Comparative Effectiveness, Regulation, and the Evidence Base of Evidence-Based Healthcare

John Mullahy, University of Wisconsin - Madison

Available at: https://works.bepress.com/johnmullahy/4/
Comparative Effectiveness, Regulation, and the Evidence Base of Evidence-Based Healthcare

John Mullahy

University of Wisconsin-Madison,
University College Dublin,
and NBER

jmullahy@wisc.edu

PRELIMINARY, INCOMPLETE, AND ROUGH FIRST DRAFT

July 2011

Comments Welcome, Please Do Not Circulate
Abstract

There is some consensus that comparative effectiveness (CE) research holds potential to effect improvements in health outcomes and patient and societal well-being as it contributes to the evidence base of evidence-based healthcare (even in the absence of comparative cost-effectiveness research). For this to occur, the evidence base -- which itself is costly to amass -- must affect treatment decisions in ways that are likely ex ante to be health and welfare enhancing. This in turns place some onus on the producers of the data and data summaries that are the foundation of the evidence base to provide results that are informative with respect to the treatment decisions at hand. Yet the manner in which such data and data summaries are produced often depends critically on regulations that govern various aspects of such production processes. As such, there arise bottom-line linkages involving the features of regulatory language that influence the production of data and data summaries and the patient and societal well-being that may be affected by treatments that depend on the evidence base that depends on the production of data and data summaries that depends on particular regulatory language.

To address these issues, this paper explores: the kinds of data and statistical summaries of data that are commonly produced in CE analysis; the rationales for the production of particular forms of data and data summaries; the channels through which regulations and other policies are likely to influence the production of data and data summaries; and ultimately, therefore, how the evidence base and the treatment decisions and patient outcomes thereon dependent will depend on the structure of regulations.

In light of several prominent U.S. Food and Drug Administration (FDA) regulations, it is argued here that regulations that influence the nature and extent of the production of data and data summaries with narrowly focused purpose (e.g. satisfying statutory requirements to evaluate the safety and effectiveness of new drugs) may not necessarily produce the sorts of findings that contribute to the evidence base in ways that patients and their providers (as well as other parties, like healthcare payers) might find most informative. This is true even if the data and data summaries produced (say) by sponsors of new drug application have ideal statistical properties from the perspective of FDA evaluators who then are well positioned to assess safety and effectiveness as the evaluators have defined them (or have had defined for them by statute) for their purposes. Data and data summaries produced for one purpose may or may not serve well the purposes of others who may have standing in particular decisionmaking settings.
1. Comparative Effectiveness and Evidence-Based Healthcare

Comparing the effectiveness of competing healthcare treatments or interventions has become a centerpiece of current healthcare research and policy initiatives. While such comparative effectiveness (CE) analysis has a long tradition in the evaluation sciences -- albeit often called something other than "comparative effectiveness analysis" -- the 2010 Patient Protection and Affordable Care Act (PPACA) has afforded CE analysis a center stage position in current healthcare reform efforts. The notion that CE analysis can make important contributions to bolstering the evidence base of evidence-based healthcare and consequently improving health outcomes for patients is widely -- though, as will be seen below, not universally -- accepted and is certainly a pillar supporting the PPACA's emphasis on "comparative clinical effectiveness" (CCE) and "patient-centered outcomes research" (PCOR). In situations where competing treatments¹ A and B are being compared, how can one reasonably object to the notion answering the question "Is A better than B?" or the question "Is A more effective than B?" will improve patients' health and, more broadly, patients' and social well-being?

What is comparative effectiveness analysis? Various formal definitions have been offered. The definition provided by the U.S. Congressional Budget Office (U.S. CBO, 2007) is useful.

As applied in the health care sector, an analysis of comparative effectiveness is simply a rigorous evaluation of the impact of different options that are available for treating a given medical condition for a particular set of patients. Such a study may compare similar treatments, such as competing drugs, or it may analyze very different approaches, such as surgery and drug therapy. The analysis may focus only on the relative medical benefits and risks of each option, or it may also weigh both the costs and the benefits of those options. In some cases, a given treatment may prove to be more effective clinically or more cost-effective for a broad range of patients, but frequently a key issue is determining which specific types of patients would benefit most from it. Related terms include cost-benefit analysis, technology assessment, and evidence-based medicine, although the latter concepts do not ordinarily take costs into account. (U.S. CBO, 2007)

As made explicit towards the end of this selection, much of the scientific CE enterprise is in an important sense little more than old wine in new skins -- analysts across many disciplines have long and deep traditions of studying "Is A better than B?" questions. Yet recent public policy initiatives like the codification of PCOR, CCE, and related notions in the PPACA as well as the bolus investment in CE

¹ The term "treatment" is used here as a catch-all for any intervention having the objective of improving or maintaining health. Such interventions may be clinical (in which case treatment encompasses prevention, screening, and diagnosis) or may be non-clinical.
research that occurred under the 2009 American Recovery and Reinvestment Act (ARRA) have given rise to new structures (e.g. the Patient-Centered Outcomes Research Institute (PCORI)), to reorientations of healthcare research and training, and to generally invigorated efforts to pursue such questions.\(^2\) And while it is easy and not entirely inappropriate to focus much attention on interventions that fall under the purview of U.S. Food and Drug Administration (FDA) regulation (drugs, devices, biologicals, etc.) in considering CE investigations, the scope of CE analysis -- or even more specifically comparative clinical effectiveness as demarcated in the PPACA -- is broader and extends to clinical procedures, clinician providers per se, hospitals, and other domains where the comparison of effectiveness is reasonable, interesting, and consequential to assess and where particular initiatives have arisen to ostensibly inform such comparisons.\(^3\)

Evidence-based healthcare (EBHC) has long been heralded as a rational approach for providers and patients to jointly formulate treatment plans for patients' health problems. For present purposes, the key linkages between EBHC and CE analysis involve the manner in which the outputs of CE analysis contribute to the evidence base which in turn contributes to influencing -- or, as will be discussed below, not influencing -- patient-provider treatment decisions. The "evidence base" per se is not formulaically defined, but rather may encompass a variety of forms of research findings: summary reports on prominent individual studies; formal statistical meta-analyses; systematic literature reviews; etc. While of utmost practical importance, the issue of the particular composition of the evidence base is not addressed in the present paper.

Rather, this paper's main concern is that the contribution of CE analysis to the evidence base is influenced by a variety of incentives that may or may not be compatible with a goal of providing patients and providers with the evidence that would be most useful for purposes of making decisions about competing treatment options. Such incentives may arise in a variety of contexts, but of primary interest here are those that are defined by regulations and related public policies. The main idea behind what follows is that producers of data and data summaries\(^4\) -- e.g. individuals or research teams conducting clinical trials, designing electronic medical records databases, or engaging in other analogous activities --

\(^2\) Figure 1 depicts the time trends in the number of PubMed publication citations containing the expression "comparative effectiveness" in the title.

\(^3\) See, for instance, the AHRQ Effective Health Care program (http://www.effectivehealthcare.ahrq.gov) and the CMS Hospital Compare program (http://www.hospitalcompare.hhs.gov).

\(^4\) Precise definitions of these terms will be offered below. For now, it should suffice to think of "data" as samples of individual-level or raw data and "data summaries" as statistics describing such data, e.g. sample means, percentiles, proportions, etc.
are confronted with a variety of regulatory requirements or guidances that influence the kinds of data they produce and the manner in which they summarize it. Since such summaries are what typically and ultimately contribute\(^5\) to the evidence base, how such regulation-based incentives may influence these contributions become worthy of study. Before pursuing this line of inquiry, however, it is useful to embed it in broader perspectives of CE analysis and its purported merits and shortcomings.

A. Background

CE analysis -- even if not always the specific "CE" nomenclature -- was prominent in the activities of federal agencies like the Agency for Healthcare Research and Quality (AHRQ) prior to the current Administration's emphasis on healthcare reform. Indeed CE analysis has long been viewed as an important component of policy strategies that might address healthcare quality-cost dilemmas (CBO, 2007). As noted above, CE analysis emerged prominently in policy discussions in 2009 during the ARRA funding cycle and its prominence has continued in light of various provisions of the PPACA. During this time, highly publicized reports like one issued by the Institute of Medicine (IOM, 2009) that suggested prioritization of CE research activities enhanced awareness of CE-related issues in healthcare. From a policy perspective, it is also noteworthy that CE considerations spread from domains where they had traditionally been prominent (e.g. AHRQ) to areas of health and medical policy where such considerations were typically less prominent (e.g. the National Institutes of Health) (Lauer and Collins, 2010).

Perhaps in part because CE analysis is generically familiar to economists as a form of economic evaluation, it is not surprising that for the past few years economists have been examining a variety of market and welfare consequences of CE research (Basu, 2011; Basu and Philipson, 2010; Chandra et al., 2011; Jena and Philipson, 2008, 2009), with much focus understandably on CE analysis of treatments under the purview of FDA regulation. Indeed, long before current discussions of CE analysis could even have been imagined, Peltzman was examining how FDA regulations -- the 1962 Drug Amendments, specifically -- had consequences for considerations of effectiveness (Peltzman, 1973); more on this below.

\(^5\) Note that such contributions can be more or less direct (e.g. in the form of reported analyses of secondary or observational data on treatments and outcomes) or indirect (e.g. using results from randomized trials to support an FDA new drug application (NDA) in which case an FDA decision to reject the NDA will affect the relevant downstream evidence base by \textit{de facto} exclusion of certain treatments from comparison consideration).
B. Comparative Effectiveness Analysis: Promises and Concerns

A clear and probably realistic statement of the goals of data and data summary production in effectiveness and CE analysis is offered by Tunis et al., 2010, who urge the conduct of clinical studies "that would provide decision-makers with a reasonable level of confidence that the technology improves health outcomes." From this perspective, treatment decisions will be elastic with respect to such emerging evidence in directions that are beneficial to patients. How to accommodate and respect patients' values and perspectives in CE analysis and the downstream decisionmaking processes is an area of active research and policy activity (e.g. Wu, 2010; and, tangentially, U.S. FDA, CDER/CBER/CDRH, 2009). Against such involvement of patients in the decisionmaking process are juxtaposed less sanguine reports of patients' reservations about evidence-based healthcare decisionmaking (e.g. Carman et al., 2010; Gerber et al., 2010).

