

3. Health Impact Measurement

Health impact assessment is at the core of the HIF. This chapter introduces metrics of health impact and a variety of methods for performing assessment. A substantial and well funded assessment branch will be essential for this purpose. The chapter also explores foreseeable difficulties in assessment, and how these can be anticipated in the HIF's design.

INTRODUCTION

Measuring the health impact of medicines is an essential task of the HIF. It must be able to make health impact assessments that are reasonably consistent across diseases and countries. We recognize that there is no perfect metric for health or disease and no perfect algorithm for health impact assessment, and that any such assessment will inevitably rely on imperfect data. Perfection, however, is not the relevant standard. What matters is that pharmaceutical firms should have strong new incentives to deliver health improvements – and no strong new incentives to try to capture HIF rewards without health impact. HIF assessment must be sound enough so that the best strategy for firms to capture HIF rewards is to deliver health improvements. With a substantial investment in data collection and analysis, much larger than any national health system's to date, the HIF would be in a position to make its assessments sufficiently consistent and reliable to ensure that payments were allocated fairly between registrants on the basis of health impact, and would thus provide meaningful incentives to innovators to develop products with large health impact.

The HIF is not alone in seeking to measure how drugs affect health. Because of the enormous cost of health care, the measurement of health impact is becoming more important to insurers and especially to governments, which seek to reduce expenditures and to improve health care by relying more systematically on epidemiological evidence. Thus, there has been a

recent flowering of health-technology assessment programs, such as the Canadian Agency for Drugs and Technologies in Health (CADTH), the Swedish Council for Health Technology Assessment (SBU), the German Agency of Health Technology Assessment at the German Institute for Medical Documentation and Information (DAHTA@DIMDI) and similar agencies in other European countries. In the United States, Drug Effectiveness Review Project (DERP) is actively conducting comparative reviews which are being used for formulary decisions. These comparative reviews, however, should not be seen as the standard by which the HIF would assess health impact, as the HIF would review how drugs were used in actual practice in different countries, and would reassess health impact over time as new data became available. Registrants would also have strong incentives to provide data on utilization in order to bolster their case for higher rewards.

Pharmaceutical manufacturers and other health technology companies perceive to an increasing extent the importance of demonstrating that their products are therapeutically effective and therefore worth their high cost. This is leading them to engage more actively in assessing therapeutic effectiveness from an early stage and to incorporate this information in their pricing decisions.

Health impact is already being factored into decisions as to whether drugs should be listed in formularies and made eligible for reimbursement under insurance, and estimates of therapeutic effectiveness are being used to help determine the price points at

28 THE HEALTH IMPACT FUND

which new products will be sold. It is a natural step from there to make the payment for the product depend on actual health impact. This chapter explains how this might be done.

The plea of impossibility offers itself at every step, in justification of injustice in all its forms.
Jeremy Bentham

MEASURES OF HEALTH IMPACT

Since it is necessary to aggregate health impact into a single unit of measurement, the choice of metric is very important. A variety of factors are relevant.¹

Quality-Adjusted Life Years (QALYs)

Arguably, the simplest measure of health is “life years” with each additional year of life saved through a given intervention being given an equal weight. Life-years may not be a satisfactory measure in situations where health is substantially compromised because of a disease, condition, or the medicine itself.

To account for quality differences in health, the standard metric is the QALY or “Quality-Adjusted Life Year.”² A QALY is a standardized measure of health impact in which a year in perfect health is given a value of one and a year in poorer health is given a value between zero and one. QALYs account for the fact that a year in good health is worth more to people than a year in poor health. Thus, QALYs can simultaneously capture changes in morbidity and changes in mortality, and combine these into a single metric. In addition, they can be used to measure impacts on different aspects of health in the same scale.

An important part of the QALY metric is that there are weights for different health states. The derivation of how much a given health state should be worth is not trivial, since that fundamentally depends on individual preferences. A common solution to this problem is to use multi-attribute health status classifications whose values have already been evaluated in various populations. There are several widely used systems including the Health Utilities Index and the EQ-5D. These classification systems essentially pro-

vide a standardized way to grade a given health status between zero and one.

