NBER WORKING PAPER SERIES

IMPROVING CLINICAL GUIDELINES AND DECISIONS UNDER UNCERTAINTY

Charles F. Manski

Working Paper 23915
http://www.nber.org/papers/w23915

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
October 2017

Improving Clinical Guidelines and Decisions under Uncertainty
Charles F. Manski
NBER Working Paper No. 23915
October 2017
JEL No. C4,D81,I10

ABSTRACT

This paper discusses how limited ability to assess patient risk of illness and predict treatment response may affect the welfare achieved by adherence to clinical practice guidelines and by decentralized clinical practice. I explain why predictive ability has been limited, calling attention to imperfections in clinical judgment and to questionable methodological practices in the research that supports evidence-based medicine. I discuss recent econometric research that can improve the ability of guideline developers and clinicians to predict patient outcomes. Recognizing that uncertainty will continue to afflict medical decision making, I apply basic decision theory to suggest reasonable decision criteria with well-understood welfare properties.

Charles F. Manski
Department of Economics
Northwestern University
2211 Campus Drive
Evanston, IL 60208-2600
and NBER
cfmanski@northwestern.edu

1. Introduction

1.1. Adherence to Guidelines or Exercise of Judgment?

Medical textbooks and training have long offered clinicians guidance in patient care. Such guidance has increasingly become institutionalized through issuance of clinical practice guidelines (CPGs). The material made available by the National Guideline Clearinghouse at Agency for Healthcare Research and Quality (2017) gives a sense of the current scale and scope. The Clearinghouse provides periodically updated summaries of several thousand evidence-based guidelines issued by numerous health organizations.

Dictionaries typically define a "guideline" as a suggestion or advice for behavior rather than a mandate. Two among many definitions of CPGs are given by Hoyt (1997) and Institute of Medicine (IOM) (2011). Hoyt writes that CPGs are (p. 32): "official statements of practice groups, hospitals, organizations, or agencies regarding proper management of a specific clinical problem or the proper indications for performing a procedure or treatment." The IOM committee 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."

Neither Hoyt nor the IOM define CPGs as mandates. However, clinicians often have strong incentives to comply with guidelines, making adherence close to compulsory. A patient's 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. Less dramatically, adherence to guidelines 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 to guidelines, arguing that CPGs developers have superior evidence-based knowledge of treatment response than do clinicians. Hoyt (1997) states that the purpose of CPGs is to achieve (p. 32): "reduction in unnecessary

variability of care." Indeed, a prominent argument for adherence to CPGs has been to reduce "unnecessary" or "unwarranted" variation in clinical practice. 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." Institute of Medicine (2011) states (p. 26): "Trustworthy CPGs have the potential to reduce inappropriate practice variation." Another IOM report (Institute of Medicine, 2013) states (p. 2-15): "geographic variation in spending is considered inappropriate or 'unacceptable' when it is caused by or results in ineffective use of treatments, as by provider failure to adhere to established clinical practice guidelines."

These and many similar quotations exemplify a widespread belief that adherence to guidelines is socially preferable to decentralized clinical decision making. Yet there has been an absence of welfare analysis to support this belief. There are two broad reasons why patient care adhering to guidelines may differ from the care that clinicians provide. First, the developers of guidelines may differ from clinicians in their ability to predict how decisions affect patient outcomes. Second, guideline developers and clinicians may differ in how they evaluate patient outcomes. Welfare comparison of adherence to guidelines and decentralized decision making requires careful consideration of both factors.

This paper addresses the first factor; that is, how limited ability to assess patient risk of illness and to predict treatment response may affect the welfare achieved by adherence to guidelines and by decentralized clinical practice. I explain why predictive ability has been limited, calling attention to imperfections in clinical judgment and to questionable methodological practices in the research that supports evidence-based medicine. I discuss recent econometric research that can improve the ability of guideline developers and clinicians to predict patient outcomes. Recognizing that uncertainty will continue to afflict medical decision making, I apply basic concepts of decision theory to suggest reasonable decision criteria with well-understood welfare properties. The remainder of this Introduction summarizes the paper.

1.2. Summary of the Paper

To standardize valuation of patient outcomes, I suppose throughout the paper that treatment response is individualistic and that clinicians and guideline developers share the utilitarian objective of maximizing patient welfare. These assumptions imply that social and private preferences coincide. To simplify discussion of adherence to guidelines, I suppose that there exists one CPG making recommendations in a specified clinical setting. Hence, the meaning of adherence to guidelines is clear.[1]

To provide a baseline for discussion of uncertainty in decision making, I first consider patient care when clinicians have substantial knowledge of patient outcomes. After introducing the idea of personalized medicine, Section 2 examines an idealized setting of patient care with rational expectations that has been studied in normative economic research on personalized medicine. It is assumed that clinicians observe specified patient covariates, know the objective distribution of outcomes that occur when patients with the observed covariates are given alternative treatments, and choose treatments that maximize patients' expected utility. The assumption of rational expectations, while strong, does not require that clinicians be able to predict patient outcomes with certainty; that is, it does not assume perfect foresight.

In this setting, the optimal treatment rule divides into groups having the same observed covariates. All patients in a covariate group are given the care that yields the highest within-group mean welfare. Utilitarian welfare increases as more covariates are observed. An appendix gives abstract derivations of these well-known results. The text derives the optimal rule in an important class of clinical decision problem: choice between surveillance and aggressive treatment of patients at risk of developing a disease.

When clinicians have rational expectations, adherence to a CPG cannot outperform decentralized

[1] In practice, multiple organizations may issue guidelines that make conflicting recommendations. To cite just one striking example, consider surveillance of women at risk of breast cancer. The guideline issued by the National Comprehensive Cancer Network recommends periodic performance of a clinical breast examination, but the one issued by the American Cancer Society recommends against it. See National Comprehensive Cancer Network (2017) and Oeffinger *et al.* (2015).

practice and may perform less well. If a CPG conditions its treatment recommendations on all of the patient covariates that clinicians observe, it can do no better than reproduce clinical decisions. If the CPG makes recommendations conditional on a subset of the clinically observable covariates, adhering to the CPG may yield inferior welfare because the guideline does not personalize patient care to the extent possible. Thus, there is no informational argument for adherence to CPGs if clinicians have rational expectations. I observe that it is common for CPGs to condition their recommendations on a subset of the clinically observable patient covariates. I illustrate with guidelines for breast cancer screening.

Section 3 considers patient care when clinicians have imperfect judgment. A substantial body of empirical psychological research has strongly questioned the realism of clinical rational expectations. Many studies have compared the accuracy of evidence-based statistical predictions of health outcomes with ones made by subjective clinical judgment. The consensus has been that the former outperforms the latter when predictions are made using the same patient covariates. Moreover, the gap in performance persists even when clinicians condition their judgements on additional covariates that are not used in statistical predictors.

I summarize the psychological research and consider its implications for welfare comparison of adherence to CPGs and decentralized clinical practice. The research does not suffice to conclude that one system is superior to the other, but it does imply that society faces a delicate choice between alternative second-best approaches to patient care. Adherence to evidence-based CPGs may be inferior to the extent that CPGs condition on fewer patient covariates than do clinicians, but it may be superior to the extent that imperfect clinical judgment generates sub-optimal clinical decisions.

Section 4 questions whether the developers of CPGs make correct predictions of health outcomes conditional on the patient covariates used in the guidelines. I argue that the judgment of CPG developers may be imperfect due to questionable methodological practices in research on health outcomes. One questionable practice is wishful extrapolation of findings from randomized trials to clinical practice. Important cases include extrapolation from study populations to patient populations, from experimental

treatments to treatments used in practice, and from surrogate outcomes to outcomes of health interest. Another questionable practice is use of hypothesis testing to compare treatments and to choose when to report findings.

Section 5 argues that evidence-based research can inform patient care more effectively than it does at present. Studies should quantify how identification problems and statistical imprecision jointly affect the feasibility of making credible predictions of important health outcomes. Identification is usually the dominant problem. To demonstrate, I summarize aspects of my research on partial identification of risk assessments and treatment response. This includes work on identification of response to treatments for hypertension in a trial with missing data, on identification of response to diagnostic testing and treatment, and on credible ecological inference for personalized medicine.

Recognizing that medical knowledge regarding many clinical problems is incomplete, Section 6 considers patient care as a problem of decision making under uncertainty. To lay foundations for study of patient care under uncertainty, I first discuss basic concepts of decision theory in abstraction and then summarize some applications. I emphasize that there is no uniquely optimal way to make decisions under uncertainty, but there are various reasonable ways. I juxtapose several prominent decision criteria: maximization of subjective expected welfare, the maximin criterion, and the minimax-regret criterion.

I observe that, when applied to patient care with a utilitarian welfare function, these criteria may yield a recommendation that differs from the widespread admonition of CPG advocates to reduce "unwarranted variation" in clinical practice. To the contrary, they may prescribe diversification of treatment. Diversification means random assignment of observationally similar patients to different treatments. The rationale for diversification is that it prevents occurrence of gross errors that might occur if all patients were inadvertently given an inferior treatment. The possibility of learning enhances the advantage of diversifying treatment choice, the reason being that diversification yields randomized experiments.

2. Optimal Personalized Care Assuming Rational Expectations

2.1. Degrees of Personalized Medicine

The term *personalized medicine* is sometimes taken to mean health care that is literally specific to the individual, as in this definition by Ginsburg and Willard (2009, p. 278): "Personalized medicine is . . . . . health care that is informed by each person's unique clinical, genetic, genomic, and environmental information." However, patient data to support complete personalization is rarely available. Hence, the term is commonly used to mean care that varies with observed patient covariates. President's Council of Advisors on Science and Technology (2008) states (p. 7):

> " 'Personalized medicine' refers to the tailoring of medical treatment to the specific characteristics of each patient. In an operational sense, however, personalized medicine does not literally mean the creation of drugs or medical devices that are unique to a patient. Rather, it involves the ability to classify individuals into subpopulations that are uniquely or disproportionately susceptible to a particular disease or responsive to a specific treatment."

Thus, personalized medicine is a matter of degree rather than an all-or-nothing proposition.

Consider probabilistic risk assessment or prediction of treatment response. A patient may ask her clinician a seemingly straightforward questions such as "What is the chance that I will develop disease X in the next five years?" or "What is the chance that treatment Y will cure me?" Yet these questions do not have unique answers. Personalized probabilities of disease development or treatment outcomes depend on the patient covariates used to condition the predictions.

Clinicians often observe patient covariates 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. An apt example is the evidence-based predictor of risk of breast cancer used in some guidelines for breast cancer screening.

2.1.1. The Breast Cancer Risk Assessment Tool

The Breast Cancer Risk Assessment (BCRA) Tool of the National Cancer Institute (2011) gives an evidence-based probability that a woman will develop breast cancer conditional on eight personal covariates: (1) history of breast cancer or chest radiation therapy for Hodgkin Lymphoma (yes/no); (2) presence of a BRCA mutation or diagnosis of a genetic syndrome associated with risk of breast cancer (yes/no/unknown); (3) current age, in years; (4) age of first menstrual period (7-11, 12-13, ≥ 14, unknown); (5) age of first live birth of a child (no births, < 20, 20-24, 25-29, ≥30, unknown); (6) number of first-degree female relatives with breast cancer (0, 1, >1, unknown); (7) number of breast biopsies (0, 1, > 1, unknown); and (8) race/ethnicity (White, African American, Hispanic, Asian American, American Indian or Alaskan Native, unknown).

The reason that the tool assesses risk conditional on these covariates and not others is that the tool uses a modified version of the "Gail Model," based on the empirical research of Gail *et al.* (1989). The Gail *et al.* article estimated probabilities of breast cancer for white women who have annual breast examinations, conditional on covariates (1) through (7). Scientists at the National Cancer Institute later modified the model to predict invasive cancer within a wider population of women.

