SOURCES OF OUTCOMES FOR HEALTH RESEARCH

How useful are clinical trial results?

Rebecca McKibbin
Yale University

1. Introduction

Many economic problems are examined by defining a utility function over some measurable inputs – typically prices and quantities. When studying questions related to health, such as how to value medical research, a key input into the utility function is health. Unlike price and quantity inputs, health is difficult to measure. Empirical research on the value of medical research and on health care related questions more broadly is limited by the availability of data on health.

Data is limited for three reasons: (1) defining a measurement of health that can be mapped into utility with reasonable confidence; (2) the long time horizon over which health changes are realized; and (3) the availability of health measures in datasets that contain other economic information. Typically, health is divided into two components: quantity and quality of life. The former is measurable but often not realized until far into the future. This means that data needs to be collected over a long period of time and a long delay in conducting the analysis of an intervention. The latter is a similar concept to utility or well-being, which is subjective and difficult to capture quantitatively.

In this white paper I examine the methods currently available to measure health and where data can be obtained. Researchers are faced with the choice between limiting research to areas where mortality is an interesting and measurable outcome or relying on relationships between interventions and outcomes found in medical research (clinical trials). While clinical trials can be useful, I also review the evidence on the reliability of clinical trial results, an important research area in itself, and discuss areas where more research is needed.

2. Current methods for measuring health outcomes

Measuring health is much like trying to measure utility. When consumers purchase health care, what gives them utility is the effect it has on quality of life and length of life. The most common methods of measuring health are either to survey people on their utility (self-reported) or to consider only the effect on mortality. This differs from other areas of economics where it would be considered unusual to try to directly measure utility via survey or to only consider the number of years that utility will be received. Typically, using functional form assumptions on how prices and quantities affect utility, micro economists either derive sufficient statistics to make inference about welfare or use measurements of the inputs plus the functional form assumptions to compute whether utility is higher in one scenario or under another. An important barrier to using such a utility framework with prices and quantities as inputs in health is that health care markets are characterized by many failures. In particular, insurance shields consumers from the true marginal cost of medical care and physicians act as imperfect agents for consumers. This means that observed choices of consumers do not necessarily reveal their preferences in a simple utility framework. In the remainder of this section I describe the most common methods of measuring health. A potentially valuable area of future research would be to determine a way to set up a utility function over measurable inputs. This would mean that in many applications there could be conclusions to be drawn without needing to directly measure health outcomes.

2.1 QALYS

The most widely used measure of health in health evaluations is the quality-adjusted life year (QALY). The QALY combines mortality and quality of life into a single measure of value computed as:

$$\text{QALY}=\text{TIME x Utility}$$

Time is measured as the number of years that a person lives or has the condition and Utility is measured on a scale of 0 to 1 where 0 is death and 1 is perfect health.[i] This is a cardinal measure of utility and so there is meaning attached to the relative size of changes in utility.

The utility index used in the construction of the QALY is computed using response to surveys. An example of such a survey tool is the EQ-5D (there are also many others and developing more sophisticated versions of the QALY is an active area of research). The EQ-5D splits health into five dimensions: mobility; self-care; usual activities; pain/discomfort; and anxiety/depression. Respondents are asked to self-rate their severity level for each component either on a 3 or 5 point scale. The scores are adjoined into a 5 digit number, which is the health state. This splits people up into different health states, however, these states need to be ranked against each other so that utility is an index between 0 and 1. There are several different approaches to this. One method is to ask respondents to rate their health out of 100, this is known as the visual analogue scale method. The time trade-off method asks respondents to indicate the number of remaining life years in full health at which the respondent is indifferent between the longer period of impaired health and the shorter period of full health. Finally, there is the standard gamble method, where respondents complete a discrete choice task where they choose between sets of health-time states.[ii]

A number of issues have been raised with the use of QALYs as a quantitative measure of health. Firstly, it is constructed using a ranking of different health states. However, the conversion into an index means that the size of changes in the index do not have a meaningful interpretation. Secondly, because it is constructed using aggregated health states, it is not very useful for looking at changes in quality of life within diseases. For example, many cancer patients will fall into the same health state given the high level characteristics used to define a health state. However, there are many ways in which people can be made more comfortable and have higher quality of life even though they are sick. This type of change cannot be captured by the QALY. Thirdly, there is the problem of who the respondents are in the construction of the QALY weights. Often, they come from small and unrepresentative populations. Moreover, respondents are asked to evaluate health states that they have never experienced, which means that ex post preferences might different from ex ante. That is, a person who has never experienced the loss of a limb might think they would prefer to live three years less than live without the limb. However, once they lose the limb they may realize that they can adapt to life without a limb and would not give up three years of life to get the limb back.[iii]

2.2 Self-reported health

Another approach to measuring health is by asking people to rate their health on a scale. Self-reported health is similar in concept to the QALY in that it is attempting to directly quantify utility. It suffers the same shortcomings as the QALY. Since it is measured by asking respondents to rate their health on a discrete scale, the size of changes in the scale does not mean anything, and it is difficult to interpret differences across people. Moreover, the crudeness of the measure means that small changes in quality of life will not be detected. The major benefit of self-reported health is that it can be collected using a single survey question and hence is lower cost to collect. A number of economic datasets with micro data include data on self-reported health including the CPS March Supplement, the SIPP and NLSY.