Because people generally prefer health gains to occur sooner rather than later, it may be desirable to discount future impacts on health when measuring the health impact of a given medicine. This also requires one to choose a discount rate.

Given the various systems for valuing the future and for ranking different health states, QALYs in different studies are often not directly comparable. Clearly, for the purpose of comparing the health impact of different medicines in different countries, a single metric would have to be chosen for use by the Health Impact Fund.

There has been considerable academic debate over the discount rate, the quality-adjustments, and even the weighting for different ages, and there is no uniquely correct measure of any of these values in the measurement of QALYs. Thus, the HIF would simply have to make some well-informed, public choices, which would form part of the basis of how payments were allocated to the registrants.

Disability-Adjusted Life Years (DALYs)

DALYs were developed by the World Health Organization for the purpose of estimating the global burden of disease. DALYs are conceptually similar to QALYs but differ in some significant ways. Most importantly, DALY weights were determined by a group of public health experts, rather than through population-level assessments (see Drummond et al. 2005, 187).

Other Approaches

There are other approaches to measuring health impacts of a given intervention, such as Healthy Year Equivalents and Saved-Young-Life Equivalents, which are discussed in Drummond et al. (2005, ch. 6). While these have arguable benefits compared to QALYs, it is important that they are comparable in their approach. They have not, however, been as extensively used as QALYs.

MEASURING HEALTH IMPACT

In this chapter, we do not prescribe any particular metric; however, the HIF will need to choose one, and for the present we assume that it is QALYs. The HIF then needs to make an estimate of the number of incremental QALYs achieved because of the use of a given medicine globally rather than the baseline technology.³ This is properly the field of pharmacoepidemiology. Developing such an estimate is obviously challenging and this section examines a number of approaches which can be used.

The problem of determining what a medicine is worth is a familiar one in health insurance. Insurers are required to determine whether a product will be covered, and may have to bargain over the price. If they are to do this, they need to assess the value of the product for health, and in general rely on less comprehensive information about the product's effectiveness than would be available to the HIF. Thus, while the problems of health assessment initially appear overwhelming, it is important to recognize that they are not unique to the HIF system, but are common in insurance markets.

The determination of what medicine works best is also an important clinical question: there has therefore been significant interest in establishing a mechanism to determine what clinical interventions are most effective in what circumstances. The HIF's needs in terms of identifying the health impact of specific drugs are therefore very much aligned with society's interests in learning about what drugs patients should be consuming. The recent report from the Committee on Reviewing Evidence to Identify Highly Effective Clinical Services (Board on Health Care Services, 2008) therefore proposes that the US government should "fund and manage systematic reviews of clinical effectiveness" to enable better health care decision-making. To a large extent, such a program would be over-lapping in its goals and function with the health impact assessment mechanism proposed for the HIF, although of course the mandate for the HIF would be limited to assessment of the drugs actually registered with the HIF compared to the relevant baseline.

The HIF's needs in terms of identifying the health impact of specific drugs are therefore very much aligned with society's interests in learning about what drugs patients should be consuming.

Crude Aggregation

We begin by considering what it means to estimate health impact. The health impact of medicine "A" can be estimated as

$$(Q_A - Q_B) \frac{n_A}{d}$$

where

Q_A is the average QALY impact of the medicine on each affected patient, as estimated in clinical trials; Q_B is the average QALY impact of the baseline treatment on each affected patient, as estimated in clinical trials; n_A is the number of units of the medicine A sold or distributed; and d is the average number of units per patient.

The aggregation suggested above is extremely crude in various respects and unlikely to provide an accurate estimate of the true health impact of a medicine, as discussed below.

Clinical Trials Data Do Not Describe Effectiveness in the Population

It is well known that efficacy in a clinical trial does not typically reflect actual epidemiological impact (see Revicki and Frank 1999; Oster et al. 1995). There are a number of reasons.

