PERSONALIZED PATIENT CARE UNDER UNCERTAINTY

Charles F. Manski

Department of Economics and Institute for Policy Research

Northwestern University

Duke University

September 10-11, 2018

Schedule

September 10, 2018

Registration and Coffee (9:30 - 10:00 AM)

Opening Remarks (10:00 - 10:15 AM)

Lecture 1 (10:15 - 11:45 AM): Introduction, Clinical Guidelines and Clinical Judgment

Lunch (11:45 AM - 1:00 PM)

Lecture 2 (1:00 - 2:30 PM): Wishful Extrapolation from Clinical Trials to Patient Care

Break (2:30 - 3:00 PM)

Lecture 3 (3:00 - 4:30 PM): Credible Use of Evidence to Inform Patient Care

September 11, 2018

Coffee (8:30 - 9:00 AM)

Lecture 4 (9:00 - 10:30 AM): Reasonable Care under Uncertainty, Reasonable Treatment Choice with Trial Data

Break (10:30 - 11:00 AM)

Lecture 5 (11:00 AM - 12:30 PM): A Public Health Perspective on Reasonable Care, Managing Uncertainty in Drug Approval

Lunch (12:30 - 1:30 PM)

Conclusion and Closing Discussion (1:30 - 3:00 PM)

Introduction

It is useful to distinguish three broad branches of decision analysis:

normative, descriptive, prescriptive.

Normative analysis seeks to establish ideal properties of decision making.

Descriptive analysis seeks to understand and predict how decision makers behave.

Prescriptive analysis seeks to improve the performance of actual decision making.

Prescriptive analysis draws on normative thinking to define the term "improve" and on descriptive research to characterize actual decisions.

My concern is prescriptive analysis that seeks to improve patient care.

My focus is decision making under uncertainty.

By "uncertainty," I do not just mean that clinicians make probabilistic rather than deterministic predictions of patient outcomes.

I mean that available evidence and medical knowledge may not suffice to yield precise probabilistic predictions.

A patient may ask:

"What is the chance that I will develop disease X in the next five years?"

"What is the chance that treatment Y will cure me?"

A credible response may be a range, say "20 to 40 percent" or "at least 50 percent."

Decision theorists may use the terms "deep uncertainty" and "ambiguity." I encompass them within the broader term "uncertainty."

Uncertainty has sometimes been acknowledged verbally, but it has generally not been addressed in research on evidence-based medicine.

I think this a huge omission.

Surveillance or Aggressive Treatment

I pay particular attention to choice between surveillance or aggressive treatment of patients at risk of potential disease.

Consider women at risk of breast cancer.

Surveillance = periodic mammograms and clinical exams.

Aggressive treatment = preventive drug treatment or mastectomy.

Or consider surveillance or drug treatment for patients at risk of heart disease.

Or consider surveillance or chemotherapy for patients treated for localized cancer who are at risk of metastasis.

Uncertainty often looms large when one contemplates the risk of disease development and the outcome of aggressive treatments.

Decision making may require resolution of a tension between benefits and costs.

Aggressive treatment may be more beneficial to the extent that it reduces the risk of disease development or the severity of disease that does develop.

It may be more costly to the extent that it generates health side effects and financial costs beyond those associated with surveillance.

Evolution of My Research Program

I have no formal training in medicine. The contributions that I may be able to make concern the methodology of evidence-based medicine.

This lies within the expertise of econometricians, statisticians, and decision analysts.

Research on treatment response and personalized risk assessment shares a common objective: Probabilistic prediction of some patient outcome conditional on patient attributes.

Development of methodology for probabilistic conditional prediction has long been a core concern of statistics and econometrics.

Probabilistic Conditional Prediction = *regression*, *actuarial prediction*, *statistical prediction*, *machine learning*, *predictive analytics, AI*.

Statistical imprecision and **identification problems** affect empirical (evidence-based) research that uses sample data to predict population outcomes.

Statistical theory characterizes the inferences that can be drawn about a study population by observing a sample.

Identification analysis studies inferential difficulties that persist when sample size grows without bound.

I focus mainly on identification problems, which often are the dominant difficulty.

The logic of empirical inference is summarized by the relationship:

assumptions + data ⇒ conclusions.

There is a tension between the strength of assumptions and their credibility, that I have called (Manski, 2003):

The Law of Decreasing Credibility: The credibility of inference decreases with the strength of the assumptions maintained.

I have argued against making precise predictions with *incredible certitude*.

Probabilities of future events may be partially rather than point identified.

That is, research may be able to credibly bound the probability that an event will occur but not make credible precise probabilistic predictions, even with large data samples.

My research on partial identification began in the late 1980s.

Whereas my early research focused on prediction per se, I have over time investigated the implications for decision making.

Example: How might one choose between treatment A and B when one cannot credibly identify the sign of the average treatment effect?

Elementary decision theory suggests a two-step process for choice under uncertainty.

(1). Eliminate dominated actions.

(2). Choose an undominated action.

There is no **optimal** way to choose among undominated alternatives.

There are only various **reasonable** ways.

Patient care is ripe for study as a problem of decision making under uncertainty.

I have sought to learn enough about evidence-based medicine to make original contributions.

The results include studies of

* diagnostic testing and treatment under uncertainty (Manski, 2009, 2013b)
* personalized medicine with partial assessment of health risks (Manski, 2018a)
* analysis and design of randomized trials (Manski, 2004; Manski and Tetenov, 2016)
* drug approval (Manski, 2009)
* vaccination policy with partial knowledge of disease transmission (Manski, 2010, 2017b).

Manski (2018b) is a review article.

I am now in process of bringing this work together in a book.

A book provides the space to present major themes and to show how they become manifest in various specific contexts.

A book enables one to speak to a broader audience than is possible when writing research articles on particular topics.

I hope that this book will prove useful to a spectrum of readers.

* clinicians who make decisions about patient care.
* patients who participate in their own care.
* medical researchers.
* biostatisticians who assist medical researchers.
* health economists.

Some will correctly view my work as critical of existing methodologies:

* biostatisticians who have used the statistical theory of hypothesis testing to advise medical researchers on the design and analysis of randomized trials.

* guideline developers who have argued that evidence-based medicine should rest solely or predominately on evidence from randomized trials.

I hope that these readers will view my work as providing constructive suggestions.

To make the lectures and book accessible, the exposition is almost entirely verbal.

I give references to the technical articles that provide the full analysis.

Clinical Guidelines and Clinical Judgment

Medical texts and training have long offered clinicians guidance in patient care.

Guidance has increasingly become institutionalized through issuance of clinical practice guidelines (CPGs).

Dictionaries typically define a "guideline" as a suggestion or advice for behavior rather than a mandate. Institute of Medicine (IOM) (2011) writes (p. 4):

> "Clinical practice guidelines are statements that include recommendations intended to optimize patient care that are informed by a systematic review of evidence and an assessment of the benefits and harms of alternative care options."

Although CPGs are not mandates, clinicians often have strong incentives to comply with guidelines, making adherence close to compulsory.

A health insurance plan may require adherence to a CPG as a condition for reimbursement of the cost of treatment.

Adherence may furnish evidence of due diligence that legally defends a clinician in the event of a malpractice claim.

Adherence provides a rationale for care decisions that might otherwise be questioned by patients, colleagues, or employers.

The medical literature contains many commentaries exhorting clinicians to adhere, arguing that guideline developers have superior knowledge than do clinicians.

Wennberg (2011) defines "unwarranted variation" as variation that (p. 687): "isn't explained by illness or patient preference."

The UK National Health Service gives its *Atlas of Variation in Healthcare* (2015) the subtitle "Reducing unwarranted variation to increase value and improve quality."

IOM (2011) states: "Trustworthy CPGs have the potential to reduce inappropriate practice variation."

These and similar quotations exemplify a widespread belief that adherence to guidelines is socially preferable to decentralized clinical decision making.

There are two reasons why patient care adhering to guidelines may differ from the care that clinicians provide:

(1) The developers of guidelines may differ from clinicians in their ability to predict how decisions affect patient outcomes.

(2) Guideline developers and clinicians may differ in how they evaluate patient outcomes.

I focus mainly on (1).

Variation in Guidelines

Commentaries exhorting adherence to guidelines implicitly assume that guideline developers agree on appropriate patient care.

Consensus holds in some cases, but the guidelines issued by distinct health organizations often differ from one another.

Variation in care is inevitable if clinicians adhere to different guidelines.

Example:

Consider the role of the clinical breast examination (CBE) in breast cancer screening.

Guidelines issued by the National Comprehensive Cancer Network (NCCN), the American Cancer Society (ACS), and the U.S. Preventive Services Task Force (USPSTF) differ markedly.

The NCCN considers the CBE to be a core element of breast cancer screening.

The ACS takes a firmly negative position.

The USPSTF is neutral.

Degrees of Personalized Medicine

In principle, *personalized medicine* means health care specific to the individual.

In practice, the term means care that varies with some observed patient attributes.

Thus, personalized medicine is a matter of degree.

Clinicians often observe patient attributes beyond those used to predict outcomes in evidence-based risk assessments and studies of treatment response.

Hence, clinicians often can personalize patient care to a greater degree than do evidence-based CPGs.

Prediction of Cardiovascular Disease

The ASCVD Risk Estimator of the American College of Cardiology predicts the probability a person will develop atherosclerotic cardiovascular disease (ASCVD) in the next ten years.

This online tool conditions the reported probability on:

Patient Demographics: current age, sex, race

Current Labs/Exams: total, HDL, and LDL cholesterol; systolic blood pressure.

Personal History: history of diabetes, hypertension treatment, smoker, on a statin, on aspirin therapy.

It does not condition on obesity, job stress, and exercise.

These attributes are observable by clinicians and may be thought relevant risk factors.

The Breast Cancer Risk Assessment Tool

The Breast Cancer Risk Assessment (BCRA) Tool of the National Cancer Institute gives an evidence-based probability that a woman will develop breast cancer conditional on eight attributes:

(1) history of breast cancer or chest radiation therapy for Hodgkin Lymphoma.
(2) a BRCA mutation or genetic syndrome associated with risk of breast cancer.
(3) current age.
(4) age of first menstrual period.
(5) age of first live birth of a child.
(6) number of first-degree female relatives with breast cancer.
(7) number of breast biopsies.
(8) race/ethnicity.

The BCRA Tool has become widely used and is an important input to the CPG for breast cancer screening issued by the NCCN.

The reason that the Tool assesses risk conditional on eight attributes and not others is that it uses a modified version of the "Gail Model," based on Gail *et al.* (1989).

The Gail *et al.* article estimated probabilities of breast cancer for white women who have annual breast screening, conditional on attributes (1) through (7).

The BCRA Tool does not condition on further patient attributes that may be associated with risk of cancer, including:

* the number and ages of a woman's first-degree relatives.

* the prevalence of breast cancer among second-degree relatives.

* membership in ethnic groups who have differential risk of a BRCA mutation.

* behavioral attributes such as excessive drinking of alcohol.

Predicting Unrealistically Precise Probabilities

A user of the ASCVD Risk Estimator or the BCRA Tool receives a precise probability of disease development.

Statistical imprecision and identification problems make these risk assessments uncertain.

Without discussion of uncertainty, clinicians and patients may mistakenly believe that precise probabilistic risk assessments are accurate.

If uncertainty is not quantified, those who recognize the presence of uncertainty cannot evaluate the degree to which assessments may be inaccurate.

Consider statistical imprecision in the BCRA Tool.

The Gail *et al.* article calls attention to statistical imprecision in its predictions.

It describes a general procedure for estimating confidence intervals.

It reports illustrative computations of 95% confidence intervals for two women with different attributes.

The confidence intervals vary considerably in width.

The BCRA Tool does not report confidence intervals.

The NCI website that houses the Tool makes no mention of statistical imprecision.

Optimal Care Assuming Rational Expectations

The BCRA Tool and ASCVD Risk Estimator exemplify a common question.

Evidence enables assessment of risk of disease or treatment response conditional on certain patient attributes. CPG recommendations condition on these attributes.

Clinicians observe additional attributes that may be informative predictors of outcomes.

Available evidence does not show how outcomes vary with the additional attributes.

How should medical decision making proceed?

Economists have studied this question in an idealized setting of individualistic patient care with *rational expectations*.

A clinician chooses how to treat each patient in a population. The objective is to maximize a *utilitarian* welfare function. That is, the clinician aims to do what is best for each patient.

The care received by one patient affects only that individual and not other members of the population. This is realistic when considering non-infectious diseases.

Rational expectations means that the clinician knows the actual probability distribution of health outcomes that may occur if a patient with specified observed attributes is given a specified treatment.

With rational expectations, optimization of patient care has a simple solution.

Patients should be divided into groups having the same observed attributes.

All patients in a group should be given the care that yields the highest within-group mean welfare.

Thus, it is optimal to differentially treat patients with different attributes if different treatments maximize their within-group mean welfare.

Patients with the same observed attributes should be treated uniformly.

The value of maximum welfare increases as more patient attributes are observed.

Optimal Choice Between Surveillance and Aggressive Treatment

Aggressive treatment may be more beneficial to the extent that it reduces the risk of disease development or the severity of disease that does develop.

It may be more costly to the extent that it generates health side effects and financial costs beyond those associated with surveillance.

Given certain assumptions, aggressive treatment is the better option if the probability of disease development exceeds a patient-specific threshold.

Surveillance is the better option otherwise.

Psychological Research Comparing Evidence-Based Prediction and Clinical Judgment

If clinicians have rational expectations, there is no utilitarian argument for CPGs.

Psychological research concludes that clinicians do not have rational expectations.

Psychologists have found that evidence-based prediction consistently outperforms clinical judgment when the predictions are made using the same patient attributes.

The gap in performance persists even when clinical judgment uses additional attributes as predictors.

Notable early contributions include Sarbin (1943, 1944), Meehl (1954), Goldberg (1968).

Dawes, Faust, and Meehl (1989) distinguish evidence-based (aka actuarial or statistical) prediction and clinical judgment as follows:

> "In the clinical method the decision-maker combines or processes information in her or her head. In the actuarial or statistical method the human judge is eliminated and conclusions rest solely on empirically established relations between data and the condition or event of interest."

Comparing the two in circumstances where a clinician observes patient attributes that are not utilized in available evidence-based prediction, they state:

> "Might the clinician attain superiority if given an informational edge? For example, suppose the clinician lacks an actuarial formula for interpreting certain interview results and must choose between an impression based on both interview and test scores and a contrary actuarial interpretation based on only the test scores. The research addressing this question has yielded consistent results . . . . Even when given an information edge, the clinical judge still fails to surpass the actuarial method; in fact, access to additional information often does nothing to close the gap between the two methods."