One of the main concerns about CE analysis articulated in the medical literature as well as in popular media (e.g. occasional Wall Street Journal editorials and op-eds) is that the evidence base that is available to inform treatment decisions of patients and providers may not be well suited to such decision situations. This concern is often articulated in the context of treatment effect heterogeneity although frequently subgroup effectiveness analysis is offered as an antidote (see Eddy et al., 2011; Epstein and Teagarden, 2010; Garber and Tunis, 2009; and Tinetti and Studenski, 2011, for various aspects of such concerns). Tinetti and Studenski, 2011, and Bassler et al., 2008a,b, take positive perspectives on the value of accommodating subgroup heterogeneity including baseline risk differentials. Bassler et al., 2008a, in discussing how evidence-based medicine (EBM) can be used to benefit individual patients in clinical practice, assert that "...EBM emphasizes that treatment effects are usually sufficiently similar across patients with the same condition to allow application to the individual after patient-specific baseline risk is considered." Yet even AHRQ (AHRQ, 2011) appears to at least implicitly concede that accommodating population heterogeneity in CE analysis will not be easy: "Every patient is different -- different circumstances, different medical history, different values."

Closer to the specific focus of this paper are observations or concerns that the data and data summaries produced in CE analysis that become part of the evidence base may have too much asked of them. The same information will often be expected to: inform regulatory decisions; assist patients and providers in arriving at health maintaining or enhancing treatment decisions; accommodate healthcare payers' interests in knowing about high-value care options; satisfy biostatisticians' desires for particularly structured datasets; and others. Frictions may naturally arise, as data that are ideal for one objective/purpose (regulation) may be at odds with data that are optimal for other purposes (statistical
properties) and yet other purposes (patient-provider treatment decisions). Regulators, payers, biostatisticians, economists, clinicians, and patients may all examine the same data on the subject- or patient-level outcomes arising from administration of treatments A and B (from, say, a randomized trial) and arrive at different conclusions about which is better or by how much one is better than the other. It is not that one group is correct and the others wrong; it is instead that different features of these outcome distributions may matter more to one group than the others (Vanness and Mullahy, 2011). See Eichler et al., 2010, Normand and McNeil, 2010, and Robinson, 2010, for discussions of various aspects of these conflicts. For our purposes, the aspect of this of greatest interest is extent to which regulators -- who are positioned to establish or enforce the incentives that give rise to data and data summaries -- consider the information demands of patients, providers, payers, analysts, and others when such incentives are set.

CE analysis affords many opportunities to improve patients' health and well being, but constraints are nontrivial. Demets and Califf, 2011, offer a forceful assessment:

> Recent emphasis on evidence-based medicine, patient-centered outcomes research, and learning and accountable health care systems underscores the fact that most clinical trials fail to provide the evidence needed to inform medical decision making. However, the serious implications of this deficit are largely absent from public discourse...

O'Muircheartaigh, 2006, takes a broader perspective and provides a valuable warning message that merely because data and data summaries have been made available it does not follow that useful information has been made available: "It's not that these things are not known. The question is: Are these things that you want to know?"

C. Regulation, the Production of Evidence, and Healthcare Decisionmaking: This Paper's Perspective

If the central consideration with respect to how CE analysis informs the evidence base is how treatment decisions and consequent patient health outcomes and well-being respond to evidence, then at least two fundamental questions arise:

(1a) How do different forms of evidence affect treatment decisions and, therefore, downstream well-being of patients and others?

(1b) How do policies, regulations, etc. influence the nature of the evidence on CE?
If it turns out for (a) that treatment decisions are inelastic with respect to existing and imaginable forms of evidence, then the costly CE and EBHC enterprises are largely wasteful from a societal perspective. Conversely, if it turns out that treatment decisions are actually or potentially responsive to evidence and that such decisions can be more or less ex ante optimal depending on the particular forms of evidence that are available, then exploration of (1b) is of utmost importance. The focus of this paper is on (1b) and on topics related to (1b) although efforts to better understand (1a) are of central importance and merit serious assessment (Avorn and Fischer, 2010) (the paper will return to this topic in its concluding remarks).

The logic that guides the paper’s analysis is the following:

(2a) By affecting treatment decisions, evidence-based healthcare has potential to improve patient outcomes and well-being;
(2b) Treatment decisions that are made depend on the nature of the available evidence base;
(2c) The evidence base underlying evidence-based healthcare arises for the most part from studies that have analyzed the effectiveness and/or CE of various treatments targeting particular disease areas;
(2d) The constituent studies are based on data that have been produced (e.g. in randomized trials, from electronic medical records, etc.) and summarized statistically;
(2e) The nature of the data that have been produced and the manner in which they are summarized statistically are influenced by incentives or dictates that are established by regulation, reimbursement, and other policies;
(2f) Therefore, regulations and policies that influence the production of data and/or its statistical summaries can ultimately be evaluated insofar as their impacts on patient outcomes and well-being are concerned.

To the extent that this logic is compelling, then this paper’s focus on (2e) may provide some useful insights with respect to the evidence base of evidence-based healthcare. It is perhaps worth noting at the outset that this paper will explore what might be viewed by readers as a set of relatively narrow

---

6 This argument ignores some potentially important dynamic considerations that may arise from changes in the evidence base: modifications of reimbursement formulae, that may then have supply-side effects that influence treatment choices; changes in R&D strategies; investments in modifications of healthcare systems; etc.
technical and methodological issues involving the production of data and data summaries and how regulations and other policies may influence them. Indeed, it may or may not turn out from a practical perspective that devils lurk in details. However, the paper proceeds as if they do.\textsuperscript{7}

Importantly, it is reasonable to suggest that methodological considerations in CE analysis and development of the evidence base for care are far from resolved. Among other features of Section 6301 of PPACA is the creation not only of the PCORI but also of the PCORI Methodology Committee that is charged "to develop and improve the science and methods of comparative clinical effectiveness research." This seems important \textit{prima facie} evidence that work remains to be done in the CE methodology area; it is hoped that this paper makes some contribution to advancing this agenda.

\textbf{D. Plan for This Paper}

This paper is fundamentally about the data and the measurement of health outcomes that arise in investigations of how interventions or treatments affect health. The focus is on the production of data and data summaries and evidence and on how public policies and regulations shape their production and their particular structures and on how such evidence that is ultimately produced affects patients' and others' well being. Data that are analyzed in CE analysis do not materialize exogenously but rather are the output of decisions made subject to a variety of incentives, regulatory or otherwise.

By design, the paper ranges from technical statistical considerations of data structures all the way around to considerations of how specific aspects of Title 21 of the Code of Federal Regulations ("Food and Drugs") establish incentives that ultimately determine or influence the nature of the data and data summaries that become eligible to be included in evidence-based healthcare's evidence base.

The roadmap for the remainder of the paper is as follows. Section 2 offers several brief, diverse examples to motivate considerations of how data, data summaries, evidence, and treatment choices are or may be intertwined in real world treatment settings. Section 3 engages in specific discussion about the nature of the production of data and of data summaries and how these contribute to healthcare's evidence base. Section 4 tackles the first-order questions of what is meant by "effectiveness" or "comparative effectiveness", particularly in contexts where treatment effects may be heterogeneous in populations. This discussion is intended to provide the substantive foundation for how to answer the question: "Is A better than B?" Section 5 draws on specific regulatory examples from the U.S. FDA and other areas to illustrate how these considerations of data and data summary production do or may

\textsuperscript{7} The empirical work that is in progress is designed to shed some light on such issues.
affect the evidence base. Section 6 assesses some implications of the analysis for future policy and regulatory decisions. Section 7 offers some concluding remarks.

**E. What the Paper is Not About**

At this juncture it is also useful to note briefly what the paper does not concern. First, the issues here do not involve considerations of subgroups or subgroup analysis (e.g. Lagakos, 2006; Tinetti and Studenski, 2011). It is presumed that after conditioning on any imaginable covariates (including, e.g., genetic information) that there will still be within-subpopulation heterogeneity in outcomes and as will be seen such heterogeneity is a fundamental aspect of many of the issues of concern here (Vanness and Mullahy, 2011; Greenfield et al., 2007). Consequently the issues raised here all apply, whether conditional-on-x or unconditionally. There is no argument against the notion that considerations of across-subgroup heterogeneity of treatment effectiveness are enormously important in practice, but such considerations only add additional layers of complexity for the task at hand in this paper.

Second, identification of treatment outcome marginal distributions is not of concern here. For whatever outcome definitions are selected by data producers (this selection is important on its merits) it is presumed that there is a sampling strategy that generates a set of observations on such defined outcomes corresponding to treatments A and B such that the observed within-sample distributions of such outcomes statistically converge in distribution to the true population (or sub-population) marginal distributions. Issues of treatment self-selection, attrition, adherence, non-representativeness/generalizability, etc., while of critical importance in CE applications -- particularly but not exclusively those involving so-called observational data -- are not at stake here.\(^8\)

Third, issues of "statistical significance" are not addressed here. Unarguably issues of statistical significance will affect the evidence base available to decisionmakers in many instances and for a variety of reasons (positive publication bias; weighting in meta-analysis; the extent to which statistical significance influences regulatory decisions; etc.), and it is certainly not uninteresting to address issues concerning type-1 vs. type-2 decision errors by patients, regulators, and others (Peltzman, 1973). Moreover there is an active and long-running debate as to the proper role of statistical significance in

---

\(^8\) Indeed, for most purposes here it is reasonable to think of the data-production setting as one in which outcomes arising from administration of treatments A and B are being compared as if the data are being produced in a well-controlled randomized trial. One may thus choose to interpret any observed differences between the outcome distributions arising under A and B as counterfactual contrasts that would be manifested under uniform factual/counterfactual applications of A and B in a population (see Manski, 2011b, for an alternative to uniform application).
regulatory and policy settings (see, e.g., Ziliak and McCloskey, 2007), with the unanimous decision in a recent U.S. Supreme Court case Matrixx vs. Siracusano (U.S. Supreme Court, 2010) providing an interesting foundation for future policy deliberations.