First, trial participants systematically vary from the population. They tend not to have complicating co-morbidities, and they are only included if they exactly meet the characteristics identified in the trial protocol. In addition, physicians may prescribe the product for patients for whom the clinical indications are not very clear, or where the diagnosis is not

30 THE HEALTH IMPACT FUND

complete. In many developing countries, accurate diagnoses are difficult to obtain owing to a shortage of qualified physicians, and patients may self-medicate, since a prescription from a physician is neither available nor necessary to purchase the medicine.

Second, in a clinical trial, participants are typically motivated or even required to follow the strict trial protocols including taking the medicine at the approved times and frequency. In the population, patients are often non-compliant and fail to follow the prescription accurately. Frequently, patients will skip doses or stop taking the medicine if they feel better or worse.

Third, physicians in a clinical trial tend to be more attentive to their patients and patients are typically monitored weekly.

These differences will generally lead to differences between the estimates of effectiveness from clinical trials and in the general population. Evidently, the problem is confounded if $(Q_A - Q_B)$ is not estimated directly but through multiple different tests, where drug A is compared to placebo in one trial and the baseline therapy is compared to placebo in a separate trial.

Clinical Trials Data on Averages May Not Reflect the Value of Diversity

For many diseases and conditions, it is difficult for new medicines to show in clinical trials that they are unambiguously better than previous treatments. However, for given individuals, it sometimes appears that one drug may be more effective than another, perhaps because of unobserved differences between patients with similar symptoms. In such situations, clinical trials may fail to demonstrate the true value of having more than one treatment for a condition. That is, in terms of the estimating process above, $(Q_A - Q_B)$ may be relatively small or even zero as measured in a clinical trial, and yet product A may be more effective for a given individual than the baseline therapy.

Incentives for Quantity

If the number of units actually taken per patient is not known, then the reward, given the measurement

system above, would be based on the number of units distributed, rather than the number consumed appropriately. This would obviously give firms incentives to exaggerate the number of patients actually treated successfully. At the extreme, the manufacturer might collude with a wholesaler to fraudulently claim higher sales volumes than actually occurred – or, more familiarly, firms might use various incentives to aggressively promote their product to physicians, who would then overprescribe the product to patients. In either case, the innovator could obtain a reward for a health impact not realized.

Not-so-crude Aggregation

The discussion above suggests that the HIF should not use naïve aggregation of unit sales times estimated superiority as demonstrated in clinical trials, since this is likely to lead to biased and inconsistent estimates of health impact. However, that does not mean that the approach generally is unworkable.

First, if the HIF is to use data from clinical trials to help establish the degree of superiority of a given medicine over the baseline, it should augment that data with supplementary evidence from observational studies and pragmatic or practical trials which use data from normal clinical practice. It is clear that in many cases such supplementary evidence cannot be available when the product is first commercialized, and that at that time the only data must be from clinical trials. Therefore additional data on ease of compliance, characteristics of possible patients and their similarity to patients in the clinical trial, and evidence on selective superiority of the relevant product, should be provided as early as possible. In due course, epidemiological evidence on the effectiveness of the product in the population should be provided. It could be that payments by the HIF in the first few years could be made partly conditional on observed effectiveness. Since registrants would be paid on the basis of demonstrated health impact, they would have an incentive to try to design data collection systems related to their products which would create information about use and effect.

Second, the HIF should be aware of the incentives for registrants to expand sales volumes to inflate the

estimated impact of the product. To minimize this problem, the HIF should require extensive reporting of sales volumes to it directly from wholesalers, with evidence from wholesalers on which retailers purchased the medicines. This would enable the HIF to conduct audits on how the units were dispensed (as discussed below). Essentially, this is similar to the need for insurance companies to make sure that claimed sales actually took place before payment is made.

Third, the HIF could conduct or require, where feasible, population-level studies to determine the impact of certain products. Such population-level studies are in general likely to be rather expensive, and only relevant for products which are very widely consumed, but in those cases may be particularly important. Mortality data indicating cause of death and other data from hospitals and clinics indicating incidence and prevalence in the population could also be used to assist in identifying the impact of a given therapy.