The BCRA Tool personalizes predicted risk of breast cancer in multiple respects, but it does not condition on further patient covariates that a clinician can observe and that may be associated with risk of cancer. For example, when considering the number of first-degree relatives with breast cancer (item 6), the BCRA Tool does not take into account the number and ages of a woman's first-degree relatives, which logically should matter when interpreting the response to the item. Nor does it condition on the prevalence of breast cancer among second-degree relatives, a consideration that figures prominently in another risk assessment model due to Claus, Risch, and Thompson (1994). When considering race/ethnicity (item 8), the BCRA Tool groups all white woman together and does not distinguish ethnic subgroups such as Ashkenazi Jews, who are thought to have considerably higher risk of a BRCA mutation than other white subgroups, a

potentially important matter when the answer to item (2) is "unknown." Moreover, the BCRA Tool does not condition on behavioral covariates such as excessive drinking of alcohol, which has been associated with substantial increased risk of breast cancer (Singletary and Gapstur, 2001).

2.2. Formalizing Optimal Care

The BCRA Tool exemplifies a common question in patient care. Evidence from medical research enables assessment of risk of disease or treatment response conditional on certain patient covariates, so guidelines make recommendations conditional on these covariates. Clinicians also observe additional covariates that may be informative predictors of patient outcomes, but the available evidence does not show how outcomes vary with the additional covariates. How should medical decision making proceed?

When considering this question formally, I will assume that clinicians and CPG developers want to maximize a utilitarian welfare function that sums up the benefits and costs of treatment across the population. I also assume that treatment is individualistic; that is, the care received by one patient affects only that individual and not other members of the population. This assumption is usually realistic when considering non-infectious diseases.

When treatment is individualistic and welfare is utilitarian, the problem of optimizing patient care has a simple well-known solution. Patients should be divided into groups having the same observed covariates. All patients in a covariate group should be given the care that yields the highest within-group mean welfare. Thus, it is optimal to differentially treat patients with different observed covariates if different actions maximize their within-group mean welfare. Patients with the same observed covariates should be treated uniformly. The value of maximum welfare increases as more patient covariates are observed.

These findings have long been known in the literature on maximization of expected utility with rational expectations and is often stated without attribution. I do not know who first proved it, but a

relatively early version is given in Good (1967). Manski (2007) and Kadane, Schervish, and Seidenfeld (2008) give later statements and proofs in different contexts. The result has been applied in the economic literature on medical decision making by Phelps and Mushlin (1988), Meltzer (2001), and Basu and Meltzer (2007), among others. Appendix A presents an abstract derivation of the optimal treatment rule, paraphrasing Manski (2007, Sec. 11.4). Section 2.3 characterizes optimal treatment in a simple and important medical decision problem that I will use to illustrate broad ideas throughout the paper.

2.3. Choice Between Surveillance and Aggressive Treatment of Patients at Risk of Disease

A common medical decision is choice between periodic surveillance and aggressive treatment of patients at risk of potential disease. I have already mentioned one case, screening women for breast cancer. Others are choice between surveillance and drug treatment for patients at risk of heart disease or diabetes. Yet others are choice between surveillance and aggressive treatment of patients who have been treated for localized cancer and are at risk of metastasis. A semantically distinct but logically equivalent decision is choice between diagnosis of patients as healthy or ill. With diagnosis, the clinician is uncertain not whether a patient will develop the disease in the future but whether the patient is ill at present.

Choice between surveillance and aggressive treatment often requires 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. However, aggressive treatment may be more costly to the extent that it generates health side effects and financial costs beyond those associated with surveillance.

Here is a simple formalization of the decision problem. Let treatment $t = A$ denote surveillance and $t = B$ denote aggressive treatment. Let $y(A)$ and $y(B)$ be potential binary outcomes, with $y = 1$ denoting that the patient will develop the disease and $y = 0$ otherwise. Let (x, w) be the patient covariates observed by a

clinician, x being the subset used to make evidence-based predictions. Let $P_{xw}(t) \equiv P[y(t) = 1|x, w]$ be the probability that a patient with observed covariates (x, w) will develop the disease if the patient receives treatment t. In the case of diagnosis, $P_{xw}(t)$ is the probability that the patient is currently ill and does not vary with t. A clinician has rational expectations if he knows $P_{xw}(A)$ and $P_{xw}(B)$.

The utility of each care option depends on whether a patient will or will not develop the disease. Let $u_{xw}(y, t)$ denote the expected utility of treatment t to a patient with covariates (x, w) in the presence of disease outcome y. The clinician chooses a care option without knowing the disease outcome. The expected utility of each treatment t without knowledge of the disease outcome is

$$(1)\quad E[u(t)|x, w] = P_{xw}(t)\cdot u_{xw}(1, t) + [1 - P_{xw}(t)]\cdot u_{xw}(0, t).$$

Maximization of expected utility yields the treatment rule

$$(2)\quad \text{Choose A if } P_{xw}(A)\cdot u_{xw}(1, A) + [1 - P_{xw}(A)]\cdot u_{xw}(0, A) \geq P_{xw}(B)\cdot u_{xw}(1, B) + [1 - P_{xw}(B)]\cdot u_{xw}(0, B),$$
$$\text{Choose B if } P_{xw}(B)\cdot u_{xw}(1, B) + [1 - P_{xw}(B)]\cdot u_{xw}(0, B) \geq P_{x}(A)\cdot u_{xw}(1, A) + [1 - P_{xw}(A)]\cdot u_{xw}(0, A).$$

A clinician with rational expectations can implement the optimal treatment rule. In general, this rule cannot be implemented by a CPG that predicts disease development and assesses patient expected utility conditional only on covariates x rather than (x, w).

Decision rule (2) is simple as is, but it is instructive to discuss further simplifications that occur when treatment operates on disease in different ways. In some clinical settings, aggressive treatment may prevent disease development whereas surveillance does not; thus, $P_{xw}(B) = 0$ and $P_{xw}(A) > 0$. In other settings, treatment does not affect the risk of disease but may affect the severity of illness when it occurs; thus, $P_{xw}(B) = P_{xw}(A)$, but $u_{xw}(1, A)$ may differ from $u_{xw}(1, B)$. These conditions hold, for example, when clinicians

diagnose whether patients are currently ill and choose accordingly between surveillance and aggressive treatment. I show below that, in settings of both types, calculation of a simple threshold probability of disease suffices to determine the optimal treatment.

2.3.1. Aggressive Treatment Prevents Disease

Suppose that $P_{xw}(B) = 0$. Then (2) reduces to

(3) Choose A if $P_{xw}(A)\cdot u_{xw}(1, A) + [1 - P_{xw}(A)]\cdot u_{xw}(0, A) \geq u_{xw}(0, B)$,

Choose B if $u_{xw}(0, B) \geq P_{xw}(A)\cdot u_{xw}(1, A) + [1 - P_{xw}(A)]\cdot u_{xw}(0, A)$.

Hence, the optimal treatment depends on the magnitude of $P_{xw}(A)$ relative to the threshold value that equalizes the expected utility of the two treatments:

$$(4) \qquad P^*_{xw}(A) \equiv \frac{u_{xw}(0, A) - u_{xw}(0, B)}{u_{xw}(0, A) - u_{xw}(1, A)} .$$

It is generally reasonable to expect that surveillance yields higher expected utility when a patient will remain healthy rather than develop the disease; that is, $u_{xw}(0, A) > u_{xw}(1, A)$. Then surveillance is optimal when $P_{xw}(A) \leq P^*_{xw}(A)$ and aggressive treatment is optimal when $P_{xw}(A) \geq P^*_{xw}(A)$.

Inspection of (4) shows that $P^*_{xw}(A) \leq 0$ if $u_{xw}(0, A) - u_{xw}(0, B) \leq 0$; that is, if surveillance yields lower expected utility than aggressive treatment when a patient will not develop the disease. Then aggressive treatment is the better option whatever the patient's probability of disease development may be. Contrariwise, $P^*_{xw}(A) \geq 1$ if $u_{xw}(0, B) \leq u_{xw}(1, A)$; that is, if the expected utility of aggressive treatment in the absence of disease is less than that of surveillance in the presence of disease. Then surveillance is always better. When $0 < P^*_{xw}(A) < 1$, the better care option varies with the probability of disease development.

2.3.2. Aggressive Treatment Reduces the Severity of Disease

Suppose that $P_{xw}(A) = P_{xw}(B) \equiv P_{xw}$. Then (2) reduces to

(5) Choose A if $P_{xw} \cdot u_{xw}(1, A) + (1 - P_{xw}) \cdot u_{xw}(0, A) \geq P_{xw} \cdot u_{xw}(1, B) + (1 - P_{xw}) \cdot u_{xw}(0, B)$,

Choose B if $P_{xw} \cdot u_{xw}(1, B) + (1 - P_{xw}) \cdot u_{xw}(0, B) \geq P_{xw} \cdot u_{xw}(1, A) + (1 - P_{xw}) \cdot u_{xw}(0, A)$.

Hence, the optimal treatment depends on the magnitude of P_{xw} relative to the threshold value that equalizes the expected utility of the two treatments:

$$(6) \qquad P^*_{xw} = \frac{u_{xw}(0, A) - u_{xw}(0, B)}{[u_{xw}(0, A) - u_{xw}(0, B)] + [u_{xw}(1, B) - u_{xw}(1, A)]} \quad .$$

It often is reasonable to suppose that surveillance yields higher expected utility when a patient will not develop the disease and that aggressive treatment yields higher utility when a patient will develop the disease; that is, $u_{xw}(0, A) > u_{xw}(0, B)$, and $u_{xw}(1, B) > u_{xw}(1, A)$. Then $0 < P^*_{xw} < 1$. Treatment A is optimal if $P_{wx} \leq P^*_{xw}$ and B is optimal if $P_{xw} \geq P^*_{xw}$.

3. Treatment with Imperfect Clinical Judgment

3.1. Empirical Research Comparing Statistical Prediction and Clinical Judgment

If it were reasonable to think that clinicians have rational expectations, there would be no utilitarian argument to develop and publish CPGs. However, a body of empirical psychological research comparing evidence-based statistical predictions with ones made by clinical judgment has concluded that the former

consistently outperforms the latter when the predictions are made using the same patient attributes. The gap in performance persists even when clinical judgment uses additional attributes as predictors.

This research began in mid-twentieth century, some notable early contributions including Sarbin (1943, 1944), Meehl (1954), and Goldberg (1968). To describe the conclusions of the literature, I rely mainly on the influential review article of Dawes, Faust, and Meehl (1989). See Camerer and Johnson (1997) and Groves *et al.* (2000) for further review articles.

Dawes *et al.* distinguish actuarial prediction and clinical judgment as follows (p. 1668):

> "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 actuarial prediction, they state (p. 1670):

> "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."

Seeking to explain this empirical finding, they discuss an example in which the additional observed attribute is that a patient has a broken leg and then write (p. 1670-1671):

> "The broken leg possibility is easily studied by providing clinicians with both the available data and the actuarial conclusion and allowing them to use or countervail the latter at their discretion. The limited research examining this possibility, however, all shows that greater overall accuracy is achieved when clinicians rely uniformly on actuarial conclusions and avoid discretionary judgments . . . . When operating freely, clinicians apparently identify too many 'exceptions,' that is, the actuarial conclusions correctly modified are outnumbered by those incorrectly modified. If clinicians were more conservative in overriding actuarial conclusions they might gain an advantage, but this

conjecture remains to be studied adequately."

Here and elsewhere, Dawes, Faust, and Meehl caution against use of clinical judgment to subjectively predict disease risk or treatment response conditional on patient attributes that are not utilized in evidence-based assessment tools or research reports. 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 exceptions such as broken legs.

