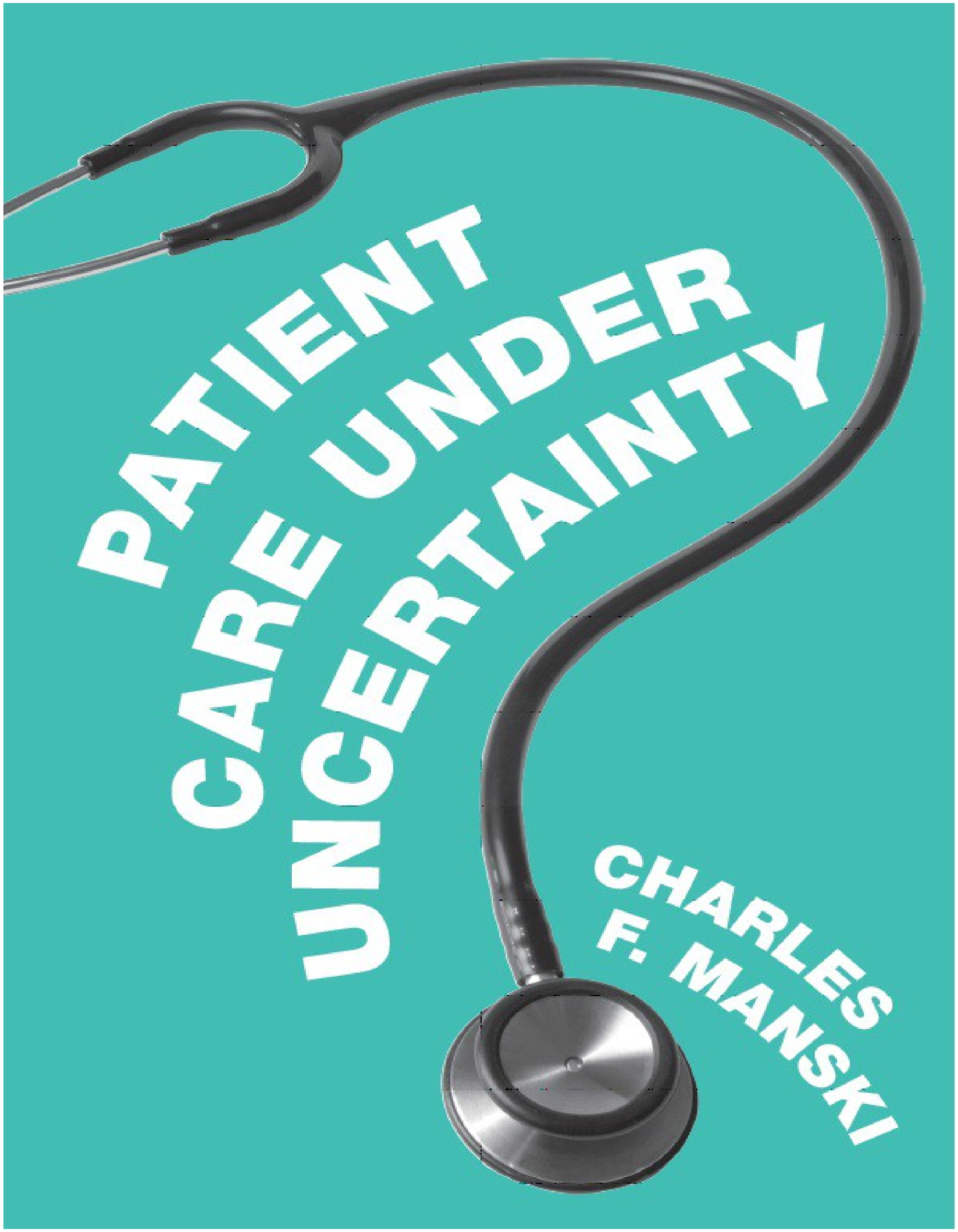

PATIENT CARE UNDER UNCERTAINTY

CHARLES F. MANSKI

Introduction

There are three broad branches of decision analysis: normative, descriptive, prescriptive.

Normative analysis seeks to establish ideal properties of decision making.

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

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

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

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

My focus is medical decision making under uncertainty.