They attribute the weak performance of clinical judgment to clinician failure to adequately grasp the logic of the prediction problem and to their use of decision rules that place too much emphasis on unusual patient attributes.

I have found no explicit reference to this psychological research in my reading of medical commentaries advocating adherence to CPGs, nor in the broader literature concerning practice of evidence-based medicine.

I have found passages in the literature on evidence-based medicine that praise rather than criticize exercise of clinical judgment.

Welfare Comparison of Adherence to Guidelines and Clinical Judgment

The psychological literature challenges the realism of assuming that clinicians have rational expectations.

However, it does not imply that adherence to CPGs would yield greater welfare than decision making using clinical judgment.

The literature has not addressed all welfare-relevant aspects of clinical decisions.

Psychologists have studied the relative accuracy of risk assessments made by evidence-based predictors and by clinicians, but they have not studied the relative accuracy of evaluations of patient preferences over health outcomes.

Psychological research has seldom examined the accuracy of probabilistic risk assessments and diagnoses.

It has been more common to assess the accuracy of point predictions.

Study of the logical relationship between probabilistic and point prediction shows that data on the latter at most yields wide bounds on the former.

Manski (1990a) considers a forecaster who is asked to give a yes/no point prediction that an event will occur.

Observation that the forecaster states "yes" or "no" only implies that he judges the probability to be in the interval [½, 1] or [0, ½] respectively.

It is not possible at present to conclude that imperfect clinical judgment makes adherence to CPGs superior to decentralized decision making.

The findings of the psychological literature do imply that welfare comparison is a delicate matter of choice between alternative systems for patient care.

Adherence to evidence-based CPGs may be inferior if CPGs condition on fewer patient attributes than do clinicians.

It may be superior if imperfect clinical judgment yields sub-optimal decisions.

How these opposing forces interplay may depend on the specifics of the setting.

Wishful Extrapolation of Trial Findings to Patient Care

The fact that predictions are evidence-based does not ensure that they use evidence effectively.

Multiple questionable methodological practices have long afflicted research on health outcomes and may afflict the development of guidelines.

I focus on predictions made with evidence from randomized trials.

Trials have long enjoyed a favored status within medical research on treatment response and are often called the "gold standard" for such research.

Guideline developers value trial evidence more than observational studies. They sometimes use only trial evidence, excluding observational studies.

The appeal of trials is that, with sufficient sample size and complete observation of outcomes, they deliver credible findings on treatment response in the study population.

However, extrapolation of findings from trials to clinical practice can be difficult.

Researchers and guideline developers often use untenable assumptions to extrapolate.

I have called this *wishful extrapolation*.

From Study Populations to Patient Populations

Study populations in trials often differ from patient populations.

It is common to perform trials studying treatment of a specific disease only on subjects who have no co-morbidities or who have specific co-morbidities.

Guideline developers sometime caution about the difficulty of using trial findings to make care recommendations for patients with co-morbidities.

The problem is well-stated by Wong *et al.* (2017) in an article presenting updated guidelines for treatment of melanoma:

"Creating evidence-based recommendations to inform treatment of patients with additional chronic conditions, a situation in which the patient may have two or more such conditions—referred to as multiple chronic conditions (MCC)—is challenging. Patients with MCC are a complex and heterogeneous population, making it difficult to account for all of the possible permutations to develop specific recommendations for care. In addition, the best available evidence for treating index conditions, such as cancer, is often from clinical trials whose study selection criteria may exclude these patients to avoid potential interaction effects or confounding of results associated with MCC. As a result, the reliability of outcome data from these studies may be limited, thereby creating constraints for expert groups to make recommendations for care in this heterogeneous patient population."

Another source of difference is that a study population consists of persons with specified demographic attributes who volunteer to participate in a trial.

Volunteers are those who respond to financial and medical incentives to participate.

The study population differs materially from the relevant patient population if subjects and non-subjects have different distributions of treatment response.

Treatment response in the latter group is not observed.

It may be wishful extrapolation to assume that treatment response in trials performed on **volunteers** with **specified demographic attributes** who **lack co-morbidities** is the same as what would occur in actual patient populations.

Trials of Drug Treatments for Hypertension

Evidence from dozens of trials was utilized by the Eighth Joint National Committee (JNC 8), which promulgated the 2014 guidelines for management of high blood in the United States (James *et al.*, 2014).

The JNC 8 highlighted trials by Beckett *et al.* (2008), SHEP Cooperative Research Group (1991), and Staessen *et al.* (1997).

Participants differed in age, countries of residence, and the rules governing exclusions for co-morbidities. They differed in their blood pressure levels before being randomized into treatment. The SHEP and Staessen trials, but not the Beckett trial, restricted eligibility to persons with *isolated systolic hypertension*, a condition in which systolic blood pressure is higher than desirable but diastolic blood pressure is in the normal range.

Campbell and the Primacy of Internal Validity

To justify trials performed on study populations that may differ substantially from patient populations, researchers often cite Donald Campbell, who distinguished between the internal and external validity of studies of treatment response (Campbell and Stanley, 1963).

A study has have *internal validity* if it has credible findings for the study population.

It has *external validity* if an invariance assumption permits credible extrapolation.

The appeal of randomized trials is their internal validity.

Wishful extrapolation is an absence of external validity.

Campbell argued that studies should be judged primarily by their internal validity and secondarily by their external validity.

This perspective has been used to argue for the primacy of experimental research over observational studies, whatever the study population may be.

The Campbell position is well grounded if treatment response is homogeneous.

Then researchers can learn about treatment response in easy-to-analyze study populations and clinicians can confidently extrapolate findings to patient populations.

However, homogeneity of treatment response seems the exception rather than the rule.

Hence, it may be wishful to extrapolate from a study population to a patient population.

From Experimental Treatments to Clinical Treatments

Treatments in trials often differ from those that occur in clinical practice.

This is particularly so in trials comparing drug treatments. Drug trials are double-blinded, neither the patient nor the clinician knowing the assigned treatment.

A double-blinded drug trial reveals the distribution of response in a setting where patients and clinicians are uncertain what treatment a patient is receiving.

It does not reveal what response would be when patients and clinicians know what drug is being administered and can react to this information.

Consider drug treatments for hypertension.

Patients react heterogeneously to the various drugs available for prescription.

A clinician treating a particular patient may sequentially prescribe alternative drugs, trying each for a period in an effort to find one that performs satisfactorily.

Sequential experimentation is not possible in a trial. The standard protocol prohibits the clinician from knowing what drug a subject is receiving and from using judgment to modify the treatment.

Blinding is also problematic for clinical interpretation of noncompliance with assigned treatments.

From Surrogate Outcomes to Health Outcomes

A serious measurement problem often occurs in trials with short durations, which measure surrogate outcomes rather than ones of intrinsic health interest.

Example: Treatments for heart disease may be evaluated using data on cholesterol levels and blood pressure rather than data on heart attacks and life span.

The most lengthy trials for drug approval, *phase 3 trials*, typically run for two to three years.

Credible extrapolation from surrogate outcomes to outcomes of interest can be challenging.

Considering approval of cancer drugs in the United States and Europe, Prasad (2017) writes:

> "Although we are approving cancer drugs at a rapid pace, few come to market with good evidence that they improve patient centred outcomes. If they do, they often offer marginal benefits that may be lost in the heterogeneous patients of the real world. Most approvals of cancer drugs are based on flimsy or untested surrogate endpoints, and postmarketing studies rarely validate the efficacy and safety of these drugs on patient centred endpoints. Add to this that the average cancer drug costs in excess of $100 000 (£75 000; €85 000) per year of treatment, and the conclusion seems that the regulatory system is broken."

From Hypothesis Tests to Treatment Decisions

A longstanding practice has been to use trial data to test a specified null hypothesis against an alternative and to use the outcome of the test to compare treatments.

A common procedure when comparing two treatments is to view one as the status quo and the other as an innovation.

The usual null hypothesis is that the innovation is no better than the status quo and the alternative is that the innovation is better.

If the null hypothesis is not rejected, it is recommended that the status quo be used in practice. If the null is rejected, the innovation is recommended.

It has been standard to fix the probability of rejecting the null hypothesis when it is correct, the probability of a Type I error.

Sample size determines the probability of rejecting the alternative when it is correct, the probability of a Type II error. Power is one minus the probability of a type II error.

The convention has been to choose a sample size that yields specified power at a value of the effect size deemed clinically important.

International Conference on Harmonisation (1999) provides this guidance for drug trials:

> "Conventionally the probability of type I error is set at 5% or less or as dictated by any adjustments made necessary for multiplicity considerations; the precise choice may be influenced by the prior plausibility of the hypothesis under test and the desired impact of the results. The probability of type II error is conventionally set at 10% to 20%."

Manski and Tetenov (2016) give several reasons why hypothesis testing may yield unsatisfactory results for medical decisions:

Use of Conventional Asymmetric Error Probabilities

It has been standard to fix the probability of Type I error at 5% and the probability of Type II error at 10-20%. The theory of hypothesis testing gives no rationale for selection of these conventional error probabilities. It gives no reason why a clinician concerned with patient welfare should find it reasonable to make treatment choices that have a substantially greater probability of Type II than Type I error.

Inattention to Magnitudes of Losses When Errors Occur

A clinician should care about the magnitudes of the losses to patient welfare that arise when errors occur. A given error probability should be less acceptable when the welfare difference between treatments is larger. The theory of hypothesis testing does not take this into account.

Limitation to Settings with Two Treatments

A clinician often chooses among several treatments and many clinical trials compare more than two treatments. Yet the standard theory of hypothesis testing only contemplates choice between two treatments. Statisticians have struggled to extend it to deal sensibly with comparisons of multiple treatments.

Example

A terminal form of cancer may be treated by a status quo treatment or an innovation.

It is known that mean patient life span with the status quo treatment is 1 year.

Medical researchers see two possibilities for the effectiveness of the innovation.

It may yield a mean life span of 1/3 year or 5 years.

A trial is performed to learn the effectiveness of the innovation.

The trial data are used to perform a test comparing the innovation and the status quo. The null is that the innovation is no more effective than the status quo. The alternative is that the innovation is more effective.

The probability of a Type I error is set at 0.05 and that of a Type II error is 0.20. The result is used to choose between treatments.

A Type I error occurs with probability 0.05 and reduces mean patient life span by 2/3 of a year (1 year minus 1/3 year).

A Type II error occurs with frequentist probability 0.20 and reduces mean patient life span by 4 years (5 years minus 1 year).

Use of the test to choose the treatment implies that society is willing to tolerate a large (0.20) chance of a large welfare loss (4 years) when making a Type II error, but only a small (0.05) chance of a small welfare loss (2/3 of a year) when making a Type I error.

The theory of hypothesis testing does not motivate this asymmetry.

Wishful Meta-Analyses of Disparate Studies

The problems discussed above relate to analysis of findings from single trials.

Further difficulties arise when researchers and guideline developers attempt to combine findings from multiple trials.

It is easy to understand the impetus for combination of findings.

Decision makers must interpret the mass of information provided by research.

The hard question is how to interpret this information sensibly.

Combination of findings is sometimes performed by *systematic review* of a set of studies. This is a subjective process similar to exercise of clinical judgment.

Statisticians have proposed *meta-analysis* in an effort to provide an objective methodology for combining the findings of multiple studies.

Meta-analysis was originally developed to address a purely statistical problem.

Suppose that multiple trials have been performed on the same study population, each drawing an independent random sample.

The most precise way to use the data combines them into one sample.

Suppose that the raw data are unavailable. Instead, multiple parameter estimates are available, each computed with the data from a different sample.

Meta-analysis proposes methods to combine the multiple estimates. The usual proposal is to compute a weighted-average of the estimates, the weights varying with sample size.

The original concept of meta-analysis is uncontroversial, but its applicability is limited.

It is rare that multiple independent trials are performed on the same population.

It is more common for multiple trials to be performed on distinct populations that may have different distributions of treatment response.

Administration of treatments and measurement of outcomes may vary across trials.

Meta-analysis are performed often in such settings, computing weighted averages of estimates for distinct study populations and trial designs.

Averages computed with subjective weights are called *Bayesian weighted averages*.

It may not be clear how to interpret a weighted average of the estimates.

Meta-analyses often use a *random-effects* model (DerSimonian and Laird, 1986).

The model assumes that each of the multiple estimates pertains to a distinct parameter value drawn at random from a population of potential parameter values.

Thus, a weighted average of the estimates is interpreted to be an estimate of the mean of all potential parameter values.

A Meta-analysis of Outcomes of Bariatric Surgery

Buchwald *et al.* (2008) use a random-effects model to combine the findings of 134 studies of the outcomes of bariatric surgery.

These included 5 randomized trials enrolling 621 patients, 28 nonrandomized but somehow otherwise controlled trials enrolling 4613 patients, and other "uncontrolled case series" with 16860 patients. The studies were performed around the world. They followed patients for different periods of time. They measured weight loss in multiple ways.

To summarize the findings, the authors write:

> "The mean (95% confidence interval) percentage of excess weight loss was 61.2% (58.1%-64.4%) for all patients."

The estimated mean of 61.2% considers the 134 studies to be a sample drawn from a population of potential studies.

The Misleading Rhetoric of Meta-Analysis

The relevance to clinical practice of a weighted average of estimates is often obscure.

DerSimonian and Laird consider each of the trials considered in a meta-analysis to be drawn at random "from a population of possible studies."

They do not explain what is meant by a population of possible studies, nor why the published studies should be considered a random sample from this population.

Even if these concepts are meaningful, they do not explain how a mean outcome across a population of possible studies connects to what should matter to a clinician, namely the distribution of health outcome across the relevant population of patients.

Medical researchers have struggled to explain how clinicians should use the findings.

The Algebraic Wisdom of Crowds

Empirical researchers who study prediction in various fields have reported that the mean of a set of predictions is more accurate than are the individual predictions.

Formally, they report that the prediction error of the mean prediction is smaller than the average prediction error across the individual predictions.

Surowiecki (2004) calls this phenomenon the "wisdom of crowds." A review article by Clemen (1989) put it this way (p. 559):

> "The results have been virtually unanimous: combining multiple forecasts leads to increased forecast accuracy. This has been the result whether the forecasts are judgmental or statistical, econometric or extrapolation. Furthermore, in many cases one can make dramatic performance improvements by simply averaging the forecasts."

Citing this literature, one may be tempted to hope that, however deficient the logic of meta-analysis may be, the methodology may work in practice.

I strongly caution against this.