Psychological research published after Dawes, Faust, and Meehl (1989) has largely corroborated the conclusions reached there, albeit occasionally with caveats. For example, Groves *et al.* (2000) conclude their meta-analysis of the literature as follows (p. 25):

> "This study confirms and greatly extends previous reports that mechanical prediction is typically as accurate or more accurate than clinical prediction. However, our results qualify overbroad statements in the literature opining that such superiority is completely uniform; it is not. In half of the studies we analyzed, the clinical method is approximately as good as mechanical prediction, and in a few scattered instances, the clinical method was notably more accurate.
>
> Even though outlier studies can be found, we identified no systematic exceptions to the general superiority (or at least material equivalence) of mechanical prediction. It holds in general medicine, in mental health, in personality, and in education and training settings. It holds for medically trained judges and for psychologists. It holds for inexperienced and seasoned judges."

It is natural to ask how psychological research on clinical judgment has affected the practice of medicine. Curiously, I have found no explicit reference to it in my reading of medical commentaries advocating adherence to CPGs. Nor have I found reference to it in the broader literature concerning practice of evidence-based medicine. Perhaps medical writers are unaware of the psychological research. Perhaps they are aware but have dismissed its relevance.

I can find passages in the literature on evidence-based medicine that, contrary to the psychological literature, praise rather than criticize exercise of clinical judgment. For example, in a well-cited commentary, Sackett (1997) calls for integration of evidence-based research with clinical judgment, writing (p. 3):

> "The practice of evidence-based medicine means integrating individual clinical expertise with the best available external clinical evidence from systematic research. By individual clinical expertise we mean the proficiency and judgment that we individual clinicians acquire through clinical experience and clinical practice. . . . . By best available external clinical evidence we mean clinically relevant research. . . . .Good doctors use both individual clinical expertise and the best available external evidence, and neither alone is enough."

3.2. Second-Best Welfare Comparison of Adherence to Guidelines and Clinical Judgment

The psychological literature strongly challenges the realism of assuming that clinicians have rational expectations. However, this literature does not per se imply that adherence to CPGs would yield greater welfare than decentralized decision making using clinical judgment. For specificity, I will again consider choice between surveillance and aggressive treatment.

One issue is that the psychological literature has not addressed all welfare-relevant aspects of clinical decisions. Section 2 showed that the optimal decision is determined by the disease probabilities $P_{xw}(\cdot)$ and the expected utilities $u_{xw}(\cdot, \cdot)$. Psychologists have studied the relative accuracy of risk assessments and diagnoses made by statistical predictors and by clinicians. However, they have not similarly studied the relative accuracy of evaluations of patient preferences over (illness, treatment) outcomes. Thus, the literature has generated findings that may be informative about the accuracy of statistical and clinical assessments of $P_{xw}(\cdot)$ but not $u_{xw}(\cdot, \cdot)$.

A second issue is that 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, 1990). For example, assume that a forecaster uses a symmetric loss function to translate a probabilistic risk assessment into a yes/no point prediction that a patient will develop

a potential disease. Then 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. Thus, analysis of the accuracy of point predictions does not reveal much about the accuracy of statistical and clinical assessment of the disease probabilities $P_{xw}(\cdot)$.

A third issue is that inaccuracy of assessments of $P_{xw}(\cdot)$ and/or $u_{xw}(\cdot, \cdot)$ does not necessarily imply sub-optimality of decisions. A decision is optimal if it agrees with the inequalities in (2). Decisions made with inaccurate clinical judgment may nevertheless respect the inequalities.

In light of these issues, 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 only imply that welfare comparison is a delicate matter of choice between alternative second-best systems for patient care. Adherence to evidence-based CPGs may be inferior to the extent that CPGs condition on fewer patient covariates than do clinicians, but it may be superior to the extent that imperfect clinical judgment yields sub-optimal decisions. How these opposing forces interplay depends on the specifics of the setting. I discuss one case below.

3.3. Surveillance or Aggressive Treatment of Women at Risk of Breast Cancer

Consider the common clinical decision between surveillance and aggressive treatment of women at risk of breast cancer. In this setting, surveillance means that a woman receives a breast examination and mammogram periodically, typically annually or biannually depending on age. Aggressive treatment encompasses several options.

One is more frequent surveillance. This does not affect the risk of disease development, but it may reduce the severity of disease outcomes by enabling earlier diagnosis and treatment of the tumor. A potential side effect is an increased risk of cancer caused by the radiation from mammograms.

Other options for aggressive treatment include strategies for reduction of the risk of disease development. These include changes to diet, administration of a drug such as taximofen, and prophylactic mastectomy. Each strategy may have side effects, most obviously in the case of prophylactic mastectomy.

The analysis of Section 2.3 suggests that, ceteris paribus, some form of aggressive treatment is the better option if the risk of breast cancer is sufficiently high and surveillance is better otherwise. Some CPGs use the BRCA Tool to assess risk and recommend aggressive treatment if the predicted probability of invasive cancer in the next five years is above a specified threshold. In particular, the National Comprehensive Cancer Network (NCCN) (2017) recommends annual surveillance if the predicted probability is below 0.017 and choice of some form of aggressive treatment if the probability is higher. A guideline issued by the American Society of Clinical Oncology (ASCO) recommends consideration of a pharmacological intervention when the predicted probability is above 0.166 (Visvanathan *et al.* 2009).

A clinician could use judgment to assess risk conditional on a richer set of patient covariates than are used in the BCRA Tool. He could also a personalized threshold probability to make the treatment decision, as shown in Section 2.3, rather than apply the value 0.017 or 0.166 to all patients. However, clinical judgment may be imperfect. It is not known whether adherence to the NCCN or ASCO guideline yields better or worse patient care than does decentralized clinical decision making.

4. Questionable Methodological Practices in Evidence-Based Medicine

The psychological literature discussed in Section 3 has questioned the judgment of clinicians, but it has not similarly questioned the accuracy of the predictions used in evidence-based guideline development. The fact that predictions are evidence-based does not ensure that they use the available evidence effectively. Multiple questionable methodological practices have long afflicted research on health outcomes and may

deleteriously affect the development of guidelines. This fact further complicates welfare comparison of adherence to guidelines and decentralized clinical practice.

I focus here on predictions made with evidence from randomized clinical trials. Trials have long enjoyed a favored status within medical research on treatment response and are often called the "gold standard" for such research. The influential Cochrane system for grading the quality of evidence ordinarily reserves its highest rating for evidence from randomized trials (Higgins and Green, 2011, Sec. 12.2.1). The drug approval process of the U.S. Food and Drug Administration (FDA) ordinarily considers only experimental evidence when making decisions on drug approval.

Guideline development acts accordingly, valuing trial evidence more than observational studies. Indeed, guideline developers sometimes choose to use only trial evidence, entirely excluding observational studies from consideration. An example is found in an article reporting a new evidence-based CPG for management of high blood pressure (James *et al.*, 2014). The authors write (p. 508): "The panel limited its evidence review to RCTs because they are less subject to bias than other study designs and represent the gold standard for determining efficacy and effectiveness."

Section 4.1 cautions against wishful extrapolation of trial findings to clinical practice. Section 4.2 criticizes the use of hypothesis testing to interpret the sample data produced by trials. While I focus on the use of evidence from trials, I do not mean to absolve observational studies. Some of the questionable practices discussed below are commonplace there as well.

4.1. Wishful Extrapolation of Trial Findings to Clinical Practice

Guideline developers have used trial data to predict treatment response whenever such evidence is available. The well-known appeal of trials is that, given sufficient sample size and complete observation of outcomes, they deliver credible findings about treatment response within the study population. However,

it is also well-known that extrapolation of findings from trials to clinical practice can be difficult.

Researchers and guideline developers often use untenable assumptions to extrapolate. I have referred to this practice as *wishful extrapolation* (Manski, 2013a). A particularly common manifestation of wishful extrapolation assumes that the treatment response that would occur in clinical practice is the same as that observed in trials. I discuss below multiple reasons why this assumption may be suspect.

4.1.1. Study Populations and Patient Populations

The study populations in trials often differ from the patient populations that clinicians treat. Trial designs often mandate important differences between these populations. A common practice has been to perform trials concerned with treatment of a specific disease only on subjects who have no co-morbidities. However, patients treated in practice may suffer from multiple conditions. Clinicians may then need to choose complexes of interacting treatments rather than treat diseases in isolation from one another.

Another source of difference between study and clinical populations is that a study population consists of patients who volunteer to participate in a trial. Volunteers respond to financial and medical incentives to participate. A financial incentive may be receipt of free treatments. A medical incentive is that participation in a trial opens the possibility of receiving a treatment that is not otherwise available.

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 often is wishful extrapolation to assume that treatment response observed in trials performed on volunteers who lack co-morbidities is the same as what would occur in the patient populations that clinicians treat in practice.

Campbell and the Primacy of Internal Validity

Seeking to justify analysis of trials performed on study populations that may differ substantially from

the populations that clinicians treat, researchers in public health and the social sciences often cite Donald Campbell, who distinguished between the internal and external validity of studies of treatment response (Campbell and Stanley, 1963; Campbell, 1984). A study is said to have *internal validity* if its findings for the study population are credible. It has *external validity* if an invariance assumption permits credible extrapolation. In this terminology, the appeal of randomized trials is their internal validity. Wishful extrapolation is an absence of external validity.

Campbell discussed both forms of validity, but he argued that studies should be judged first by their internal validity and only secondarily by their external validity. This perspective has been used to argue for the universal primacy of experimental research over observational studies, whatever the study population may be. The reason given is that properly executed randomized experiments have high internal validity. The Campbell perspective has also been used to argue that the best observational studies are those that most closely approximate randomized experiments.

Campbell's view has been endorsed by Rosenbaum (1999), who recommends that observational studies of human subjects aim to approximate the conditions of laboratory experiments. Rosenbaum, like Campbell, downplays the importance of having the study population be similar to the population of interest, writing (page 259): "Studies of samples that are representative of populations may be quite useful in describing those populations, but may be ill-suited to inferences about treatment effects."

From the perspective of treatment choice, the Campbell-Rosenbaum position is well grounded if treatment response is homogeneous. Then researchers can aim to learn about treatment response in easy-to-analyze study populations and clinicians can be confident that research findings can be extrapolated to patient populations of interest. In human populations, however, homogeneity of treatment response seems the exception rather than the rule. To the degree that treatment response is heterogeneous, it may be wishful to extrapolate findings from a study population to a patient population of interest, as optimal treatments in the two may differ. Hence, I see no general reason to value internal validity above external validity.

4.1.2. Experimental Treatments and Treatments of Interest

The treatments assigned in trials often differ from those that would be assigned in clinical practice. This is particularly so in trials comparing drug treatments, one of which may be a placebo. These trials are normally double-blinded, neither the patient nor his physician knowing the assigned treatment. Hence, a trial reveals the distribution of response in a setting where patients and physicians are uncertain what drug a patient is receiving. It does not reveal what response would be in a clinical setting where patients and physicians would know what drug is being administered and would be able to react to this information.

Blinding is particularly problematic for clinical interpretation of the noncompliance and attrition that often occur in drug trials. When a trial subject chooses not to comply with the specified trial protocol or to drop out of the trial, he makes this decision knowing only the probability that he is receiving each drug, not the actuality. Patient behavior may differ in clinical settings where the patient and clinician know what drug is being administered.

It has been common in study of trial data to perform intention-to-treat analysis, which examines the outcomes of assignment into a treatment group rather than the outcomes of receipt of treatment. Noncompliance is logically impossible in intention-to-treat analysis because subjects have no ability to modify their treatment assignments. This fact may tempt one to think that compliance need not be a concern in study of trial data. This temptation should be resisted. Intention to treat analysis does not predict how patients would behave in clinical practice, when they know the treatments that their physicians have prescribed for them.

4.1.3. Surrogate Outcomes and Outcomes of Interest

A serious measurement problem occurs when trials have short durations. Clinicians and patients often want to learn long-term outcomes of treatments, but short trials reveal only short-run outcomes. Credible extrapolation from the surrogate outcomes measured in short trials to the long-term outcomes of

interest can be highly challenging.

