Northwestern | INSTITUTE FOR POLICY RESEARCH

WP-17-21

Reasonable Patient Care Under Uncertainty

Charles F. Manski
Board of Trustees Professor in Economics
IPR Fellow
Northwestern University

Version: February 2018
Forthcoming in *Health Economics*

DRAFT

ABSTRACT

This paper discusses how limited ability to predict illness and treatment response may affect the welfare achieved in patient care. The discussion covers both decentralized clinical decision making and care that adheres to clinical practice guidelines. Manski explains why predictive ability has been limited, calling attention to questionable methodological practices in the research that supports evidence-based medicine. He summarizes research on identification whose objective is to yield credible prediction of patient outcomes. Recognizing that uncertainty will continue to afflict medical decision making, Manski applies basic decision theory to suggest reasonable decision criteria with well-understood welfare properties. Previous research on medical decision making has largely embraced Bayesian decision theory. He summarizes research studying the minimax-regret criterion, which seeks uniformly near-optimal decisions.

1. Introduction

This paper discusses how limited ability to predict illness and treatment response may affect the welfare achieved in patient care. The discussion covers both decentralized clinical decision making and care that adheres to clinical practice guidelines. I explain why predictive ability has been limited, calling attention to questionable methodological practices in the research that supports evidence-based medicine. I summarize research on identification whose objective is to yield credible prediction of 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. Previous research on medical decision making has largely embraced Bayesian decision theory. I summarize research studying the minimax-regret criterion, which seeks uniformly near-optimal decisions.

To provide a baseline for consideration of care under uncertainty, I first address decision making when clinicians have substantial knowledge of patient outcomes. 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 when patients with these covariates are given alternative treatments, and choose treatments that maximize patients' expected utility. Treatment response is assumed to be individualistic.

In this setting, the optimal treatment rule divides patients into groups having the same observed covariates. All patients in a group are given the care that yields the highest within-group mean welfare. Mean welfare increases as more covariates are observed. To illustrate, I derive the optimal rule when the problem is choice between surveillance and aggressive treatment of patients at risk of developing a disease.

Commentators on patient care often exhort clinicians to adhere to clinical practice guidelines (CPGs). Adherence to a CPG cannot outperform decentralized care if treatment response is individualistic and utilitarian clinicians have rational expectations. If a CPG conditions its 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. CPGs often condition their recommendations on a subset of the observable patient covariates. I illustrate with guidelines for breast cancer screening.

Section 3 considers patient care when clinicians have imperfect judgment. Empirical psychological research has 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, even when clinicians condition their judgments on patient covariates not used in statistical predictors.

I summarize the psychological research and consider its implications for welfare comparison of adherence to CPGs and decentralized care. 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. It may be superior to the extent that imperfect clinical judgment generates sub-optimal clinical decisions.

Section 4 challenges the accuracy of evidence-based predictions of patient outcomes, calling attention to multiple questionable methodological practices in research on health outcomes. Of particular concern 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.

Moving from critique to prescription, section 5 shows how evidence-based research can inform patient care more effectively than it does at present. Studies should quantify how identification problems

and statistical imprecision affect credible prediction of health outcomes. Focusing on identification, I summarize research on partial identification of risk assessment and treatment response. This includes work on identification of response to treatments in trials with missing data, on identification of response to diagnostic testing and treatment, and on credible ecological inference for personalized risk assessment.

Recognizing that medical knowledge is incomplete, Section 6 considers patient care as a problem of decision making under uncertainty. By "uncertainty," I do not merely mean that clinicians make probabilistic rather than deterministic predictions of patient outcomes. My concern is decision making when, as a result of identification problems and statistical imprecision, the available credible evidence and medical knowledge do not suffice to yield precise probabilistic predictions.

For example, a patient may ask her clinician a seemingly straightforward question 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 the clinician may not be able to provide precise answer to these questions. It may be that a credible response is a range, say "20 to 40 percent" or "at least 50 percent." Decision analysts sometimes use the terms "deep uncertainty" and "ambiguity" to describe such scenarios, but I simply use "uncertainty."

To lay foundations, I first discuss basic concepts of decision theory. 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. Study of the minimax-regret criterion is especially fruitful, because maximum regret is interpretable as uniform nearness to optimality.