The wisdom of crowds is not an empirical regularity.

It is an algebraic result that holds when a convex loss function measures prediction error. With such a loss function, the result is a consequence of Jensen's Inequality. (McNees, 1992; Manski, 2011, 2016).

The result holds regardless of whether one combines predictions by their simple mean or a weighted average.

It holds regardless of the quality of the individual predictions that are combined.

Credible Use of Evidence to Inform Patient Care

Evidence-based research can better inform patient care if it seeks to provide knowledge that promotes effective decision making.

Optimal care, as formalized by medical economists, chooses treatments that maximize patient welfare conditional on observed attributes.

From this perspective, studies of treatment response are useful to the degree that they reveal how welfare varies with treatments and attributes.

Methodological research should aim to determine the information that studies provide when evidence is combined with credible assumptions.

Statistical imprecision and identification problems limit the informativeness of studies.

Statistical imprecision stems from small sample size.

Identification problems are the difficulties that persist when sample size grows without bound.

Statistical imprecision may be vexing, but identification is the more fundamental challenge.

Identification of Treatment Response

Unobservability of Counterfactual Treatment Outcomes

Perhaps the most fundamental identification problem in analysis of treatment response stems from the unobservability of counterfactual treatment outcomes.

Observation of a study population can reveal the outcomes that patients realize with the treatments they actually receive. It cannot reveal that outcomes that patients would have experienced with other treatments.

Optimal treatment choice requires comparison of potential outcomes under alternative treatments. Hence, the available evidence is inherently incomplete.

The unobservability of counterfactual outcomes is a matter of universal logic. The logic holds in randomized trials and observational studies alike.

The problem is not resolvable by increasing sample size.

It is a basic aspect of empirical inference that can be mitigated only by making credible assumptions that relate observed and counterfactual outcomes.

Researchers often draw a sharp distinction between observational studies and trials, asserting that unobservability of counterfactual outcomes poses a problem for interpretation of the former but not the latter.

The foundation for this distinction is that a particular assumption is credible in trials with perfect compliance, but may not be otherwise.

Assume that the patients in a large study population are randomly assigned to treatments and that all comply with their assigned treatments. Then the distribution of treatment response in the overall study population is replicated in the sub-population who receive each treatment.

This does not hold precisely in trials of finite size, but statistical imprecision diminishes as sample size increases.

The assumption of random treatment assignment does not render observable the counterfactual outcomes of individual patients.

It provides a credible basis to make probabilistic predictions of outcomes under alternative treatments for patients in the study population.

One observes the realized distribution of outcomes for the patients assigned to a specified treatment and uses this distribution to predict the outcomes that would occur if this treatment were to be assigned to other patients.

The credibility of these predictions underlies the common remark that trials are the "gold standard" for analysis of treatment response.

Now consider observational studies.

It may not be credible to assume random treatment allocation when clinicians purposefully choose the treatments that patients receive.

Nor may it be credible to make another assumption that enables probabilistic prediction of counterfactual outcomes.

Hence, it is common to view the unobservability of counterfactual outcomes as a problem that afflicts observational studies.

Trials with imperfect compliance share with observational studies the difficulty that it may not be credible to assume that treatment allocation is random.

Much research on evidence-based medicine avoids consideration of imperfect compliance by performing intention-to-treat analysis.

Such analysis does not solve the identification problem generated by noncompliance.

It just makes it appear in a different form, namely that of an extrapolation problem.

The trial enables one to predict outcomes when patients are offered treatments in the setting of a trial.

To inform patient care, we want to predict outcomes when patients are offered treatments in clinical practice.

Extrapolation Problems

Extrapolation is an identification problem, not a matter of statistical imprecision. Increasing the sample sizes of trials would increase the precision of findings, but it would not improve the credibility of extrapolations.

Extrapolation from observational studies may also be problematic, but extrapolation from observational studies tends to pose a less severe problem than does extrapolation from trials.

The study populations in trials are volunteers and exclude persons with co-morbidities. Observational studies often collect data from broad study populations that are similar to the populations that clinicians treat.

Drug trials use a double-blind protocol that prevents knowing what treatments subjects receive. Observational studies enable one to learn about patient care as it occurs in practice.

Studying Identification

Although identification problems are ubiquitous in analysis of treatment response, they are given little formal attention in the textbooks and articles that have provided the standard methodological framework for evidence-based medical research.

These sources focus on statistical imprecision and provide at most verbal discussions related to identification.

As a result, medical researchers, guideline developers, and clinicians may have little quantitative sense of how identification problems affect prediction.

Study of identification has been a core concern of research in econometrics from the beginnings of the field in the 1930s through the present.

For about fifty years, the literature focused on identification of systems of linear simultaneous equations, most famously applied to analyze market transactions.

This early research can be used to analyze treatment response with observational data, when treatments vary in the magnitude of their dose.

It is not applicable to comparison of qualitatively different treatments, which has been the prevalent concern of evidence-based medicine.

Hence, it is not surprising that econometricians and medical researchers long found little reason to communicate with one another.

The prospects for beneficial communication have grown substantially from the 1990s onward, as econometricians have increasingly studied identification of response to qualitatively different treatments, using data from observational studies and trials.

The recent econometric research has multiple branches, whose differing objectives affect the features of treatment response that authors want to learn.

Some research aims to learn so-called "causal effects" as a subject of intrinsic interest.

Some research studies treatment response in order to inform future treatment decisions.

The latter objective motivates my research.

When realistic data are combined with credible assumptions, research may yield informative bounds on treatment response but typically does not yield exact findings.

Thus, treatment response is *partially identified* rather than *point-identified*.

Econometric research aims to characterize the identified bounds in a tractable manner.

This done, guideline developers and clinicians should be able to use the findings.

Data from both observational studies and trials may be informative, so analysis of both should be encouraged.

Identification with Missing Data on Patient Outcomes or Attributes

Missing data on patient outcomes and attributes is a frequent occurrence.

Researchers commonly assume that data are missing at random.

This done, researchers often report findings only for sampled patients with complete data. Or they impute missing values and report findings for all patients, acting as if imputed values are actual values.

Assuming that data are missing at random superficially solves the identification problem.

It is rare for researchers to discuss the credibility of the assumption.

Yet there is often reason to worry that the assumption is unrealistic.

I have sought to understand how missing data may affect the conclusions drawn in empirical research.

I have found it useful to begin by first asking what one can learn about treatment response in the absence of any knowledge of the process generating missing data.

Conclusions drawn in this manner are weaker but more credible than those drawn by assuming that data are missing at random or by making some other assumption.

Thus, analysis of inference with missing data illustrates the *Law of Decreasing Credibility*.

Inference without assumptions about the nature of missing data basically is a matter of contemplating all possible configurations of the missing data.

Doing so generates the set of possible conclusions in empirical research, called the *identification region* or the *identified set*.

The practical challenge is to characterize this set of possible conclusions in a tractable way, so applied researchers can use the findings.

Analysis is simple when only outcome data are missing (Manski, 1989, 1990).

The simplest case occurs when the objective is to learn the success probability for a treatment that may fail or succeed. The smallest and largest success probabilities are determined by conjecturing that all missing outcomes are failures or successes.

The same reasoning holds when the objective is to learn the mean or median life span that a treatment yields.

The smallest and largest values are determined by conjecturing that all patients with missing outcomes die immediately or live as long as humanly possible.

Analysis is more complex when some sample members may have missing attribute data. Horowitz and Manski (1998, 2000) study these settings.

Missing Data in a Trial of Treatments for Hypertension

Horowitz and Manski (2000) analyzed identification of treatment response when a trial is performed but some patient outcome or attribute data are missing.

Focusing on cases in which outcomes are binary (success or failure), we derived sharp bounds on success probabilities without imposing any assumptions about the distribution of the missing data.

This analysis contrasts sharply with the conventional practice in medical research of assuming that missing data are missing at random or have some other structure.

We applied the findings to data from a trial comparing treatments for hypertension.

Materson *et al.* (1993) reported on a trial comparing treatments for hypertension sponsored by the U.S. Department of Veteran Affairs (DVA).

Male veteran patients were randomly assigned to one of 6 antihypertensive drug treatments or to placebo.

Treatment was defined to be successful if DBP < 90 mm Hg on two consecutive measurement occasions and DBP ≤ 95 mm Hg later.

The authors performed an intention-to-treat analysis that interpreted attrition from the trial as lack of success. From this perspective there were no missing outcomes.

We obtained the trial data and used them to examine how treatment response varies with an attribute that has missing data.

This is “renin response,” taking the values (low, medium, high), which had previously been studied as a factor that might be related to successful treatment.

Renin-response was measured at the time of randomization, but data were missing for some subjects.

We removed the intention-to-treat interpretation of attrition as lack of success.

Instead, we viewed subjects who leave the trial as having missing outcome data.

Missing Data in the DVA Hypertension Trial

Treatment	Number Randomized	Observed Successes	None Missing	Missing Only y	Missing Only x	Missing (y, x)
1	188	100	173	4	11	0
2	178	106	158	11	9	0
3	188	96	169	6	13	0
4	178	110	159	5	13	1
5	185	130	164	6	14	1
6	188	97	164	12	10	2
7	187	57	178	3	6	0

Bounds on Success Probabilities Conditional on Renin Response

Renin Response	Treatment 1	2	3	4	5	6	7
Low	[0.54, 0.61]	[0.52, 0.62]	[0.43, 0.53]	[0.58, 0.66]	[0.66, 0.76]	[0.54, 0.65]	[0.29, 0.32]
Medium	[0.47, 0.62]	[0.60, 0.74]	[0.53, 0.68]	[0.50, 0.69]	[0.68, 0.85]	[0.41, 0.65]	[0.27, 0.32]
High	[0.28, 0.50]	[0.64, 0.86]	[0.56, 0.75]	[0.63, 0.84]	[0.55, 0.78]	[0.34, 0.59]	[0.28, 0.40]

Missing Data on Family Size When Predicting Genetic Mutations

Partial identification analysis can quantify the potential severity of many missing data problems that occur in research on treatment response and risk assessment.

Consider predictions of the risk that a woman with observed family history carries a gene mutation that make development of breast cancer likely.

These predictions have been used to develop precise probabilistic thresholds for recommendations that women should be referred for genetic testing.

Amir *et al.* (2010) acknowledge difficulties in obtaining data on of family histories:

> "All risk assessment models have limitations: Adoption, small family size . . . ., and lack of information about family history reduce the usefulness of all models to some degree. It is known that because of the reluctance of people to discuss their medical conditions, particularly those involving cancer, generations of family medical history are lost to present-day patients who are receiving care in the era of genetic testing. . . . . There is therefore a need to improve methods for collecting and acknowledging family history even while risk models continue to have their accuracy improved. . . . .

It is not clear how clinicians might use this verbal caution.

Partial identification analysis can quantify the implications of the missing data problem.

Partial Personalized Risk Assessment

An extreme case of missing attribute data occurs when evidence-based predictions condition on a subset of the patient attributes that clinicians observe. Then data on the attributes not used in evidence-based research is entirely missing.

Clinicians have used subjective judgment to predict outcomes conditional on all observed patient attributes, but psychologists have found these judgments fallible.

Manski (2018a) shows that this identification problem can be mitigated if the predictions made by evidence-based studies can be combined with auxiliary data that reveal the distribution of the missing attributes across the patient population.

Predicting Mean Remaining Life Span

A common problem in health risk assessment is to predict remaining life span conditional on observed patient attributes.

Life tables from the Centers for Disease Control provide actuarial predictions of life span in the U. S. conditional on (age, sex, race).

The life tables do not predict life span conditional on other patient attributes that clinicians observe. For example, clinicians readily observe patient blood pressure.

A clinician may want to predict the mean life span for a patient with a certain (age, sex, race, blood pressure), but the life tables alone do not provide an evidence-based prediction.

Manski (2018a) shows that one can bound mean remaining life span conditional on (age, sex, race, blood pressure) if one combines the evidence in the life tables with data on the distribution of blood pressure among persons with the specified (age, race, sex).

Such data exist in the National Health and Nutrition Examination Survey (NHANES).

Consider 50-year-old non-Hispanic (NH) males whose race is either black or white.
I use a standard binary classification of high blood pressure (HBP).

The life tables show that mean remaining life span for 50-year-old NH black and white males are 26.6 and 29.7 years.

The NHANES data show that the prevalence of HBP among 50-year-old NH black and white males are 0.426 and 0.334.

Combining the data yields these sharp bounds:

$$18.1 \leq \text{mean life years for (age 50, NH black male, not HBP)} \leq 35.4,$$

$$14.3 \leq \text{mean life years for (age 50, NH black male, HBP)} \leq 38.5,$$

$$23.8 \leq \text{mean life years for (age 50, NH white male, not HBP)} \leq 36.4,$$

$$15.6 \leq \text{mean life years for (age 50, NH white male, HBP)} \leq 42.0.$$

Tighter bounds can be obtained if the data are combined with assumptions restricting the distribution of life span conditional on the observed attributes.

Strong enough assumptions point-identify mean life span, but these lack credibility.

There is a substantial middle ground between making no assumptions and making assumptions strong enough to yield point identification.

Manski (2018a) reports tighter bounds with *bounded-variation* assumptions.

These restrict the magnitudes of risk assessments and the degree to which they vary with patient attributes, enabling clinicians to express quantitative judgments in a structured way.

Assume that persons with HBP have lower life expectancy than ones without HBP.

Assume that black males have between 0 and 2.5 years lower life expectancy than white males conditional on blood pressure.

Combining these assumptions with the data yields these bounds:

$$29.4 \leq \text{mean life years for (age 50, NH black male, not HBP)} \leq 35.4,$$
$$14.7 \leq \text{mean life years for (age 50, NH black male, HBP)} \leq 22.9,$$
$$31.9 \leq \text{mean life years for (age 50, NH white male, not HBP)} \leq 36.4,$$
$$16.3 \leq \text{mean life years for (age 50, NH white male, HBP)} \leq 25.4.$$

Credible Inference with Observational Data

When cautioning against study of observational data, medical researchers and guideline developers commonly use qualitative rather than quantitative terms.

The Cochran Handbook states (Higgins and Green, 2011): "the extent, and even the direction, of the bias is difficult to predict."

The authors of the JNC 8 guidelines for treatment of hypertension state (James *et al.*, 2014):

> "The panel limited its evidence review to RCTs because they are less subject to bias than other study designs."

These verbal cautions may reflect a perception that it is not possible to characterize quantitatively the severity of the identification problem created by the unobservability of counterfactual outcomes.

They may also reflect a perception that the only way to use observational data to learn about treatment response is to compute point estimates with unknown biases.