Trials for drug approval by the FDA provide a good illustration. The most lengthy, called *phase 3 trials*, typically run for only two to three years. When trials are not long enough to observe the health outcomes of real interest, the practice is to measure surrogate outcomes and base drug approval decisions on their values. For example, treatments for heart disease may be evaluated using data on patient cholesterol levels and blood pressure rather than data on heart attacks and life span. Thus, the trials used in drug approval may only reveal the distribution of surrogate outcomes in the study population, not the distribution of outcomes of real health interest.

Some researchers have called attention to the difficulty of extrapolating from surrogate outcomes to health outcomes of interest. For example, Fleming and Demets (1996), who review the prevalent use of surrogate outcomes in phase 3 trials evaluating drug treatments for heart disease, cancer, HIV/AIDS, osteoporosis, and other diseases, write (p. 605): "Surrogate end points are rarely, if ever, adequate substitutes for the definitive clinical outcome in phase 3 trials."

4.1.4. Wishful Aggregation of Findings in Meta-Analyses

The issues discussed above concern extrapolation of findings from single trials. Further issues arise when researchers attempt to combine findings from multiple trials. A common practice is to perform a meta-analysis.

Meta-analysis was originally proposed to address a purely statistical problem. One wants to estimate as well as possible some parameter characterizing a study population. For example, the parameter of interest may be the mean outcome that would occur if all members of the population were to receive a specified treatment. Suppose that K independent trials drawing random samples of sizes $N_1, \ldots, N_K$ have been performed on the same study population. If the raw data on the trial outcomes are available, the most precise way to estimate the parameter combines the samples into one of size $\sum_{k=1,..,K} N_K$ and computes the estimate

using all the data. Suppose, however, that the raw data are unavailable, making it infeasible to combine the samples. Instead, K parameter estimates are available, each computed with the data from a different sample. Meta-analysis proposes methods to combine the K estimates so as to achieve as precise an estimate of the parameter as possible. The usual proposal is to compute a weighted-average of the estimates, the weights varying with sample size.

While the original concept of meta-analysis is uncontroversial, its applicability is limited. It is rarely the case that multiple independent trials are performed on the same population. It is much more common for multiple trials to be performed on distinct study populations that may have different distributions of treatment response. The protocols for administration of treatments and measurement of outcomes may vary across trials as well. Meta-analysis are performed often in such settings, computing weighted averages of estimates for distinct study populations and trial designs.

The obvious problem is that it may not be clear how to define and interpret a weighted average of the K separate estimates. Meta-analyses sometimes answer these questions through the lens of a random-effects model. The model assumes that each of the K estimates pertains to a distinct parameter value drawn at random from a population of potential parameter values. Then a weighted average of the K estimates is interpreted to be an estimate of the mean of all potential parameter values. See, for example, DerSimonian and Laird (1986). This approaches yields a well-defined estimand under the maintained assumptions, but the relevance of this estimand to clinical practice may be obscure.

4.2. Misplaced Use of Hypothesis Testing

Leaving aside all of the issues that arise in extrapolating trial findings to clinical practice, there remains the familiar statistical problem of interpreting the samples of treatment outcomes generated by trials. 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. Hypothesis testing is also used to decide what findings to report in research articles. This section critiques these practices.

4.2.1. Using Hypothesis Tests to Compare Treatments

A common procedure when comparing two treatments in a trial is to view one of them 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 treatment be used in clinical practice. If the null is rejected, it is recommended that the innovation be the treatment of choice. This type of test is institutionalized in the FDA drug approval process, which calls for comparison of a new drug with a placebo or a previously approved treatment. Approval of the new drug normally requires rejection of the null hypothesis of zero average treatment effect in two independent trial (Fisher and Moyé, 1999).

The standard practice has been to perform a test that fixes the probability of rejecting the null hypothesis when it is correct, called the probability of a Type I error. Then sample size determines the probability of rejecting the alternative hypothesis when it is correct, called the probability of a Type II error. The power of a test is defined as one minus the probability of a type II error. The convention has been to choose a sample size that yields specified power at some value of the effect size deemed clinically important. For example, International Conference on Harmonisation (1999) has provided guidance for the design and conduct of trials evaluating pharmaceuticals, stating (p. 1923):

> "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) observe that there are several reasons why hypothesis testing may yield unsatisfactory results for medical decisions. These include the following:

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. In particular, 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 more than the probabilities of Type I and II error. He should care as well 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, but 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.

4.2.2. Using Hypothesis Tests to Choose When to Report Findings

Beyond its use to choose between treatments, hypothesis testing is also used to determine when research articles should report trial findings conditional on observed patient covariates. Section 2.2 showed that optimal patient care segments patients into covariate groups and maximizes expected utility within each group. Clinicians commonly have much information — medical histories, diagnostic test findings, and

demographic attributes —about the patients they treat. Yet the medical journal articles that report on trials typically present trial findings aggregated to broad demographic groups.[2]

Conventional ideas about what constitutes adequate statistical precision for an empirical finding to be of interest have been strongly influenced by the theory of hypothesis testing. Conditioning on covariates generally reduces the statistical precision of estimates of treatment effects, to the point where findings become "statistically insignificant." Aiming to avoid publication of statistically insignificant results ex ante, researchers often report findings only for groups whose sample sizes are large enough to perform tests with conventional Type I and II error probabilities. Moreover, researchers sometimes selectively report findings that are statistically significant ex post by standard criteria. This reporting practice has been recognized to generate publication bias (Ioannidis, 2005; Wasserstein and Lazar, 2016).[3]

If researchers want to inform patient care, they should not view statistical insignificance as a reason to refrain from studying and reporting observable heterogeneity in treatment response. Clinicians should be concerned with the quantitative variation of outcomes with treatments and covariates. Hypothesis tests do not address this question. Subject to considerations of subject confidentiality and space constraints, research journals should encourage publication of findings conditional on observed patient covariates. When journal

---

[2] For example, Crits-Christoph *et. al.* (1999) report on a trial placing 487 cocaine-dependent patients in one of four treatment groups, each designated treatment combining group drug counseling (GDC) with another form of therapy. The article provides much descriptive information on subject covariates including measures of race, sex, age, education, employment status, type and severity of drug use, psychiatric state, and personality. Yet the article does not report treatment outcomes conditional on any of these patient covariates. Indeed, its Conclusion section makes no reference to the possibility that treatment response might vary with covariates, stating simply (page 493): "Compared with professional psychotherapy, a manual-guided combination of intensive individual drug counseling and GDC has promise for the treatment of cocaine dependence."

[3] Beyond concern for statistical significance, there may be other reasons why studies of treatment response report only aggregated findings. Researchers may think that treatment response is homogenous across patients; then reporting findings conditional on patient covariates has no value. Or concern for the confidentiality of subjects' identities may inhibit reporting covariate data. Or editorial restrictions on the lengths of journal articles may prevent researchers from reporting useful findings.

space constraints prevent publication of all potentially informative findings, researchers should report them on the internet or through other means.

5. Using Evidence to Inform Clinical Practice

The questionable evidential practices described in Section 4 have become prevalent because medical research has adhered to methodological guidelines developed in classical statistics, whose motivations are remote from the decision making concerns of clinical practice. Research on treatment response has used the statistical theory of randomized experiments pioneered by R. A. Fisher (Fisher, 1935), whose objectives are to test hypotheses regarding and estimate the magnitudes of treatment effects. Fisher's theory does not directly address the problem of treatment choice. Nevertheless, researchers and guideline developers have used it for that purpose.

Evidence-based studies of treatment response can better inform clinical practice if they seek to provide knowledge that promotes effective decision making. Section 2 formalized an optimal treatment rule as one that assigns each patient a treatment that maximizes expected utility conditional on the person's observed covariates. From this perspective, studies of treatment response are useful to the degree that they reveal how expected utility varies with treatments and covariates. It is unrealistic to think that evidence-based studies can provide all the information that clinicians would like to have. The task of methodological research should be to illuminate the information that different types of studies can credibly supply.

To begin, one should recognize that studies commonly experience both statistical imprecision and identification problems. Small sample sizes limit the precision of inference. Identification problems are the inferential difficulties that persist even when sample size grows without bound. I focus here on identification

problems, which typically are the dominant difficulty. I will discuss sampling imprecision from the perspective of treatment choice in Section 6.

5.1. Credible Identification Analysis

The unobservability of counterfactual treatment outcomes creates a fundamental identification problem when attempting to draw conclusions from observational studies, where treatment selection may be related to treatment response in an unknown way. Identification problems also complicate inference from trials, which typically do not attain the ideal that persons have in mind when they refer to them as the "gold standard" for research. As discussed in Section 4, the subjects in trials are volunteers who meet specific criteria and, hence, may not be representative of patient populations. Moreover, trials may have non-compliance, attrition, and measure surrogate outcomes rather than ones of real interest. An unfortunate characteristic of traditional empirical research on treatment response has been that it gives clinicians and guideline developers little sense of how identification problems limit inference. Whether the data are obtained from a trial or an observational study, researchers commonly report point estimates that may have fragile foundations.

In a research program studying identification with realistic data and credible assumptions, I have found that such research may yield informative bounds on treatment effects but not their precise values. Early contributions include Manski (1990, 1997) and Manski and Pepper (2000). Manski (2007) gives a broad textbook exposition. Manski (2013, 2017) and Horowitz and Manski (2000) study specific identification problems that arise in clinical practice. The research considers inference from observational studies as well as trials. I show that both can be informative to some degree.

Recall the formal description of optimal treatment choice given in Appendix A. My research has sought to characterize what one would learn about the vector $E[y(\cdot)|x] \equiv \{E[u(t)|x = \xi], t \in T, \xi \in X\}$ of

expected utilities that determine the optimal treatment rule if the sample sizes in available studies were to grow without bound and if the study evidence were to be combined with credible assumptions. The canonical finding is that the expected utilities would be revealed to lie in some informative set, called the *identification region*, but would not be pinned down precisely. Thus, the expected utilities are typically *partially identified* rather than *point-identified*.

The practical task of identification analysis is to characterize identification regions in a tractable manner, to enable clinicians and guideline developers to make use of the findings. The remainder of this section describes three cases, the first using data from a trial and the latter two using observational evidence.

5.2. Identification of Response to Treatments for Hypertension from a Trial with Missing Data

Horowitz and Manski (2000) analyzed identification of mean treatment response when a trial is performed but some outcome or covariate data are missing. We supposed that outcomes are binary and derived sharp bounds on $E[y(\cdot)|x]$ 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, described below.

Materson *et al.* (1993) presented findings from a trial comparing treatments for hypertension sponsored by the U.S. Department of Veteran Affairs (DVA). Male veteran patients at 15 DVA hospitals were randomly assigned to one of 6 antihypertensive drug treatments or to placebo: hydrochlorothiazide ($t = 1$), atenolol ($t = 2$), captopril ($t = 3$), clonidine ($t = 4$), diltiazem ($t = 5$), prazosin ($t = 6$), placebo ($t = 7$). The trial had two phases. In the first, the dosage that brought diastolic blood pressure (DBP) below 90 mm Hg was determined. In the second, it was determined whether DBP could be kept below 95 mm Hg for a long time. Treatment was defined to be successful if DBP < 90 mm Hg on two consecutive measurement

occasions in the first phase and DBP ≤ 95 mm Hg in the second. Treatment was deemed unsuccessful otherwise. Thus the outcome of interest was binary, with y = 1 if the criterion for success is met and y = 0 otherwise. Materson *et al.* (1993) recommended that clinicians treating hypertension should consider this medical outcome variable as well as patient's quality of life and the cost of treatment.

The Materson *et al*. (1993) article examined how treatment response varies with the race and age of the patient. There were no missing data on the race and age covariates. 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 outcome data either. Horowitz and Manski (2000) obtained the trial data and used them to examine how treatment response varies with another covariate that does have missing data. This was the biochemical indicator "renin response," taking the values x = (low, medium, high), which had previously been studied as a factor that might be related to successful treatment (Freis, Materson, and Flamenbaum 1983). Renin-response was measured at the time of randomization, but data were missing for some subjects in the trial. Horowitz and Manski also 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. The pattern of missing covariate and outcome data is shown in Table 1 of Horowitz and Manski (2000), reproduced here.