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

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

A patient may ask:

> "What is the chance that I will develop disease X in the next five years?"
>
> "What is the chance that treatment Y will cure me?"

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

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

Evolution of My Research Program

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

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

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

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

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

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

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

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

Unobservability of counterfactual treatment outcomes

External Validity: extrapolation from study populations to patient care

Imperfect Data Quality: missing data, surrogate outcomes, measurement errors

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

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

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

I am concerned with the implications for decision making.

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

There is no **optimal** way to choose. There are **reasonable** ways.

Book Contents

Wishful Extrapolation from Research to Patient Care

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

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

I focus on predictions made with evidence from randomized trials.

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

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

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

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

Researchers and guideline developers often use untenable assumptions to extrapolate.

I have called this *wishful extrapolation*.

From Study Populations to Patient Populations

Study populations in trials often differ from patient populations.

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

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

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

Participation in a trial may be restricted to persons in certain age categories who reside in certain locales.

Among such persons, volunteers are those who respond to financial and medical incentives to participate.

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

From Experimental Treatments to Clinical Treatments

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

This is particularly so in trials comparing drug treatments.

Drug trials are double-blinded, neither the patient nor the clinician knowing the assigned treatment.

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

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

Consider drug treatments for hypertension.

Patients react heterogeneously to the various drugs available for prescription.

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

Sequential experimentation is not possible in a blinded trial.

The standard protocol prohibits the clinician from knowing what drug a subject is receiving and from using judgment to modify the treatment.

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

From Surrogate Outcomes to Health Outcomes

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

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

The longest trials for FDA drug approval, *phase 3 trials*, run for two to three years.

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

Wishful Meta-Analyses of Disparate Studies

The problems discussed above concern analysis of findings from a single trial.

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

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

Decision makers must somehow interpret the mass of information provided by evidence-based research.

The hard question is how to interpret this information sensibly.

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

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

Suppose that multiple trials have been performed on the same population, each drawing an independent random sample. The best way to use the data combines them into one sample.

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

Meta-analysis proposes methods to combine the estimates.

A common proposal computes a weighted average, weighting estimates by sample size.

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

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

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

The protocols for administration of treatments and the 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.

Then it is not clear how to interpret a weighted average of the estimates.

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

The model considers trials to be drawn at random "from a population of possible studies."

Then each trial estimates a parameter drawn at random from a population of possible parameters. A weighted average estimates the mean of these parameters.

The relevance to clinical practice is obscure.

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

They do not explain how a population of possible studies connects to what matters to a clinician—the distribution of health outcome across the relevant population of patients.

Treatment Choice with Trial Data:
Statistical Decision Theory Should Supplant Hypothesis Testing

(Manski, *The American Statistician*, (2019*)*

A common procedure when comparing two treatments in a randomized trial is to view one as the status quo, the other as an innovation, and use a classical hypothesis test to choose between them.

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 is not rejected, use of the status quo in clinical practice is recommended.

If the null is rejected, use of the innovation is recommended.

The convention has been to fix the probability of a Type I error.   Sample size determines the probability of a Type II error.

International Conference on Harmonisation (1999) provides 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 (*PNAS*, 2016) give several reasons why hypothesis testing may yield unsatisfactory results for medical decisions and other treatment choices.

1. Use of Conventional Asymmetric Error Probabilities

It has been standard to fix the probabilities of Type I and II errors at 5% and 10-20%.

Testing theory gives no rationale for selection of these error probabilities.

It does not explain 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.

2. *Inattention to Magnitudes of Losses to Welfare When Errors Occur*

A clinician should care about the magnitudes of the losses to patient welfare that arise when errors occur.

A given error probability should be less acceptable when the welfare difference between treatments is larger.

Testing theory does not take this into account.

3. Limitation to Settings with Two Treatments

A clinician often chooses among several treatments.

Many trials compare more than two treatments.

Standard testing theory only contemplates choice between two treatments.

Statisticians have struggled to extend it to compare multiple treatments.

Example of Issues 1 and 2

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

Mean patient life span with the status quo treatment is known to be one year. Researchers see two possibilities for the innovation. It may be less effective than the status quo, yielding mean life span of 1/3 of a year, or it may be more effective, with mean life span of 5 years.