Finally, considerations of cost-effectiveness are not (specifically) of interest here. While it is conceptually appropriate to consider "cost" as one dimension of the outcome space in a CE analysis, issues of cost have been accorded a special -- and, at best, secondary -- place in CE and outcomes research in the current healthcare reform environment. Whether CE research can deliver its touted improvements in healthcare quality and costs without considerations of comparative cost-effectiveness remains to be seen, but many prominent observers are skeptical (Neumann and Weinstein, 2010).

It should be emphasized that since the paper addresses topics of concern to a wide variety of contributors to the CE enterprise -- biostatisticians, economists, regulators, payers, clinicians, patients, and others -- it almost certainly fails to do justice to particular issues of concern to any of these groups.

2. Data, Evidence, and Treatment Choices: Four Examples

This section provides brief discussion of four examples wherein data, data summaries, evidence, and treatment choices are or may be enmeshed in real world treatment settings and where it is not always obvious that the data readily available to decisionmakers is what decisionmakers would ideally most want to know.

A. Surgery vs. Chemotherapy for Ovarian Cancer

Vergote et al., 2010, randomized to either primary debulking surgery (PDS) or neoadjuvant chemotherapy (NACT) 632 eligible patients with stage IIIIC or IV epithelial ovarian carcinoma, fallopian-tube carcinoma, or primary peritoneal carcinoma. Figure 2 is adapted from the Vergote et al. paper. Imagine initially that the length of the post-intervention followup period was 18 months and that the corresponding survival data were as displayed in the top panel of figure 2. With only this information available to decisionmakers, regulators, payers, etc., many would favor NACT over PDS based on CE criteria. Actual followup extended considerably beyond 18 months, however, with the results as depicted in the bottom panel of figure 2. Presented with these data, many standard criteria for evaluating comparative effectiveness (e.g. difference in marginal median or marginal mean survival

---

9 Analogously, imagine that the difference between survival outcomes at 18 months was sufficiently large and significant to merit early stopping for benefit (Bassler et al., 2010) based on criteria that may have been specified in the trial's protocol.
time) would suggest a preference for PDS over NACT. Yet such a preference is not \textit{a priori} unambiguous, either from a patient preference perspective or from alternative statistical criteria to which decisionmakers might appeal to assess which treatment is "better", as will be discussed later in the paper in section 5.

\textbf{B. Progression-Free Survival, Erlotinib (Tarceva) vs. Placebo}

On December 16, 2009, the FDA Oncology Drugs Advisory Committee (ODAC) evaluated erlotinib (Tarceva) in a Supplemental NDA for first-line maintenance therapy in patients with locally advanced or metastatic non-small cell lung cancer (U.S. FDA ODAC, 2009). One component of the evidence presented to the ODAC was a report of progression-free survival from a randomized trial of erlotinib against placebo. These data are displayed in figure 3. When judged by standards of a log-rank test\textsuperscript{10} there were significant differences between treatment and placebo. However, median survival times are virtually indistinguishable. Which of these criteria -- or others -- should guide regulatory decisionmaking clearly should depend on decisionmakers' objectives.

\textbf{C. Treatments for Clinically Localized Prostate Cancer}

Exhibit 1 displays some of the information presented in the AHRQ \textit{Clinician's Guide for Treatments for Clinically Localized Prostate Cancer} (AHRQ, 2008). The \textit{Clinician's Guide} reports outcomes ten years after diagnosis (see the bar chart on the right of the exhibit). After ten years, expectant management relative to radical prostatectomy demonstrated higher rates of distant metastases (.25 vs. .15) and death from any cause (.32 vs. .27). However, an examination of the full data reported in Bill-Axelson et al., 2005, shows that the five-year differences between these outcomes are essentially indistinguishable. It is imaginable that patients, providers, payers, and others could have some interest in the five-year parameters even though these are not in the "evidence base" made available to clinicians in the AHRQ \textit{Clinician's Guide}.

\begin{footnotesize}
\textsuperscript{10}A log-rank test is a common approach for testing for differences of two survivor functions. The log-rank statistic can be computed as \( R = \sum_{t=2}^{T} \frac{n_{o_t}y_{o_t} - n_{t}y_{t}}{\sqrt{\sum_{t=1}^{T} \left( \frac{n_{o_t}y_{o_t} - n_{t}y_{t}}{n_{t} - 1} \right)^2}} \).
\end{footnotesize}
In July 1982, I learned that I was suffering from abdominal mesothelioma, a rare and serious cancer usually associated with exposure to asbestos. When I revived after surgery, I asked my first question of my doctor and chemotherapist: "What is the best technical literature about mesothelioma?" She replied, with a touch of diplomacy (the only departure she has ever made from direct frankness), that the medical literature contained nothing really worth reading.

The literature couldn't have been more brutally clear: mesothelioma is incurable, with a median mortality of only eight months after discovery. I sat stunned for about fifteen minutes, then smiled and said to myself: so that's why they didn't give me anything to read. Then my mind started to work again, thank goodness.

... The problem may be briefly stated: What does "median mortality of eight months" signify in our vernacular? I suspect that most people, without training in statistics, would read such a statement as "I will probably be dead in eight months" - the very conclusion that must be avoided, since it isn't so, and since attitude matters so much. I was not, of course, overjoyed, but I didn't read the statement in this vernacular way either.

... Variation is the hard reality, not a set of imperfect measures for a central tendency. Means and medians are the abstractions. Therefore, I looked at the mesothelioma statistics quite differently and not only because I am an optimist who tends to see the doughnut instead of the hole, but primarily because I know that variation itself is the reality. I had to place myself amidst the variation. When I learned about the eight-month median, my first intellectual reaction was: fine, half the people will live longer; now what are my chances of being in that half.

... The distribution was indeed, strongly right skewed, with a long tail (however small) that extended for several years above the eight month median. I saw no reason why I shouldn't be in that small tail, and I breathed a very long sigh of relief. My technical knowledge had helped. I had read the graph correctly. I had asked the right question and found the answers. I had obtained, in all probability, the most precious of all possible gifts in the circumstances - substantial time. I didn't have to stop and immediately follow Isaiah's injunction to Hezekiah - set thine house in order for thou shalt die, and not live. I would have time to think, to plan, and to fight.

The common link among these four examples is that there are different ways of looking at the same data. This is an obvious point. Why this may be important, however, is that such different perspectives may lead decisionmakers to arrive at different treatment choices AND that it may not always be straightforward for decisionmakers to access features of the distributions of outcomes that
would potentially be of interest to them: the prostate cancer Clinician's Guide reports only the ten-year rates; Prof. Gould had to undertake his own research to discover the structure of the mesothelioma survival distribution. Moreover, as will be explored below in section 5, there are circumstances where decision-relevant data and data summaries are not available because of the manner in which regulations effectively set incentives for particular forms of data and data summary production.

3. Evidence: The Production of Data and of Data Summaries to Inform Healthcare Decisionmaking

A. Producers of Data and Data Summaries

This paper suggests that producers of the data and data summaries that may ultimately contribute to the evidence base are responsive to incentives -- including those induced by regulations -- in the manner in which they produce such data and its summaries. Issues of measurement, some subtle, are of central concern here so some care with terminology is warranted. To fix ideas, "data production" may entail: generating observations on outcomes during the course of a new randomized trial for ovarian cancer treatments; inputting patient case information in an innovative EMR system; conducting annual surveys that contribute to new versions of publicly available data sets like MEPS, MCBS, etc.

In this taxonomy, "data summary production" may be: estimation of 24-month survival probabilities from a randomized trial; monthly reporting of quality metrics generated from EMR systems; econometric estimation hospital outcome models using MEPS; etc. Finally, the contribution to evidence base by that such data and and data summaries may provide arises from: reports of prominent single studies; meta-analyses of data summaries; systematic reviews of the literature; etc.\textsuperscript{11} Unless a data producer simply reports the entirety of a produced dataset without statistical summarization, there will ultimately be some form(s) of data summary(ies) (aggregation) that is(are) reported by the producer and that may subsequently become the basis of regulatory decisions and then perhaps contributions to the evidence base.