In fact, the unobservability of counterfactual outcomes affects inference from observational studies in a simple way.

The problem is simply that counterfactual outcome data are missing.

Bounds with no Knowledge of Counterfactual Outcomes

Manski (1990) begins by asking what one can learn about treatment response in the absence of knowledge of the process generating counterfactual outcomes.

The identification region is obtained by contemplating all logically possible values of the counterfactual outcomes.

The practical findings are bounds on success probabilities and on mean and median outcomes under specified treatments.

These bounds quantify what may be learned from observational data without assumptions.

Consider inference on the success probability for a treatment with a binary outcome or on the mean or median outcome with a continuous outcome.

Lower bounds on these measures of treatment response are determined by conjecturing that all counterfactual outcomes take the smallest values that can possibly occur. Upper bounds are determined analogously.

Consider the average treatment effect (ATE) comparing treatments A and B.

The lower bound on the ATE is the lower bound on the mean outcome with treatment B minus the upper bound on the mean outcome with A.

The upper bound on the ATE is analogous.

The bound on the ATE is particularly simple when the outcome is binary and (A, B) are the only feasible treatments.

Then every patient under study receives one of the two treatments and not the other.

The width of the bound on the ATE necessarily equals 1. This is half the width of the logical bound [-1, 1] that would hold if the data from the observational study were not available.

Thus, an observational study performed with no knowledge of counterfactual outcomes takes one half way toward learning the exact value of the ATE.

Sentencing and Recidivism

Manski and Nagin (1998) analyze sentencing and recidivism of juvenile offenders in the state of Utah.

The structure of this matter is analogous to patient care, with judges in the role of clinicians and juvenile offenders in the role of patients.

The feasible treatments are alternative sentencing options.

The outcome of interest is recidivism; that is, future offending.

Judges in Utah have had the discretion to order various sentences for juvenile offenders. Some offenders have been given sentences with no residential confinement and others have been sentenced to confinement. These are akin to surveillance and aggressive treatment.

A possible alternative policy would be to replace judicial discretion with a mandate that all offenders in Utah be confined. Another would be to mandate that no offenders be confined.

We supposed that the outcome of interest is whether an offender commits a new offense in the two years following sentencing. No new offense indicates that treatment succeeds and commission of a new offense indicates that it fails.

We obtained data on the sentences received and the recidivism outcomes realized by all male offenders in Utah who were born from 1970 through 1974 and who were convicted of offenses before they reached age 16.

11 percent of the offenders were sentenced to confinement. 23 percent of these persons did not offend again in the two years following sentencing.

89 percent were sentenced to non-confinement. 41 percent of these persons did not offend again.

Assuming that judges sentence randomly, the data imply that the success probability for confinement is 0.23 and for non-confinement is 0.41. Hence, ATE = -0.18.

However, one may think it not credible to assume that judges sentence randomly.

We compute bounds that assume no knowledge of counterfactual outcomes.

The bound on the success probability for confinement is [0.03, 0.92], with width 0.89.

The bound on the success probability for non-confinement is [0.36, 0.47], with width 0.11.

Thus, the data reveal much more about recidivism with mandatory non-confinement than with mandatory confinement.

The bound on the ATE is [-0.44, 0.56], with width 1.

One would like to learn more about the ATE than a bound of width 1.

This can be achieved by combining the data with assumptions. We present bounds obtained with a variety of assumptions.

We conjecture two different models of judicial decision making that imply distinct restrictions on counterfactual outcomes.

We also bring to bear assumptions that use *instrumental variables*.

These assumptions posit that the sub-populations of offenders who reside in different judicial districts of Utah respond to sentencing similarly but face different sentencing selection rules.

Identification of Response to Diagnostic Testing and Treatment

A common prelude to treatment is to order a diagnostic test to learn more about a patient.

Suppose that a patient with symptoms presents to a clinician, who initially observes demographic traits and medical history.

The clinician may prescribe a treatment immediately. Or he may order a test that yields further information about the patient and then choose a treatment.

The clinical decision has several aspects. Should the test be ordered? What treatment should be chosen in the absence of the test? What treatment should be chosen when the test is ordered and the result observed?

Optimal Testing and Treatment

Phelps and Mushlin (1988) initiated study of this sequential decision problem using the rational-expectations optimization framework.

The value of ordering a diagnostic test is that doing so reveals a patient attribute that the clinician would not observe otherwise, namely the test result.

The potential usefulness of testing is expressed by the *expected value of information*, defined by Meltzer (2001) as:

> "the change in expected utility with the collection of information."

The expected value of information is necessarily non-negative and is positive if the result affects the optimal treatment.

It follows that a clinician should always order a test if performing the test has no direct negative effect on patient utility.

However, performing a test may negatively affect utility. For example, biopsies, CT scans, and colonoscopies are invasive and expensive. Hence, a test should be performed only if the expected value of information outweighs the direct utility cost.

Phelps and Mushlin assumed that clinicians have the knowledge needed to optimize testing and treatment. They characterized optimal testing and treatment given this knowledge.

Identification of Testing and Treatment Response with Observational Data

One might obtain the knowledge assumed by Phelps and Mushline by performing an ideal randomized trial.

A trial with multiple arms, one for each possible testing and treatment decision, could yield the knowledge of test results and treatment response needed to optimize.

However, performance of this ideal trial is rare.

Often the only available evidence is observational data generated by the testing and treatment decisions that occur in clinical practice. Then it may be unrealistic to suppose that clinicians have the knowledge that Phelps and Mushlin assumed.

Manski (2013b) characterizes the partial knowledge obtained when one has observational data on a study population and uses various assumptions that restrict counterfactual testing results and treatment outcomes.

I suppose that the diagnostic test has two possible results, positive or negative. I suppose that there are two feasible treatments, A being surveillance and B being aggressive treatment.

A common practice is to choose aggressive treatment if and only if a diagnostic test is performed and the result is positive. The chosen treatment is surveillance if the test result is negative or if the patient is not tested.

I call this practice *aggressive treatment with positive testing* (ATPT).

I study identification when the available evidence is observation of a study population that adheres to the ATPT practice.

One can observe test results for patients who are tested. One can observe health outcomes

* under treatment A for patients who are not tested and for patients who are tested and have a negative test result
* under treatment B for patients who are tested and have a positive test result.

One cannot observe test results for patients who are not tested. One cannot observe health outcomes

* under treatment B for patients who are not tested and for patients who are tested and have a negative test result
* under treatment A for patients who are tested and have a positive test result.

The observational evidence yields informative bounds on some of the quantities that determine optimal patient care.

I initially derive bounds without making assumptions that restrict counterfactual testing and treatment outcomes.

I then show what more can be learned if the evidence is combined with several assumptions that may be credible in some settings.

Combining Multiple Predictions

I have cautioned against use of meta-analysis to combine the predictions of disparate studies.

This leaves open how to combine multiple predictions credibly.

Suppose that one obtains N predictions of a patient outcome.

Each prediction uses appropriate data and a plausible prediction model, but the predictions disagree with one another. It may be that one or none is accurate.

There is no logical reason to form an average prediction.

One can only conclude that the predictions pose N possible futures, with others perhaps possible as well.

Combining Multiple Breast Cancer Risk Assessments

Explicit statement of this conclusion is rare, but there are some notable cases.

The Gail Model is the most prominent model that predicts future development of breast cancer in women with specified personal attributes, but it is not the only such model.

The review article of Amir *et al.* (2010) considers four other models: the Claus Model, the BRCAPRO Model, the Jonker Model, the IBIS Model, and the BOADICEA Model.

These models differ from one another in multiple respects, including the patient attributes used to condition predictions, the mathematical forms of the models, and the study data used to estimate model parameters.

As a consequence, they often yield different probabilistic predictions when applied to a woman with specified attributes.

Domchek *et al.* (2003) compare the Gail and Claus Models and find:

> "Concordance of the two models is only fair, with the greatest discrepancies seen with nulliparity, multiple benign breast biopsies, and a strong paternal or first-degree family history."

When multiple models yield disparate probabilistic predictions, uncertainty is inevitable.

A clinician who insists on having a precise probabilistic prediction may want to choose among those that are available.

Amir *et al.* (2010) attempt to assist such a clinician by presenting a flowchart that recommends how a clinician should choose among the five models that the authors compare.

The recommended choice depends on certain patient attributes. The authors write:

> "It is clear that some models are better than others in certain circumstances."

Other writers do not agree with the attempt by Amir *et al.* to provide a guideline for clinicians to choose among the models. In an editorial commenting on the Amir *et al.* article, Gail and Mai (2010) write:

> In our opinion, the flowchart in figure 3 of Amir et al. should only be regarded as a preliminary attempt to synthesize a complex literature and should not be used as a true guide to action. In fact, the lack of independent assessments of calibration of these models . . . . is a serious deficiency in the confirmatory research needed to show that these models yield reliable risk estimates."

Two articles approach combination of predictions in a manner that I think appropriate.

In their review article comparing the Gail and Claus Models, Domchek *et al.* (2003) reject the notion that a clinician must choose one or the other. Instead, they suggest that the two models (p. 600) "may provide helpful ranges" of probabilistic predictions.

Mandelblatt *et al.* (2009) use multiple models to generate a range of predictions of breast cancer development and mortality under alternative strategies for mammography screening. They write:

> "Each model has a different structure and assumptions and some varying input variables, so no single method can be used to validate results against an external gold standard. . . . Overall, using 6 models to project a range of plausible screening outcomes provides implicit cross-validation, with the range of results from the models as a measure of uncertainty."

Reasonable Care under Uncertainty

Clinicians and guideline developers should view patient care as a problem of decision making under uncertainty.

Precedents for this conclusion exist in the literature on medical decision making.

Institute of Medicine (2011) calls attention to the assertion by the Evidence-Based Medicine Working Group that:

> "clinicians must accept uncertainty and the notion that clinical decisions are often made with scant knowledge of their true impact."

Some CPGs use a rating system to rank the strength of recommendations by the certainty that they are correct. The James *et al.* (2014) article summarizing guidelines for treatment of hypertension describes its rating system this way:

A Strong Recommendation: There is high certainty based on evidence that the net benefit is substantial.

B Moderate Recommendation: There is moderate certainty based on evidence that the net benefit is moderate to substantial or there is high certainty that the net benefit is moderate.

C Weak Recommendation: There is at least moderate certainty based on evidence that there is a small net benefit.

Perhaps the most compelling evidence that CPGs recognize uncertainty is that they change their recommendations as new research accumulates.

Verbal recognition of uncertainty has not led guideline developers to examine patient care formally as a problem of decision making under uncertainty.

The IOM (2011) report on guideline development expresses skepticism about decision analysis, stating:

> "A frontier of evidence-based medicine is decision analytic modeling in health care alternatives' assessment. . . . . Although the field is currently fraught with controversy, the committee acknowledges it as exciting and potentially promising, however, decided the state of the art is not ready for direct comment."

I think that formal analysis of patient care under uncertainty has much to contribute to guideline development and to patient care.

Formalizing Uncertainty: States of Nature

The standard formalization of decision under uncertainty supposes that a decision maker must choose among a set of feasible actions.

In patient care, the actions are the feasible alternative treatments.

A decision maker faces uncertainty if the welfare achieved by an action depends on an unknown *state of nature*; a feature of the environment that is incompletely known.

The starting point for decision theory is to suppose that the decision maker lists all the states of nature that he believes could possibly occur.

This list, called the *state space*, expresses partial knowledge.

In patient care, the state of nature may encompass a patient's current health status, the manner in which disease will progress in this patient, and the patient's response to alternative treatments.

When considering patient outcomes, one may find it useful to define states of nature in deterministic personal terms or in probabilistic group terms.

A personal state of nature is a patient-specific outcome. (e.g., whether or not a patient is ill.)

A group state is the distribution of outcomes for patients with specified observed attributes. (e.g., the fraction of patients with specified observed attributes who are ill.)

Optimal and Reasonable Decisions

The fundamental difficulty of decision making under uncertainty is clear even in a simple setting with two feasible actions and two states of nature.

Suppose that one action yields higher welfare in one state of nature and the other action yields higher welfare in the other state.

Then the decision maker does not know which action is better.

Thus, optimization is impossible.

Let the two feasible actions be surveillance and aggressive treatment.

A personal state of nature may be the presence or absence of disease.

It is common for surveillance to yield higher welfare in the absence of disease and for aggressive treatment to be better in the presence of disease.

Then a clinician who does not know whether disease is present cannot determine whether surveillance or aggressive treatment is better.

Decision theory suggests a two-step process when facing uncertainty.

The first step is to eliminate *dominated* treatments; those which are definitely inferior to others.

Formally, an action is dominated if some other treatment is at least as good in all states of nature and superior in some state.

The second step is to choose among the actions that are not dominated.

This is subtle because there is no optimal way to choose among undominated alternatives.

There are only various reasonable ways, each with its own properties.

The term "reasonable" inevitably has multiple interpretations. In a monograph on statistical decision theory, Ferguson (1967) wrote:

> "It is a natural reaction to search for a 'best' decision rule, a rule that has the smallest risk no matter what the true state of nature. Unfortunately, *situations in which a best decision rule exists are rare and uninteresting*. For each fixed state of nature there may be a best action for the statistician to take. However, this best action will differ, in general, for different states of nature, so that no one action can be presumed best over all."

He went on to write: "A *reasonable* rule is one that is better than just guessing."

Once one accepts that there are multiple reasonable ways to make decisions under uncertainty, the admonition of the medical literature that clinicians should adhere to CPGs loses some of its force.

Commentaries often exhort clinicians to adhere to CPGs in order to reduce "unnecessary" or "unwarranted" variation in clinical practice.

This prescription for clinical practice is justified when medical knowledge suffices to distinguish optimal treatments from dominated ones.

It is not justified when clinicians choose among undominated treatments.

Different reasonable criteria for decision making may yield different treatment choices.

Reasonable Decision Criteria

Decision theorists have distinguished three primary situations regarding information that a decision maker may or may not have beyond specification of the state space. They have studied decision criteria suited to each situation.

Decisions with Rational Expectations

The decision maker knows the probabilistic process generating observed outcomes; that is, he has *rational expectations*.

Then it is optimal to give each patient in the group the treatment that maximizes the mean outcome within the group.

Maximization of Subjective Expected Utility

The decision maker does not know the probabilistic process generating outcomes. Instead he introspects and places a subjective probability distribution on outcomes.

Decision theorists call such a decision maker *Bayesian*.

The usual prescription for decision making with a subjective distribution is to maximize subjective expected utility.