A trial is performed to learn the effectiveness of the innovation. The data are used to perform a test comparing the innovation and the status quo. The error probabilities are set at 0.05 and 0.20. The result is used to choose between the treatments.

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

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

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

The theory of hypothesis testing does not motivate this asymmetry.

Principles of Statistical Decision Theory

Wald (1950) considered the uses of sample data to make decisions under uncertainty.

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

He recommended evaluation of statistical decision functions as procedures, specifying how a decision maker would use whatever data are realized. Thus, the theory is frequentist.

He proposed evaluate of a statistical decision function by the distribution of loss that it yields across realizations of the sampling process. He focused attention on mean sampling performance.

He prescribed a three-step decision process.

1. Specify the state space (parameter space), which indexes the set of values of unknown quantities that the decision maker deems possible.

2. Eliminate inadmissible statistical decision functions.

   A decision function is inadmissible (weakly dominated) if there exists another one that yields at least as good sampling performance in every possible state of nature and strictly better performance in some state.

3. Use some criterion to choose an admissible statistical decision function. Leading criteria are maximization of subjective expected welfare (Bayes rule), maximin, and minimax regret.

Early applications of the Wald theory focused on prediction rather than treatment choice.

An econometric literature on treatment choice has developed since the early 2000s.

See Manski (2004, 2005, 2007), Manski and Tetenov (2007, 2014, 2016, 2019), Hirano and Porter (2009), Schlag (2006), Stoye (2009, 2012), Tetenov (2012), and Kitagawa and Tetenov (2018).

A statistical decision function uses the data to choose a treatment allocation, so such a function has been called a *statistical treatment rule* (STR).

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

The state space specifies the feasible distributions of treatment response.

The objective has been maximization of a social welfare function that sums treatment outcomes across the population.

The literature mainly studies the minimax-regret criterion.

Treatment Choice with Existing Trial Data

Consider a trial with two treatments and a population of observationally identical patients.

Suppose that a health planner must assign treatment A or B to each member of patient population J.

Each patient $j \in J$ has response function $y_j(\cdot): T \rightarrow Y$ mapping treatments $t \in T$ into individual outcomes $y_j(t) \in R$. Let P denote the distribution of treatment response in the population.

The members of the population may respond heterogeneously to treatment, but they are observationally identical to the planner.

For $\delta \in [0, 1]$, the planner can allocate fraction δ of patients to B and $1 - \delta$ to A.

The planner wants to choose δ to maximize an additive welfare function

$$U(\delta, P) \quad = \quad E[y(A)]\cdot(1 - \delta) + E[y(B)]\cdot\delta \quad = \quad \alpha\cdot(1 - \delta) + \beta\cdot\delta \quad = \quad \alpha + (\beta - \alpha)\cdot\delta,$$

where $\alpha \equiv E[y(A)]$ and $\beta \equiv E[y(B)]$.

$\beta - \alpha$ is the average treatment effect.

It is optimal to set $\delta = 1$ if $\beta - \alpha > 0$ $\delta = 0$ if $\beta - \alpha < 0$.

The problem of interest is treatment choice when incomplete knowledge of P makes it impossible to determine the sign of $\beta - \alpha$.

Sample data are available, with sample space Ψ and sampling distribution Q.

An STR $\delta(\cdot): \Psi \to [0, 1]$ maps the data into a treatment allocation. The welfare realized with data ψ is the random variable

$$U(\delta, P, \psi) \quad = \quad \alpha + (\beta - \alpha)\cdot\delta(\psi).$$

The state space $[(P_s, Q_s), s \in S]$ is the set of (P, Q) pairs that the planner deems possible.

Expected welfare in state s is

$$W(\delta, P_s, Q_s) \quad = \quad \alpha_s + (\beta_s \; - \alpha_s)\cdot E_s[\delta(\psi)],$$

where $E_s[\delta(\psi)] \equiv \int_\Psi \delta(\psi)dQ_s(\psi)$.

The Bayes, maximin, and MMR rules are

Bayes rule: $$\max_{\delta \in [0, 1]} \int_S W(\delta, P_s, Q_s)d\pi(s),$$

where π is a subjective distribution on the state space.