Cast in technical terms, the production of data amounts to generating empirical marginal

\textsuperscript{11} With focus on a specific objective (that is considered below in section 5), Zarin et al., 2011, offer a complementary taxonomy of measures that encompasses many of the issues considered here. These authors provide a hierarchy whose highest level is "Domain" (e.g. anxiety), followed by "Specific Measurement" (e.g. Hamilton Anxiety Rating Scale), followed by "Specific Metric" (e.g. change from baseline), and concluding with "Method of Aggregation" (e.g. proportion of participants with decrease above 50%). See exhibit 2 for details. The Zarin et al. "Method of Aggregation" corresponds in essence to the idea of a "data summary" as discussed above.
distribution functions \( \hat{F}_{jN}(y) \) of sample size \( N_j \) of outcomes under treatments \( j=A,B \), which, as noted above, are assumed to converge in distribution to the corresponding population marginals \( F_j(y) \).\(^{12}\) The production of data summaries then involves using the \( \hat{F}_{jN}(y) \) to compute empirical statistics

\[
\hat{\theta}_{jN} = t(\hat{F}_{jN}(y))
\]

or, more broadly, empirical statistical functionals (Lehmann, 1998)

\[
\hat{s}_{jN} = S[\hat{F}_{jN}(y)]
\]

which would typically purport to converge in large samples to corresponding population parameters.\(^{13}\)

Without risking much loss of generality, these functionals will be described as simple parameters \( \theta_j \), where \( j=A,B \) indexes the outcome distribution and \( p \) indexes generally elements of a set of functionals or parameters that may be of interest (mean, quantile, tail probability, etc.). For instance, \( \theta_A^p \) might be the population median of \( F_A(y) \) while \( \theta_B^p \) might be defined by an upper-tail probability defined by

\[
\int_{\theta_B^p}^{\infty} dF_B(y) = .05
\]

This paper is less concerned with specific values of the index \( p \) but particularly concerned with situations where different decisionmakers are focusing on different \( p \)'s.

The premise to be explored henceforth is that the manner and magnitude of data and data summary production is subject to incentives that arise from market signals as well as regulatory requirements.\(^{14}\) Specifically, the seemingly subtle but ultimately central question is which measurements, metrics, and aggregations are undertaken and, therefore, privileged relative to others that might be used as ultimate contributors to the evidence base. At a minimum, some combination of

\(^{12}\) It would certainly be possible and often quite relevant to accommodate multivariate outcomes \( y = [y_1, \ldots, y_M] \) in this discussion. For simplicity this is not pursued here.

\(^{13}\) For readers unfamiliar with the concept of a functional, a straightforward interpretation for our purposes is that a functional assigns a single value to the entirety of the range of values of a function defined over its full domain. (Technically, a functional is a real-valued function on a vector space of functions.) Here the relevant domains will be the possible levels of health outcomes, the functions will be the probability distributions of those outcomes, and the functionals (in this case called "statistical functionals") might, as such, be any decision-relevant property of such probability distributions: means, medians, quantiles, tail probabilities, etc. Consistent with standard measurement theory (e.g. Stephens, 1946), not all functionals are appropriate descriptors for all measurement structures, e.g. means or other moments of ordinal-scale outcomes. This issue arises again in one of the examples discussed below in section 5.

\(^{14}\) Philipson has written extensively on various aspects of the markets for data (e.g. Philipson 1997a, 1997b; Malani and Philipson, 2011); some insights from this work are relevant in the contexts of data and data summary production of concern here.
the the definition of "y", the sample sizes $N_j$, the comparator definition ("B" if treatment "A" is the focus treatment) and the specification(s) of the $S[.]$ are subject to choice by the data and data summary producers.

If provider and patient treatment decisions depend on the nature of evidence available at the time of decision, then such concerns are real and potentially welfare-relevant. It is thus instructive to consider why in any particular regulatory and treatment context particular quantities are favored and whether those quantities that are so privileged during the data and data summary production, regulatory, and evidence agglomeration phases are those that speak most closely to patients and providers preferences and their ultimate treatment choices. Section 5 describes several regulatory areas where these issues arise and -- at least in some instances -- can be shown to matter.

**B. "Evidence Elasticity" of Treatment Decisions**

The fundamental premise of evidence-based healthcare is that changes in evidence base will effect changes in treatment decisions, presumably in such a way that effects better outcomes and well-being for patients and others of sufficient magnitude and breadth that it justifies the costs incurred to bring about changes in the evidence base (through data production, data summary production, regulation, etc.). In the current "personalized medicine" environment, there is considerable effort devoted to understanding and implementing policies to deliver decision-relevant information to patients and providers (see, e.g., Epstein and Teagarden, 2010, and Garber and Tunis, 2009).

Patients and healthcare providers are the ultimate consumers of data and data summaries as they turn -- or may not turn -- to the evidence base to inform choices among treatment options they confront. Others -- e.g. regulators, payers, professional societies, etc. -- are intermediary consumers of data and data summaries as they engage in activities that shape the treatment options available to patients and providers via product approvals, reimbursements, guidelines, and other phenomena that may affect downstream treatment choices.

A simple framework that conveys the essential elements of this process endows patients with a set of treatment options at time $t$ $Z_t|J_{t-1}\equiv\{z_{t1},...,z_{tn}\}$ that depends on the data, data summaries, and evidence (J) of type $\alpha$ available at time $t-1$ (e.g. FDA marketing approval or rejection affects whether a particular treatment will be in $Z_t$). Patients at time $t$ select treatments from $Z_t$ according to 

$$\arg\max_{z_t\in Z_t}\left\{V(z_t|J_t)-c(z_t|J_{t-1})\right\},$$

where $V(z_t|J_t)$ represents anticipated wellbeing from treatment $z_t$. 

14
given what evidence exists at the time of decisionmaking, and \( c(z_t | p_{t-1}) \) represents the costs (in utility units) to the patient of selecting \( z_t \), which costs might in general depend on evidence at time \( t-1 \) due, e.g., to payers cost-sharing decisions made on the basis of that evidence (e.g. formulary tier status based on value or outcome evidence).

The welfare implications of different forms of data or evidence can thus be seen to arise from several sources. If instead of type-\( \alpha \) data and evidence the information available to all decisionmakers (regulators, payers, patients, etc.) at times \( t-1 \) and \( t \) was of type-\( \beta \) \( (\beta_{t-1}, \beta_t) \) then: (a) different treatment option sets \( Z_{t-1} | \beta_t \) may be available; (b) different cost structures \( c(z_t | \beta_t) \) may arise; and (c) different anticipated wellbeing from treatment \( V(z_t | \beta_t) \) may be manifested. Any or all of these changes due to different data and evidence actualities may then result in changes in both patient treatment decisions and downstream patient wellbeing. Thus when assessing the merits of regulations and other policy tools that affect the production of data and data summaries that in turn inform the evidence base, it is useful to recognize that real changes in treatment decisions and patient (and, presumably, societal) wellbeing are at stake.

It should be noted that these decisionmaking scenarios share much in common with aspects of the Value of Information (VOI) literature which, among other things, evaluates the costs and benefits of increments to decisionmakers' information sets, often modeled in terms of updates in priors about treatment-related health outcomes. See Basu and Meltzer, 2007, Garber and Meltzer, 2009, and Meltzer, 2001, for a variety of approaches to VOI research.

4. What is "Effectiveness"? What is "Comparative Effectiveness"

A. Characterizing Effectiveness

In light of the great variety of disease states and treatment options that exist, it is perhaps unsurprising that precise general definitions of "effectiveness" that are not vacuous are difficult to pin down in either regulatory language or in the clinical/scientific literature. Subtopics (a), (c), (d), and (e) in the Appendix provide some of the regulatory language used to characterize "effectiveness" in several domains under the FDA's regulatory authority. The regulatory language for biological product approvals is instructive:

Effectiveness means a reasonable expectation that, in a significant proportion of the target population, the pharmacological or other effect of the biological product, when used under adequate directions, for use and warnings against unsafe use, will serve a
clinically significant function in the diagnosis, cure, mitigation, treatment, or prevention of disease in man. (21 CFR Sec. 601.25)

Such language obviously -- and presumably intentionally -- permits considerable flexibility in interpretation and application to particular instances ("reasonable expectation"; "significant proportion"; "target population"; "clinically significant function"). In light of these broad characterizations, why then do some effectiveness or outcome measures -- data and data summaries -- come to be privileged (explicitly, implicitly; intentionally or unintentionally) in subsequent CE analyses and ultimately in the evidence base relative to others?

B. Using Stochastic Dominance Relationships in CE Evaluations

At the population level, stochastic dominance relationships (Levy, 1998) are useful standards for comparison of two outcome distributions \( F_A(y) \) and \( F_B(y) \) defined on the univariate outcome \( y \). The contrasts may arise between observably different units (individuals, subpopulations, etc.) or may entail counterfactual comparisons as when \( F_A(y) \) and \( F_B(y) \) are marginals from a population joint distribution \( F(y_A,y_B) \), where \( y_A \) and \( y_B \) are potentials outcomes, only one of which is observed as \( y \) under a binary treatment rule \( T \) as \( y=(1-T)y_A + Ty_B \).

Following Davidson and Duclos, 2000, let \( D^+_j(y)=F_j(y) \), \( j=A,B \), and \( D^+_3(y)=\int_0^y D^+_{i-1}(u)du \) for integer \( s>1 \). Then \( F_B(y) \) stochastically dominates \( F_A(y) \) at integer order \( s>0 \) ("\( F_B(y) \triangleright_s F_A(y) \") iff \( D^+_A(y) \triangleright D^+_B(y) \). The strongest relationship between \( F_A(y) \) and \( F_B(y) \) is zero-order dominance (\( \triangleright_0 \); Castagnoli, 1984) in which \( F_B(y) \triangleright_0 F_A(y) \) if \( \min(\text{support}(F_A(y))) > \max(\text{support}(F_A(y))) \). For present purposes, dominance order provides a natural means for structuring a hierarchy of relationship strengths between \( F_A(y) \) and \( F_B(y) \) that range from \( \triangleright_0 \) (\( F_B(y) \triangleright_0 F_A(y) \)) to \( \triangleright_s \) for increasing \( s \) (\( F_B(y) \triangleright_s F_A(y) \)) and analogously for situations where \( F_A(y) \) stochastically dominates \( F_B(y) \) at various

\[\text{When } F_A(y) \text{ and } F_B(y) \text{ are counterfactual marginals from the population joint df } F(y_A,y_B), \triangleright_0 \text{ corresponds to the extreme situation where all individual-level treatment effects must have the same sign. See figure 4 for a graphical depiction of } \triangleright_0 \text{ and } \triangleright_1.\]
orders. Two points are noteworthy. First, $F_A(y)$ and $F_B(y)$ needn't have any dominance relationship between them. Second, $\succ_0$ and $\succ_1$ both imply that $E(g(y_0)) > E(g(y_1))$ for all monotone increasing $g(.)$ (Levy, 1998); among other things, this result can be particularly useful in the analysis of ordinal-scale outcomes as will be seen below in one of the examples discussed in section 5.