The asserted subjective distribution may not accurately describe the actual process generating patient outcomes.

Hence, maximizing subjective expected utility may not optimize patient care.

Decisions under Ambiguity: The Maximin and Minimax-Regret Criteria

The decision maker asserts no knowledge beyond that the true state of nature lies within the specified state space.

Decision theorists refer to this as a situation of *ambiguity* or *deep uncertainty*.

A clinician may view ambiguity in terms of personal or group states of nature.

Consider prediction of life span. The clinician may know that if a specific patient were treated in a certain manner, life span would be between 2 months and 5 years.

Or the clinician may know that if a group of patients with specified attributes were all treated in this manner, mean life span would be between 6 months and 3 years.

Identification problems generate ambiguity, viewed in terms of group states of nature.

One wants to learn the distribution of outcomes for a group of patients.

Combining the available data with credible assumptions, one concludes that the distribution lies in some set of possible distributions, the identification region.

Thus, the identification region for an outcome distribution is its state space.

When making a choice under ambiguity, a reasonable way to act is to use a decision criterion that achieves adequate performance in all states of nature.

The two most commonly studied are the maximin and minimax-regret (MR) criteria.

Maximin Criterion

The maximin criterion chooses an action that maximizes the minimum welfare that might possibly occur across all states of nature.

Thus, a clinician makes a worst-case prediction for the outcome of each treatment and then chooses the treatment with the least-bad worse-case prediction.

Let life span be the outcome of interest.

A clinician using the maximin criterion might choose the treatment that yields the largest value for minimum life span (a personal outcome) or for minimum mean life span (a group outcome).

Minimax-Regret Criterion

Consider the performance of a specified treatment in a given state of nature.

Compute the loss in welfare that would occur if one were to choose this treatment rather than the one that is best in this state. This is *regret*.

The decision maker must choose without knowing the true state. To achieve adequate performance in all states, he computes the maximum regret of each treatment across all states.

The criterion chooses a treatment that minimizes maximum regret; that is, maximum distance from optimality.

Example: Choose between A and B. Life span is the outcome of interest.

Suppose that patient response to A and B is known to vary with the presence or absence of some mutation. It is not known if the patient has the mutation.

A yields life span 5 years with the mutation and 2 months without it.

B yields life span 4 months with the mutation and 3 years without it.

Then A is optimal with the mutation and B is optimal without it.

In the state with the mutation, regret is zero under A and 4.67 years under B.

In the state without the mutation, regret is zero under B and 2.83 years under A.

Maximum regret is 2.83 years for treatment A and 4.67 years for B.

Hence, A is the MR choice.

The maximin and MR criteria are sometimes confused with one another.

Whereas maximin considers only the worst outcome that an action may yield, MR considers the worst outcome relative to what is achievable.

Savage (1951) introduced minimax regret. He distinguished it sharply from maximin. He wrote that the maximin criterion is “ultrapessimistic,” while minimax regret is not.

Maximum regret quantifies how uncertainty—lack of knowledge of the true state of nature—potentially diminishes the quality of decisions.

Reasonable Choice Between Surveillance and Aggressive Treatment

Let aggressive treatment be the better option if the risk of disease exceeds a computable patient-specific threshold and surveillance be better otherwise.

Consider decision making when a clinician does not know a patient's precise risk of disease but can bound it.

Formally, a patient is described by attributes and risk of disease is the fraction of patients with these attributes who would become ill if treated in a specified manner.

Risk of disease is a group state of nature whose identification region is described by a bound.

Knowledge of a bound on the risk of disease suffices to determine that surveillance is optimal if the upper bound is less than the patient-specific threshold.

Similarly, aggressive treatment is optimal if the lower bound on risk of disease is greater than the threshold.

Knowledge of the bound does not determine the optimal treatment if the threshold lies within the bound.

Consider risk of breast cancer.

The NCCN guideline calls for aggressive treatment if a woman's probability of developing the disease in the next five years exceeds 0.017.

A clinician who observes patient attributes beyond those used in the BCRA Tool may not know the patient's risk of cancer conditional on her observed attributes but may be able to bound the risk.

Surveillance is optimal if the upper bound is less than 0.017 and aggressive treatment is optimal if the lower bound is greater than 0.017.

The optimal treatment is indeterminate if the value 0.017 lies within the bound.

When the optimal treatment is indeterminate, treatment choice depends on the criterion used.

A clinician maximizing subjective expected utility forms a subjective mean for the patient's unknown risk of disease and acts as if the true risk equals the subjective mean.

A clinician making a maximin decision acts as if the true risk of disease equals the upper bound on risk.

One making a minimax-regret decision acts as if the true risk of equals the midpoint of the bound on risk.

Reasonable Treatment Choice with Trial Data

Consider use of sample data to inform decision making.

I focus on trial data and consider only statistical imprecision, as has been the case in the statistical literature on trials.

Trial data aim to provide information about the outcome distributions that would occur if groups of patients with specified attributes would to be treated in a specified manner.

Trials of finite size cannot definitively reveal outcome distributions.

Statistical theory characterizes the information that trials provide.

Medical researchers have applied concepts of statistical theory whose foundations are distant from treatment choice.

It has been common to use hypothesis tests to compare treatments.

The Wald (1950) development of statistical decision theory provides a coherent framework for use of trial data to make treatment decisions.

A body of recent research applies statistical decision theory to determine treatment choices that achieve adequate performance in all states of nature, in the sense of maximum regret.

This provide an appealing practical alternative to use of hypothesis tests.

Principles of Statistical Decision Theory

Wald considered the general problem of using sample data to make decisions.

He posed the task as choice of a *statistical decision function*, which maps potential data into a choice among the feasible actions.

He recommended ex ante evaluation of statistical decision functions as procedures, chosen prior to realization of the data, specifying how a decision maker would use whatever data may be realized. Thus, the theory is frequentist.

Expressing the objective as minimization of loss, he proposed evaluation of a statistical decision function by its mean loss across realizations of the sampling process, which he called *risk*.

In the presence of uncertainty, he prescribed a three-step decision process:

(1) Specify the state space.

(2) Eliminate inadmissible statistical decision functions. *Inadmissible* is a synonym for dominated. A decision function is inadmissible if there exists another that yields at least as good mean sampling performance in every state of nature and strictly better mean performance in some state.

(3) Use some reasonable criterion to choose an admissible statistical decision function.

Maximization of subjective expected utility has been studied extensively.

Statistical decision theorists call these Bayes decisions. This focuses attention on the Bayesian process of transforming a *prior* subjective distribution, determined before observing the sample data, into a *posterior* distribution after observing the data.

The Bayesian prescription is sometimes asserted to be antithetical to frequentist statistics, but Wald provided a clear frequentist perspective on Bayes decisions.

He showed that minimization of Bayes risk, a frequentist decision criterion, yields the same decisions as occur if one performs Bayesian inference, combining the prior distribution with the data to form a posterior subjective distribution, and then chooses an action to minimize the posterior mean of expected loss.

Bayesian decision making may be compelling when one feels able to place a credible subjective prior distribution on the state space.

However, Bayesians have long struggled to provide guidance on specification of priors and the matter continues to be controversial.

A spectrum of views regarding Bayesian analysis of randomized trials is expressed by the authors and discussants of Spiegelhalter, Freedman, and Parmar (1994).

The controversy suggests that inability to express a credible prior is common in actual decision settings.

When one finds it difficult to assert a credible subjective distribution, a reasonable way to act is to use a decision criterion that yields adequate performance in all states.

The maximin and minimax-regret criteria provide two ways to formalize this idea.

Wald's version of maximin chooses a decision function that minimizes maximum risk across all states; hence, Wald called it minimax rather than maximin.

One may apply the minimax-regret criterion with sample data.

One measures the regret of a specified statistical decision function in a given state of nature by the difference between the minimum risk achievable in that state and the risk obtained with the specified decision function.

Some History, Post Wald

The Wald framework has breathtaking generality.

In principle, it enables comparison of all statistical decision functions whose risk functions exist.

It enables comparison of alternative sampling processes as well as decision rules. It uses no asymptotic approximations. It applies whatever information the decision maker may have.

One might anticipate that it would play a central role in modern statistics, but this has not occurred. A surge of important extensions and applications followed in the 1950s.

However, this period of rapid development came to a close by the 1960s, with the exception of Bayesian statistical decision theory.

Why did statistical decision theory lose momentum long ago?

One reason may have been the technical difficulty of the subject.

Wald's ideas are easy to describe, but applying them can be analytically and computationally demanding.

Another reason may have been diminishing interest in decision making as the motivation for analysis of sample data.

Modern statisticians tend to view their objectives as estimation and hypothesis testing rather than decision making.

The near absence of the subject in modern journals and textbooks is unfortunate.

Recent Work on Statistical Decision Theory for Treatment Choice

Bayesian statistical decision theory has long been available as a methodology to design trials and to choose treatments with trial data.

However, Bayesians have struggled to provide guidance on specification of priors.

Recent research avoids specification of priors and instead studies treatment choice using maximum regret to measure performance across states.

Contributions include Manski (2004, 2005, 2007a, 2007b), Schlag (2006), Hirano and Porter (2009), Stoye (2009, 2012), Tetenov (2012), Manski and Tetenov (2016), and Kitagawa and Tetenov (2018).

The recent research adopts a public health perspective.

It supposes that the objective of treatment choice is to maximize a welfare function that sums treatment outcomes across a population of patients who may have heterogeneous treatment response.

A statistical decision function uses the data to choose an allocation of patients to treatments.

Such a function has been called a *statistical treatment rule* (*STR*).

The mean sampling performance of an STR across repeated samples is its *expected welfare.*

In Wald's terminology, expected welfare is the negative of risk.

Practical Appeal

Researchers have usually studied treatment choice with trial data on bounded outcomes.

The minimax-regret criterion behaves more reasonably than the maximin criterion.

Numerical methods may be required to precisely minimize maximum regret.

To simplify, researchers have often studied the maximum regret of the *empirical success* (ES) rule, which chooses the treatment with the highest observed trial average outcome.

This simple rule either exactly or approximately minimize maximum regret in common settings with two treatments when sample size is moderate.

In contrast, the maximin rule ignores the trial data, whatever they may be.

Conceptual Appeal

Maximum regret quantifies how lack of knowledge diminishes the quality of decisions.

An STR with small maximum regret is uniformly near-optimal across all states.

Maximum regret is well-defined with multiple treatments and when patients have heterogeneous attributes.

The concept is especially transparent when there are two treatments and the members of the patient population are observationally identical.

Suppose there are two feasible treatments, A and B.

In a state where A is better, the regret of an STR is the product of the probability across repeated samples that the rule commits a Type I error (choosing B) times the magnitude of the loss in expected welfare that occurs when choosing B.

In a state where B is better, regret is the probability of a Type II error (choosing A) times the magnitude of the loss in expected welfare when choosing A.

Thus, regret considers Type I and II error probabilities symmetrically and it measures the magnitudes of the losses that errors produce.

Example

Consider again the example in which a hypothesis test is used as an STR to choose between a status quo treatment for cancer and an innovation.

There are two states: the innovation yields mean life span of 1/3 year or 5 years.

In the first state, the regret of the "test rule" equals 1/30 year; a 0.05 chance of a Type I error times a 2/3 year reduction in mean life span with improper choice of the innovation.

In the second state, the regret of the test rule equals 4/5 year; a 0.20 chance of a Type II error times a 4 year reduction in mean life span with improper choice of the status quo.

Thus, the maximum regret of the test rule is 4/5 year.

One could seek an STR that has smaller maximum regret.

A simple option would be to reverse the conventional probabilities of Type I and Type II errors; that is, use a test with a 0.20 chance of Type I error and 0.05 chance of Type II error.

In the first state, the regret of this STR equals 2/15 year; a 0.20 chance of Type I error times a 2/3 year reduction in mean life span with improper choice of the innovation.

In the second state, the regret of rule equals 1/5 year; a 0.05 chance of Type II error times a 4 year reduction in mean life span with improper choice of the status quo.

Thus, the maximum regret of the unconventional test rule is 1/5 of a year.

There may exist other STRs that perform even better.

Designing Trials to Enable Near-Optimal Treatment Choice

The above research concerns use of existing trial data to make treatment choices.

A complementary question asks how trials should be designed to inform treatment choice.

The convention has been to choose sample sizes that yield specified statistical power.

A better idea would be to choose sample size to enable near-optimal treatment choice.

Using Power Calculations to Choose Sample Size

The use of power calculations to set sample sizes derives from the presumption that trial data will be used to test a hypothesis.

The convention has been to choose a sample size that yield specified power at a value of the effect size deemed clinically important.

Trials with samples too small to achieve conventional error probabilities are called "underpowered" and are criticized as scientifically useless and medically unethical.

Sample Size Enabling Near-Optimal Treatment Choice

An ideal objective for the design of trials would be to collect data that enable subsequent implementation of an optimal treatment rule in the patient population.

Optimality is not achievable with finite sample size, but near-optimal rules—ones with small maximum regret—exist when trials are large enough.

Manski and Tetenov (2016) trial design that enables near-optimal treatment choices.

We show that, given any $\varepsilon > 0$, *ε-optimal* rules exist when trials have large enough sample size.

An ε-optimal rule has expected welfare, across repeated samples, within ε of the welfare of the best treatment in every state. Equivalently, it has maximum regret no larger than ε.

We give simple sufficient conditions on sample sizes that ensure existence of ε-optimal treatment rules when there are multiple treatments and outcomes are bounded.

We report exact results for the case of two treatments and binary outcomes.

Use of near-optimality to set sample sizes requires specification of a value for ε.

We suggest that a possible way to specify ε is to relate it to the *minimum clinically important difference* (MCID) in the average treatment effect comparing alternative treatments.

Many writers call an average treatment effect clinically significant if its magnitude is greater than a specified value deemed minimally consequential in clinical practice.

Research articles reporting trial findings sometimes pose particular values of MCIDs when comparing alternative treatments for specific diseases.

Findings with Binary Outcomes, Two Treatments, and Balanced Designs

Determination of sample sizes that enable near-optimal treatment is simple in settings with binary outcomes, two treatments, and a balanced design which assigns the same number of subjects to each treatment group.

We provide exact computations of the minimum sample size that enables ε-optimality with three treatment rules, for various values of ε.

ε	ES Rule	One-Sided 5% z-Test	One-Sided 1% z-Test
0.01	145	3488	7963
0.03	17	382	879
0.05	6	138	310
0.10	2	33	79
0.15	1	16	35

Reconsidering Sample Size

Our findings yield a broad conclusion that sample sizes determined by clinically relevant near-optimality criteria are much smaller than ones set with conventional power calculations.