Maximin rule: $$\max_{\delta \in [0, 1]} \min_{s \in S} W(\delta, P_s, Q_s).$$

Minimax-regret rule: $$\min_{\delta \in [0, 1]} \max_{s \in S} [\max(\alpha_s, \beta_s) - W(\delta, P_s, Q_s)].$$

Measuring Performance by Maximum Regret

Practical Appeal

MMR behaves more reasonably than maximin in the context of treatment choice.

When a trial has a balanced design and outcomes take a bounded range of values, it has been found that the MMR rule is well approximated by the *empirical success* (ES) rule, which chooses the treatment with the highest observed average outcome in the trial.

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

Conceptual Appeal

Maximum regret quantifies how lack of knowledge of the true state diminishes the quality of decisions. An STR with small maximum regret is uniformly near-optimal across all states.

The concept is especially transparent when there are two treatments, say A and B.

In a state where A is better, the regret of an STR is the product of its probability of a Type I error (choosing B) and the magnitude of the loss in expected welfare that occurs when choosing B.

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

A simple case occurs when there are two states of nature and when Type I and Type II errors yield equal losses in expected welfare, say L. Then the maximum regret of an STR is L times the maximum of its probabilities of Type I and Type II errors.

Suppose that sample size has been chosen to give 0.05 probability of Type I error and 0.20 probability of Type II error, using a conventional test. Consider the STR that uses this test to choose a treatment. The maximum regret of this "test rule" is L times 0.20.

One can obtain a test rule with smaller maximum regret by enlarging the critical region for the test. Enlarging the critical region increases the probability of Type I error and reduces that of Type II error.

Maximum regret decreases until one enlarges the critical region to the degree that it equalizes the probabilities of Type I and Type II errors.

Designing Trials to Enable Near-Optimal Treatment Choice

(Manski and Tetenov, *PNAS*, 2016)

Statistical power calculations have been used to choose sample size in RCTs.

An ideal objective is to collect data that enable implementation of an *optimal* rule––one whose expected welfare equals the welfare of the best treatment in every state of nature.

Optimality is not achievable in general, but *ε-optimal* rules do exist when trials have large enough sample size.

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

Implementation of the idea requires specification of a value for ε.

The necessity to choose an effect size of interest when designing trials already arises in conventional practice, where the trial planner must specify the alternative hypothesis to be compared with the null.

A possibility is to base ε on the *minimum clinically meaningful difference* (MCMD) in the average treatment effect comparing alternative treatments.

Medical writers call an average treatment effect clinically significant if its magnitude is greater than ε for a value of ε deemed minimally consequential in clinical practice.

We consider trials that draw predetermined numbers of subjects at random within groups stratified by covariates and treatments.

The analytical findings are simple sufficient conditions on sample sizes that ensure existence of ε-optimal treatment rules when outcomes are bounded.

These conditions are obtained by application of Hoeffding (*JASA*, 1963) and other large deviations inequalities to evaluate the performance of empirical success rules.

We provide exact computations of minimal sample sizes enabling ε-optimality that hold when there are two treatments and outcomes are binary.

Findings with Binary Outcomes, Two Treatments, and Balanced Designs

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

We compute the minimum sample size that enables ε-optimality when a clinician uses one of three different treatment rules, for various values of ε.

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

Based on our exact calculations and analytical findings with large-deviations inequalities, we conclude that sample sizes determined by clinically relevant near-optimality criteria tend to be much smaller than ones set by conventional statistical power criteria.

Reduction of total sample size can lower the cost of executing trials, the time needed to recruit subjects, and the complexity of managing trials across 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.

Trial Size for Near-Optimal Choice Between Surveillance and Aggressive Treatment: Reconsidering MSLT-II

(Manski and Tetenov, *The American Statistician*, 2019)

We develop and apply a refined version of the analysis in Manski and Tetenov (P*NAS*, 2016) to trials that compare aggressive treatment of patients with surveillance.

Internists choose between prescription of pharmaceuticals and surveillance when treating patients at risk of heart disease or diabetes. Oncologists choose between surveillance and aggressive treatments such as surgery or chemotherapy when treating cancer patients at risk of metastasis.

Aggressive treatment may be appealing to the extent that it better prevents onset or reduces the severity of illness. Surveillance may be attractive to the extent that it avoids side effects that may occur with aggressive treatment.