I summarize work that uses the minimax-regret criterion to address sampling imprecision when choosing treatments with data from existing randomized trials and when choosing sample size for new trials. In both contexts, the research yields tractable decision rules with transparent welfare properties. In the trial design context, analytical findings complemented by numerical calculations yield a broad conclusion that sample sizes which make maximum regret reasonably small tend to be smaller than ones set using

conventional statistical power criteria.

I also reassess the widespread admonition of CPG advocates to reduce "unwarranted variation" in clinical practice. Viewing treatment choice from a public health perspective, my research shows that a health planner using the minimax-regret criterion to allocate a population to two treatments under uncertainty always chooses to diversify. 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.

Some readers may correctly view this paper as critical of methodologies they have advocated. These include biostatisticians who have used the theory of hypothesis testing to advise clinical researchers on the design and analysis of randomized trials. They include researchers and guideline developers who have argued that evidence-based medicine should rest either solely or predominately on evidence from randomized trials, disregarding or downplaying evidence from observational studies. They also include decision analysts who have argued that patient care should always apply Bayesian decision theory. I hope that these readers will make the effort to understand the bases for my criticisms and that they will view the applications of econometrics and decision theory discussed here as constructive suggestions.

2. Optimal Personalized Care Assuming Rational Expectations

2.1. Degrees of Personalized Medicine

The term *personalized medicine* is sometimes defined to mean health care that is literally specific

to the individual, as in this definition by Ginsburg and Willard (2009, p. 278), adopted by American Medical Association (2010): "Personalized medicine is . . . . . health care that is informed by each person's unique clinical, genetic, genomic, and environmental information." Yet evidence to support complete personalization is rarely available. Hence, the term is commonly used to mean care that varies with some individual characteristics. President's Council of Advisors on Science and Technology (2008, p. 7) states:

> " '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. Clinicians classify patients into groups based on observed medical history and the results of screening and diagnostic tests. Personalized probabilities of disease development or treatment outcomes depend on the patient covariates used to condition the predictions.

2.2. Optimal Personalized Care

When studying optimal personalized care, medical economists typically assume that clinicians want to maximize a utilitarian welfare function; that is, one respecting patient preferences. They also usually assume that treatment is individualistic; that is, the care received by one patient affects does not affect other members of the population. This assumption is realistic when considering non-infectious diseases. I maintain both assumptions throughout this paper, but recognize that some circumstances warrant study of alternatives.

When treatment is individualistic and welfare is utilitarian, optimization of care has a 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 treatments 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 are often stated without attribution. I do not know who first proved the basic results, but a relatively early version is given in Good (1967). The results have been applied in the economic literature on medical decision making by Phelps and Mushlin (1988), Claxton (1999), Meltzer (2001), Basu and Meltzer (2007), and Manski (2013a), among others.

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

To illustrate, I characterize optimal personalized care when the problem is choice between periodic surveillance and aggressive treatment of patients at risk of potential disease. A prominent case that I will discuss later is screening women for breast cancer. Others are choice between surveillance and treatment for patients at risk of heart disease or diabetes. Yet others are choice between surveillance and adjuvant 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, it 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 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, with x being a subset used in a clinical guideline. Let $P_{xw}(t) \equiv P[y(t) = 1|x, w]$ be the probability that a patient with 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 y. The expected utility of treatment t without knowledge of y is

$$(1) \qquad P_{xw}(t)\cdot u_{xw}(1, t) + [1 - P_{xw}(t)]\cdot u_{xw}(0, t).$$

Maximization of expected utility yields the optimal 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 assesses expected utility conditional only on covariates x rather than (x, w).

Decision rule (2) is simple, but it is instructive to discuss further simplifications that occur when treatment operates on disease in certain 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 disease development 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 $u_{xw}(1, B)$. 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 $P_{xw} \cdot u_{xw}(1, B) + (1 - P_{xw}) \cdot u_{xw}(0, B) \geq P_{xw} \cdot u_{xw}(1, A) + (1 - P_{xw}) \cdot u_{xw}(0, A)$.

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

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

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 $P_{xw} \geq P^*_{xw}$.