Reduction of total sample size can lower the cost of executing trials, the time necessary to recruit subjects, and the need to perform trials across multiple centers.

Reduction of sample size per treatment arm can make it feasible to perform trials that increase the number of arms and thus yield information about more treatment options.

A Public Health Perspective on Reasonable Care

Treatment Diversification

A useful perspective when considering public health is to hypothesize a planner who treats a population of patients.

The planner's objective is to maximize utilitarian social welfare across the relevant patients, adding up their individual outcomes.

A planner with rational expectations would make the same decisions as would clinicians optimizing treatment of individual patients.

The public health perspective differs from that of the clinician under uncertainty.

Suppose that the feasible treatments are A and B.

A clinician treating an individual patient can only choose one treatment or the other.

A planner can diversify treatment of observationally similar patients, allocating some fraction to A and the remainder to B.

Treatment diversification is useful because it enables a planner to avoid gross errors that would occur if all patients were given an inferior treatment.

Treatment diversification is analogous to financial diversification.

A portfolio is diversified if an investor allocates positive fractions of wealth to different investments.

Diversification enables an investor facing uncertain asset returns to limit the potential negative consequences of placing 'all eggs in one basket.'

The finance literature shows that an investor seeking to maximize subjective expected utility diversifies if utility is a sufficiently concave function of the investment return and the probability distribution of returns has sufficient spread.

Treatment diversification by a health planner can be studied in the same manner.

Manski (2007a, 2009) takes a different approach, studying treatment allocation using the minimax-regret criterion.

The central result is that when there are two undominated treatments, a planner using the minimax-regret criterion always chooses to diversify.

The fraction of patients assigned to each treatment depends on the available knowledge of treatment response.

Example: Treating X-Pox

Suppose that a new disease called x-pox is sweeping a community. It is impossible to avoid infection. If untreated, infected persons always die.

Researchers propose two treatments, say A and B. The researchers know that one treatment is effective, but they do not know which one. They know that administering both treatments in combination is fatal.

There is no time to experiment to learn which treatment is effective. Everyone must be treated right away.

A public health agency must decide how to treat the community. The agency wants to maximize the survival rate of the population.

It can select one treatment and administer it to everyone. Then the entire population will either live or die.

It can give one treatment to some fraction of the community and the other treatment to the remaining fraction. Then the survival rate will be one of the two chosen fractions.

If half the population receives each treatment, the survival rate is fifty percent.

What might the agency reasonably do?

There are two states of nature. Treatment A is effective in one state and B is effective in the other.

All treatment allocations are undominated.

The planner might use the expected welfare, maximin, or minimax regret criterion to choose a treatment allocation.

A planner who places subjective probabilities on the states of nature and evaluates a treatment allocation by its subjective expected welfare would assign everyone to the treatment with the higher subjective probability of being effective.

The maximin and the minimax-regret criteria both prescribe that the planner should assign half the population to each treatment.

Although the maximin and minimax-regret criteria deliver the same treatment allocation in this example, the two criteria are not the same in general.

To see this, amend the description of the x-pox problem by adding a third state of nature, say u, in which neither treatment is effective.

Adding this third state does not affect the choices made by a planner who maximizes expected welfare or minimizes maximum regret.

All treatment allocations solve the maximin problem. The reason is that there now exists a possible state on nature in which everyone dies, regardless of treatment.

Adaptive Diversification

Consider a planner who makes treatment decisions in a sequence of periods, facing a new group of patients each period.

The planner may observe the outcomes of early decisions and use this evidence to inform treatment later on.

Diversification is advantageous for learning treatment response because it generates randomized experiments.

As evidence accumulates, the planner can revise the fraction of patients assigned to each treatment in accord with the available knowledge.

I have called this *adaptive diversification*.

A simple approach to multi-period treatment choice is to use the *adaptive minimax-regret* (*AMR*) criterion.

In each period, this criterion applies the static MR criterion using the information available at the time.

It is adaptive because successive cohorts may receive different allocations as knowledge of treatment response increases over time.

The AMR criterion treats each cohort as well as possible, in the MR sense, given the available knowledge.

It does not ask the members of one cohort to sacrifice for future cohorts.

Nevertheless, diversification enables learning about treatment response.

The fractional allocations produced by the AMR criterion are randomized experiments, so one may ask how AMR differs from conventional trials.

The AMR criterion randomizes treatment of all observationally similar patients.

The treatment groups in trials are typically small fractions of the patient population.

Trials draw subjects from pools of persons who volunteer to participate and who meet specific conditions, such as the absence of co-morbidities.

The Practicality of Adaptive Diversification

Implementation in Centralized Health-Care Systems

Adaptive diversification may be feasible in centralized health-care systems where a planning entity chooses treatments for a broad patient population.

Examples: the Military Health System in the United States, the National Health Service in the United Kingdom, private health maintenance organizations.

An open question is whether the relevant planners and patient populations would accept the idea.

A possible objection is that diversification violates a version of the ethical principle calling for "equal treatment of equals."

Fractional allocations are consistent with this principle in the *ex ante* sense that observationally identical people have the same probability of receiving a particular treatment.

They violate it in the *ex post* sense that observationally identical persons ultimately receive different treatments.

Fractional allocations are consistent with prevailing standards of medical ethics, which permit randomized trials under conditions of clinical equipoise.

Should Guidelines Encourage Treatment Variation under Uncertainty?

Suppose that controlled adaptive diversification is not feasible.

We can nevertheless question whether the medical community should continue to discourage treatment variation.

CPGs could encourage clinicians to recognize that treatment choice may reasonably depend on how one interprets the available evidence and on the decision criterion that one uses.

The result could then be natural treatment variation that yields some of the benefits of adaptive diversification.

Managing Uncertainty in Drug Approval

The approval process of the Food and Drug Administration determines whether a drug can legally be sold within the United States.

To obtain approval, a pharmaceutical firm must provide information on treatment response through performance of randomized trials that compare the new drug with an existing treatment or a placebo.

The FDA makes a (yes/no) approval decision after reviewing the findings.

Type I and II Errors in Drug Approval

FDA evaluation of new drugs occurs with partial knowledge of treatment response.

Approval decisions are susceptible to two types of errors.

Type I errors occur when new drugs that are inferior to accepted treatments are approved because they appear superior when evaluated using available information.

Type II errors occur when new drugs that are superior to accepted treatments are disapproved because they appear inferior when evaluated using available information.

Type II errors commonly are permanent. After a drug is disapproved, use ceases and no further data on treatment response are produced.

Some Type I errors eventually are corrected through the FDA's post-market surveillance program, which analyzes data on the outcomes experienced when the drug is used in clinical practice.

The post-market surveillance program only aims to detect adverse side effects of approved drugs, not to assess their effectiveness in treating the conditions for which they are intended.

Errors due to Statistical Imprecision and Wishful Extrapolation

A well-recognized potential source of errors in drug approval is the statistical imprecision of empirical findings from trials with finite samples of subjects.

The FDA limits the frequency of statistical errors by requiring that trial sizes suffice to perform hypothesis tests with specified power.

There is much reason to question the use of power calculations to choose trial sizes.

Nevertheless, power calculations ensure that trial sizes are large enough to make statistical error a relatively minor concern.

The dominant determinants of errors in drug approval are extrapolation problems.

The approval process essentially assumes that treatment response in the relevant patient population will be similar to response in the study population.

It assumes that response in clinical practice will be similar to response with double-blinded treatment assignment.

It assumes that drug effectiveness measured by outcomes of interest will be similar to effectiveness measured by surrogate outcomes.

These assumptions are unsubstantiated, but they have become enshrined by long use.

FDA Rejection of Formal Decision Analysis

In 2012, Congress responded to controversy about the drug approval process by requiring the FDA to implement a "structured risk-benefit assessment framework."

The Food and Drug Administration Safety and Innovation Act of 2012 (Public Law 112-144) amended the Federal Food Drug and Cosmetic Act by requiring the FDA to

> "implement a structured risk-benefit assessment framework in the new drug approval process to facilitate the balanced consideration of benefits and risks, a consistent and systematic approach to the discussion and regulatory decision-making, and the communication of the benefits and risks of new drugs."

After enactment of the legislation, the FDA published a plan intended to fulfill the Congressional requirement.

The agency responded skeptically to critics of the prevailing approval process who have recommended that the FDA use formal decision analysis.

The agency described and defended its preferred approach to drug approval.

It rejected quantitative/formal decision analysis and argued for continuation of the "structured qualitative" approach that it has used in the past. The agency stated:

"In the last few years, as other disciplines such as decision science and health economics have been applied to drug regulatory decision-making, there has been much discussion among regulators, industry, and other stakeholders regarding “qualitative” versus “quantitative” approaches to benefit-risk assessment.

Some hold the view that a quantitative benefit-risk assessment encompasses approaches that seek to quantify benefits and risks, as well as the weight that is placed on each of the components such that the entire benefit-risk assessment is quantitative. This approach is typical of quantitative decision modeling.

It usually requires assigning numerical weights to benefit and risk considerations in a process involving numerous judgments that are at best debatable and at worst arbitrary. The subjective judgments and assumptions that would inevitably be embodied in such quantitative decision modeling would be much less transparent, if not obscured, to those who wish to understand a regulator's thinking.

Furthermore, application of quantitative decision modeling seems most appropriate for decisions that are largely binary. Many benefit-risk assessments are more nuanced and conditional based on parameters that could be used to effectively manage a safety concern in the post-market setting. There is significant concern that reliance on a relatively complex model would obscure rather than elucidate a regulator's thinking.

These concerns have led FDA to the conclusion that the best presentation of benefit-risk considerations

involves focusing on the individual benefits and risks, their frequency, and weighing them appropriately. FDA believes that this can be accomplished by a qualitative descriptive approach for structuring the benefit-risk assessment that satisfies the principles outlined earlier in this section, while acknowledging that quantification of certain components of the benefit-risk assessment is an important part of the process to support decision-making.

FDA considers it most important to be clear about what was considered in the decision, to be as quantitative as possible in characterizing that information, and to fully describe how that information was weighed in arriving at a conclusion.

Quantitative assessments certainly underpin the qualitative judgments of FDA's regulatory decisions, but FDA has adopted a structured qualitative approach that is designed to support the identification and communication of the key considerations in FDA's benefit-risk assessment and how that information led to the regulatory decision."

Adaptive Partial Drug Approval

Adaptive diversification of drug treatment after FDA approval of new drugs could reduce the impact of Type I errors in drug approval, diversifying treatment choice and producing new post-market information on treatment response.

Adaptive diversification after drug approval would not reduce Type II errors.

Clinicians can only choose among the drugs that are approved by the FDA.

To reduce both Type I and II errors, the FDA could replace its present approval process with one of adaptive partial drug approval.

The permitted use of a new drug now has a sharp discontinuity at the date of the FDA approval decision.

Beforehand, a small fraction of the patient population receives the new drug in trials.

Afterwards, use is unconstrained if approval is granted and zero if approval is not granted.

An adaptive approval process would eliminate this discontinuity and instead permit use of a new drug to vary smoothly as evidence accumulates.

I proposed such an approval process in Manski (2009). Others have made similar proposals, without an explicit decision-theoretic motivation. See Eichler *et al.* (2012).

Adaptive Limited-Term Sales Licenses

A version of adaptive approval would use limited-term sales licenses.

The approval process would begin, as at present, with a pharmaceutical firm performing preclinical testing followed by Phase 1 and 2 trials. The changes would appear in the subsequent Phase 3 trials and in the FDA decision process.

The duration of Phase 3 trials would be lengthened sufficiently to measure health outcomes of real interest, not just surrogate outcomes.

The present binary approval decision following Phase 3 would be replaced by an adaptive process that monitors the trial while in progress and that periodically grants limited-term sales licenses.

An adaptive limited-term sales license would permit a firm to sell no more than a specified quantity of the new drug over a specified time period.

Subject to this bound, treatment allocation would be determined by the decentralized pricing decisions of the pharmaceutical firm, coverage decisions of insurers, and treatment decisions of physicians and patients.

On each iteration of the decision, the maximum quantity of drug that the firm is permitted to sell would be set by the FDA with the assistance of an expert advisory board, similar to those now used in drug approval.

A possibility would be to use the AMR criterion.

When the lengthened Phase 3 trial is complete and the outcomes of health interest have been observed, the FDA would make a longer-term approval decision.

If the drug is deemed safe and effective, the firm would be permitted to sell it with no quantity restriction. Further use would be prohibited otherwise.

The FDA would retain the right to rescind approval should new evidence warrant.

Post-market surveillance would be necessary because lengthening Phase 3 trials to measure health outcomes of interest may not suffice to determine with certainty whether the innovation is superior to the status quo.

Open Questions

An adaptive drug approval process using limited-term sales licenses is one way that the FDA might improve the present approval process.

Alternatively, the agency could empower some health-care providers to prescribe the drug prior to final approval or could restrict treatment to patients with specified characteristics.

An open question is how pharmaceutical firms respond to the incentives that any approval process puts in place.

We don't know much about how the approval process affects the decisions of firms to perform basic research, initiate trials, and submit New Drug Applications.

Conclusion

I initially asked: Should clinicians adhere to guidelines or exercise judgment?

Guidelines call for uniform treatment of observationally similar patients. The argument weakens when clinicians choose care under uncertainty.

1. Guidelines issued by different health organizations may disagree with one another. Then clinicians have to use judgment to determine which guideline to follow.

2. Clinicians observe patient attributes that are not considered in guidelines. Then clinicians can personalize care to a greater degree than is possible with guidelines.

3. Predictions of treatment response used in evidence-based guideline development rest on questionable methodological practices.

Thus, clinicians may reasonably interpret available evidence in different ways and may reasonably use different decision criteria to choose treatments.

The rationale for treatment variation strengthens when one considers patient care as a public health problem.

Then adaptive diversification of treatment can be valuable.

Separating the Information and Recommendation Aspects of Guidelines

Manski (2013a) proposed separating two tasks of guideline development that have commonly been performed in conjunction with one another.

One is to characterize medical knowledge. The other is to make care recommendations.

I continue to think that these two tasks should be separated.

Having guideline development groups characterize knowledge can improve clinical practice.

Medical research is vast, it continues to grows rapidly, and it requires expertise to interpret.

It is not feasible for individual clinicians to keep up on their own.

Synthesis by expert panels seems essential.

The problem is that current approaches to synthesis of medical research are less informative than they should be.

A huge deficiency is over-attention to internal validity and neglect of external validity.

This asymmetry has manifested itself in quantification of statistical imprecision without quantification of identification problems.