Table 1: 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

Horowitz and Manski (2000) used the identification analysis to estimate sharp bounds on the success probabilities $\{P[y(t) = 1 \mid x], t = 1, \ldots, 7\}$ without imposing assumptions on the distribution of missing data.

Rather than report the bounds on the success probabilities directly, the article reported the implied bounds on the average treatment effects $\{P[y(t) = 1 | x] - \{P[y(7) = 1 | x], t = 1, \ldots, 6\}$, which measure the efficacy of each treatment relative to the placebo. Table 2 shows the estimates of the bounds on the success probabilities themselves, which have previously been reported in Manski (2008).

Table 2: Bounds on Success Probabilities Conditional on Renin Response

Renin	Treatment						
Response	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]

To focus on the identification problem, suppose that the estimates are the actual bounds rather than finite-sample estimates. Observe that even though the findings are bounds rather than precise success probabilities, many bounds are sufficient narrow to enable one to conclude that certain treatments are dominated; that is, definitely inferior to others. For patients with low renin response, treatments 1, 2, 3, 4, 6, and 7 are all dominated by treatment 5, which has the greatest lower bound (.66). For patients with medium renin response, treatments 1, 3, 6, and 7 are dominated by treatment 5, which again has the greatest lowest bound (.68). For patients with high renin response, treatments 1, 6, and 7 are dominated by treatment 2, which has the greatest lowest bound (.64). Thus, without imposing any assumptions on the distribution of missing data, a clinician can reject treatments 1, 6, and 7 for all patients, reject treatment 3 for patients with medium renin response, and determine that treatment 5 is optimal for patients with low renin response.

5.3. Identification of Response to Diagnostic Testing and Treatment

Manski (2013) studies identification of response to diagnostic testing and treatment. I consider a common scenario in which a patient presents to a clinician, who obtains initial evidence on health status. The clinician may prescribe a treatment immediately or he may first order a diagnostic test that yields further evidence on health status. In the latter case, he prescribes a treatment after observation of the test result. 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?

Phelps and Mushlin (1988) studied this decision using an extension of the optimization framework described in Section 2. My analysis begins by observing that clinicians and guideline developers often do not have all of the knowledge needed to solve the optimization problem. I characterize the partial knowledge generated when one can observe a study population where decisions adhere to a common clinical practice. I then show how combining these observational data with various credible assumptions yields further knowledge but still not enough to solve the optimization problem. I summarize below.

5.3.1. Optimal Testing and Treatment

Phelps and Mushlin (1988) assumed that clinicians have rational expectations and maximize expected utility. The value of ordering a diagnostic test is that doing so reveals a patient covariate that the clinician would not observe otherwise, namely the test result. Appendix A showed that the expected value of information is necessarily non-negative and is positive if the result affects the optimal treatment. It follows that the clinician should always order the test if performing the test has no negative effect on patient utility. However, performing a test often does negatively affect utility. For example, biopsies, CT scans, and colonoscopies are invasive and expensive procedures. Hence, the test should be performed only if the

expected informational benefit outweighs the utility cost. Phelps and Mushlin recognized that the clinician faces a dynamic programming problem and they characterized the solution.

Manski (2013) poses a version of this problem as the prelude to identification analysis. As in Section 2, a clinician's objective is to maximize patients' expected utility. I assume that patients always comply with the clinician's decisions. Let x denote the initially observed covariates of a patient. There are two feasible treatment, t = A and t = B. Let s = 1 or 0 indicate whether the clinician orders the test. Let r denote the test result. I assume that the result is binary, r = p (positive) or r = n (negative). An example would be a biopsy, which either finds malignant cells (r = p) or does not (r = n).

The actions that the clinician may choose and the knowledge of patient covariates accompanying each action may be expressed as a decision tree. The clinician chooses s = 0 or s = 1with knowledge of x. If he chooses s = 0, he chooses t = A or t = B with knowledge of x. If he chooses s = 1, he chooses t = A or t = B with knowledge of (x, r). The clinician aggregates the benefits and harms of making a particular testing decision s and treatment choice t for a given patient into a scalar welfare measure y(s, t).

Let f(r|x) denote the fraction of patients with covariates x who would have test result r if they were to be tested. Let E[y(s, t)|x] be the mean welfare if all patients with covariates x were to receive (s, t). Let E[y(s, t)|x, r] be the mean welfare if all patients with covariates x and test result r were to receive (s, t). It can be shown that the optimal testing and treatment decisions are

(7a) Choose s = 1 if $\sum_{r \in \{p, n\}} f(r|x)[\max\{E[y(1, A)|x, r], E[y(1, B)|x, r]\}] \geq \max\{E[y(0, A)|x], E[y(0, B)|x]\}$,

choose s = 0 otherwise.

(7b) If s = 0, choose t = B if $E[y(0, B)|x] \geq E[y(0, A)|x]$ and choose t = A otherwise.

(7c) If s = 1 and r = p, choose t = B if $E[y(1, B)|x, p] \geq E[y(1, A)|x, p]$ and choose t = A otherwise.

(7d) If $s = 1$ and $r = n$, choose $t = B$ if $E[y(1, B)|x, n] \geq E[y(1, A)|x, n]$ and choose $t = A$ otherwise.

The treatment allocations in (7b-14d) are transparent. Whatever testing decision the clinician makes and whatever test result occurs, he should choose a treatment that maximizes expected utility conditional on the observed covariates and test result. The testing allocation in (7a) is more subtle. It is always optimal to test if testing never has a negative direct effect on patient utility; that is, if $y(1, t) \geq y(0, t)$ for all treatments and patients. A decision not to test can be optimal only if testing sometimes negatively affects utility. This can happen if testing is invasive, costly, or harms patients by delaying treatment. Criterion (7a) gives the circumstances in which the information value of testing outweighs its negative utility effect.

5.3.2. Partial Identification of Response to Testing and Treatment with Observational Data

Criteria (7) show that determination of the optimal testing and treatment decisions for a patient with initial covariates x requires sufficient knowledge of $f(r|x)$, $E[y(0, t)|x]$, and $E[y(1, t)|x, r]$ for $t \in \{A, B\}$ and $r \in \{n, p\}$. In principle, one might obtain this knowledge by performing a randomized trial. An ideal trial with four arms, one for each value of (s, t), would reveal the distribution of test results and treatment response. However, an ideal trial often is infeasible. Often the only available evidence are observational data generated by the testing and treatment decisions that occur in clinical practice.

I analyze identification of response to testing and treatment when $t = A$ is surveillance and $t = B$ is aggressive treatment. A common clinical 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 clinical practice *aggressive treatment with positive testing* (ATPT). I assume that the available evidence are observational data in a study population that adheres to the ATPT practice.

The identification problem arises from the unobservability of counterfactual testing and treatment outcomes. To focus attention on this core difficulty, I abstract from others that may arise in practice. I suppose that one observes the entire study population rather than a sample, that the study population has the same composition as the population to be treated, and that outcomes are bounded on [0, 1].

Recall that, to optimize care, the clinician wants to learn $\{E[y(0, t)|x], E[y(1, t)|x, r], f(r|x)\}$ for $t \in \{A, B\}$ and $r \in \{n, p\}$. Given the ATPT practice, the evidence reveals nothing about $E[y(0, B)|x]$, $E[y(1, B)|x, n]$, and $E[y(1, A)|x, p]$. This is because ATPT mandates $t = A$ when $s = 0$ and when $(s = 1, r = n)$, while it mandates $t = B$ when $(s = 1, r = p)$.

On the other hand, the evidence yields informative bounds on $E[y(0, A)|x]$, $E[y(1, A)|x, n]$, $E[y(1, B)|x, p]$, and $f(r|x)$. Let $z = 1$ if a person in the study population was tested and $z = 0$ if he was not. Define $g(n, x) \equiv f(r = n|x, z = 1)P(z = 1|x) + P(z = 0|x)$ and $g(p, x) \equiv f(r = p|x, z = 1)P(z = 1|x) + P(z = 0|x)$. The results are these bounds:

(8a) $E[y(0, A)|x, z = 0]P(z = 0|x) \leq E[y(0, A)|x] \leq E[y(0, A)|x, z = 0]P(z = 0|x) + P(z = 1|x)$.

(8b) $E[y(1, A)|x, n, z = 1]f(r = n|x, z = 1)P(z = 1|x)/g(n, x) \leq E[y(1, A)|x, n] \leq$
$\{E[y(1, A)|x, n, z = 1]f(r = n|x, z = 1)P(z = 1|x) + P(z = 0|x)\}/g(n, x)$.

(8c) $E[y(1, B)|x, p, z = 1]f(r = p|x, z = 1)P(z = 1|x)/g(p, x) \leq E[y(1, B)|x, p] \leq$
$\{E[y(1, B)|x, p, z = 1]f(r = p|x, z = 1)P(z = 1|x) + P(z = 0|x)\}/g(p, x)$.

(8d) $f(r = n|x, z = 1)P(z = 1|x) \leq f(r = n|x) \leq f(r = n|x, z = 1)P(z = 1|x) + P(z = 0|x)$.

A corresponding bound on $f(r = p|x)$ follows immediately from the one that holds for $f(r = n|x)$.

The above bounds are derived without making any assumptions about the outcomes of counterfactual testing and treatment decisions. I show that more can be learned if the observational data are combined with assumptions that may be credible in some settings. These results are summarized below.

Random Testing Assumption

Each of the five quantities that is partially identified with no assumptions about counterfactual outcomes becomes point identified if testing is random conditional on x. Random testing may occur through performance of a randomized trial of testing, which is not prohibited by the ATPT practice. Random testing implies these equalities:

(9a) $E[y(0, A)|x] = E[y(0, A)|x, z = 0]$.

(9b) $E[y(1, A)|x, n] = E[y(1, A)|x, n, z = 1]$.

(9c) $E[y(1, B)|x, p] = E[y(1, B)|x, p, z = 1]$.

(9d) $f(r = n|x) = f(r = n|x, z = 1)$.

The quantities on the right-hand side of these equations are observable. The assumption of random testing does not, however, help to identify $E[y(0, B)|x]$, $E[y(1, B)|x, n]$, and $E[y(1, A)|x, p]$.

Test Result as a Monotone Instrumental Variable

Patients with negative results on a diagnostic test are often thought to be healthier than ones with positive results. Hence, a clinician may find it credible to predict that patients with negative test results have better future prospects, on average, than do patients with positive results. Formally, the clinician may find it credible to assume $E[y(s, t)|x, n] \geq E[y(s, t)|x, p]$ for specified values of (s, t). Then test result is a *monotone instrumental variable* (MIV). See Manski and Pepper (2000).

The MIV assumption for (s, t) = (1, A) implies that E[y(1, A)|x, p] is no larger than the basic upper bound on E[y(1, A)|x, n] obtained without assumptions. If one also assumes random testing, then E[y(1, A)|x, p] is no larger than the known value of E[y(1, A)|x, n]. The MIV assumption for (s, t) = (1, B) implies that E[y(1, B)|x, n] is no smaller than the basic lower bound on E[y(1, B)|x, p]. If one also assumes random testing, then E[y(1, B)|x, n] is no smaller than the known value of E[y(1, B)|x, p].

Monotone Response to Testing

A clinician may believe that testing cannot directly improve welfare but may decrease it. For example, he may think that testing has no therapeutic effect but may be invasive or costly. Formally, the clinician may find it credible to assume the inequality $y(0, t) \geq y(1, t)$ for specified values of t and for all patients. This is a *monotone-treatment-response* (MTR) assumption. See Manski (1997).

The MTR assumption for t = B yields a lower bound on E[y(0, B)|x], the bound depending on what other assumptions are imposed. A simple finding emerges if one combines the MTR assumption and the MIV assumption for (s, t) = (1, B). Then E[y(0, B)|x] is no smaller than the basic lower bound on E[y(1, B)|x, p]. If one also assumes random testing, E[y(0, B)|x] is no smaller than the known value of E[y(1, B)|x, p].