Returning to the ovarian cancer treatment example discussed in section 2, it is instructive to reposition the question of whether PDS or NACT is the "better" or "comparatively more effective" intervention. Standard criteria like difference in median survival time might suggest a preference for PDS over NACT. Using dominance criteria, however, it can be demonstrated\(^\text{16}\) that $F_{\text{NACT}}(y) \succ F_{\text{PDS}}(y)$ for orders $s \geq 7$. There is no a priori reason why median survival time advantage should necessarily be privileged over seventh-order dominance in such decisionmaking,\(^\text{17}\) even though the former is certainly more likely to be encountered in the literature and regulatory arena than the latter.

C. Empirical Comparative Effectiveness Analysis

The empirical enterprise of CE analysis boils down to evaluating the "Is A better than B?" question via meaningful contrasts of data summaries that have been produced. The technical basis of these contrasts can be interpreted as comparisons of statistics $\theta_{i,N_i}$ or empirical statistical functionals $s_{i,N_i}$ defined above. These comparison may be as simple as differences between treatment arms of

\(^{16}\) Davidson and Duclos note that the $D_i(y)$ can be obtained mechanically as

$$D_i(y) = \frac{1}{(s-1)!} \int_0^y (x-y)^{s-1} dF_i(y)$$

and that this quantity can be estimated from the empirical marginal df $\hat{F}_{i,N_i}(y)$ as

$$\hat{D}_i = \frac{1}{N_i(s-1)!} \sum_{i=1}^{N_i} (x-y_i)^{s-1} \times 1(y_i < x),$$

where $N_i$ is the sample size.

\(^{17}\) Suppose that surviving long enough to attend the wedding of one's only child, scheduled eighteen months from today's intervention, weighs heavily in one's preferences. It is certainly plausible that the treatment offering the higher probability of 18-month survival -- NACT in this instance -- would be selected rationally.
means or proportions or as complex as may be appropriate to inform particular aspects of patient preferences that may be salient at the time of treatment decisions (recall Gould’s story from section 2).

A more subtle but potentially more welfare-relevant version of such comparisons would contrast "value functionals" (Hirano and Porter, 2009) defined on the $\hat{F}_{jN}(y)$, e.g. $\mathbb{V}[\hat{F}_{A,N}(y)] - \mathbb{V}[\hat{F}_{B,N}(y)]$ in which the links between particular features of patient or decisionmaker preferences and particular features of the distributions of outcomes are necessarily made explicit. In other cases, the relevant contrasts or comparisons of the outcome data cannot be explicitly cast as differences between data summaries but instead are more general juxtapositions of the entireties of the respective marginal outcome distributions, e.g. $C(\hat{F}_{A,N}(y), \hat{F}_{B,N}(y))$, where $C(.)$ is some contrast relationship defined on the two marginals. This general structure encompasses comparisons like stochastic dominance, the log-rank test criterion, and others.

Thus while there are reasonably defensible methods of comparison that can be deployed to answer the "Is A better than B?" question in empirical CE analysis, these all still rest on the data and data summaries that have been produced to inform the evidence base. Since in many instances there are alternatives to the measures and summaries actually used that could have been used, it remains to be demonstrated by particular measures and summaries are privileged over others. While the next section of the paper cannot provide a definitive and cross-cutting answer to this question, the examples provided therein should help to provide some basis for addressing these issues. By illustrating how a variety of data and data summary approaches issues arise as a consequence of regulatory and related policy structures, it is hoped that at a minimum the privileging issue is viewed as credible and potentially important. Perhaps beyond this, the discussion might shed light on directions for studying systematically linkages that are identified between and among regulations, data, evidence, treatment, and wellbeing.

5. Regulation and the Production of Data, Data Summaries, and Evidence: Examples and Case Studies

This section considers several regulatory settings in which the regulatory requirements have consequences for the composition of the evidence base via their impacts on the production of data and/or of data summaries. In each instance, particular case studies are described that illustrate the actual or potential importance of the regulatory language for the evidence base that ultimately emerges.
In the economics literature, the notion of linking regulatory activities to the evidence base available to decisionmakers has a long history dating back at least to Peltzman, 1973. In that work, Peltzman was concerned with the 1962 FDA Amendments in which effectiveness considerations came to be part of FDA’s regulatory mandate (beyond the safety considerations that had been in effect prior to 1962). Peltzman focused on the question of the efficiency of regulating the availability of effectiveness information as opposed to permitting consumers of medical products (drugs) to learn about effectiveness experientially and then form their own assessments.\(^{18}\) While Peltzman's analysis has a different orientation that the topics considered here, it is interesting that the basic ideas about regulation, data production, and how evidence moves treatment were anticipated in Peltzman's work nearly 40 years ago.

For present purposes, the particular regulatory domains discussed here are selected both because they are interesting and consequential on their own merits as well as being suggestive of other regulatory areas where analogous issues might be suspected to arise.\(^ {19}\) It is hoped that these examples contribute to convincing the reader that the evidence base of evidence-based healthcare is inexorably linked to regulatory rules and activities and, therefore, that assessments of how to improve the evidence base of evidence-based healthcare might be informed in part by scrutiny of the incentives set in place by such processes.

**A. Outcome Measures Suggested by FDA Guidance Documents**

The FDA issues so-called Guidance Documents on a wide variety of issues that fall under the its regulatory purview. In FDA’s language:

Guidance documents represent the Agency's current thinking on a particular subject. They do not create or confer any rights for or on any person and do not operate to bind FDA or the

\(^{18}\) The centerpiece of Peltzman’s analysis is the demand structure reproduced in figure 5 in which demand curve ADM represents the demand the consumer would have based on priors about the effectiveness of the treatment whereas demand curve GHEN represents the consumer’s valuation of the product after experience reveals that the product is less effective than had been anticipated initially. The "overuse" deadweight loss at price B (HDE) is what must be weighed against other costs and inefficiencies that may arise due to regulatory requirements for sponsor-producers to generate evidence on effectiveness and have that evidence be weighed by FDA regulators.

\(^{19}\) Moreover, while these regulatory contexts are statutorily distinct, there turn out to be some interesting practical overlaps in several instances. For instance, the outcome measures reported to the clinical trials registry ClinicalTrials.gov (example 5a) may be shaped by FDA guidances on outcome or endpoint measurement (example 5b) and such measurement guidances may in some instances concern the applicability of surrogate or biomarker measures (example 5c).
public. An alternative approach may be used if such approach satisfies the requirements of the applicable statute, regulations, or both. (http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/default.htm)

While such Guidances may not have the force of regulatory authority, in many instances they have influence on behavior that is tantamount to that of regulatory authority.

One of the many areas upon which Guidance Documents touch is in the measurement of endpoints or outcomes for a variety of disease areas or treatment domains. By suggesting particular modes of measurement and data and data summary production for (among others) sponsors of NDAs, such Guidances de facto create significant incentives to follow the stated recommendations and potentially costly disincentives to diverge from them. Why particular metrics and methods are selected remains an area of current research, but those that are specified have features and impacts that are real and consequential, as the following two examples will make clear.

Weight Management

The FDA Guidance on the development of drugs and biologicals for weight management (FDA CDER/CBER, 2007) recommends measurement and summary of primary efficacy endpoints (outcomes) in Phase-3 trials via two metrics: "The difference in mean percent loss of baseline body weight in the active-product versus placebo-treated group" (mean effect); and "The proportion of subjects who lose at least 5 percent of baseline body weight in the active-product versus placebo-treated group" (categorical effect). Figure 6 depicts these criteria using artificial sample outcome distributions for treatment and control groups. The shaded area between the two curves is the mean effect, whereas the difference between the heights of the curves at the 5% body weight loss abscissa. Subsequent effectiveness research (e.g. Smith et al., 2010; Gadde et al., 2011) has followed these recommendations in clinical studies.

Why such particular measures and summaries are specified is under study, but regardless of why the measures matter. In October 2010, Lorqess (lorcaserin hydrochloride; Arena Pharmaceuticals) was denied FDA approval in part because of failure to demonstrate adequate effectiveness advantages by the standards of the Guidance criteria. The Metabolic Drugs Advisory Committee that counsels the FDA on such decisions concluded in its September 2010 assessment:

When gauged by the standards of the Division’s 2007 draft guidance for Developing Products for Weight Management, the mean weight loss associated with the lorcaserin 10 mg QD and BID dose was about 3% greater than the mean weight loss with placebo. Therefore lorcaserin did not satisfy the guidance’s mean efficacy criterion. However, the
lorcaserin 10 mg BID dose did, by a slim margin, satisfy the categorical efficacy criterion. (FDA, Metabolic Drugs Advisory Committee, 2010)

Alternative criteria that could imaginably been utilized may or may not have arrived at the same conclusion regarding effectiveness compared with placebo. For Arena Pharmaceuticals and for overweight patients in search of effective treatments, however, the Guidance criteria mattered.

*Pain Assessment and Irritable Bowel Syndrome*

Among other dimensions of the disorder, FDA’s Guidance on clinical evaluation of products of treatment of Irritable Bowel Syndrome (IBS) (FDA, CDER, 2010) focuses on pain symptoms related to the disorder. In assessing effectiveness, this Guidance suggests utilizing co-primary endpoints involving stool frequency and pain intensity. For pain intensity, the recommended definition of treatment response is "Decrease in weekly average of worst abdominal pain in past 24 hours score of ≥ 30% compared with baseline" (italics in original). For the particular method of assessing pain intensity, the Guidance suggests:

We recommend evaluating abdominal pain intensity by using an 11-point (i.e., 0 to 10) numeric rating scale that asks patients daily to rate their worst abdominal pain over the past 24-hours. This type of pain assessment has been used to assess pain in somatic, visceral, and neuropathic chronic pain conditions.