2.4. Adherence to Guidelines or Decentralization of Care?

Medical textbooks and training have long offered clinicians guidance in patient care. Such guidance has increasingly become institutionalized through issuance of clinical practice guidelines. 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 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 patient care. There may be good public-health reasons to adhere to guidelines in resource-constrained health systems, as argued by Wailoo *et al.* (2004), or when treating infectious diseases. However, adherence to a CPG cannot outperform decentralized care if treatment is individualistic, welfare is utilitarian, and clinicians have rational expectations. 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.

It is common for CPGs to condition their recommendations on a subset of the clinically observable patient covariates. Consider, for example, prediction of development of cardiovascular disease (CVD). The online tool at American College of Cardiology (2017) predicts ten-year and lifetime risk of CVD conditional on patient (age, sex, race, several cholesterol levels, systolic blood pressure, history of diabetes and smoking, current treatment status). Other patient covariates that are not used but that may help predict CVD include obesity, job stress, and exercise. I discuss another risk assessment tool at greater length below.

2.4.1. The Breast Cancer Risk Assessment Tool

An apt illustration of how available evidence affects risk assessment is the Breast Cancer Risk Assessment (BCRA) Tool of the National Cancer Institute (2011). This tool has become widely used in clinical practice (Susan G. Komen, 2016). It is an important input to the CPG for breast cancer screening issued by the National Comprehensive Cancer Network (2017).

The BCRA Tool gives an evidence-based probability that a woman will develop breast cancer conditional on eight 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, $\geq$ 14, unknown); (5) age of first live birth of a child (no births, $<$ 20, 20-24, 25-29, $\geq$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 BCRA Tool assesses risk conditional on these covariates and not others is that it 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 many respects, but it does not condition on further observable patient covariates that may be associated with risk of cancer. When considering the number of first-degree relatives with breast cancer (item 6), the Tool does not consider the number and ages of a woman's first-degree relatives, which 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 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 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 increased risk of breast cancer (Singletary and Gapstur, 2001).

3. Treatment with Imperfect Clinical Judgment

3.1. Empirical Research Comparing Statistical Prediction and Clinical Judgment

Section 2 showed that there is no reason to develop CPGs if treatment is individualistic, welfare is utilitarian, and clinicians have rational expectations. However, 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 covariates. The gap in performance persists even when clinical judgment uses additional covariates as predictors. The psychological findings provide a potential rationale for clinicians to adhere to evidence-based CPGs.

This research began in mid-twentieth century, notable early contributions including Sarbin (1943, 1944), Meehl (1954), and Goldberg (1968). To describe the conclusions of the literature, I rely mainly on the 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 covariates 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."

Here and elsewhere, Dawes, Faust, and Meehl caution against use of clinical judgment to subjectively predict disease risk or treatment response conditional on patient covariates 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.

Psychological research published after Dawes, Faust, and Meehl (1989) has largely corroborated the conclusions reached there; see, for example, Groves *et al*. (2000). 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 in the broader literature concerning practice of evidence-based medicine. I have found passages in the literature on evidence-based medicine that, contrary to the psychological literature, praise rather than criticize exercise of clinical judgment. See, for example, Sackett (1997).

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

The psychological literature 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 compared the accuracy of risk assessments and diagnoses made by statistical predictors and by clinicians, but they have not compared the 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 point predictions. Study of the logical relationship between probabilistic and point prediction shows that data on the latter at most yields wide

bounds on the former (Manski, 1990a). 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 $P_{xw}(\cdot)$.

In light of these and other 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 choice between surveillance and aggressive treatment of women at risk of breast cancer. Here surveillance usually means that a woman receives a breast exam and mammogram periodically, 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 BCRA 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. 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. As far as I am aware, it is not known whether adherence to the NCCN or ASCO guideline yields better or worse patient outcomes than does decentralized clinical decision making.

4. Questionable Methodological Practices in Evidence-Based Medicine

The psychological literature discussed in Section 3 has questioned the subjective judgment of clinicians, but it has not similarly questioned the accuracy of evidence-based predictions. 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.

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).

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

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, 2013b). 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 may be 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 actual patient populations.