Guideline panels recognize uncertainty only qualitatively and shun use of decision theory when considering how to cope with uncertainty.

FDA drug approval process manifests the same deficiencies.

I question when guidelines should make recommendations for patient care under uncertainty.

Making recommendations asks guideline developers to aggregate the benefits and harms of care into a scalar measure of welfare.

It requires them to specify a decision criterion to cope with partial knowledge.

Care recommendations may be contentious if perspectives vary.

Moreover, having all clinicians adhere to the same guidelines may be sub-optimal from a public health perspective.

It does not recognize the attraction of diversification as a means of avoiding gross errors in treatment. It does not exploit the opportunity for learning that diversification provides.

Educating Clinicians in Care under Uncertainty

An alternative to having guidelines make care recommendations would be to enhance the ability of clinicians to make reasonable patient-care decisions under uncertainty.

It would be useful to introduce medical students to core concepts of uncertainty quantification and decision analysis as part of their basic education.

An important part of the solution will be to bring specialists in risk assessment and decision analysis into the clinical team who jointly contribute to patient care.

To instruct basic medical students and develop specialists in patient care under uncertainty will require that medical schools create and implement appropriate curricula.

References

Academy of Medical Sciences (2015), *Stratified, Personalised or P4 Medicine: a New Direction for Placing the Patient at the Centre of Healthcare and Health Education*, www.acmedsci.ac.uk/viewFile/56e6d483e1d21.pdf, accessed July 4, 2016.

Agency for Healthcare Research and Quality (2017), https://www.guideline.gov/, accessed August 18, 2017.

AIM at Melanoma Foundation (2018), "Stage I Melanoma," https://www.aimatmelanoma.org/stages-of-melanoma/stage-i-melanoma/, accessed January 27, 2018.

Altman, D. (1980), "Statistics and Ethics in Medical Research: III How Large a Sample?" *BMJ*, 281, 1336-1338.

Amir, E., O. Freedman, B. Seruga, and D. Evans (2010), "Assessing Women at High Risk of Breast Cancer: A Review of Risk Assessment Models," *Journal of the National Cancer Institute*, 102, 680-691.

American College of Cardiology (2017), ASCVD Risk Estimator Plus, http://tools.acc.org/ASCVD-Risk-Estimator-Plus/#!/calculate/estimate/, accessed October 14, 2017.

Basu, A. and D. Meltzer (2007), "Value of information on preference heterogeneity and individualized care," *Medical Decision Making*, 27, 112-27.

Beckett, N. R. Peters, A. Fletcher, *et al.* (2008), "HYVET Study Group. Treatment of Hypertension in Patients 80 Years of Age or Older," *New England Journal of Medicine*, 358, 1887-1898.

Berger, J. (1985), *Statistical Decision Theory and Bayesian Analysis*, Second Edition, Springer: New York.

Burchwald, H., Y. Avidor, E. Braunwald, M. Jensen, W. Pories, K. Fahrbach, and K. Schoelles (2008), "Bariatric Surgery: A Systematic Review and Meta-analysis," *Journal of the American Medical Association*, 292, 1724-1737.

Camerer, C. and E. Johnson (1997), "The Process-Performance Paradox in Expert Judgment: How Can Experts Know so Much and Predict so Badly," in *Research on Judgment and Decision Making*, W. Goldstein and R. Hogarth (editors), Cambridge: Cambridge University Press.

Campbell, D. (1984), "Can We Be Scientific in Applied Social Science?," *Evaluation Studies Review Annual*, 9, 26–48.

Campbell, D. and J. Stanley (1963), *Experimental and Quasi-Experimental Designs for Research*, Chicago: Rand McNally.

Canner, P. (1970), "Selecting One of Two Treatments When the Responses Are Dichotomous," *Journal of the American Statistical Association*, 65, 293-306.

Caulley, L., C. Balch, M. Ross, and C. Robert (2018), "Management of Sentinel-Node Metastasis in Melanoma," *New England Journal of Medicine*, 378, 85-88.

Chen, S. and G. Parmigiani (2007), "Meta-Analysis of *BRCA1* and *BRCA2* Penetrance," *Journal of Clinical Oncology*, 25, 1329-1333.

Cheng, Y., F. Su, and D. Berry (2003), "Choosing Sample Size for a Clinical Trial Using Decision Analysis," *Biometrika*, 90, 923-936.

Claus, E, N. Risch, and W. Thompson (1994), "Autosomal Dominant Inheritance of Early-onset Breast Cancer. Implications for Risk Prediction," *Cancer*, 73, 643-651.

Clemen, R. (1989), "Combining Forecasts: A Review and Annotated Bibliography," *International Journal of Forecasting,* 5, 559-583.

Crits-Christoph, P., L. Siqueland, J. Blaine, A. Frank, L. Luborsky, L. Onken, L. Muenz, M. Thase, R. Weiss, D. Gastfriend, G. Woody, J. Barber, S. Butler, D. Daley, I. Salloum, S. Bishop, L. Najavits, J. Lis, D. Mercer, M. Griffin, K. Moras, and A. Beck, (1999), "Psychosocial Treatments for Cocaine Dependence," *Archives of General Psychiatry*, 56, 493–502.

Davis, C., H. Nasi, E. Gurpinar, E. Poplavska, A. Pinto, and A. Aggarwal (2017), "Availability of Evidence of Benefits on Overall Survival and Quality of Life of Cancer Drugs Approved by European Medicines Agency: Retrospective Cohort Study of Drug Approvals 2009-13," *BMJ*, 359, doi:10.1136/bmj.j4530.

Dawes, R., R. Faust, and P. Meehl (1989), "Clinical Versus Actuarial Judgment," *Science*, 243, 1668-1674.

DeGroot, M. (1970), *Optimal Statistical Decisions*, New York: McGraw-Hill.

DerSimonian, R. and N. Laird (1986), "Meta-Analysis in Clinical Trials," *Controlled Clinical Trials*, 7, 177-188.

Domchek, S., A. Eisen, K. Calzone, J. Stopfer, A. Blackwood, and B. Weber (2003), "Application of Breast Cancer Risk Prediction Models in Clinical Practice," *Journal of Clinical Oncology,* 21, 593-601.

Duncan, O. and B. Davis (1953), "An Alternative to Ecological Correlation," *American Sociological Review*, 18, 665-666.

Eggermont A. and R. Dummer (2017), "The 2017 Complete Overhaul of Adjuvant Therapies for High-risk Melanoma and its Consequences for Staging and Management of Melanoma Patients," *European Journal of Cancer*, 86, 101-105.

Eichler, H. Oye, L. Baird, E. Abadie, J. Brown, C. Drum, J. Ferguson, S. Garner, P. Honig, M. Hukkelhoven, J. Lim, R. Lim, M. Lumpkin, G. Neil, B. O'Rourke, E.Pezalla, D.Shoda, V. Seyfert-Margolis, E. Sigal, J. Sobotka, D. Tan, T. Unger, and G. Hirsch (2012), "Adaptive Licensing: Taking the Next Step in the Evolution of Drug Approval," *Clinical Pharmacology & Therapeutics*, 91, 426-437.

Faries, M. (2018), "Completing the Dissection in Melanoma: Increasing Decision Precision," *Annals of Surgical Oncology*, https://doi.org/10.1245/s10434-017-6330-4.

Faries, M., J. Thompson, A. Cochran, R. Andtbacka, N. Mozzillo, J. Zager, T. Jahkola, T. Bowles, A. Testori, P. Beitsch, H. Hoekstra, M. Moncrieff, C. Ingvar, M. Wouters, M. Sabel, E. Levine, D. Agnese, M. Henderson, R. Dummer, C. Rossi, R. Neves, S. Trocha, F. Wright, D. Byrd, M. Matter, E. Hsueh, A. MacKenzie-Ross, D. Johnson, P. Terheyden, A. Berger, T. Huston, J. Wayne, B. Smithers, H. Neuman, S. Schneebaum, J. Gershenwald, C. Ariyan, D. Desai, L. Jacobs, K. McMasters, A. Gesierich, P. Hersey, S. Bines, J. Kane, R. Barth, G. McKinnon, J. Farma, E. Schultz, S. Vidal-Sicart, R. Hoefer, J. Lewis, R. Scheri, M. Kelley, O. Nieweg, R. Noyes, D. Hoon, H. Wang, D. Elashoff, and R. Elashoff (2017), "Completion Dissection or Observation for Sentinel-Node Metastasis in Melanoma," *New England Journal of Medicine*, 376, 2211-222.

Ferguson, T. (1967), *Mathematical Statistics: A Decision Theoretic Approach*, Academic Press:San Diego.

Fisher, L. and L. Moyé (1999), "Carvedilol and the Food and Drug Administration Approval Process: An Introduction," *Controlled Clinical Trials*, 20, 1-15.

Fisher, R. (1935), *The Design of Experiments.* London: Oliver and Boyd.

Fleming, T. and D. Demets (1996), "Surrogate End Points in Clinical Trials: Are We Being Misled?" *Annals of Internal Medicine,* 125, 605-613.

Freis, E., B. Materson, and W. Flamenbaum (1983), "Comparison of Propranolol or Hydorchlorothiazide Alone for Treatment of Hypertension, III: Evaluation of the Renin-Angiotensin System," *The American Journal of Medicine*, 74, 1029-1041.

Gail, M, L. Brinton, D. Byar, D. Corle, S. Green, C. Shairer, and J. Mulvihill (1989), "Projecting Individualized Probabilities of Developing Breast Cancer for White Females Who Are Being Examined Annually," *Journal of the National Cancer Institute*, 81,1879-86.

Gail, M. and P. Mai (2010), "Comparing Breast Cancer Risk Assessment Models," *Journal of the National Cancer Institute*, 102, 665-668.

Gassenmaier, M., T. Eigentler, U. Keim, M. Goebeler, E. Fiedler, G. Schuler, U. Leiter, B. Weide, E. Grischke, P. Martus, and C. Garbe (2017), "Serial or Parallel Metastasis of Cutaneous Melanoma? A Study of the German Central Malignant Melanoma Registry," *Journal of Investigative Dermatology,* 137, 2570-2577.

Ginsburg, G. and H. Willard (2009), "Genomic and Personalized Medicine: Foundations and Applications," *Translational Research*, 154, 277-287.

Go, A., D. Mozaffarian, V. Roger, E. Benjamin, J. Berry, W. Borden, D. Bravata, S. Dai, E. Ford, C. Fox, S. Franco, H. Fullerton, C. Gillespie, S. Hailpern, J. Heit, V. Howard, M. Huffman, B. Kissela, S. Kittner, D. Lackland, J. Lichtman, L. Lisabeth, D. Magid, G. Marcus, A. Marelli, D. Matchar, D. McGuire, E. Mohler, C. Moy, M. Mussolino, G. Nichol, N. Paynter, P. Schreiner, P. Sorlie, J. Stein, T. Turan, S. Virani, N. Wong, D. Woo, and M. Turner; on behalf of the American Heart Association Statistics Committee and Stroke Statistics Subcommittee (2013), "Heart disease and stroke statistics—2013 update: a report from the American Heart Association," *Circulation*, 127, e6-e245.

Goldberg, L. (1968), "Simple Models or Simple Processes? Some Research on Clinical Judgments," *American Psychologist*, 23, 483-496.

Good, I. (1967), "On the Principle of Total Evidence," *The British Journal for the Philosophy of Science*, 17, 319-321.

Groves, W., D. Zald, B. Lebow, B. Snitz, and C. Nelson (2000)," Clinical Versus Mechanical Prediction: A Meta-Analysis," *Psychological Assessment*, 12, 19-30.

Halpern, S., J. Karlawish, and J. Berlin (2002), "The Continued Unethical Conduct of Underpowered Clinical Trials," *Journal of the American Medical Association*, 288, 358-362.

Higgins J. and S. Green (editors) (2011), *Cochrane Handbook for Systematic Reviews of Interventions*, Version 5.1.0, The Cochrane Collaboration, http://handbook-5-1.cochrane.org/, accessed August 31, 2017.

Hodges, E. and E. Lehmann (1950), "Some Problems in Minimax Point Estimation," *Annals of Mathematical Statistics*, 21, 182-197.

Horowitz, J. and C. Manski (1995), “Identification and Robustness with Contaminated and Corrupted Data,” *Econometrica*, 63, 281-302.

Horowitz, J., and C. Manski (2000), “Nonparametric Analysis of Randomized Experiments with Missing Attribute and Outcome Data,” *Journal of the American Statistical Association*, 95, 77-84.

Hoyt, D. (1997), "Clinical Practice Guidelines," *American Journal of Surgery*, 173, 32-34.

Institute of Medicine (2011), *Clinical Practice Guidelines We Can Trust*, Washington, DC: National Academies Press.

Institute of Medicine (2013), *Variation in Health Care Spending: Target Decision Making, Not Geography*, Washington, DC: The National Academies Press.

International Conference on Harmonisation (1999) ICH E9 Expert Working Group. Statistical principles for clinical trials: ICH harmonized tripartite guideline. *Statistics in Medicine,* 18, 1905-1942.

Ioannidis, J. (2005), "Why Most Published Research Findings are False," *PLoS Medicine,* 2, e124.

James, P, S. Oparil, B. Carter, W. Cushman, C. Dennison-Himmelfarb, J. Handler, D. Lackland, M. LeFevre, T. MacKenzie, O. Ogedegbe, S. Smith Jr, L. Svetkey, S. Taler, R. Townsend, J.Wright Jr, A. Narva, and E. Ortiz (2014), "Evidence-Based Guideline for the Management of High Blood Pressure in Adults Report From the Panel Members Appointed to the Eighth Joint National Committee (JNC 8)," *Journal of the American Medical Association,* 311, 507-520.

Kasumova, G., A. Haynes, and G. Boland (2017), "Lymphatic versus Hematogenous Melanoma Metastases: Support for Biological Heterogeneity without Clear Clinical Application," *Journal of Investigative Dermatology,* 137, 2466-2468.

Karmali, K., D. Goff, H. Ning, and D. Lloyd-Jones (2014), "A Systematic Examination of the 2013 ACC/AHA Pooled Cohort Risk Assessment Tool for Atherosclerotic Cardiovascular Disease," *Journal of the American College of Cardiology,* 64, 959-968.

Kitagawa, T. and A. Tetenov (2018), "Who Should be Treated? Empirical Welfare Maximization Methods for Treatment Choice," *Econometrica*, forthcoming.