Clinical assessment of pain intensity is notoriously complicated (Dworkin et al., 2005; IOM, 2011). Use of such 11-point scales or 0-100 range visual analog scales (VAS) in data production and of sample averages in data summary production is common in this diagnostic and treatment area (e.g. Gilron et al., 2005; Clegg et al., 2006). From a psychometric perspective, however, both of these measures are unambiguously ordinal-scale measures (although some argue that interval-scale properties might apply). Averages and percentage changes of ordinal-scale measures are not obviously descriptive of real phenomena. Analytical methods do exist that would permit the evaluation of the CE of competing interventions in some circumstances -- for instance, zero- or first-order stochastic dominance involving two distributions of ordinal measures does indicate unambiguous preference -- but methods are not obviously favored in practice. In light of such considerations, the particular 30% characterization of treatment response in conjunction with the measurement issues raised by use of an ordinal intensity measure suggests caution. It may ultimately be that the evidence that emerges from studies that utilize such constructs are important to and influential on patients and providers. This is a
matter that merits investigation, however, rather than one that can be taken for granted.

B. Clinical Trial Registries and ClinicalTrials.gov

In early 2000, the main U.S. clinical trials registry ClinicalTrials.gov came online amid concerns about post hoc cherry picking to report only favorable trial results (see Dahm, et al., 2009), informing patients about ongoing and planned trials in disease areas that might be of relevance, and other concerns. The notion that pre-trial registration of important features of a trial would reduce the scope for such gaming and would lead to a strong evidence base for healthcare is compelling. Indeed, in autumn of 2004, the International Committee of Medical Journal Editors (ICMJE) announced that the journals it represented would no longer (after July 1, 2005) publish the findings of clinical trials that had not been preregistered in ClinicalTrials.gov or similar registries in other nations (DeAngelis et al., 2004). Soon thereafter, Title VIII of the 2007 Food and Drug Administration Amendments Act required trials involving treatments that fall under FDA regulatory purview to register in ClinicalTrials.gov, with the added proviso that trial results would also be required to be reported in the database.

Figure 7 displays the time trend of trial registrations and the impact the ICMJE and FDAAA 2007 actions had on the time trend. As of June 25, 2011, the ClinicalTrials.gov registry contained 28,472 registered Phase-2 trials (856 of which reported results), and 19,052 Phase-3 trials (1,402 of which reported results). Tables 1 and 2 display a few of the prominent data fields from several recently-registered selected Phase-3 trials for Obesity and Diabetes, respectively.

With respect to the requirement to register various trial details, registrants are supposed to adhere to criteria established by the WHO in what is known as the WHO Trial Registration Data Set (or "minimal data set"), or a closely-related ClinicalTrials.gov version of the WHO criteria (See Tse et al., 2009). Regarding registration of primary and secondary outcomes (endpoints), the WHO criteria call for the following:

Primary Outcome(s)
Outcomes are events, variables, or experiences that are measured because it is believed that they may be influenced by the intervention. The Primary Outcome should be the outcome used in sample size calculations, or the main outcome(s) used to determine the effects of the intervention(s). Most trials should have only one primary outcome.

For each primary outcome provide:
The name of the outcome (do not use abbreviations)
The metric or method of measurement used (be as specific as possible)
The timepoint(s) of primary interest
Example:

Outcome Name: Depression  
Metric/method of measurement: Beck Depression Score  
Timepoint: 18 weeks following end of treatment

In theory, the primary and secondary outcome registration requirements under FDAAA 2007 have the potential to distort trial sponsors' behavior and, consequently, impact the evidence based. If sponsors must report on all outcomes studied rather than cherry picking, then decisions about which outcomes on which to produce data and data summaries may imaginably be influenced by the registration requirements. More subtly, registrants may be able to provide vague information about the outcomes being studied that permits some post hoc maneuverability (witness some of the entries in Tables 1 and 2).

Finally, and perhaps most subtly, the statistical approaches and data summaries (means, percentiles, etc.) used to analyze the produced data are rarely reported, and these too provide opportunities for possibly undesirable post-trial flexibility. Zarin et al., 2011, have analyzed non-Phase-1 trial registration data from ClinicalTrials.gov and discovered a high degree of unsatisfactory registration practice, with under 40% of their sample satisfactorily registering specific metrics and/or methods of aggregation (see Exhibit 2; also see Mathieu et al., 2009). Under FDAAA 2007, statistical methodologies are required to be reported in the ClinicalTrials.gov results section, and full trial protocols (which presumably include planned statistical analyses) are slated for inclusion in the database as well.

Even if proper registration of all other data elements is enforced strongly, the extra degrees of freedom afforded by slack enforcement of registration for various dimensions of primary and secondary outcomes and the statistical analyses thereof may be used advantageously by sponsor-registrants. Sponsors will presumably engage in registration practices that suit best their objectives of product approval, marketing, and reimbursement, subject to the apparently somewhat vague constraints imposed by FDAAA 2007. Awareness that such practices may ultimately impact the nature of the evidence that emerges from the trials conducted is warranted.

FORTHCOMING:

[C. Surrogate Endpoints, Clinical Endpoints, and Accelerated Approval]  
[D. Adverse Event Reporting and Labeling Changes]

---

20 There are some commonalities here with the issues that arise in the multi-target regulation or multiple policy instrument areas (see Bennear and Stavins, 2007).
6. Public Policy Considerations and Directions for Research

The topics discussed here raise considerations perhaps worthy of policy attention and future research. Several ideas are discussed briefly here.

A. Data Standardization and Measurement of Universal Outcomes

Etheridge, 2010, has advocated for standardization of data elements in clinical trials and effectiveness studies, while Tinetti and Studenski, 2011 have recently called for measurement and use of universal health outcomes in research and clinical care. While such appeals may be attractive from the perspectives of bookkeeping and tidying up of the evidence base, it is not at all obvious that standardization or appeal to universal health outcomes (QALYs?) would necessarily result in evidence that is more useful to patients and providers as they confront treatment choices. Indeed, standardization of data and data summaries using standards that are not usefully informative could be highly counterproductive. Measurement of universal outcome measures that do not speak to features of patients’ health that are salient at the time of treatment decisions will likely provide little sway at the point of decision. Such proposals may have merits in other contexts, but tradeoffs involved in their adoption should be appreciated and studied.

B. Reporting Empirical Methodologies in ClinicalTrials.gov

The findings by Zarin et al., 2011, of substandard reporting of various data categories in ClinicalTrials.gov raises concerns about both the integrity of the registration and results reporting in this registry as well as the impacts such reporting practices may ultimately have for the evidence base. The failure of registrants to report the statistical approaches to be used when analyzing data -- much less provide clarity about the definitions of the outcomes being analyzed -- leaves open the possibility that various forms of methodological data mining could be undertaken even if the statistical methods were required to be revealed before unblinding. The FDAAA 2007 requirements that original trial protocols also be submitted to the registry would provide some insurance against such gaming, but guaranteeing that information about methodologies to be used be transparent and easily accessible may prove difficult.

FORTHCOMING
[C. Guidance on Patient-Reported Outcomes]

D. PCORI Methodology Committee

The PCORI Methodology Committee faces a challenging and important mandate. Many issues -- some of which have been raised here -- contributing to the ultimate success or failure of the CE enterprise entail thorny methodological issues. To advance the CE enterprise in a positive direction, the Methodology Committee should pursue a creative agenda that enhances both the science and the practice of CE analysis. It may be useful to review the Methodology Committee’s charge in the PPACE:

Section 6301 of PPACA

“(C) FUNCTIONS.—Subject to subparagraph (D), the methodology committee shall work to develop and improve the science and methods of comparative clinical effectiveness research by, not later than 18 months after the establishment of the Institute, directly or through subcontract, developing and periodically updating the following:

“(i) Methodological standards for research. Such methodological standards shall provide specific criteria for internal validity, generalizability, feasibility, and timeliness of research and for health outcomes measures, risk adjustment, and other relevant aspects of research and assessment with respect to the design of research. Any methodological standards developed and updated under this subclause shall be scientifically based and include methods by which new information, data, or advances in technology are considered and incorporated into ongoing research projects by the Institute, as appropriate. The process for developing and updating such standards shall include input from relevant experts, stakeholders, and decisionmakers, and shall provide opportunities for public comment. Such standards shall also include methods by which patient subpopulations can be accounted for and evaluated in different types of research. As appropriate, such standards shall build on existing work on methodological standards for defined categories of health interventions and for each of the major categories of comparative clinical effectiveness research methods (determined as of the date of enactment of the Patient Protection and Affordable Care Act).

E. Research Plans

Planned empirical analysis [TO BE REVISED]

1. ClinicalTrials.gov: Changes in nature of registrations pre-post ICMJE or FDAAA-2007 (Phase 2,3; specific high-prevalence diseases) (statistical methods; outcome measures; etc.)

2. ClinicalTrials.gov: Examine concordance between registered (primary) outcomes and outcomes reported in ClinicalTrials.gov Results database and in published versions (Phase 2,3)
3. FDA Guidances: Examine post guidance concordance of outcome measures with guidance recommendations; examine pre-post guidance differences in outcome measures

7. Summary and Conclusions

As is true with any applied evaluation science, CE analysis done well is not a simple matter. There is an ideal paradigm wherein relevant data and data summaries and solid empirical methodologies join forces to build an evidence base that enhances patient and social welfare by improving treatment choices made by patients and their providers. But since applied CE analysis is an applied science, tradeoffs between conceptual rigor and informative empirical research are likely to be inevitable. Managing these tradeoffs so that empirical CE analysis rests on a solid conceptual foundation will be important.