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 the clinician knowing the assigned treatment. Hence, a trial reveals the distribution of response in a setting where patients and clinicians are uncertain what drug a patient is receiving. It does not reveal what response would be in the usual clinical setting where patients and physicians know what drug is being administered and can 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. Compliance may differ when 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 cannot 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 prescribed for them.

4.1.3. Surrogate Outcomes and Outcomes of Interest

A serious measurement problem often 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. Extrapolation from the surrogate outcomes measured in short trials to long-term outcomes can be challenging.

Trials for drug approval by the Food and Drug Administration (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. It is common 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 may be 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 more common for multiple trials to be performed on distinct 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 the issues that arise in extrapolating trial findings to clinical practice, there remains the familiar statistical problem of interpreting the samples of patient 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 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 practice. If the null is rejected, it is recommended that the innovation be the treatment of choice. This type of test is institutionalized in FDA drug approval, 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 in two trials (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, the probability of a Type I error. Then sample size determines the probability of rejecting the alternative when it is correct, 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 journal articles that report on trials typically present trial findings aggregated to broad demographic groups.

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 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."

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).

If researchers want to inform patient care, they should not view statistical insignificance as a reason to refrain from reporting observable heterogeneity in treatment response. Clinicians should be concerned with the variation of outcomes with treatments and covariates. Hypothesis tests do not address this question. Research journals should encourage publication of findings conditional on observed patient covariates. When 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 methodological practices described in Section 4 have become prevalent because medical research has largely adhered to classical statistical theory, whose concerns are remote from patient care. 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, medical 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. I focus here on identification problems. I discuss sampling imprecision from the perspective of treatment choice in Section 6.

5.1. Credible Identification Analysis

It is well known that 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 advocates 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 from a trial or an observational study, researchers report point estimates that may have fragile foundations.

Studying identification with realistic data and credible assumptions, research may yield informative bounds on treatment response but not precise findings. Early contributions include Manski (1990b, 1997) and Manski and Pepper (2000). Manski (2007) gives a broad textbook exposition. Horowitz and Manski (2000), Manski (2013a, 2017), and Mullahy (2017) study specific identification problems that arise in

clinical practice. The research considers inference from observational studies as well as trials. Both can be informative to some degree.

My research has sought to characterize what one can learn about the vector of expected utilities that determine the optimal treatment rule when sample size grows without bound and the study evidence is combined with credible assumptions. The canonical finding is that the expected utilities are revealed to lie in some informative set, called the *identification region*, but they are not pinned down precisely. Thus, the expected utilities are *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. Focusing on cases in which outcomes are binary, we derived sharp bounds on mean outcomes without imposing any assumptions about the distribution of the missing data. This analysis contrasts sharply with the conventional practice in medical research of assuming that missing data are missing at random or have some other structure.

We applied the findings to data from a trial comparing treatments for hypertension. Materson *et al.* (1993) reported on a trial comparing treatments for hypertension sponsored by the U.S. Department of Veteran Affairs (DVA). Male veteran patients 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 these 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 their identification analysis to estimate sharp bounds on the success probabilities $\{P[y(t) = 1|x], t = 1, \ldots, 7\}$ 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. Credible Ecological Inference for Personalized Risk Assessment

Manski (2017) studies the identification problem faced by a clinician who observes more patient covariates than are used in an evidence-based predictor 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:

$$(7) \qquad P(y \mid x = k) = P(w = j \mid x = k)P(y \mid x = k, w = j) + P(w \neq j \mid x = k)P(y \mid x = k, w \neq j).$$

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 5.3.1 summarizes findings on the former case, while Section 5.3.2 considers the latter.

5.3.1. 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 (7). When y is binary, the identification region is the interval

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

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

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 may observe. For concreteness, 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].

5.3.2. 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), but 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 covariates, enabling clinicians to express quantitative judgments in a structured way. To illustrate, I continue with prediction of life span.

Assume that persons with HBP have lower life expectancy than those without HBP. Also assume that black males have between 0 and 2.5 years less remaining life expectancy than white males conditional on blood pressure. 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|age 50, NH black male, not HBP) $\in$ [29.4, 35.4], E(y|age 50, NH black male, HBP) $\in$ [14.7, 22.9],
E(y|age 50, NH white male, not HBP) $\in$ [31.9, 36.4], E(y|age 50, NH white male, HBP) $\in$ [16.3, 25.4].