The need for a refined version of the analysis arises because our earlier work studied settings in which there is only a primary health outcome of interest, without secondary outcomes.

An important aspect of choice between surveillance and aggressive treatment is that the latter may have side effects.

The prevailing approach to choosing sample size in trials has been to focus entirely on the primary outcome of a treatment, without considering secondary outcomes.

When aggressive treatment may have serious side effects, it is more reasonable to consider how the primary outcome and side effects jointly determine patient welfare.

This requires new analysis.

As a case study, we reconsider a recent trial comparing *nodal observation* and *lymph node dissection* when treating patients with early-stage cutaneous melanoma at risk of metastasis.

Nodal observation is surveillance of lymph nodes by ultrasound scan, a procedure that has negligible side effects. Lymph node dissection is a surgical procedure in which the lymph nodes in the relevant regional basin are removed.

Dissection is commonly viewed as an aggressive treatment. A particularly concerning side effect is lymphedema, which may reduce patient quality of life substantially.

Choice between nodal observation and lymph node dissection is a common decision faced in early treatment of melanoma, breast cancer, and other forms of localized cancer.

The Multicenter Selective Lymphadenectomy Trial II (MSLT-II) compared dissection and observation for melanoma patients who had recently undergone sentinel lymph-node biopsy and who had obtained a positive finding of malignancy.

The primary outcome was defined to be melanoma-specific survival for three years following the date of randomization.

Findings were reported in Faries *et al*. (*NEJM*, 2017).

Our concern is choice of sample size in the trial.

Using a conventional statistical power calculation, the investigators assigned 971 patients to dissection and 968 to observation.

Choosing the MSLT-II Sample Size to Enable Near-Optimal Treatment

We assume a simple patient welfare function.

Welfare with nodal observation equal 1 if a patient survives for three years and 0 otherwise.

Welfare with dissection depends on whether a patient experiences lymphedema.

When a patient does not experience lymphedema, welfare with dissection equals 1 if the patient survives for three years and 0 otherwise.

When a patient experiences lymphedema, welfare is lowered by a specified fraction *h*, whose value expresses the harm associated with lymphedema. A patient who experiences lymphedema has welfare $1 - h$ if he survives and –*h* if he does not survive.

Let nodal observation be treatment A. Let y(A) = 1 if a patient survives and y(A) = 0 otherwise. Mean patient welfare with observation is the survival probability P[y(A) = 1].

Let lymph node dissection be treatment B. Let y(B) = 1 if a patient survives and y(B) = 0 otherwise. Let s(B) = 1 if a patient experiences lymphedema and s(B) = 0 otherwise. Mean patient welfare with dissection is P[y(B) = 1] − h·P[s(B) = 1].

Observation is optimal if P[y(A) = 1] > P[y(B) = 1] − h·P[s(B) = 1].

The MSLT-II trial yields information about probabilities of survival and lymphedema.

The sample size determines the extent of the information. For any positive ε, a sample of size N per treatment arm enables ε-optimal treatment if N is sufficiently large.

Findings

We compute sample sizes that enable ε-optimal treatment for any values of *h* and ε. We assume that treatment choice will be made with the empirical success rule.

We perform computations using two methods to determine maximum regret.

One applies simulated annealing and the other uses a normal approximation to the finite-sample distribution of empirical success.

Table 1: Near-optimality (maximum regret) of ES rules

(N is the number of subjects per treatment arm)