5.4. Credible Ecological Inference for Personalized Medicine

Manski (2017) studies the identification problem faced by a clinician who observes more patient covariates than are used in evidence-based predictors of health outcomes. As in Section 2, suppose there exists an objectively correct evidence-based probabilistic prediction that conditions on patient covariates x. Moreover, a clinician observes further patient covariates w. Let y denote a patient outcome of interest, perhaps indicating whether the patient will develop a specified disease or remaining life span. Suppose the clinician want to choose a care option that maximizes expected utility conditional on the observed covariates.

To accomplish this, a clinician treating a patient with covariates (x = k, w = j) wants to know the "long" probability distribution P(y|x = k, w = j) that predicts outcomes conditional on this value of (x, w). However, the evidence-based predictor only reveals the "short" distribution P(y|x = k) that conditions just on x.

To understand the identification problem, I begin with the Law of Total Probability, which relates the short and long predictive distributions:

$$(10) \qquad P(y|x = k) = P(w = j|x = k)P(y|x = k, w = j) + P(w \neq j|x = k)P(y|x = k, w \neq j).$$

Knowledge of P(y|x = k) alone reveals nothing about P(y|x = k, w = j). Any distribution P(y|x = k, w = j) satisfies the equation if P(w = j|x = k) = 0. Partial conclusions may be drawn if one has evidence revealing P(y|x = k) and P(w = j|x = k), provided that the latter is positive. The problem of identification of P(y|x, w) given knowledge of P(y|x) and P(w|x) is called the *ecological inference* problem.

The basic version of the problem considers identification without structural assumptions that restrict P(y|x, w). Tighter conclusions may be drawn if one combines knowledge of P(y|x) and P(w|x) with such assumptions. Sections 5.4.1 summarizes findings on the former case, while Section 5.4.2 considers the latter.

5.4.1. Prediction without Structural Assumptions

The joint identification region for P(y|x = k, w = j) and P(y|x = k, w ≠ j) given knowledge of P(y|x) and P(w|x) is the set of pairs of long distributions that satisfy the Law of Total Probability (10). When y is binary, the identification region is the interval

$$(11) \quad P(y = 1|x = k, w = j) \in [0, 1]$$

$$\cap \left[\frac{P(y = 1|x = k) - P(w \neq j|x = k)}{P(w = j|x = k)}, \frac{P(y = 1|x = k)}{P(w = j|x = k)}\right].$$

This result was sketched by Duncan and Davis (1953). A proof is given in Horowitz and Manski (1995).

When y is real-valued, there is no simple characterization of the identification region for P(y|x, w), but Horowitz and Manski (1995) derive tractable expressions for the identification regions of the mean and quantiles of P(y|x, w). Manski (2017) uses prediction of life span to illustrate. I summarize here.

Predicting Life Span

A common problem in health risk assessment is to predict remaining life span conditional on observed patient covariates. Let y denote remaining life span. 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 covariates that clinicians commonly observe. Let x classify 50-year-old males into one of two races, non-Hispanic (NH) black or white. Let w classify persons into those with or without high blood pressure (HBP).

The life tables show that E(y|age 50, NH black male) = 26.6 and E(y|age 50, NH white male) = 29.7. Data in the National Health and Nutrition Examination Survey (NHANES) enable estimation of P(w|x). I use the age-aggregated probabilities P(HBP|NH black male) = 0.426 and P(HBP|NH white male) = 0.334. Combining the life table and NHANES data yields these sharp bounds on E(y|age, race, sex, blood pressure):

E(y|age 50, NH black male, not HBP) $\in$ [18.1, 35.4], E(y|age 50, NH black male, HBP) $\in$ [14.3, 38.5],
E(y|age 50, NH white male, not HBP) $\in$ [23.8, 36.4], E(y|age 50, NH white male, HBP) $\in$ [15.6, 42.0].

5.4.2. Prediction with Bounded-Variation Assumptions

Tighter predictions may be feasible with structural assumptions. The literature has developed approaches that impose strong assumptions which point-identify P(y|x, w). However, these typically lack credibility.

There is a substantial middle ground between making no structural assumptions and making assumptions strong enough to yield point identification. *Bounded-variation* assumptions flexibly 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. To illustrate, I continue prediction of life span.

Predicting Life Span with Bounded-Variation Assumptions

Assume that persons with HBP have lower life expectancy than those without HBP. Thus,

$$0 \leq E(y|\text{age 50, NH white male, not HBP}) - E(y|\text{age 50, NH white male, HBP}),$$

$$0 \leq E(y|\text{age 50, NH black male, not HBP}) - E(y|\text{age 50, NH black male, HBP}).$$

Also assume that black males have up to 2.5 years less life expectancy than white males conditional on blood pressure. That is,

$$0 \leq E(y|\text{age 50, NH white male, not HBP}) - E(y|\text{age 50, NH black male, not HBP}) \leq 2.5,$$

$$0 \leq E(y|\text{age 50, NH white male, HBP}) - E(y|\text{age 50, NH black male, HBP}) \leq 2.5.$$

Combining these assumptions with the bounds on E(y|x, w) that were obtained using only knowledge of P(y|x) and P(w|x) yields these bounded-variation bounds:

$$E(y|\text{age 50, NH black male, not HBP}) \in [29.4, 35.4], \quad E(y|\text{age 50, NH black male, HBP}) \in [14.7, 22.9],$$

$$E(y|\text{age 50, NH white male, not HBP}) \in [31.9, 36.4], \quad E(y|\text{age 50, NH white male, HBP}) \in [16.3, 25.4].$$

These bounds are highly informative. In particular, they reveal that the life expectancy of 50-year-old black males without HBP is at least (29.4 – 22.9 = 6.5) years higher than that of those with HBP. For 50-year-old white males, the corresponding disparity is also at least (31.9 – 25.4 = 6.5) years.

6. Reasonable Medical Decisions under Uncertainty

6.1. Recognizing Uncertainty

Section 3 cited extensive psychological research which concludes that clinicians have imperfect judgment. Sections 4 cited multiple questionable methodological practices in the evidence-based research that guideline developers rely on. The identification analysis cited in Section 5 shows that evidence from trials or observational studies combined with credible assumptions typically do not yield precise probabilistic predictions of patient outcomes but may yield informative bounds.

I conclude that it is often unrealistic to think that either clinicians or guideline developers have rational expectations regarding disease development and treatment outcomes. That is, they often do not have sufficient knowledge to make objectively correct probabilistic predictions conditional on observed patient covariates. Hence, they cannot determine optimal treatments in the sense of Section 2. Instead, they should view patient care as a problem of decision making under uncertainty.

Precedents for this conclusion can readily be found in the literature on medical decision making. For example, considering treatment of cancer, Mullins *et al.* (2010) observes that (p. 59): "there is considerable uncertainty surrounding the clinical benefits and harms associated with oncology treatments." Institute of Medicine (2013) calls attention to the assertion by the Evidence-Based Medicine Working Group that (p. 33):

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

Many CPGs use a rating system to rank the strength of recommendations by the certainty that they are correct. For example, the James *et al.* (2014) article summarizing guidelines for treatment of hypertension describes its rating system this way (p. 510):

> "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 guideline developers recognize uncertainty is that CPGs regularly change their recommendations as new research accumulates. To cite just one of numerous examples, a sequence of randomized trials over the past twenty years have improved knowledge regarding the usefulness of sentinel lymph node biopsy as a diagnostic test and completion lymph node dissection as a prophylactic treatment for potential visceral metastasis of melanoma (e.g., McGregor, 2013; Faries *et al.* 2017). Guidelines regarding these procedures have in the past and continue to evolve accordingly.

Curiously, recognition of uncertainty has not led guideline developers to examine patient care formally as a problem of decision making under uncertainty. In fact, the influential Institute of Medicine (2011) report on guideline development expresses skepticism about decision analysis, stating (p. 171):

> "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."

The report does not explain the basis for this assessment.

I find the IOM perspective surprising. The foundations of decision analysis were largely in place by the middle of the 20th Century. Applications to medical decision making have been performed since at

least the 1980s, albeit typically with rational expectations assumptions. Medical research makes much use of biological science, technology, and quantitative statistical methods. Why then should CPG development acknowledge uncertainty only verbally?

Formal analysis of patient care under uncertainty has much to contribute to guideline development and to clinician decision making. Section 6.2 reviews basic principles of decision theory. Sections 6.3 through 6.5 describe several recent applications to medical decision making.

6.2. Some Basic Decision Theory

The standard formalization of decision making under uncertainty supposes that a decision maker must choose among a set of feasible actions. The welfare achieved by any action depends on an unknown feature of the environment, called the *state of nature*. In the setting of Section 2, the decision maker is a clinician, the actions are the feasible alternative treatments for a patient, and welfare is the expected utility of a treatment. The decision maker lists all the states of nature that he believes could possibly occur. This list, called the *state space*, expresses partial knowledge. The larger the state space, the less the decision maker knows about the consequences of each action.

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.

Basic decision theory suggests a two-step decision process. The first and obvious step is to eliminate dominated treatments—an action is dominated if one knows that some other one is superior in all feasible states of nature. The second and more subtle step is to choose an undominated action. This is subtle because there is no optimal way to choose among undominated alternatives. There are only various reasonable ways.

6.2.1. Decision Criteria

What are reasonable ways to make an undominated choice? Perhaps best known is Bayesian decision theory, which recommends that one place a subjective probability distribution on unknowns and maximize subjective expected utility. The Bayesian perspective is compelling when one feels able to place a credible subjective distribution on the state space. However, Bayesian statisticians have long struggled to provide guidance on specification of priors and the matter continues to be controversial. See, for example, the spectrum of views regarding Bayesian analysis of randomized trials 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 achieves uniformly satisfactory results, whatever the true state of nature may be. There are multiple ways to formalize the idea of uniformly satisfactory results. The two most commonly studied are the maximin and minimax-regret (MR) criteria.

The maximin criterion chooses an action that maximizes the minimum welfare that might possibly occur. The minimax regret criterion considers each state of nature and computes the loss in welfare that would occur if one were to choose a specified action rather than the one that is best in this state. This quantity, called *regret*, measures the nearness to optimality of the specified action in the state of nature. The decision maker must choose without knowing the true state. To achieve a uniformly satisfactory result, he computes the maximum regret of each action; that is, the maximum distance from optimality that the action would yield across all possible states of nature. The MR criterion chooses an action that minimizes this maximum distance from optimality.

The maximin and MR criteria are sometimes confused with one another, but they yield the same choice only in certain special cases. The former chooses an action that maximizes the minimum welfare that might possibly occur. The latter chooses an action that minimizes the maximum loss to welfare that can

possibly result from not knowing the welfare function. Thus, whereas the maximin criterion considers only the worst outcome that an action may yield, MR considers the worst outcome relative to what is achievable in a given state of nature. Savage (1951) distinguished the maximin criterion sharply from MR, writing that the former criterion is "ultrapessimistic" while the latter is not.

6.2.2. Statistical Decision Theory

The above description of decision criteria suffices when uncertainty stems purely from identification problems, but an extension is necessary when one uses sample data to inform decision making. Then one chooses an action contingent on the data that are observed.

The Wald (1950) development of statistical decision theory considers the decision problem ex ante, before the data are observed. Then the decision maker's task is to select a *statistical decision function*; that is, a rule specifying how the chosen action will vary with the data. Wald proposed evaluation of statistical decision functions by their mean performance across repetitions of the sampling process. This grounds the Wald theory in frequentist statistical thinking. See Ferguson (1967) and Berger (1985) for comprehensive expositions. When statistical decision theory has been apply to the problem of treatment choice, a statistical decision function has been called a *statistical treatment rule* (Manski, 2004).