At this juncture assessing how to move high quality CE analysis in meaningful ways into the discourse between patients and providers would appear to be high-order priority (see Avorn and Fisher, 2010). Understanding patients’ surely heterogeneous preferences and how they link to the underlying data and data summary structures of the evidence base is of considerable importance if the considerable resources invested in the CE enterprise are not to be squandered. While situations in which phenomena like 7th order stochastic dominance are encountered are likely to be few and far between, it seems essential not to take for granted the seemingly idiosyncratic but nonetheless very real nuances of patients’ preferences. Caring more than anything about the prospects of surviving long enough to participate in a child’s wedding ceremony eighteen months in the future is a very real feature of a patient’s preferences that CE analysis should strive to inform well.

Acknowledgments

Thanks are owed to Anirban Basu and to participants at seminars at Georgetown, Chicago, Coimbra, Milano-Bicocca, Wisconsin, York, and the 2011 iHEA Meetings in Toronto for helpful comments on earlier phases of this work. Bruce Hansen suggested some useful ways to think about the structural breaks evident in figure 5. Discussions with David Kindig and Steve Ziliak have been influential on my thinking about a variety of issues in CE analysis. Partial financial support has been provided by the RWJ Health & Society Scholars Program at UW-Madison. This paper is dedicated to the memory of Roger Sherman.
References


Institute of Medicine, Committee on Comparative Effectiveness Research Prioritization. 2009. *Initial

Institute of Medicine. 2011. Relieving Pain in America: A Blueprint for Transforming Prevention, Care, Education and Research. Washington: Institute of Medicine, Committee on Advancing Pain Research, Care, and Education.


O'Muircheartaigh, C. 2006. Aims of Public Policy Address. (Harris School of Public Policy, Univ. of Chicago).


Appendix: Selections from CFR Title 21 (Food and Drugs)

[a. Characterizing "Effectiveness" for New Drug Applications]

Sec. 314.125 Refusal to approve an application.

... 
(b) FDA may refuse to approve an application for any of the following reasons:

... 
(5) There is a lack of substantial evidence consisting of adequate and well-controlled investigations, as defined in Sec. 314.126, that the drug product will have the effect it purports or is represented to have under the conditions of use prescribed, recommended, or suggested in its proposed labeling.

[b. Measurement, Methodology, and Analysis]

Sec. 314.126 Adequate and well-controlled studies.

(a) The purpose of conducting clinical investigations of a drug is to distinguish the effect of a drug from other influences, such as spontaneous change in the course of the disease, placebo effect, or biased observation. The characteristics described in paragraph (b) of this section have been developed over a period of years and are recognized by the scientific community as the essentials of an adequate and well-controlled clinical investigation. The Food and Drug Administration considers these characteristics in determining whether an investigation is adequate and well-controlled for purposes of section 505 of the act. Reports of adequate and well-controlled investigations provide the primary basis for determining whether there is "substantial evidence" to support the claims of effectiveness for new drugs. Therefore, the study report should provide sufficient details of study design, conduct, and analysis to allow critical evaluation and a determination of whether the characteristics of an adequate and well-controlled study are present.

(b) An adequate and well-controlled study has the following characteristics:

(1) There is a clear statement of the objectives of the investigation and a summary of the proposed or actual methods of analysis in the protocol for the study and in the report of its results. In addition, the protocol should contain a description of the proposed methods of analysis, and the study report should contain a description of the methods of analysis ultimately used.

(2) The study uses a design that permits a valid comparison with a control to provide a quantitative assessment of drug effect.

(3) The method of selection of subjects provides adequate assurance that they have the disease or condition being studied, or evidence of susceptibility and exposure to the condition against which prophylaxis is directed.

... 
(6) The methods of assessment of subjects' response are well-defined and reliable. The protocol for the study and the report of results should explain the variables measured, the methods of observation, and criteria used to assess response.

(7) There is an analysis of the results of the study adequate to assess the effects of the drug. The report of the study should describe the results and the analytic methods used to evaluate them, including any appropriate statistical methods.
Sec. 330.10 Procedures for classifying OTC drugs as generally recognized as safe and effective and not misbranded, and for establishing monographs.

For purposes of classifying over-the-counter (OTC) drugs as drugs generally recognized among qualified experts as safe and effective for use and as not misbranded drugs, the following regulations shall apply:

(a) Procedure for establishing OTC drug monographs--

   (4) Standards for safety, effectiveness, and labeling. The advisory review panel, in reviewing the data submitted to it and preparing its conclusions and recommendations, and the Commissioner, in reviewing the conclusions and recommendations of the panel and the published proposed, tentative, and the final monographs, shall apply the following standards to determine general recognition that a category of OTC drugs is safe and effective and not misbranded:

   ...  
   (ii) Effectiveness means a reasonable expectation that, in a significant proportion of the target population, the pharmacological effect of the drug, when used under adequate directions for use and warnings against unsafe use, will provide clinically significant relief of the type claimed....

Sec. 601.25 Review procedures to determine that licensed biological products are safe, effective, and not misbranded under prescribed, recommended, or suggested conditions of use.

For purposes of reviewing biological products that have been licensed prior to July 1, 1972, to determine that they are safe and effective and not misbranded, the following regulations shall apply.

...  
(d) Standards for safety, effectiveness, and labeling. The advisory review panel, in reviewing the submitted data and preparing the panel's conclusions and recommendations, and the Commissioner of Food and Drugs, in reviewing and implementing the conclusions and recommendations of the panel, shall apply the following standards to determine that a biological product is safe and effective and not misbranded.

...  
(2) Effectiveness means a reasonable expectation that, in a significant proportion of the target population, the pharmacological or other effect of the biological product, when used under adequate directions, for use and warnings against unsafe use, will serve a clinically significant function in the diagnosis, cure, mitigation, treatment, or prevention of disease in man.
[e. Medical Devices and “Effectiveness”]

Sec. 860.7 Determination of safety and effectiveness.

... 
(e)(1) There is reasonable assurance that a device is effective when it can be determined, based upon valid scientific evidence, that in a significant portion of the target population, the use of the device for its intended uses and conditions of use, when accompanied by adequate directions for use and warnings against unsafe use, will provide clinically significant results.

[f. Clinical Endpoints; Surrogate Endpoints]

[Subpart H Accelerated Approval of New Drugs for Serious or Life-Threatening Illnesses]

Sec. 314.510 Approval based on a surrogate endpoint or on an effect on a clinical endpoint other than survival or irreversible morbidity.

FDA may grant marketing approval for a new drug product on the basis of adequate and well-controlled clinical trials establishing that the drug product has an effect on a surrogate endpoint that is reasonably likely, based on epidemiologic, therapeutic, pathophysiologic, or other evidence, to predict clinical benefit or on the basis of an effect on a clinical endpoint other than survival or irreversible morbidity. Approval under this section will be subject to the requirement that the applicant study the drug further, to verify and describe its clinical benefit, where there is uncertainty as to the relation of the surrogate endpoint to clinical benefit, or of the observed clinical benefit to ultimate outcome. Postmarketing studies would usually be studies already underway. When required to be conducted, such studies must also be adequate and well-controlled. The applicant shall carry out any such studies with due diligence.
Figure 1

*pubmed.gov* Citations, "comparative effectiveness" in Article Title, 2005Q1-2010Q2
(Red Line Indicates Passage of ARRA in Feb. 2009)
Survival after Primary Debulking Surgery (PDS) or Neoadjuvant Chemotherapy (NACT) for Stage IIIc or IV Ovarian Cancer (Adapted from Fig. 2B in Vergote et al., 2010)

(a)

(b)
Figure 3
Progression-Free Survival, Erlotinib (Tarceva) vs. Placebo
From U.S. FDA ODAC, 2009 (Line Segment Indicating 0.5 PFS Added by Author)

Kaplan-Meier Curves of Progression-free Survival in Overall Population (Full Analysis Set)

Log-rank Test:
\( p \)-Value: < .0001

Hazard Ratio: 0.71
95\% CI: [0.62:0.82]
Figure 4
Stochastic Dominance of Empirical Distributions: Zero-Order and First-Order Dominance

$F_A(y)^{>}_0 F_B(y)$

$F_A(y)^{>}_1 F_B(y)$
Figure 5
Demand Functions Based on Effectiveness Information (Adapted from Peltzman, 1973)
Figure 6
Depiction of Primary Efficacy Endpoints for Weight Management Products based on FDA Guidance (U.S. FDA CDER/CBER 2007)
(Dark Blue and Light Blue Curves Depict Prototypical Survivor Functions for Control and Treatment Interventions, Respectively)
Figure 7
ClinicalTrials.gov All New Submissions (Quarterly)

(a) Red Line Indicates Effective Date of ICMJE Registration Requirements;
Purple Line Indicates Effective Date of FDAAA 2007 Regulations

(b) Trend Lines Based on: Pre-ICMJE Requirement Period (Gray);
Post-ICMJE but Pre-FDAAA 2007 Period (Light Blue);
Post-FDAAA 2007 Period (Green)
Figure 8
Probability of Readmission or Mortality of Low-Risk Patients with Heart Failure under Usual Care or Usual Care with Nurse Management (Adapted from Fig. 2 in DeBusk et al., 2004)
Exhibit 1
AHRQ Clinician’s Guide for Treatments for Clinically Localized Prostate Cancer

**Treatments for Clinically Localized Prostate Cancer**

This guide summarizes clinical evidence comparing the effectiveness and safety of treatments for clinically localized prostate cancer. It discusses expectant management and three active treatments (radical prostatectomy, radiation therapy, and hormone therapy). This guide does not cover nutritional supplements. It also does not cover some newer treatments (cryotherapy, high-intensity focused ultrasound, and laparoscopic or robotic-assisted prostatectomy) for which there is little research about comparative effectiveness. This guide does not address strategies to prevent or screen for prostate cancer or strategies to treat advanced prostate cancer.