Fourth, the HIF could use information from the Global Burden of Disease (GBD) project, to help ensure that its estimates across countries were consistent with the measured burden of diseases and conditions. The GBD project, managed by the Institute for Health Metrics and Evaluation, is a major effort to perform a complete systematic assessment of the data on all diseases and injuries, and to produce comprehensive and comparable estimates of the burden of diseases, injuries, and risk factors, around the world.

Finally, it is important to remember that the HIF is intended to be an option, so that in cases where a firm has a product which it believes is effective, but for which the clinical trials and other epidemiological evidence does not show a substantial effect, the firm can exploit its usual rights under the patent system. The HIF is designed to reward products which have high demonstrated health impact.

THE COST OF HEALTH IMPACT ASSESSMENT

Health impact assessment would be expensive, given the need to assess a variety of medicines globally. There would, of course, be some economies of scale from assessing many medicines at the same time, and efficiencies from assessing the same medicine year

after year. However, a reasonable perspective is that if the HIF had an annual budget of \$6 billion, it could spend about \$600 million on administration and assessment, with the bulk being devoted to assessment. This would make it by far the largest health assessment agency in the world. For comparison's sake, NICE (the UK's National Institute for Clinical Excellence) has a budget of approximately \$50 million. NICE publishes around 25 technology appraisals, 12 clinical guidelines and 60 pieces of interventional procedures guidance each year (NICE 2004). The HIF would have, assuming a stock of about 20 medicines registered at any time, a requirement to evaluate the impact of those medicines around the world, which would be a much more difficult process than that undertaken by NICE. However, there could be considerable external benefits from such an assessment process, including primarily that it would enable better prescribing as the relative therapeutic benefits of different products were better understood.

The HIF would be by far the largest health assessment agency in the world.

A budget of \$600 million, spent on roughly 20 medicines at any given time, yields an average budget per year per drug of \$30 million. How would this be spent? Part would be allocated to evaluating clinical evidence. Current estimates of the cost of trials can be found in Holve and Pittman (2008), who estimate that *head-to-head studies* range in price from approximately \$2.5 million for relatively small studies to \$20 million for large studies. Such studies, of course, would not be conducted every year; some such studies could be performed by the registrants, though the HIF could also commission its own independent studies where needed. *Observational studies* range in cost from \$1.5 million to \$4 million. The HIF would require observational studies in different settings, though not every year, so this could be quite costly. However, it is likely that observational studies would be less expensive in developing countries. *Systematic reviews* of evidence tend to cost up to around \$0.3 million. The HIF would also require a substantial *auditing* function to ensure that the products were be-

32 THE HEALTH IMPACT FUND

ing distributed and used in ways consistent with the findings of the observational studies. Finally, there would be a significant *overhead* component related to obtaining the functions of the technical branch and other operational branches, which could be shared across products.

Errors using inadequate data are much less than those using no data at all.

Charles Babbage

FORESEEABLE DIFFICULTIES

Location-dependent QALYs

QALYs are essentially meant to be based on the preferences of individuals. It is likely that health preferences and circumstances differ systematically across countries, so that, for example, being confined to a wheelchair may have very different impacts in the Netherlands and in Nepal. However, unless such preferences are accounted for in the QALY system used, the QALY will fail to give proper weights to health states in different countries.

Inadequate Data on Drug Use

An important obstacle to estimating the health impact of different medicines is the availability of good data. This is, of course, an obstacle in general to the practice of evidence-based medicine. For example, it is estimated that of the more than two trillion dollars spend on health care in the United States annually, less than one-tenth of one percent is devoted to learning what works best (Institute of Medicine, 2008). There is probably a good case to made for a general increase in expenditures on learning what is effective and when. This, of course, applies particularly to the HIF, which would require better data than is commonly available to make consistent estimates of health impact.