Statistical decision theory may be used to study the decision criteria described in Section 6.2.1. In each case, one evaluates a criterion by the mean welfare that it yields across repeated samples. Bayes decisions contingent on sample data are often studied without reference to the Wald framework, but they are subsumed within it when one views a Bayesian ex ante as someone who uses a particular sample-dependent decision rule.

6.3. Clinical Decision Making Recognizing the Ecological Inference Problem

To illustrate patient care under uncertainty stemming from an identification problem, consider again the ecological inference problem described in Section 5.4. The basic lesson was that a clinician who observes more covariates than are used in evidence-based predictors may draw credible partial conclusions about the long outcome distribution P(y|x, w) but not learn it precisely. How might such a clinician reasonably choose patient care?

To address this question in a specific setting, Manski (2017) considers the setting of Section 2.3.1. To recall, the choice is between surveillance and aggressive treatment. Aggressive treatment prevents disease. Given a patient with covariates (x, w), surveillance is optimal when $P_{xw}(A) \leq P^*_{xw}(A)$ and aggressive treatment is optimal when $P_{xw}(A) \geq P^*_{xw}(A)$, where $P^*_{xw}(A)$ is the threshold probability defined in (4).

Consider decision making when a clinician does not know $P_{xw}(A)$ but can use available evidence and credible assumptions to conclude that $P_{xw}(A) \in [P_{xwL}, P_{xwH}]$, where P_{xwL} and P_{xwH} are known lower and upper bounds. The clinician can still maximize expected utility if $P^*_{xw}(A)$ is not interior to $[P_{xwL}, P_{xwH}]$. Then t = A is sure to be optimal if $P_{xwH} \leq P^*_{xw}(A)$ and t = B is sure to be optimal if $P^*_{xw}(A) \leq P_{xwL}$. However, he cannot maximize expected utility if $P^*_{xw}(A)$ is interior to $[P_{xwL}, P_{xwH}]$. Then there exist feasible values of $P_{xw}(A)$ that make only A optimal and other values that make only B optimal.

6.3.1. Treatment Choice with Alternative Decision Criteria

The Bayesian prescription places a subjective distribution on $P_{xw}(A)$ and maximizes subjective expected utility. Let π_{xw} denote the subjective mean that a Bayesian clinician holds for $P_{xw}(A)$. A Bayesian clinician acts as if $P_{xw}(A) = \pi_{xw}$.

The maximin criterion evaluates each action by the worst expected utility that it may yield and it chooses an action with the least-bad worst expected utility. The worst feasible expected utilities under

options A and B occur when $P_{xw}(A)$ equals its upper bound P_{xwH}. Hence, the clinician acts as if $P_{xw}(A) = P_{xwH}$. The maximin choice is A if $P_{xwH} \leq P^*_{xw}(A)$ and B if $P_{xwH} \geq P^*_{xw}(A)$.

The minimax-regret criterion evaluates each action by the worst reduction in expected utility that it may yield relative to the highest expected utility achievable. Let P_{xwM} denote the midpoint of the interval $[P_{xwL}, P_{xwH}]$. Manski (2017) shows that the MR choice is the same as a clinician maximizing expected utility would make if he were to know that the probability of illness is $P_{xw}(A) = P_{xwM}$.

6.3.2. Rethinking Care with Evidence-Based Prediction

The psychological literature on clinical judgment does not recommend any of the decision criteria discussed here. It recommends that the clinician suppress knowledge of patient covariates w and act as if $P_{xw} = P(y = 1|x)$. This is inappropriate if $P(y = 1|x)$ does not lie in the interval $[P_{xwL}, P_{xwH}]$. The recommendation of the psychological literature is rationalizable if $P(y = 1|x)$ is a possible value of $P_{xw}(A)$. However, one could similarly recommend acting as if $P_{xw}(A)$ is any element of $[P_{xwL}, P_{xwH}]$.

Decision making with the maximin or MR criterion is equivalent to acting as if $P_{xw}(A)$ takes particular values in $[P_{xwL}, P_{xwH}]$, namely P_{xwH} for maximin and P_{xwM} for MR. Singling out these values has a firmer justification because they yield choices that are uniformly satisfactory in the maximin or MR sense.

My negative conclusion regarding acting as if $P_{xw} = P(y = 1|x)$ does not contradict the conclusion of psychological research that evidence-based prediction outperforms clinical judgment. Psychologists may be correct that clinicians fail to grasp the logic of the prediction problem, generating an empirical finding in favor of evidence-based prediction. What the analysis does suggest is that it may be possible to improve on both evidence-based prediction and subjective clinical judgment by formalizing clinical judgment.

6.4. Minimax-Regret Treatment Choice with Trial Data

To illustrate patient care under uncertainty stemming from sampling imprecision, I discuss research studying use of the MR criterion to choose treatments with data from a classical randomized trial. By a classical trial, I mean one that has none of the extrapolation problems discussed in Section 3.1. Rather, the trial yields precisely the type of data that one would like to have to predict patient outcomes under alternative treatments. The only difficulty is imprecision because the trial sample size is finite. I first discuss research on treatment choice using existing trial data and then work on choice of sample size when designing trials.

6.4.1. Treatment Choice Using Existing Trial Data

Modern study of MR treatment choice using trial data was initiated in Manski (2004) and developed further in Schlag (2006) and Stoye (2009, 2012) inter alia. Common to this body of work is the supposition that the decision maker's objective is to maximize a welfare function that sums treatment outcomes across the population, as in Section 2. For example, the objective may be to maximize the five-year survival rate in a population of cancer patients or mean life span in a population with a chronic disease.

The MR criterion is applicable in general settings with multiple treatments, but it is easiest to explain when there are two treatments, say A and B. Consider a state of nature in which treatment A is better. The regret (that is, nearness to optimality) of a specified treatment rule in this state is the product of the probability across repeated samples that the rule commits a Type I error (choosing B) and the magnitude of the loss in welfare that occurs when choosing B. Similarly, in a state where treatment B is better, regret is the probability of a Type II error (choosing A) times the magnitude of the loss in welfare when choosing A.

Recall the critique in Section 4.2 of the conventional use of hypothesis testing to choose a treatment. I called attention to the asymmetric attention to Type I and Type II error probabilities and the inattention to magnitudes of losses when errors occur. Evaluating treatment rules by regret overcomes both problems.

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

Research on MR treatment choice has shown that, in general, the statistical treatment rule that minimizes maximum regret must be computed numerically. However, there are good practical and analytical reasons to focus attention on the *empirical success*(ES) rule, which chooses the treatment with the highest reported average outcome in the trial. The practical appeal is that the ES rule is a simple and plausible way to use the results of a trial. The analytical reason is that the ES rule has been shown to either exactly or approximately minimize maximum regret in various empirically common settings with two treatments when sample size is moderate (Stoye, 2009, 2012).

6.4.2. Designing Trials to Enable Near-Optimal Treatment Choice

From the perspective of 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 of interest; that is, a rule for use of trial data that always selects the best treatment, with no chance of error. Optimality is too strong a property to be achievable with finite sample size but near-optimal rules—ones with small maximum regret—exist when classical trials are large enough.

Manski and Tetenov (2016) investigate 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 within ε of the welfare of the best treatment in every state of nature. Equivalently, it has maximum regret no larger than ε.

We consider trials that draw predetermined numbers of subjects at random within groups stratified by covariates and treatments. We report exact results for the special case of two treatments and binary outcomes. 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. These conditions are obtained by application of large deviations inequalities to evaluate the performance of empirical success rules.

6.5. Error Limitation and Learning by Adaptive Diversification of Treatment

I observed at the outset that a prominent argument for adherence to CPGs has been to reduce "unnecessary" or "unwarranted" variation in clinical practice. The meaning of these adjectives is clear in the rational expectations setting of Section 2, where optimization of patient care is feasible. An attribute of optimal care is that all patients with the same observed covariates receive the same treatment. Hence, variation in the care of observationally similar patients is sub-optimal.

The argument for uniform treatment of similar patients loses its potency when clinicians choose patient care under uncertainty. I observed above that there is no uniquely optimal choice among undominated actions. Uncertainty implies clinical equipoise, so treatment variation is consistent with medical ethics. Different clinicians may reasonably interpret the available evidence in different ways and may reasonably use different decision criteria to choose treatments. Thus, there is no prima facie reason to view treatment variation as unnecessary or unwarranted.

Manski (2007, 2009) uses decision theory to show that random variation in treatment of observationally similar patients is valuable under uncertainty. I develop this conclusion by considering patient care as a public health problem rather than one of treating individual patients. I consider decision making by a health planner who treats a population of patients. I show that two motives—diversification and learning—encourage a planner to randomize the treatment of observationally similar patients.

Financial diversification is a familiar recommendation for portfolio allocation. 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.' Analogously, treatment is diversified if a health planner randomly assigns observationally similar patients to different treatments. Treatment diversification enables a planner to avoid gross errors that might occur if all patients were inadvertently given an inferior treatment.

Diversification motivates random variation in clinical practice at a given point in time. Over time, such variation is even more useful because it effectively yields a population-wide trial that yields new evidence about treatment response. As time passes and evidence accumulates, a planner can revise the fraction of patients assigned to each treatment in accord with the available knowledge. I have called this idea *adaptive diversification*. Section 6.5.1 summarizes my decision theoretic argument for diversification. Section 6.5.2 discusses learning.

6.5.1. Minimax-Regret Diversification with Two Treatments

Classical decision theoretic analysis of financial portfolio allocation shows that an investor seeking to maximize expected utility chooses to diversify if utility is a sufficiently concave function of the investment return and the probability distribution of returns has sufficient spread. Treatment diversification by a Bayesian health planner can be studied in the same manner. Instead, Manski (2007, 2009) approach the health planner's problem from the minimax-regret perspective. The central result is that when there are two undominated treatments, the planner always chooses to diversify. The specific fraction of patients assigned to each treatment depends on the available knowledge of treatment response.

Consider patients who have observed covariates x. The planner's task is to allocate these patients between the treatments, say A and B. A treatment allocation is a $\delta_x \in [0, 1]$ that randomly assigns a fraction δ_x of these patients to treatment B and the remaining $1 - \delta_x$ to treatment A. Let $\alpha_x \equiv E[u(A)|x]$ and $\beta_x \equiv E[u(B)]$ be expected utility if all patients were to receive treatment A or B respectively. The optimal treatment allocation is $\delta_x = 1$ if $\beta_x \geq \alpha_x$ and $\delta_x = 0$ if $\beta_x \leq \alpha_x$.

The problem of interest is treatment choice when (α_x, β_x) is not known. To formalize the problem, let S index the feasible states of nature. Let the planner know that (α_x, β_x) lies in the set $[(\alpha_{xs}, \beta_{xs}), s \in S]$. This identification region is the set of values that the planner concludes are feasible when he combines available empirical evidence with assumptions he finds credible to maintain.

Partial knowledge is unproblematic for decision making if $(\alpha_{xs} \geq \beta_{xs}, s \in S)$ or if $(\alpha_{xs} \leq \beta_{xs}, s \in S)$; $\delta_x = 0$ is optimal in the former case and $\delta_x = 1$ in the latter. However, all $\delta_x \in [0, 1]$ are undominated if $\alpha_{xs} > \beta_{xs}$ for some values of s and $\alpha_{xs} < \beta_{xs}$ for other values. I consider this situation. The analysis is applicable whenever $[(\alpha_{xs}, \beta_{xs}), s \in S]$ is bounded. Denote the extreme feasible values as $\alpha_{xL} \equiv \min_{s \in S} \alpha_{xs}$, $\beta_{xL} \equiv \min_{s \in S} \beta_{xs}$, $\alpha_{xU} \equiv \max_{s \in S} \alpha_{xs}$, and $\beta_{xU} \equiv \max_{s \in S} \beta_{xs}$.