These bounds are highly informative. They reveal that the life expectancy of 50-year-old blacks without HBP is at least 6.5 years higher than that of those with HBP. For whites, the disparity is also at least 6.5 years.

5.4. Identification of Response to Diagnostic Testing and Treatment

I have thus far discussed treatment choice when a clinician has pre-specified knowledge of patient attributes. A common prelude to treatment choice is to learn more about a patient. This section considers a common scenario in which a patient with symptoms presents to a clinician, who initially observes some patient attributes such as demographic traits and medical history. The clinician may prescribe a treatment immediately or he may first order a diagnostic test that yields further information about the patient. 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?

5.4.1. Optimal Testing and Treatment

Phelps and Mushlin (1988) initiated study of this sequential decision problem using the rational-expectations optimization framework described in Section 2. The value of ordering a diagnostic test is that doing so reveals a patient attribute that the clinician would not observe otherwise, namely the test result. The potential usefulness of testing is expressed by the *expected value of information*, defined succinctly by Meltzer (2001) as (p. 119): "the change in expected utility with the collection of information."

It can be shown that the expected value of information is necessarily non-negative and is positive if the result affects the optimal treatment. It follows that a clinician should always order a test if performing the test has no direct negative effect on patient utility. However, performing a test may negatively affect utility. For example, biopsies, CT scans, and colonoscopies are invasive and expensive procedures. Hence, a test should be performed only if the expected value of information outweighs the direct utility cost.

Phelps and Mushlin assumed that clinicians have the knowledge needed to optimize testing and treatment. This includes knowledge of (a) expected patient utility with each treatment, in the absence of

testing; (a) the probability distribution for the test result, and (c) expected patient utility with each treatment, with knowledge of the test result. They characterized optimal testing and treatment given this knowledge.

5.4.2. Identification of Testing and Treatment Response with Observational Data

The analysis of Phelps and Mushlin is instructive, providing a clear prescription for optimal patient care when clinicians have the requisite knowledge of testing and treatment response. In principle, one might obtain this knowledge by performing an ideal randomized trial. A trial with multiple arms, one for each possible testing and treatment decision, could yield the knowledge of test results and treatment response needed to optimize. However, performance of this ideal trial is rare. Often the only available evidence is observational data generated by the testing and treatment decisions that occur in clinical practice. Then it may be unrealistic to suppose that clinicians have the knowledge that Phelps and Mushlin assumed.

Manski (2013a) characterizes the partial knowledge obtained when one combines observational data on a study population with various assumptions that restrict counterfactual testing results and treatment outcomes. I examine a common setting where the diagnostic test has two possible results, positive or negative. Clinicians commonly call a test result "positive" if it suggests illness and "negative" otherwise. I suppose that there are two feasible treatments, A being surveillance and B being aggressive treatment.

A common practice is to choose aggressive treatment if and only if a diagnostic test is performed and the result is positive. The chosen treatment is surveillance if the test result is negative or if the patient is not tested. I call this practice *aggressive treatment with positive testing* (ATPT). I study identification when the available evidence is observation of a study population that adheres to the ATPT practice.

An observational study reveals some but not all of the knowledge needed to optimize patient care for a group of patients who share the same initial attributes. One can observe test results for patients who are tested. One can observe health outcomes under (i) treatment A for patients who are not tested, (ii)

treatment A for patients who are tested and have a negative test result, and (iii) treatment B for patients who are tested and have a positive test result.

Other outcomes are counterfactual. One cannot observe test results for patients who are not tested. One cannot observe health outcomes under (i) treatment B for patients who are not tested, (ii) treatment B for patients who are tested and have a negative test result, and (iii) treatment A for patients who are tested and have a positive test result. These health outcomes are counterfactual because the ATPT practice assigns treatment B if and only if a patient is tested and has a positive test result.

Manski (2013a) shows that the observational evidence yields informative bounds on some of the quantities that determine optimal patient care. I initially derive bounds without making assumptions that restrict counterfactual testing and treatment outcomes. I then show what more can be learned if the evidence is combined with several assumptions that may be credible in some settings.