Especially in the poorest countries, it is likely to be very difficult to obtain good-quality data on the distribution and use of drugs, in part because of less well developed information and communications

systems, and in part because in those countries, drug distribution systems tend to be multi-tiered and opaque. In addition, since in the poorest countries physician shortages are endemic, correct diagnosis is less common and many patients purchase drugs directly from local pharmacies or retailers without prescriptions. Adherence to prescription protocols may be spotty. Thus, it is likely that it will be relatively difficult to obtain comprehensive data on health impacts of drugs in such settings.

Here the incentives created by the HIF for firms to monitor data and to promote effective use of their registered medicines, as discussed in chapter 7, not only would help the HIF to assess health impact, but could also be of great value in other health promotion efforts.

The HIF would have to seek out a wide variety of data sources to make the best estimates possible, including confidential information as available. It should also try to obtain input from different sources, including patients, doctors, pharmacists, etc., to enable a comprehensive picture of the use of the registered product.

The problem of inadequate data can lead to a variety of types of errors. Some errors would be random, and would be unlikely to significantly affect the expected payments for a given product. Other errors could arise systematically, with bias between diseases and countries, because of a variety of factors, such as the differing propensity of patients to report health outcomes depending on the illness. Such systematic errors would be more problematic, and would influence firms' willingness to innovate or to register their products with the HIF. A third type of error is more serious: if registrants could systematically misrepresent the health impact of their medicines. The HIF would have to undertake careful auditing of reported data by registrants to minimize the extent to which such misrepresentation influenced the allocation of payments.

Differing Interpretations of Incomplete Data

Given that the HIF will make assessments of health impact which will depend on data from a large number of countries, it is certain that data will be incom-

plete in a variety of dimensions, including the estimated therapeutic benefits of a product compared to the baseline per patient, the effectiveness of the drug in the population, the number of units distributed, and extent to which distribution reached persons with relevant indications, and the quality of diagnosis and compliance. All of these will be to varying degrees incomplete in different countries, and this will require sophisticated inference. Based on the assumptions used and the techniques for inference, estimates may differ substantially. Since a ten percent increase in the estimated health impact translates into a roughly ten percent increase in payments from the HIF, firms will have an incentive to make strong claims about the effectiveness of their products. This could lead to disagreements over what share of the HIF disbursement each firm should receive. Thus, the HIF will need to establish a transparent and unbiased methodology developed in conjunction with pharmaceutical firms and governments, *before* it begins actual assessment of health impact. (Again, here it is important to stress that though no single methodology can be ideal in every circumstance, the HIF will have to be clear and transparent about its processes so that innovators can know what to expect if they register their products with the HIF.)

An important consideration is that the HIF has to pay out a fixed sum in a given year, so that the disagreement is fundamentally between the health impact assessments of different companies, with the HIF acting as an arbitrator. Therefore, it will be in the interest of pharmaceutical firms to have a clear and fair methodology established at the beginning.

Comparative Clinical Data Failure to Demonstrate Differences

In order to make appropriate judgments about the effectiveness of one drug compared to another in the population, evidence from clinical trials can be relied on to set some baseline. However, even at the clinical trial level, the data on the superiority of one medicine over another are often unclear. For example, the Comparative Effectiveness Reviews published by the Agency for Healthcare Research and Quality (AHRQ) show that, in a variety of classes of medi-

cines, clinical trial data does not provide a basis to make claims of substantial superiority of one treatment compared to another.

Even with large numbers of trials, it is often impossible to detect significant clinical differences between competing drugs, even when these have different mechanisms of action. This suggests that registrants of new drugs that are similar to existing treatments may find it difficult to claim health impact rewards based on therapeutic superiority. This suggests that most HIF-registrations will be for genuinely novel products that bring substantial incremental benefit to patients. (The HIF would not be an attractive mechanism for products that do not provide significant advantages over pre-existing therapies.) For the HIF to be attractive for novel products with significant health impacts, it will need to be financed adequately. This helps to establish a minimal size for the HIF at several billion dollars per year, since below this level it would not be sufficient to support a portfolio of more than a few important medicines.