		h = 0		*h* = 0.1		*h* = 0.2		*h* = 0.3		*h* = 0.4		*h* = 0.5	
		Simulated	Normal	Simulated	Normal	Simulated	Normal	Simulated	Normal	Simulated	Normal	Simulated	Normal
N =		annealing	approx.	annealing	approx.	annealing	approx.	annealing	approx.	annealing	approx.	annealing	approx.
10		0.038209	0.037490	0.044905	0.039672	0.045017	0.041857	0.045794	0.044046	0.046704	0.046237	0.049236	0.048431
20		0.026947	0.026689	0.030401	0.028180	0.030479	0.029672	0.031212	0.031166	0.032803	0.032661	0.034487	0.034157
30		0.021983	0.021841	0.024046	0.023039	0.024516	0.024237	0.025874	0.025435	0.026805	0.026634	0.028039	0.027834
40		0.019029	0.018937	0.020710	0.019963	0.021105	0.020989	0.021930	0.022016	0.023172	0.023044	0.024218	0.024071
50		0.017016	0.016949	0.018182	0.017860	0.018829	0.018772	0.019865	0.019683	0.020688	0.020595	0.021621	0.021507
60		0.015530	0.015480	0.016640	0.016307	0.017170	0.017134	0.018019	0.017962	0.018859	0.018789	0.019709	0.019617
70		0.014376	0.014336	0.015228	0.015099	0.015890	0.015861	0.016708	0.016624	0.017444	0.017387	0.018227	0.018150
80		0.013447	0.013414	0.014291	0.014124	0.014861	0.014835	0.015612	0.015546	0.016306	0.016257	0.017034	0.016968
90		0.012677	0.012649	0.013369	0.013317	0.014009	0.013985	0.014690	0.014653	0.015365	0.015321	0.016048	0.015990
100		0.012025	0.012002	0.012724	0.012634	0.013287	0.013266	0.013952	0.013898	0.014570	0.014276	0.015215	0.014845
150		0.009817	0.009804	0.010330	0.010316	0.010841	0.010827	0.011359	0.011339	0.011876	0.011680	0.012395	0.012150
200		0.008501	0.008493	0.008941	0.008933	0.009384	0.009374	0.009826	0.009814	0.010274	0.010128	0.010720	0.010537
250		0.007603	0.007597	0.007995	0.007990	0.008390	0.008382	0.008786	0.008775	0.009183	0.009066	0.009580	0.009433

We focus on sample size enabling near-optimal treatment when $h = 0.2$ and $\varepsilon = 0.0085$.

Setting $h = 0.2$ supposes that suffering from lymphedema reduces welfare by 0.2.

This value is suggested by Cheville *et al*. (*Cancer*, 2010).

Setting $\varepsilon = 0.085$ follows from the way that the MSLT-II investigators performed their power calculation.

They judged a difference of 5 percentage points in survival to be a clinically meaningful loss in welfare and they judged 0.17 to be an acceptable probability of Type II error.

Regret equals the magnitude of welfare loss times the probability that the loss will occur. Thus, the investigators judged $0.17 \times 0.05 = 0.0085$ to be an acceptable level of regret.

We find that ε-optimal treatment is achievable with 244 patients assigned to each of observation and dissection.

The total sample size of 488 is much smaller than the 1939 in MSLT-II.

Conclusion

Science does not always progress monotonically. There are times when important, even fundamental, ideas are discovered and receive attention but then are neglected.

This occurred with statistical decision theory, which received considerable attention in the middle of the twentieth century but was largely forgotten thereafter.

A revival focusing on treatment choice began in the early 2000s.

I hope that methodologists and applied researchers will (re)learn statistical decision theory and use it when studying not only treatment choice but decision making with sample data more generally.

Personalized Risk Assessment

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

Yet evidence to support complete personalization is rarely available.

Hence, the term is commonly used to mean care that varies with observed patient characteristics.

Thus, personalized medicine is a matter of degree.

Clinicians often can personalize patient care to a greater degree than can evidence-based risk assessments and studies of treatment response.

Personalized Prediction of Cardiovascular Disease

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

This online tool conditions the reported probability on:

Patient Demographics: current age, sex, race

Current Labs/Exams: cholesterol, blood pressure.

Personal History: diabetes, hypertension treatment, smoking, statin or aspirin therapy.

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

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

The Breast Cancer Risk Assessment Tool

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

(1) history of breast cancer or chest radiation therapy for Hodgkin Lymphoma.

(2) a BRCA mutation or genetic syndrome associated with risk of breast cancer.

(3) current age.

(4) age of first menstrual period.

(5) age of first live birth of a child.

(6) number of first-degree female relatives with breast cancer.

(7) number of breast biopsies.

(8) race/ethnicity.

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

* number and ages of a woman's first-degree relatives.
* prevalence of breast cancer among second-degree relatives.
* membership in ethnic groups who have differential risk of a BRCA mutation.
* behavioral attributes such as excessive drinking of alcohol.