6. Reasonable Medical Decisions under Uncertainty

6.1. Recognizing Uncertainty

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

I conclude that it is often unrealistic to assume that either clinicians or guideline developers have rational expectations regarding disease development and treatment outcomes. That is, they do not have

sufficient knowledge to make objectively correct probabilistic predictions conditional on observed patient covariates. Given this, they should view patient care as a problem of decision making under uncertainty.

Ample precedents for this conclusion exist. Considering treatment of cancer, Mullins *et al.* (2010) observe that (p. 59): "there is considerable uncertainty surrounding the clinical benefits and harms associated with oncology treatments." Institute of Medicine (2011) 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 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 and completion lymph node dissection as diagnostic tests and adjuvant treatments for metastasis of melanoma (e.g., Faries, 2018). Guidelines regarding these procedures have changed in the past and continue to evolve.

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

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 this paper, the decision maker is a clinician, the actions are the feasible 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, the *state space*, expresses partial knowledge. The larger the state space, the less the decision maker knows about the outcome 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 step is to eliminate dominated treatments: an action is dominated if one knows that some other one is at least as good in all feasible states of nature and superior in some state. The second 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, each with its own properties.

The term "reasonable" inevitably has multiple interpretations. In his seminal book on statistical decision theory, Wald (1950) explained his focus on the minimax criterion in part by stating (p. 18): "a minimax solution seems, in general, to be a reasonable solution of the decision problem when an a priori distribution . . . . does not exist or is unknown to the experimenter." In a later monograph on statistical decision theory, Ferguson (1967) wrote (p. 28):

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

He went on to write (p. 29): "A *reasonable* rule is one that is better than just guessing."

6.2.1. Decision Criteria

Bayesian Decision Theory

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. Advocacy of Bayesian decision theory has been common from the middle of the twentieth century onward; see, for example, Luce and Raiffa (1957) and Berger (1985). There exists a substantial literature applying Bayesian thinking to medical decision making; see, for example, Claxton (1999) and Drummond *et al.* (2015).

The Bayesian perspective is compelling when one feels able to place a credible subjective distribution on the state space. However, a subjective distribution is a form of knowledge, and a decision

maker may not feel able to assert one. Bayesians have long struggled to provide guidance 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 subjective distribution is common in actual decision settings.

Bayesian decision theorists recognize that inability to specify a credible subjective distribution may yield poor decisions. Berger (1985) cautions that (page 121): "A Bayesian analysis may be 'rational' in the weak axiomatic sense, yet be terrible in a practical sense if an inappropriate prior distribution is used." Yet researchers who advocate application of Bayesian theory to medical decision making have tended not to reiterate Berger's caution.

Consider, for example the discussion of "Assigning Distributions to Parameters" in Drummond *et al.* (2015). The authors write that performance of a probabilistic sensitivity analysis (p. 399):

> "forces the analyst to be explicit, justifying the use of particular distributions on the basis of current evidence and the credibility of any assumptions that might be required. In fact, the choice of distribution is not at all arbitrary if standard statistical methods are followed. The choice should be informed by the nature of the parameter itself, the way the parameter was estimated, and the summary statistics reported, so the statistical uncertainty in its estimation is reflected. As in all statistical analysis, it also requires judgements about the potential bias and relevance of the available evidence."

This passage and the surrounding discussion recognize that specification of a subjective distribution requires judgment, but the authors do not explain how an analyst might justify the use of a particular distribution in practice. Nor do they caution that the chosen distribution may strongly affect the decision made.

Criteria Achieving Uniformly Satisfactory Decisions

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 applied to treatment choice, a statistical decision function has been called a *statistical treatment rule* (Manski, 2004).

Statistical decision theory may be used to study the criteria described in Section 6.2.1. In each case, one evaluates a criterion by the mean welfare that it yields across 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 discussed in Section 5.4. We found that a clinician who observes more covariates than are used in an evidence-based predictor 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 act?

Manski (2017) addresses this question in the setting of Section 2.3.2. To recall, the choice is between surveillance and aggressive treatment. Treating a patient with covariates (x, w), surveillance is optimal when $P_{xw} \leq P^*_{xw}$ and aggressive treatment is optimal when $P_{xw} \geq P^*_{xw}$, where P^*_{xw} is the threshold disease probability defined in (6).