Leiter, U., R. Stadler, C. Mauch, W. Hohenberger, N. Brockmeyer, C. Berking, C. Sunderkötter, M. Kaatz, K. Schulte, P. Lehmann, T. Vogt, J. Ulrich, R. Herbst, W. Gehring, J. Simon, U. Keim, P. Martus, and C. Garbe (2016), "Complete Lymph Node Dissection Versus No Dissection in Patients with Sentinel Lymph Node Biopsy Positive Melanoma (DeCOG-SLT): a Multicentre, Randomised, Phase 3 Trial," *The Lancet Oncology*, 17, 757-767.

Mandelblatt, J., K. Cronin, S. Bailey, D. Berry, H. de Koning, G. Draisma, H. Huang, S. Lee, M. Munsell, S. Plevritis, P. Ravdin, C. Schechter, B. Sigal, M. Stoto, N. Stout, N. van Ravesteyn, J. Venier, M. Zelen, and E. Feuer (2009), "Effects of Mammography Screening Under Different Screening Schedules: Model Estimates of Potential Benefits and Harms," *Annals of Internal Medicine*, 151, 738-747.

Manski, C. (1990a), “The Use of Intentions Data to Predict Behavior: A Best Case Analysis,” *Journal of the American Statistical Association*, 85, 934-940.

Manski, C. (1990b), “Nonparametric Bounds on Treatment Effects,” *American Economic Review Papers and Proceedings*, 80, 319-323.

Manski, C. (1995), *Identification Problems in the Social Sciences*, Cambridge: Harvard University Press.

Manski, C. (1997), “Monotone Treatment Response,” *Econometrica*, 65, 1311-1334.

Manski, C. (2004), "Statistical Treatment Rules for Heterogeneous Populations," *Econometrica*, 72, 221-246.

Manski, C. (2005), *Social Choice with Partial Knowledge of Treatment Response*, Princeton: Princeton University Press.

Manski, C. (2007a), *Identification for Prediction and Decision,* Cambridge: Harvard University Press.

Manski, C. (2007b), “Minimax-Regret Treatment Choice with Missing Outcome Data,” *Journal of Econometrics*, 139, 105-115.

Manski, C. (2008), “Studying Treatment Response to Inform Treatment Choice,” *Annales D’Économie et de Statistique*,” 91-92, 93-105.

Manski C. (2009), "Diversified Treatment under Ambiguity," *International Economic Review*, 50, 1013-1041.

Manski, C. (2011), "Interpreting and Combining Heterogeneous Survey Forecasts," in M. Clements and D. Hendry (editors), *Oxford Handbook on Economic Forecasting*, Oxford: Oxford University Press, 457-472.

Manski, C. (2013a), *Public Policy in an Uncertain World*, Cambridge, MA: Harvard University Press.

Manski, C. (2013b), "Diagnostic Testing and Treatment under Ambiguity: Using Decision Analysis to Inform Clinical Practice," *Proceedings of the National Academy of Sciences*, 110, 2064-2069.

Manski, C. (2016), "Interpreting Point Predictions: Some Logical Issues," *Foundations and Trends in Accounting*, 10, 238-261.

Manski, C. (2018a), "Credible Ecological Inference for Medical Decisions with Personalized Risk Assessment," *Quantitative Economics*, 9, 541-569.

Manski, C. (2018b), "Reasonable Patient Care under Uncertainty," *Health Economics*, https://doi.org/10.1002/hec.3803.

Manski, C. and D. Nagin (1998), “Bounding Disagreements about Treatment Effects: a Case Study of Sentencing and Recidivism,” *Sociological Methodology*, 28, 99–137.

Manski, C. and J. Pepper (2000), “Monotone Instrumental Variables: With an Application to the Returns to Schooling,” *Econometrica*, 68, 997-1010.

Manski, C. and A. Tetenov (2016), " Sufficient Trial Size to Inform Clinical Practice," Proceedings of the *National Academy of Sciences*, 113, 10518-10523.

Manski, C. and A. Tetenov (2018), "Trial Size for Near-Optimal Treatment: Reconsidering MSLT-II," Department of Economics, Northwestern University.

Materson, B., D. Reda,, W. Cushman, B. Massie, E. Freis, M. Kochar, R. Hamburger, C. Fye, R. Lakshman, J. Gottdiener, E. Ramirez, and W. Henderson (1993), “Single-Drug Therapy for Hypertension in Men: A Comparison of Six Antihypertensive Agents with Placebo,” *New England Journal of Medicine*, 328, 914-921.

McGregor, J. (2013), "Too Much Surgery and Too Little Benefit? Sentinel Node Biopsy for Melanoma as it Currently Stands," *British Journal of Dermatology*, 169,

233-235.

McNees, S. (1992), "The Uses and Abuses of 'Consensus' Forecasts," *Journal of Forecasting*, 11, 703-710.

Meehl, P. (1954), *Clinical Versus Statistical Prediction: a Theoretical Analysis and a Review of the Evidence*, Minneapolis: University of Minnesota Press.

Meltzer, D. (2001), "Addressing Uncertainty in Medical Cost-Effectiveness: Implications of Expected Utility Maximization for Methods to Perform Sensitivity Analysis and the Use of Cost-Effectiveness Analysis to Set Priorities for Medical Research," *Journal of Health Economics,* 20, 109-129.

Morton, D., D Wen, J. Wong, J. Economou, L. Cagle, F. Storm, L. Foshag, A. Cochran (1992), "Technical Details of Intraoperative Lymphatic Mapping for Early Stage Melanoma," *Archives of Surgery*, 127, 392-399.

Morton, D., J. Thompson, A. Cochran, N. Mozzillo, R. Elashoff, R. Essner, O. Nieweg, D. Roses, H. Hoekstra, C. Karakousis, D. Reintgen, B. Coventry, E. Glass, and H. Wang (2006), "Sentinel-Node Biopsy or Nodal Observation in Melanoma," *New England Journal of Medicine*, 355, 1307-1317.

Morton, D., J. Thompson, A. Cochran, N. Mozzillo, O. Nieweg, D. Roses, H. Hoekstra, C. Karakousis, C. Puleo, B. Coventry, M. Kashani-Sabet, B. Smithers, E. Paul, W. Kraybill, J. McKinnon, H.. Wang, R. Elashoff, and M.B. Faries (2014), "Final Trial Report of Sentinel-Node Biopsy versus Nodal Observation in Melanoma," *New England Journal of Medicine*, 370, 599-609.

Mullins, D., R. Montgomery, and S. Tunis (2010), "Uncertainty in Assessing Value of Oncology Treatments," *The Oncologist*, 15 (supplement 1), 58-64.

National Cancer Institute (2011), *Breast Cancer Risk Assessment Tool*, http://www.cancer.gov/bcrisktool/, accessed August 19, 2017.

National Cancer Institute (2018), *NCI Dictionary of Cancer Terms*, www.cancer.gov/publications/dictionaries/cancer-terms, accessed January 26, 2018.

National Comprehensive Cancer Network (2017), *Breast Cancer Screening and Diagnosis*, Version 1.2017, www.nccn.org/professionals/physician_gls/pdf/breast-screening.pdf, accessed March 8, 2018.

National Health Service (2015), *The NHS Atlas of Variation in Healthcare*, http://fingertips.phe.org.uk/profile/atlas-of-variation, accessed 12 May 2017.

National Institute for Health and Care Excellence (2015), "Melanoma: Assessment and Management," NICE Guideline [NG 14], https://www.nice.org.uk/guidance/ng14/chapter/1-Recommendations#staging-investigations-2, accessed January 29, 2018.

Oeffinger, K., E. Fontham, R. Etzioni, A. Herzig, J. Michaelson, Y. Shih, L. Walter, T. Church, C. Flowers, S. LaMonte, A. Wolf, C. DeSantis, J. Lortet-Tieulent, K. Andrews, D. Manassaram-Baptiste, D. Saslow, R. Smith, O. Brawley, and R. Wender (2015), "Breast Cancer Screening for Women at Average Risk: 2015 Guideline Update From the American Cancer Society," *Journal of the American Medical Association*, 314, 1599-1614.

Peltzman, S. (1973), "An Evaluation of Consumer Protection Legislation: The 1962 Drug Amendments," *Journal of Political Economy*, 81, 1049-1091.

Phelps, C. and A. Mushlin (1988), "Focusing Technology Assessment Using Medical Decision Theory," *Medical Decision Making*, 8, 279-289.

Prasad, V. (2017), "Do Cancer Drugs Improve Survival or Quality of Life?" *BMJ*, 359, doi: https://doi.org/10.1136/bmj.j4528.

President's Council of Advisors on Science and Technology (2008), "Priorities for Personalized Medicine," http://oncotherapy.us/pdf/PM.Priorities.pdf, accessed August 19, 2017.

Psaty, B., N. Weiss, C. Furberg, T. Koepsell, D. Siscovick, F. Rosendaal, N. Smith, S. Heckbert, R. Kaplan, D. Lin, T. Fleming, and E. Wagner (1999), "Surrogate End Points, Health Outcomes, and the Drug-Approval Process for the Treatment of Risk Factors for Cardiovascular Disease," *Journal of the American Medical Association*, 282, 786-790.

Sackett, D. (1997), "Evidence-Based Medicine," *Seminars in Perinatology*, 21, 3-5.

Sarbin, T. (1943), "A Contribution to the Study of Actuarial and Individual Methods of Prediction," *American Journal of Sociology*, 48, 593– 602.

Sarbin, T. (1944), "The Logic of Prediction in Psychology," *Psychological Review*, 51, 210-228.

Savage, L. (1951), "The Theory of Statistical Decision," *Journal of the American Statistical Assocation,* 46, 55-67.

Schlag, K. (2006), "Eleven – Tests Needed for a Recommendation," European University Institute Working Paper ECO No. 2006/2.

Sedgwick, P. (2014), "Clinical Significance Versus Statistical Significance," *BMJ*, 348:g2130, doi: 10.1136/bmj.g2130.

SHEP Cooperative Research Group (1991), "Prevention of Stroke by Antihypertensive Drug Treatment in Older Persons with Isolated Systolic Hypertension: Final Results of the Systolic Hypertension in the Elderly Program (SHEP)," *Journal of the American Medical Association*, 265, 3255-3264.

Singletary, K. and S. Gapstur (2001), "Alcohol and Breast Cancer: Review of Epidemiologic and Experimental Evidence and Potential Mechanisms," *Journal of the American Medical Association*, 286, 2143-2151.

Spiegelhalter, D. (2004), "Incorporating Bayesian Ideas into Health-Care Evaluation," *Statistical Science*, 19, 156-174.

Spiegelhalter D., L. Freedman, and M. Parmar (1994), "Bayesian Approaches to Randomized Trials" (with discussion), *Journal of the Royal Statistics Society Series A*, 157, 357-416.

Staessen, J. R. Fagard, L. Thijs, *et al.* (1997), "The Systolic Hypertension in Europe (Syst-Eur) Trial Investigators. Randomised Double-blind Comparison of Placebo and Active Treatment for Older Patients with Isolated Systolic Hypertension," *Lancet*, 350, 757-764.

Stein, R. (2007), "Critical Care without Consent," *Washington Post*, May 27, A01.

Stoye, J. (2009), "Minimax Regret Treatment Choice with Finite Samples," *Journal of Econometrics*, 151, 70-81.

Stoye, J. (2012), Minimax Regret Treatment Choice with Attributes or with Limited Validity of Experiments," *Journal of Econometrics*, 166, 138-156.

Surowiecki, J. (2004), *The Wisdom of Crowds*, New York: Random House.

Susan G. Komen (2016), *Estimating Breast Cancer Risk*, ww5.komen.org/BreastCancer/GailAssessmentModel.html, accessed July 9, 2016.

Temin, P. (1980), *Taking Your Medicine: Drug Regulation in the United States*, Cambridge, Mass.: Harvard University Press.

Torjesen, I. (2013), "Sentinel Node Biopsy for Melanoma: Unnecessary Treatment?" *BMJ*, 346:e8645 doi: 10.1136/bmj.e8645.

U.S. Food and Drug Administration (2013), *Structured Approach to Benefit-Risk Assessment in Drug Regulatory Decision-Making*, https://www.fda.gov/downloads/forindustry/userfees/prescriptiondruguserfee/ucm329758.pdf, accessed March 11, 2018.

U.S. Food and Drug Administration (2014), *Statistical Guidance for Clinical Trials of Nondiagnostic Medical Devices.*

U.S. Food and Drug Administration (2017a), "The FDA's Drug Review Process: Ensuring Drugs Are Safe and Effective," https://www.fda.gov/Drugs/ResourcesForYou/Consumers/ucm143534.htm, accessed March 11, 2018.

U.S. Food and Drug Administration (2017b), "Postmarketing Surveillance Programs," https://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/ucm090385.htm, accessed March 11, 2018.

U.S. Preventive Services Task Force (2009), "Screening for Breast Cancer: U.S. Preventive Services Task Force Recommendation Statement, "*Annals of Internal Medicine*, 151, 716-727.

Viscusi, K., J. Harrington, and J. Vernon (2005), *Economics of Regulation and Antitrust*, Cambridge, Mass.: MIT Press.

Visvanathan, K., R. Chlebowski, P. Hurley, N. Col, M. Ropka, D. Collyar, M. Morrow, C. Runowicz, K. Pritchard, K. Hagerty, B. Arun, J. Garber, V. Vogel, J. Wade, P. Brown, J. Cuzick, B. Kramer, and S. Lippman (2009), "American Society of Clinical Oncology Clinical Practice Guideline Update on the Use of Pharmacologic Interventions Including Tamoxifen, Raloxifene, and Aromatase Inhibition for Breast Cancer Risk Reduction," *Journal of Clinical Oncology*, 27, 3235-3258.

Wald, A. (1950), *Statistical Decision Functions*, Wiley: New York.

Wasserstein, R. and N. Lazar (2016), "The ASA's Statement on p-Values: Context, Process, and Purpose," *American Statistician* 70, 129-133.

Wennberg, J. (2011), "Time to Tackle Unwarranted Variations in Practice," *BMJ*, 342, 26 March, 687-690.

Wong, S., M. Faries, E. Kennedy, S. Agarwala, T. Akhurst, C. Ariyan, C. Balch, B. Berman, A. Cochran, K. Delman, M. Gorman, J. Kirkwood, M. Moncrieff, J. Zager, and G. Lyman (2017), "Sentinel Lymph Node Biopsy and Management of Regional Lymph Nodes in Melanoma: American Society of Clinical Oncology and Society of Surgical Oncology Clinical Practice Guideline Update," *Journal of Clinical Oncology*, http://ascopubs.org/doi/full/10.1200/JCO.2017.75.7724.