Predicting Kidney Transplant Outcomes with Partial Knowledge of HLA Mismatch

(Manski, Tambur, and Gmeiner, *PNAS*, 2019)

An important problem in organ transplantation is to predict the patient outcomes that would occur if an offered organ were to be transplanted into a specific recipient.

We focus on transplant of a kidney from a deceased donor.

Observed organ attributes include information about the quality of the organ and about genetic features of the donor related to activation of the patient's immune system, measured by Human Leukocyte Antigen (HLA) genotype.

Observed patient attributes include information about age, health, and HLA type.

Research in transplant immunology has shown that, all else equal, patient outcomes vary with the degree of HLA mismatch.

The greater the mismatch, the more likely it is that the immune system will produce antibodies that attack the graft, lowering its functionality and eventually destroying it.

HLA mismatch is a potentially powerful predictor of outcomes, despite the development of immunosuppressive therapies that seek to decrease the response of a patient’s immune system to the foreign tissue in a graft.

Immunosuppression may reduce graft rejection, but it also may harm the patient by lowering the effectiveness of the immune system in protecting against disease.

When considering whether to accept an organ, clinicians and patients would benefit from accurate prediction of outcomes conditional on observed donor/recipient attributes.

Accurate prediction means probabilistic prediction of the frequency of outcomes conditional on organ quality, patient age/health, and HLA mismatch.

A simple tool for prediction of kidney transplant outcomes is the Kidney Donor Profile Index (KDPI). See the Organ Procurement and Transplantation Network (OPTN).

KDPI does not use information on HLA mismatch. It is determined only by demographic and health attributes of the deceased donor.

Accurate prediction of outcomes conditional on HLA mismatch would enable better genetic matching of donors and recipients.

Available low-resolution data on donor and recipient HLA types do not suffice to realize the full promise of better matching, but these data have some predictive value.

The analysis in this paper aims to improve prediction.

We combine data in

* the Scientific Registry of Transplant Recipients (SRTR) on individual transplant outcomes conditional on low-resolution HLA types

* HaploStats on the distribution of refined HLA types within specific populations.

The way we combine the data uses knowledge from research on transplant immunology.

Clinicians who want to predict outcomes conditional on HLA mismatch can use findings in research articles that present estimated prediction models.

Particularly visible has been Rao et al. (2009), who developed various versions of the Kidney Donor Risk Index (KDRI), one of which was later transformed into the KDPI.

Rao et al. used a proportional hazards model to interpret SRTR data on the outcomes of deceased-donor kidney transplants.

The model predicts the probability of all-cause graft failure as a function of time since transplant and of variables describing organ quality, recipient age/health, and HLA mismatch.

All-cause graft failure means patient return to dialysis, re-transplant, or death.

The KDRI is a summary statistic that measures the risk of graft failure for a transplant with specified characteristics relative to the risk for a specified reference transplant.

The proportional hazards model assumes that this relative risk remains constant as a function of time since transplant.

In 2013, the OPTN approved a new national Kidney Allocation System (KAS) that makes official use of the KDPI, a truncated and renormalized version of the KDRI.

The KDPI truncates the KDRI by considering only organ attributes.

It does not compute relative risk conditional on recipient attributes and HLA mismatch.

Our analysis is in three parts.

We first consider outcome prediction using only the incomplete characterization of HLA provided in the SRTR data.

Our work differs methodologically from Rao et al. in that we compute nonparametric estimates of graft survival rather than estimates of a proportional hazards model.

Considering one and five-year survival, we find that lower survival probabilities are associated with higher values of recipient age, KDPI, and number of HLA mismatches.

The estimated reductions in survival probability with increases in these risk factors are much larger when predicting five-year survival than when predicting one-year survival.

Not only are the reductions larger when predicting five-year survival than one-year survival, but they are differentially so.

This finding provides evidence against the proportional hazards model.

It is consistent with research in transplant immunology showing that generation of *de novo* donor-specific antibodies following transplant, a consequence of HLA mismatch, plays a cumulative rather than time-invariant role in graft loss.

Moreover, it is known that late antibody-mediated rejection is less responsive to treatment and thus more associated with graft loss.

We next critique an HLA imputation method suggested in the transplant literature.

The idea is for clinicians who have partial knowledge of HLA mismatch to use available data on the frequency distribution of distinct HLA genotypes within ethnic/nationality groups to impute genotypes for specific donors and recipients.