**CLINICAL ISSUE**

Prostate cancer is common and primarily affects men older than 65. About 10 percent of men diagnosed with prostate cancer will have clinically localized disease (cancer confined to the prostate gland).

Since the early 2000s, the number of men screened with the prostate specific antigen (PSA) blood test and the incidence of newly diagnosed cases have increased substantially.

PSA screening detects cancer earlier and, if smaller size than other clinical methods, increased detection of localized disease has led to more frequent intervention with treatments that are potentially effective but have side effects.

**CLINICAL BOTTOM LINE**

- Men who have high Gleason scores have higher rates of cancer recurrence and death than men who have tumors with low Gleason scores. This is true no matter what treatments are used.
  
  **Level of evidence:** **B**

- Men who have radical prostatectomy for localized prostate cancer are less likely to develop distant metastases than those who receive expectant management.
  
  **Level of evidence:** **C**

- Radical prostatectomy, radiation therapy, and hormonal therapy result in more long-term side effects than expectant management. These include sexual, urinary, and bowel problems.
  
  **Level of evidence:** **C**

- Evidence is insufficient to determine whether radiation or hormonal therapy results in fewer deaths or cancer progressions than expectant management.
  
  **Level of evidence:** **C**

- Evidence is insufficient to determine whether any type of radiation or hormonal therapy results in fewer deaths or cancer recurrences than radical prostatectomy.
  
  **Level of evidence:** **C**

**Figure 2. Radical Prostatectomy Versus Expectant Management**

<table>
<thead>
<tr>
<th>Outcome 10 Years After Diagnosis (%)</th>
<th>Radical Prostatectomy</th>
<th>Expectant Management</th>
</tr>
</thead>
<tbody>
<tr>
<td>Local progression</td>
<td>19</td>
<td>25</td>
</tr>
<tr>
<td>Distant metastases</td>
<td>25</td>
<td>15</td>
</tr>
<tr>
<td>Death from prostate cancer</td>
<td>15</td>
<td>10</td>
</tr>
<tr>
<td>Death from any cause</td>
<td>32</td>
<td>27</td>
</tr>
</tbody>
</table>

1 Results are from a large, randomized controlled trial of men diagnosed between 1989 and 1999. Average age at diagnosis was 65 years, average Gleason score 5–6, and average PSA 13 mg/mL. Most cancers were not diagnosed by PSA testing. Men in both groups who had local progression received hormonal therapy. (N Engl J Med 2005;352:1977-84.)

**SOURCE**

The source material for this guide is a systematic review of 155 research publications. The review, Comparative Effectiveness of Hormone-Induced Localized Prostate Cancer (2010), was prepared by the Evidence-based Practice Center, The Agency for Healthcare Research and Quality (AHRQ) funded the systematic review and this guide. The guide was developed using feedback from clinicians who reviewed preliminary drafts.

**CONFIDENCE SCALE**

The evidence ratings in this guide are derived from a systematic review of the literature. The level of evidence is based on the overall quality and quantity of clinical evidence.

1. Evidence. Studies are consistent with each other and are based on clinical evidence. Further research is unlikely to change the conclusions.
2. Moderate evidence. Studies are consistent but further research may change the conclusions.
3. Limited evidence. There are few studies, or studies are inconsistent and weak.

2010 AHRQ

2013 AHRQ

July 2010
Exhibit 2
Zarin et al., 2011, Example Taxonomy for Clinical Trial Outcome Registration

Level 1
Domain

Level 2
Specific Measurement

Level 3
Specific Metric

Level 4
Method of Aggregation
<table>
<thead>
<tr>
<th>NCT ID</th>
<th>Title</th>
<th>Conditions</th>
<th>Interventions</th>
<th>Enrollment</th>
<th>Outcome Measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCT01258114</td>
<td>Maâthermes: Spa Treatment for Overweight and Obesity</td>
<td>Overweight</td>
<td>Obesity</td>
<td>Other: SPA treatment (ST)</td>
<td>Other: Non SPA treatment (NST)</td>
</tr>
<tr>
<td>NCT01158417</td>
<td>Resveratrol in Type2 Diabetes and Obesity</td>
<td>Type 2 Diabetes</td>
<td>Obesity</td>
<td>Insulin Resistance</td>
<td>Drug: Placebo</td>
</tr>
<tr>
<td>NCT01239550</td>
<td>Insulin Detemir in Obesity Management</td>
<td>Diabetes</td>
<td>Obesity</td>
<td>Drug: Detemir</td>
<td></td>
</tr>
<tr>
<td>NCT01344161</td>
<td>Effect of Vitamin D on Metabolic Profile in Overweight or Obese Women</td>
<td>Overweight</td>
<td>Obesity</td>
<td>Dietary Supplement: Effect of vitamin D on metabolic profile in women,</td>
<td>Dietary Supplement: Effect of vitamin D on metabolic profile in women.</td>
</tr>
<tr>
<td>NCT01332877</td>
<td>Breakfast Size and Weight Loss in Overweight/Obese Adults</td>
<td>Obesity</td>
<td>Behavioral: Enriched breakfast</td>
<td>Behavioral: Control breakfast</td>
<td></td>
</tr>
<tr>
<td>NCT01233349</td>
<td>Safety and Efficacy of Litramine in Overweight and Obese Subjects</td>
<td>Overweight</td>
<td>Obesity</td>
<td>Weight Loss</td>
<td>Device: Litramine</td>
</tr>
</tbody>
</table>
Table 2
ClinicalTrials.gov, Selected Studies First Registered after July 1, 2010. Phase-3 Trials for Diabetes (Selected Fields)

<table>
<thead>
<tr>
<th>NCT ID</th>
<th>Title</th>
<th>Conditions</th>
<th>Interventions</th>
<th>Enrollment</th>
<th>Outcome Measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCT01248143</td>
<td>Molecular and Clinical Effects of Green Tea and Fermented Papaya Preparation on Diabetes and Cardiovascular Diseases</td>
<td>Assess the Effect of Green Tea on Diabetes</td>
<td>Dietary Supplement: Green tea</td>
<td>300</td>
<td>Assess the effects of green tea and FPP on the levels of C-reactive proteins</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Assess the Effect of Fermented Papaya Pretreatment on Diabetes</td>
<td>Dietary Supplement: FPP</td>
<td></td>
<td>Assess the effect of green tea and fermented papaya preparation on development of atheroma and drug therapy outcomes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Effects of Green Tea and FPP on C-reactive Proteins</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Effects of Green Tea and FPP on Lipid Profiles in Diabetes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Effect of Green Tea and FPP on Atheroma Formation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NCT01234649</td>
<td>Combined Liraglutide and Metformin Therapy in Women With Previous Gestational Diabetes Mellitus (GDM)</td>
<td>Gestational Diabetes Mellitus</td>
<td>Drug: Metformin XR only</td>
<td>150</td>
<td>An index of insulin secretion in relation to insulin resistance (IS-SI) will be calculated</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Type 2 Diabetes Mellitus</td>
<td>Drug: Metformin XR plus liraglutide</td>
<td></td>
<td>Insulin resistance - baseline (HOMA-IR) and composite insulin sensitivity index [ISIOGTT], and pancreatic ß-cell function (corrected insulin response [CIRglupeak] and insulogenic index [GI])/(HOMA-IR)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Metabolic Syndrome</td>
<td></td>
<td></td>
<td>Cardiometabolic risk measures</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Impaired Glucose Tolerance</td>
<td></td>
<td></td>
<td>Anthropometric measurements</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Impaired Fasting Glucose</td>
<td></td>
<td></td>
<td>Development of dysglycemia</td>
</tr>
<tr>
<td>NCT01323348</td>
<td>Effect of Diabetes Education During Retinal Ophthalmology Visits on Diabetes Control</td>
<td>Diabetes</td>
<td>Behavioral: Diabetes Education</td>
<td>2000</td>
<td>Change in HbA1c</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Diabetic Retinopathy</td>
<td></td>
<td></td>
<td>Diabetes Care Knowledge</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Blood Pressure</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Body Mass Index</td>
</tr>
<tr>
<td>NCT01162876</td>
<td>A Clinical Pharmacology Study of Saxagliptin in Patients With Type 2 Diabetes Mellitus</td>
<td>Diabetes</td>
<td>Drug: saxagliptin</td>
<td>20</td>
<td>Pharmacokinetics</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Pharmacodynamics</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Safety issues</td>
</tr>
<tr>
<td>NCT01201928</td>
<td>Pulmonary Function Substudy for Subjects Enrolled in Studies MKC-TI-161, MKC-TI-162 or MKC-TI-166</td>
<td>Type 1 Diabetes Mellitus</td>
<td>Drug: Comparator administered in parent trial</td>
<td>200</td>
<td>Comparison of change from baseline to final treatment visit in pulmonary function (forced expiratory volume in 1 second, forced vital capacity, total lung capacity, and DLco)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Type 2 Diabetes Mellitus</td>
<td>Drug: Technosphere Insulin Inhalation Powder</td>
<td></td>
<td>between treatment Groups (TI vs comparator arms) using ANCOVA models.</td>
</tr>
<tr>
<td>NCT01184703</td>
<td>Low Glycemic Index Diet in Patients With Type 1 Diabetes</td>
<td>Diabetes Mellitus, Type 1</td>
<td>Other: Low GI food</td>
<td>65</td>
<td>HbA1c</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>LDL-cholesterol</td>
</tr>
</tbody>
</table>