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

interior to $[P_{xwL}, P_{xwH}]$. Then there exist feasible values of P_{xw} 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 δ_{xw} denote the subjective mean that a Bayesian clinician holds for $P_{xw}(A)$. A Bayesian clinician acts as if $P_{xw} = \delta_{xw}$.

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

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 $P_{xw} = P_{xwM}$.

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 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}. However, one could similarly recommend acting as if P_{xw} is any element of $[P_{xwL}, P_{xwH}]$.

Decision making with the Bayesian, maximin, or MR criterion is equivalent to acting as if P_{xw} takes particular values in $[P_{xwL}, P_{xwH}]$; δ_{xw} for the Bayesian, P_{xwH} for maximin, and P_{xwM} for MR. Singling out these

values has a firmer justification because they yield choices derived from established principles of decision theory.

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 clinical judgment by use of decision theory.

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 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 includes Manski (2004), 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 social welfare function that sums treatment outcomes across the population. 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 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 observed 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. However, 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 å > 0, *ε-optimal* rules exist when trials have large enough sample size. An ε-optimal rule has expected welfare, across repeated samples, within ε of the welfare of the best treatment in every state 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.

Our analytical findings, complemented by numerical calculations, yield a broad conclusion that sample sizes determined by clinically relevant near-optimality criteria tend to be much smaller than ones set conventional statistical power criteria. Reduction of sample size relative to prevailing norms can be beneficial in multiple ways. Reduction of total sample size can lower the cost of executing trials, the time necessary to recruit adequate numbers of subjects, and the complexity of managing trials across multiple centers. Reduction of sample size per treatment arm can make it feasible to perform trials that increase the number of treatment arms and, hence, yield information about a wider variety of treatment options.

6.5. Error Limitation and Learning by Adaptive Diversification of Treatment

I observed in Section 2 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. A feature 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. Recall that 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 variation in choice among undominated treatments 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. That is, I consider a health planner who treats a population of patients rather than a clinician who focuses on an individual patient. 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, variation is yet more useful because it generates a population-wide trial that yields new evidence about treatment response. As 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 gives decision theoretic arguments for diversification and 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.

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 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 $ä_x \in [0, 1]$ that randomly assigns a fraction $ä_x$ of these patients to treatment B and the remaining $1 - ä_x$ to treatment A. Let $á_x \equiv E[u(A)|x]$ and $â_x \equiv E[u(B)]$ be expected utility if all patients receive treatment A or B respectively. Social welfare with allocation $ä_x$ is $á_x(1 - ä_x) + â_x ä_x$. The optimal treatment allocation is $ä_x = 1$ if $â_x \geq á_x$ and $ä_x = 0$ if $â_x \leq á_x$.

The problem of interest is treatment choice when $(á_x, â_x)$ is not known. To formalize the problem, let S index the feasible states of nature. Let the planner know that $(á_x, â_x)$ lies in the set $[(á_{xs}, â_{xs}), s \in S]$. 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 $(á_{xs} \geq â_{xs}, s \in S)$ or if $(á_{xs} \leq â_{xs}, s \in S)$; $ä_x = 0$ is optimal in the former case and $ä_x = 1$ in the latter. However, all $ä_x \in [0, 1]$ are undominated if $á_{xs}$

$> \hat{a}_{xs}$ for some values of s and $\acute{a}_{xs} < \hat{a}_{xs}$ for other values. I consider this situation. The analysis is applicable whenever $[(\acute{a}_{xs}, \hat{a}_{xs}), s \in S]$ is bounded. Denote the extreme values as $\acute{a}_{xL} \equiv \min_{s \in S} \acute{a}_{xs}$, $\hat{a}_{xL} \equiv \min_{s \in S} \hat{a}_{xs}$, $\acute{a}_{xU} \equiv \max_{s \in S} \acute{a}_{xs}$, and $\hat{a}_{xU} \equiv \max_{s \in S} \hat{a}_{xs}$.