2.3 Mortality

A popular approach to measuring health used by economists is length of life or mortality. The idea is to study problems in the context of diseases where there is expected to be an economically relevant effect on short-term mortality and no effect on quality of life. If quality of life is unaffected by the intervention, then welfare differences will only depend on quantity differences. Examples of diseases that are studied are reducing suicides from depression or preventing heart attacks, hospital readmissions or surgical complication resulting in deaths. However, it is not valid to ignore quality of life effects if they are likely to be present just on the basis that mortality is easy to measure. In most cases, the quality of life is likely to be the more substantial component of the effect.

3. What information can clinical trials provide?

Many economic datasets do not contain information on health and rarely include metrics such as QALYs. In contrast, clinical trials are conducted to demonstrate the relationship between a medical intervention and health. By collecting data on a specific intervention, they are able to collect a wide range of outcomes, often including QALYs. Although there is variation in the design of trials, randomized control trials are a valuable source of information on the relationship between medical interventions and health outcomes.

A potential way to incorporate clinical trials into economic analysis is through two-sample IV.[iv] Two-sample IV allows consistent instrumental variables estimation when only the outcome and the instrument (Y and Z but not X) are observed in one dataset and only the endogenous variable and the instrument (X and Z but not Y) are observed in another dataset. There are two conditions required to use this approach: the intervention or instrument needs to appear in both data sets and the sample needs to be drawn over the same population. There are many applications where the first condition could be met. For example, when the instrument is the use of a particular medical treatment or when it is a change in measurable health (such as a change in bone fractures) that we want to map into quality of life. The second condition is more problematic and is complicated by the fact that many clinical studies are conducted on narrow populations that are not reflective of the group that will eventually use the drug. This is discussed further in section 4. One important caveat to using clinical trials for two-sample IV is that, at the moment, researchers rarely have access to the underlying data (which is slowly changing). This means that it would not be possible to examine whether covariates are balanced across samples, change the sample used, modify the control variables or other aspects of the specification.

4. Concerns about quality of clinical trial results

While the section above outlines the potential benefits of using clinical trials to measure health outcomes, in recent years, questions have been raised about the reliability of findings in biomedical research.[v] In this section I discuss the most common issues and the research surrounding their prevalence.

4.1 Publication bias

Publication bias refers to the situation where published findings in a literature are not representative of the research being conducted. This could occur, for example, because certain types of findings such as non-null results or certain hypothesis are favored by the publishers. Publication bias is a known issue across all kinds of scientific research: social and natural. There are two main categories for how this could occur, which are discussed in this section.

4.1.1 Selection

Selection bias occurs when studies that show a null result have a lower probability of being published than studies that find support for a hypothesis. The problem with publication bias is that the same clinical trial may have been conducted multiple times and only the one time a favorable result was found. A p-value of 0.05 is typically used as the threshold for statistical significance. This means that one in twenty results will be a statistical fluke. If you run the study multiple times and only publish the statistically significant result, you are likely publishing a result that is not real.

Researchers have used two methods to identify selection bias. The first is to simply count the number of studies that are known to have been begun and then see how many have been published.[vi,vii] A number of papers have investigated this in a variety of scientific disciplines. For example, a study of new drugs approved between 1998 and 2000 found that after 5 years more than half of trial results were unpublished.[.viii] Although since 2007 it has been mandatory to register all clinical trials conducted in the US this does not appear to have improved the reporting of null results.[ix,x] The second method for identifying selection bias is to look for a discontinuity in the distribution of p-values of published results around 0.05. I did not find evidence of this method having being used to analyze biomedical research as a field. However, many analyses of research quality are conducted inside particular disease grouping or within particular journals.