The regret of allocation δ_x in state of nature s is the difference between the maximum achievable welfare and the welfare achieved with this allocation. Maximum welfare in state s is $\max(\alpha_{xs}, \beta_{xs})$. Hence, δ has regret $\max(\alpha_{xs}, \beta_{xs}) - [\alpha_{xs} + (\beta_{xs} - \alpha_{xs})\delta]$. The minimax-regret criterion computes the maximum regret of each allocation over all states and chooses one to minimize maximum regret. Thus, the criterion is

$$(12) \quad \min_{\delta_x \in [0, 1]} \max_{s \in S} \max(\alpha_{xs}, \beta_{xs}) - [\alpha_{xs} + (\beta_{xs} - \alpha_{xs})\delta_x].$$

Manski (2007, 2009) proves that the MR criterion always diversifies treatment when the optimal treatment is not known. Let $S_x(A)$ and $S_x(B)$ be the subsets of S on which treatments A and B are superior; that is, $S_x(A) \equiv \{s \in S: \alpha_{xs} > \beta_{xs}\}$ and $S_x(B) \equiv \{s \in S: \beta_{xs} > \alpha_{xs}\}$. Let $M_x(A) \equiv \max_{s \in Sx(A)} (\alpha_{xs} - \beta_{xs})$ and $M_x(B) \equiv \max_{s \in Sx(B)} (\beta_{xs} - \alpha_{xs})$ be maximum regret on $S_x(A)$ and $S_x(B)$ respectively. The general result is

$$(13) \quad \delta_{xMR} = \frac{M_x(B)}{M_x(A) + M_x(B)}.$$

This is a diversified allocation because $M_x(A) > 0$ and $M_x(B) > 0$.

Expressions $M_x(A)$ and $M_x(B)$ simplify when $(\alpha_{xL}, \beta_{xU})$ and $(\alpha_{xU}, \beta_{xL})$ are feasible values of (α_x, β_x), as is so when the identification region is rectangular. Then $M_x(A) = \alpha_{xU} - \beta_{xL}$ and $M_x(B) = \beta_{xU} - \alpha_{xL}$. Hence,

$$(14) \quad \delta_{xMR} = \frac{\beta_{xU} - \alpha_{xL}}{(\alpha_{xU} - \beta_{xL}) + (\beta_{xU} - \alpha_{xL})}.$$

6.5.2. Adaptive Diversification

Now consider a health 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 treatment decisions and use this evidence to inform treatment later on. Diversification is advantageous for learning treatment response because it generates randomized experiments. As time passes and 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. Formally, increasing knowledge means that the state space S shrinks over time as evidence accumulates.

The AMR criterion is normatively appealing because it 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 the benefit of future cohorts. Nevertheless, the diversification of treatment performed for the benefit of the current cohort enables learning about treatment response.

The fractional allocations produced by the AMR criterion are randomized experiments, so it is natural to ask how application of AMR differs from the current design of trials. There are important differences in the fraction and composition of the population randomized into treatment. The AMR criterion randomizes treatment of all observationally similar patients. In contrast, the treatment groups in trials are typically small fractions of the patient population. For example, in trials conducted to obtain Food and Drug Administration approval of new drugs, treatment groups usually comprise no more than two to three thousand persons, whereas the patient population may contain hundreds of thousands or millions of persons. Moreover, as discussed in Section 4, trials draw subjects from pools of persons who volunteer to participate and who meet specific conditions, such as the absence of co-morbidities. Hence, trials at most reveal the distribution of treatment response within certain sub-populations of patients, not within the full population.

6.5.3. Should Guidelines Encourage Treatment Variation under Uncertainty?

Controlled implementation of adaptive diversification may be possible in centralized health care systems where there exists a planning entity who chooses treatments for a broad patient population. Examples are the Military Health System in the United States, the National Health Service in the United Kingdom, and some private health maintenance organizations. However, it would appear difficult to achieve in the American health care system, where clinicians make individual treatment decisions. Clinicians may not be willing to intentionally randomize treatment except in trials, even though uncertainty implies clinical equipoise and so makes randomization consistent with medical ethics.

Suppose that controlled adaptive diversification is not feasible in the American context. We can nevertheless question whether the medical community should continue to discourage treatment variation when treatment response is uncertain. Instead, 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 error-limitation and

learning benefits of diversification. I am not certain whether CPGs should actively encourage treatment variation under uncertainty, but I think the idea warrants consideration.

Appendix A: Optimal Personalized Treatment

A.1. The Choice Set

Let T be a set of feasible treatments. Suppose that a decision maker, called a planner for short, must choose a rule assigning a treatment to each member of a population of patients. The planner observes certain covariates $x_j \in X$ for each member j of the population. The covariate space X has finitely many elements and $P(x = \xi) > 0$ for all $\xi \in X$.

Suppose that the treatment chosen for one patient does not constrain the treatments available to other patients nor affect their outcomes. The planner can differentiate patients with different covariate values, but he cannot distinguish among patients with the same observed covariates. Then a feasible treatment rule either assigns all patients with the same observed covariates to one treatment or fractionally allocates these patients across treatments in a random manner. Formally, a feasible treatment rule is a function $\delta(\cdot, \cdot)$ that maps $T \times X$ into the unit interval and whose values sum to one across the elements of T; that is, $\sum_{t \in T} \delta(t, \xi) = 1$ for all $\xi \in X$. Let Δ denote the space of all such functions. The planner's choice set is Δ.

A subclass of Δ are the *singleton rules* that assign all patients with the same observed covariates to one treatment. Thus, $\delta(\cdot, \cdot)$ is a singleton rule if, for each $\xi \in X$, $\delta(t, \xi) = 1$ for some $t \in T$ and $\delta(s, \xi) = 0$ for all $s \neq t$. Non-singleton fractional rules randomly allocate patients with covariates ξ across multiple treatments, with assignment shares $[\delta(t, \xi), t \in T]$.

A.2. The Welfare Function and Optimal Treatment Rule

Let $u_j(t)$ denote patient welfare from assigning treatment t to patient j. A utilitarian welfare function sums up the contributions to welfare of all treatment decisions. Maximizing such a welfare function is equivalent to maximizing mean welfare in the population. For each treatment rule $\delta(\cdot, \cdot)$, the population mean welfare that would be realized if the planner were to choose rule $\delta(\cdot, \cdot)$ is

$$(A1) \quad U(\delta, P) \equiv \sum_{\xi \in X} P(x = \xi) \sum_{t \in T} \delta(t, \xi) \cdot E[u(t)|x = \xi].$$

Here $U(\delta, P)$ denotes social welfare when treatment rule $\delta(\cdot, \cdot)$ is applied to a population with distribution P of treatment response. The expression $E[u(t)|x = \xi]$ is the mean welfare realized when patients with covariates ξ receive treatment t. The fraction of the population with covariates ξ and treatment t is $P(x = \xi)\delta(t, \xi)$. The double summation on the right-hand-side of (A1) aggregates welfare across patients with different values of ξ and t.

The planner wants to solve the optimization problem

$$(A2) \quad \max_{\delta \in \Delta} \quad U(\delta, P).$$

Let S denote the unit simplex in $R^{|T|}$. The maximum is achieved if, for each $\xi \in X$, the planner chooses $\delta(\cdot, \xi)$ to solve the problem

$$(A3) \quad \max_{\delta(\cdot, \xi) \in S} \quad \sum_{t \in T} \delta(t, \xi) \cdot E[u(t)|x = \xi].$$

The maximum in (A3) is achieved by a singleton rule that allocates all patients with covariates ξ to a treatment that solves the problem

$$\text{(A4)} \qquad \max_{t \in T} E[u(t)|x = \xi].$$

Thus, the optimal treatment rule assigns all patients with covariates ξ to a treatment that yields the highest within-group mean welfare. It follows from (A1) and (A4) that the welfare achieved by an optimal rule is

$$\text{(A5)} \qquad \sum_{\xi \in X} P(x = \xi) \max_{t \in T} E[u(t)|x = \xi].$$

A.3. The Expected Value of Information

Suppose that, in addition to observing patient covariates x, the planner also observes additional covariates $w_j \in W$ for each member j of the population, where W has finitely many elements and $P(w = \omega) > 0$ for all $\omega \in W$. Repeating the derivation of Section 2.2.1shows that the welfare achieved by an optimal rule now is

$$\text{(A6)} \qquad \sum_{(\xi, \omega) \in X \times W} P(x = \xi, w = \omega) \max_{t \in T} E[u(t)|x = \xi, w = \omega].$$

The difference (A6) − (A5) is sometimes called the *expected value of information* (EVI), defined succinctly by Meltzer (2001) as (p. 119) "the change in expected utility with the collection of information."

The EVI is always non-negative and is positive if the optimal treatment varies with w conditional on at least one value of x. This implies that, assuming clinicians have rational expectations, decentralized

practice necessarily yields welfare greater than or equal to that achievable by adherence to CPGs that condition their recommendations on fewer patient covariates. To show the result, use the Law of Iterated Expectations and Bayes Theorem to rewrite (A5) and (A6) as

$$(A5') \qquad \sum_{\xi \in X} P(x = \xi) \max_{t \in T} \sum_{\omega \in W} E[u(t)|x = \xi, w = \omega] \cdot P(w = \omega|x = \xi).$$

$$(A6') \qquad \sum_{\xi \in X} P(x = \xi) \sum_{\omega \in W} P(w = \omega|x = \xi) \max_{t \in T} E[u(t)|x = \xi, w = \omega].$$

Jensen's Inequality implies that, for every value of x,

$$(A7) \qquad \sum_{\omega \in W} P(w = \omega|x = \xi) \max_{t \in T} E[u(t)|x = \xi, w = \omega] \geq \max_{t \in T} \sum_{\omega \in W} E[u(t)|x = \xi, w = \omega] \cdot P(w = \omega|x = \xi).$$

The inequality is strict if the optimal treatment varies with w conditional on the value of x.

References

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

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

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

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.

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

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.

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

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

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

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.

Ginsburg, G. and H. Willard (2009), "Genomic and Personalized Medicine: Foundations and Applications," *Translational Research*, 154, 277-287.

Goldberg, L. (1968), "Simple Models or Simple Processes? Some Research on Clinical Judgments," *American Psychologist*, 23, 483-496.

Groves, W., D. Zald, B. Lebow, B. Snitz, and C. Nelson (2000)," Clinical Versus Mechanical Prediction: A Meta-Analysis," *Psychological Assessment*, 12, 19-30.

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.

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 Covariate 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.

Kadane, J., M. Shervish, and T. Seidenfeld (2008), "Is Ignorance Bliss?" *Journal of Philosophy*, 105, 5-36.

Manski, C. (1990), "The Use of Intentions Data to Predict Behavior: A Best Case Analysis," *Journal of the American Statistical Association*, 85, 934-940.

Manski, C. (1990), "Nonparametric Bounds on Treatment Effects," *American Economic Review Papers and Proceedings*, 80, 319-323.

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. (2007), *Identification for Prediction and Decision,* Cambridge: Harvard University Press.

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. (2013), *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. (2017), "Credible Ecological Inference for Medical Decisions with Personalized Risk Assessment," Department of Economics, Northwestern University.

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.

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," *The 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.

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.

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 Comprehensive Cancer Network (2017), *Breast Cancer Screening and Diagnosis*, Version 1.2017, www.nccn.org/professionals/physician_gls/pdf/breast-screening.pdf, login required, accessed August 19, 2017.

National Health Service (2015), *The NHS Atlas of Variation in Healthcare*, http://fingertips.phe.org.uk/profile/atlas-of-variation, accessed 12 May 2017.

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.

Phelps, C. and A. Mushlin (1988), "Focusing technology assessment using medical decision theory," *Medical Decision Making*, 8, 279-289.

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.

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.

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., L. Freedman, and M. Parmar (1994), "Bayesian Approaches to Randomized Trials" (with discussion), *Journal of the Royal Statistics Society Series A*, 157, 357-416.

Stoye, J. (2009), "Minimax Regret Treatment Choice with Finite Samples," *Journal of Econometrics*, 151, 70-81.

Stoye, J. (2012), Minimax Regret Treatment Choice with Covariates or with Limited Validity of Experiments," *Journal of Econometrics*, 166, 138-156.

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.