The regret of allocation $\ddot{a}_x$ 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 $\max (\acute{a}_{xs}, \hat{a}_{xs})$. Hence, $\ddot{a}_x$ has regret $\max (\acute{a}_{xs}, \hat{a}_{xs}) - [\acute{a}_x(1 - \ddot{a}_x) + \hat{a}_x\ddot{a}_x]$. 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

$$(9) \qquad \min_{\ddot{a}_x \in [0, 1]} \max_{s \in S} \max (\acute{a}_{xs}, \hat{a}_{xs}) - [\acute{a}_x(1 - \ddot{a}_x) + \hat{a}_x\ddot{a}_x].$$

Manski (2007, 2009) prove 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, $S_x(A) \equiv \{s \in S: \acute{a}_{xs} > \hat{a}_{xs}\}$ and $S_x(B) \equiv \{s \in S: \hat{a}_{xs} > \acute{a}_{xs}\}$. Let $M_x(A) \equiv \max_{s \in Sx(A)} (\acute{a}_{xs} - \hat{a}_{xs})$ and $M_x(B) \equiv \max_{s \in Sx(B)} (\hat{a}_{xs} - \acute{a}_{xs})$ be maximum regret on $S_x(A)$ and $S_x(B)$ respectively. The general result is

$$(10) \qquad \ddot{a}_{xMR} = \frac{M_x(B)}{M_x(A) + M_x(B)}.$$

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 $(\acute{a}_{xL}, \hat{a}_{xU})$ and $(\acute{a}_{xU}, \hat{a}_{xL})$ are feasible values of $(\acute{a}_x, \hat{a}_x)$, as is so when the identification region is rectangular. Then $M_x(A) = \acute{a}_{xU} - \hat{a}_{xL}$ and $M_x(B) = \hat{a}_{xU} - \acute{a}_{xL}$. Hence,

$$(11) \qquad \ddot{a}_{xMR} = \frac{\hat{a}_{xU} - \acute{a}_{xL}}{(\acute{a}_{xU} - \hat{a}_{xL}) + (\hat{a}_{xU} - \acute{a}_{xL})}.$$

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 decisions and use this evidence to inform treatment later on. Diversification is advantageous for learning treatment response because it generates randomized experiments. As evidence accumulates, the planner can revise the fraction of patients assigned to each treatment in accord with the available knowledge. I have called this *adaptive diversification*.

A simple approach to multi-period treatment choice is to use the *adaptive minimax-regret* (*AMR*) criterion. In each period, this criterion applies the static MR criterion using the information available at the time. It is adaptive because successive cohorts may receive different allocations as knowledge of treatment response increases over time. 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.

6.6. Should Guidelines Encourage Treatment Variation under Uncertainty?

To conclude, I suggest implementation of adaptive diversification 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.

What about health care systems 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.

When adaptive diversification is not feasible, we can nevertheless question whether the medical community should continue to discourage treatment variation across clinicians. Instead, CPGs could encourage clinicians to recognize that patient care under uncertainty 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.

References

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

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

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

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.

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

Claxton, K. (1999), "The Irrelevance of Inference: a Decision-making Approach to the Stochastic Evaluation of Health Care Technologies," *Journal of Health Economics*, 18, 341-364.

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.

Drummond, M. M. Sculpher, K. Claxton, G. Stoddart, and G. Torrance (2015), *Methods for the Economic Evaluation of Health Care Programmes*, Oxford: Oxford University Press.

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

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

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.

Luce, R. and H. Raiffa (1957), *Games and Decisions*, New York: Wiley.

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

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

Manski, C. (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. (2013a), "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. (2013b), *Public Policy in an Uncertain World*, Cambridge, MA: Harvard University Press.

Manski, C. (2017), "Credible Ecological Inference for Medical Decisions with Personalized Risk Assessment," *Quantitative Economics*, forthcoming.

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," *New England Journal of Medicine*, 328, 914-921.

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.

Mullahy, J. (2017), "Individual Results May Vary: Elementary Analytics of Inequality-Probability Bounds, with Applications to Health-Outcome Treatment Effects," NBER Working Paper No. 23603.

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, 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.

Susan G. Komen (2016), *Estimating Breast Cancer Risk*, ww5.komen.org/BreastCancer/GailAssessmentModel.html, accessed July 9, 2016.

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.

Wailoo, A., J. Roberts, J. Brazier, and C. McCabe (2004), "Efficiency, Equity, and NICE Clinical Guidelines," *BMJ*, 328, 536-537.

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.