4.1.2 Inflation bias (p-hacking)

Inflation bias refers to bias in published results caused by researchers testing many specifications or outcomes and then only reporting those with significant results.[xi] For example, by dropping observations, changing the definition of the treatment of control group, trying difference covariates, ceasing the trial or the analysis as soon as a significant p-value is found. This can be detected by examining the distribution of p-values in published research. If there is bunching just under 0.05 then there is inflation bias. There is evidence of widespread inflation bias across all scientific disciplines.[xii] Depending on the probability that the underlying hypothesis being tested in true, there could potentially be a huge proportion of studies that are false positives.[xiii] For example, if 100 studies are run and only one is based on a true hypothesis, you would expect to find six positive results, one real and five false. If just the positives are published then 80% of published results would be untrue.

4.2 Other issues

There are several other issues that have been raised with clinical trials where I could not find general empirical studies estimating the extent of the problem. This is in part because much of this type of research is conducted at the individual indication level or at the journal level.

- Multiple hypothesis testing: Multiple hypothesis testing refers to the testing of many hypotheses simultaneously. The chances of finding at least one statistically significant results rises the more hypothesis that are tested.[xiv]
- Choosing the most favorable comparator rather than most relevant: There are numerous ways in which trial protocols can be setup to distort findings. A commonly discussed issue is the use of a placebo as a comparator rather than running a head to head comparison with best practice (ref).
- Exclusion of an outcome expected to be important: Another issue in trial design is the deliberate exclusion of an adverse event as an outcome so that no data will be collected on it and hence no

effect will be found. This is difficult to detect unless a serious adverse event is discovered ex-post and an investigation determines that there is evidence that researchers suspected a problem and did not act upon this knowledge.

- External Validity: many clinical trials are conducted on groups that are not representative of the final group studied or are studied on conditions that aren't representative of the conditions in real clinical practice.[xv]

5. Data sets for studying or obtaining clinical trial results

Data from clinical trials is available from a variety of sources; public and commercial. Publically available data is easy (and free) to obtain but is more difficult to work with than commercial datasets. In this section I discuss the publically available data and list commercial datasets. Information on the cost and content of commercial datasets is difficult to obtain and so I only provide the companies own description of the from data their websites.

5.1 Publically available data

Clinicaltrials.gov

Clinicaltrials.gov is the US clinical trial registry website. Trials covered by FDAAA 801 must be reported here. It was created as a result of the Food and Drug Administration Modernization Act of 1997 and was made available to the public in 2000. Since 2008, many types of trials have been required to register there. The registry can be searched manually or the entire registry can be downloaded here. Although the registration of trials covered by law should be comprehensive, there are still many issues with registration compliance with reporting of results is poor. [xvi, xvii]

PubMed

PubMed is a database of citations and abstracts from biomedicine and health research. It is possible to link clinicaltrials.gov and PubMed from 2005 onwards as all clinical trial journal articles are supposed to provide the clinical trial ID (this is not enforced though). A description of the elements available in the PubMed database can be found here . The use of this for analysis of clinical results is hindered by the fact that many of the results are contained inside the abstract category and this contains a lot of text, which would need to be parsed. The size of the PubMed database means that it cannot be downloaded in its entirety. Users must query it; information on queries is available here.

Drugs@FDA

The FDA provides a searchable database that contains the information inserts of all drug packages (the "labels"). These information inserts contain the results from the clinical trials that the manufacturer presented to the FDA in order gain marketing approval. Although only summary information is included, this can be useful because these are often the pivotal trials.

World Health Organization (WHO) International Clinical Trials Registry Platform Search Portal

The WHO maintains a list of the primary registries for clinical trials in a number of countries. It can be found at http://www.who.int/ictrp/network/primary/en/.

5.2 Commercial Datasets

Several companies have compiled databases on clinical trials and pipeline research. I include for reference the blurbs provided on the company websites.

Pharmaprojects

“Pharmaprojects provides access to over 60,000 highly detailed and fully searchable drug profiles, updated continuously with granular information on development history timelines, licensing information, molecular structure and more. The Pharmaprojects solution provides a simple web-based user interface, with advanced search capabilities for power users, updated in real-time and fully integrated with Citeline’s other products.”

Website: https://citeline.com/products/pharmaprojects/

IMSworld R&D Focus

“Monitor the progress of drugs through the R&D pipeline worldwide. This unique resource includes details of more than 9,700 drugs in active development from more than 3,000 companies”

Website: http://www.ovid.com/site/catalog/databases/1244.jsp

Adis R&D Insight

“A database for drug research and development, disease treatment and decision making, based on trusted, scientifically sound data”

Website: http://www.springer.com/gp/adis/products-services/new-adisinsight

Thomson Cortellis

“Pipeline Compare combines pipeline data from multiple commercial databases into one system where customers can search across the consolidated data. The key feature of Pipeline Compare is the sophisticated mapping process that Thomson Reuters have developed to enable customers to compare information from different sources with different terminologies & vocabularies”

Website: http://thomsonreuters.com/en/products-services/pharma-life-sciences/drug-development/pipeline-compare.html

Ceterwatch Drugs in Clinical Trials Database

“The Drugs in Clinical Trials Database contains more than 4,000 new investigational treatments currently or previously in Phase I through Phase IV trials worldwide. Updated weekly, drug profiles include indications for use, current trial initiations and results, study phase status and manufacturer contact information.”

