., Olmsted, S. et al. Developing and interpreting models to improve diagnostics in developing countries. *Nature*, S1, 3-8 (2006). Available online at http://www.nature.com/diagnostics (as of December 7, 2006).

2. Olmsted, S. S., Derose, K. P. & Beighley, C. M. Determining Access to Care and User Requirements for Diagnostic Tests in Developing Countries WR-423-HLTH (RAND Corporation, Santa Monica, 2006). Available online at http://www.rand.org/health/feature/research/0612_global.html (as of December 7, 2006).

3. J.A. Berkley, A.C. Versteeg, I. Mwangi, B.S. Lowe, and C.R.J.C. Newton . Indicators of Acute Bacterial Meningitis in Children at a Rural Kenyan District Hospital. *Pediatrics*, Vol. 114, No. 6, December 2004, pp. e713-e719.

4. Samuelson, P. A. Foundations of Economic Analysis (Harvard Univ. Press, Cambridge, 1947).