Surrogate End-points

A common method for measuring efficacy in drugs is to examine their effect on so-called “surrogate” end-points. The National Institutes of Health define a surrogate endpoint as “a biomarker intended to substitute for a clinical endpoint” (Cohn 2004). For example, the effect of a drug on cholesterol levels has been used to measure efficacy, although the real interest is in the effect of the drug on mortality and morbidity. Surrogate endpoints are used because it is less expensive and much quicker to measure biomarkers, rather than mortality. In cases where there is a strong case that the biomarker is highly correlated with health, its use for the purpose of drug approval may be justified on the basis that patients would otherwise be denied access to a useful drug. However, for the purposes of the HIF, the use of surrogate endpoints clearly raises significant problems since it would be difficult for the HIF to confidently estimate health impact on the basis of such biomarkers.

The question, whatever we spend [on health care], is whether we are getting our money's worth. In general, good information and appropriate incentives are necessary to allocate resources efficiently.

Ben S. Bernanke

“Excessive” Sales

As mentioned above, firms will have an incentive to exaggerate the number of patients helped and the average health impact on each patient, in order to increase their share of payments from the HIF. The exaggeration of the number of patients may occur in a number of different ways.

First, firms may simply report more sales than actually occurred, possibly in collusion with wholesalers. This would of course be fraudulent and presumably a firm would in these circumstances forfeit any future payments from the HIF on this product.

Second, firms might bribe wholesalers to buy more drugs than they would really want. The proposed mechanism described in chapter 2 suggests that there would be a standard price. However, if a manufacturer offered a bribe of \$2 million to a wholesaler to buy one million pills at the standard price of \$1 each, and then to distribute them at low or possibly negative prices to pharmacies, neither manufacturer nor wholesaler has an incentive to report this activity, which might be hidden through unacknowledged discounts in the price of other drugs. In this case, it becomes harder to identify such collusive activities, without confirming through pharmacy records that the products were sold.

Most insurance companies solve this problem by insuring the consumer directly, so that the manufacturer would need to collude with individual consumers to exaggerate sales, which is generally difficult. However, manufacturers interact with doctors to encourage them to write prescriptions for their products. When these interactions involve payments, subsidies, gifts, etc. to physicians, it may be seen as a form of collusion. In the case of the HIF, it will be necessary to engage in auditing of sales to ensure that pharmacies did actually dispense drugs which were

shipped to them, which in turn makes it essential to obtain records from manufacturers and distributors concerning shipments of HIF-registered products.

Interacting and substitute treatments

When treatments are not independent of each other—because they are either complements or substitutes—the assessment of health impact is complicated. The HIF could use, in such circumstances, a version of the approach employed by Evans et al (2005).⁴ This approach essentially takes account of the interactions between treatments to infer separate effects for each, in a way consistent with the discussion of synergistic effects in chapter 2.

SUMMARY

It is difficult to conduct uniform and reliable health impact assessments, especially on a global scale and over the full range of medicines. But, with substantial investment into assessment techniques and measurement, these difficulties can be solved to enable health impact assessments that would be sufficiently accurate to create effective new innovation incentives that improve significantly upon those provided by the present system. What is required for the HIF to generate fair, effective incentives is that health impact can be measured in a way that is consistent and predictable across products and countries. Measurement inaccuracies will certainly arise, but provided these are random and not too large, their effect on incentives and on payments to registrants will be small. Ideally, the measurement of health impact should be perfectly accurate, since this would provide the best possible incentives for pharmaceutical innovation. In practice, assessments need only be good enough: to make it profitable for innovators to aim to improve health, to make it unprofitable for them to try to game the system excessively, and to ensure that each registered drug's overall reward – derived from its worldwide impact over the entire reward period – is reasonable given its actual health impact.

NOTES

1. For a discussion of measuring health impact in the context of the global burden of disease, see Murray et al. (2002).
2. The following discussion draws heavily on chapter 6 of Drummond et al. (2005).
3. The baseline is the set of pharmaceuticals available two years before the medicine was introduced; see chapter 2.
4. See particularly their Methods Appendix, Boxes C and E.