Website: http://www.centerwatch.com/drug-information/pipeline/app/login.aspx

6. References

[i] Encyclopedia of Public Health: Volume 1: A - H Volume 2: I - Z edited by Wilhelm Kirch.

[ii] Weinstein M, Torrance G and A McGuire (2009) "QALYs: The basics", *Value in Health,* 12(1): S5-S9. https://www.ispor.org/meetings/Invitational/QALY/Paper2revised.PDF, Last accessed 8/30/2016.

[iii] Pettitt DA, Raza S, Naughton B and et al (2016) "The limitations of QALY: A literature review"*, Journal of Stem Cell Research & Therapy*, 6(4). http://www.omicsonline.org/open-access/the-limitations-of-qaly-a-literature-review-2157-7633-1000334.pdf, Last accessed 8/30/2016.

[iv] Angrist J and and J Pischke (2009) *Mostly harmless econometrics: an empiricist's companion,* Princeton: Princeton University Press.

[v] Ioannidis J (2005) "Why most published research findings are false", *PLoS Med,* 2(8): e124a, http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1182327/ , Last accessed 8/30/2016.

[vi] McGauran N, Wieseler B, Kreis J and et al (2010) "Reporting bias in medical research – a narrative review" *Trials,*11:37. https://trialsjournal.biomedcentral.com/articles/10.1186/1745-6215-11-37 , Last accessed 8/30/2016.

[vii] Ross JS, Mulvey GK, Hines EM, Nissen SE and HM Krumholz (2009) "Trial Publication after Registration in ClinicalTrials.Gov: A Cross-Sectional Analysis", *PLoS Med,* 6(9): e1000144. http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1000144 Last accessed 8/30/2016.

[viii] Lee K, Bacchetti P and I Sim (2008) "Publication of clinical trials supporting successful new drug applications: a literature analysis" PLoS Med, 5: e191-10.1371.

[ix]Chen R, Desai N R, Ross JS, Zhang W, Chau KH, Wayda B and et al. (2016) "Publication and reporting of clinical trial results: cross sectional analysis across academic medical centers", *BM,* 352 :i637. http://www.bmj.com/content/352/bmj.i637, Last accessed 8/30/2016.

[x] Riveros C, Dechartres A, Perrodeau E, Haneef R, Boutron I and et al. (2013) "Timing and Completeness of Trial Results Posted at ClinicalTrials.gov and Published in Journals", *PLoS Med* ,10(12): e1001566. http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1001566, Last accessed 8/30/2016.

[xi] Head ML, Holman L, Lanfear R, Kahn AT and MD Jennions (2015) "The Extent and Consequences of P-Hacking in Science", *PLoS Biol* 13(3):e1002106, http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002106#pbio.1002106.ref012 , Last accessed 8/30/2016.

[xii] Head ML, Holman L, Lanfear R, Kahn AT and MD Jennions (2015) "The Extent and Consequences of P-Hacking in Science", *PLoS Biol* 13(3):e1002106, http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002106#pbio.1002106.ref012 , Last accessed 8/30/2016.

[xiii] Ioannidis J (2005) "Why most published research findings are false", *PLoS Med,* 2(8): e124a, http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1182327/ , Last accessed 8/30/2016.

[xiv] Pan Q (2013) "Multiple Hypotheses Testing Procedures in Clinical Trials and Genomic Studies" *Frontiers in Public Health*, 1(63). http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3859974/ , Last accessed 8/30/2016.

[xv] Khorsan R and C Crawford (2014) "How to Assess the External Validity and Model Validity of Therapeutic Trials: A Conceptual Approach to Systematic Review Methodology", *Evidence-based Complementary and Alternative Medicine*, http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3963220/ , Last accessed 8/30/2016.

[xvi] Anderson ML, Chiswell K, Peterson ED, Tasneem A, Topping J and RM Califf (2015) "Compliance with results reporting at ClinicalTrials.gov" *N Engl J Med*, 372(11):1031–9. http://www.nejm.org/doi/full/10.1056/NEJMsa1409364#t=article, Last accessed 8/30/2016.

[xvii] Huser V and JJ Cimino (2013) "Linking ClinicalTrials.gov and PubMed to Track Results of Interventional Human Clinical Trials", *PLoS ONE*, 8(7):e68409, http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3706420/ , Last accessed 8/30/2016.