This done, the suggestion is for a clinician contemplating whether to accept an offered organ to act as if he or she has complete knowledge of mismatch.

We show that imputation does not improve prediction of transplant outcomes.

To the contrary, it may diminish the accuracy of predictions and generate sub-optimal transplant decisions.

The third part of our analysis considers a clinician who possesses more complete knowledge of mismatch than is available in the SRTR data.

This situation would be salutary if accurate probabilistic predictions of outcomes conditional on clinically observed mismatch were available.

However, such predictors are not currently available.

Bringing to bear econometric research on partial identification of conditional probability distributions, we combine SRTR and HaploStats data to yield credible partial predictions conditional on the clinically observed mismatch information.

Nonparametric Prediction of Graft Failure with SRTR Data

The use by Rao *et al.* of a proportional hazards model to predict graft failure adheres to the widespread medical application of this model to predict survival.

The model assumes that the hazard of graft failure at any point in time following the date of transplant is an unrestricted function of time multiplied by a parametric function of observed covariates characterizing the organ and recipient.

This function measures the risk of failure for transplants with a specified value of the covariates relative to that with a reference covariate value.

The standard practice supposes that the relative-risk function is log-linear.

Thus, the hazard rate at date T for a transplant with covariates x is $h(T|x) = h_0(T)\cdot\exp(xb)$, where $h_0(T)$ is the baseline hazard, exp(xb) is relative risk, and b is a parameter vector.

The model is simple, but its simplicity comes with a potential cost in credibility: The assumption that relative risk is a time-invariant log-linear function of covariates is strong.

An unfortunate feature of the widespread application of the proportional hazards model in medical research has been that its realism is rarely questioned.

We are aware of no biological basis to assume that the relative risk of graft failure across organ and recipient covariates is invariant as the length of time since transplant grows.

We are aware of no basis to assume that relative risk is log-linear.

Nonparametric Prediction

When ample data are available, nonparametric prediction of outcomes becomes feasible.

The SRTR Standard Analysis Files (SAF) provide ample data.

The SAF are an updated version of the data used to develop the KDRI.

They provide (donor-recipient) data with personal identifiers removed, made available by the SRTR for research purposes.

We use SAF data for deceased-donor kidney transplants performed in the years 2009--2018 to estimate nonparametrically the probability of graft survival for one or five years, conditional on specified (organ, recipient) covariates.

We use race and age to characterize the recipient.

We use the KDPI to characterize the quality of an organ.

We use the available HLA data to characterize mismatch.

For each organ and recipient, the file records two-digit types for HLA loci (A, B, DR).

Humans have two antigens at each HLA locus, from their paternal and maternal heritage.

Hence, the number of mismatches at each locus takes one of the three values {0, 1, 2}. Thus, there are 3 x 3 x 3 = 27 possible values for (A, B, DR) mismatch, with the value (0, 0, 0) indicating a perfect match and (2, 2, 2) a complete mismatch.

We condition on the specific 27-valued mismatch pattern, rather than simply counting the number of mismatches (0 to 6).

There is biological reason to think that the effect of mismatch on organ failure varies with the specific loci and genotype of mismatches, not just with the number of mismatches.

To formalize the prediction problem, let y denote the number of years following transplant when graft failure occurs.

Let x indicate the (A, B, DR) mismatch between donor and recipient. Let z denote the other organ and recipient attributes on which we condition prediction (recipient race, recipient age, organ KDPI).

In principle, we can use the SAF data to estimate the survival probability $P(y > T|z, x)$ for any number T of years.

We focus on survival for one or five years.

Data

The SRTR data extend back to 1987, but we restrict analysis to 2009 onward.

One reason is that it has been standard to use molecular rather than cruder serological methods for HLA typing from 2008 on.

Another is that immunosuppression practices have been approximately stable since the mid-2000s.

We form 27 sub-samples of data, each with a distinct value of HLA mismatch.

For illustration, we focus on transplants to the SRTR category of White recipients.

The SAF documents 48,945 kidney transplants to White recipients in 2009--2018.

We restrict attention to patients on the waiting list to receive transplant of a single kidney.

We drop cases with missing data, yielding 45,217 with complete data.

40,316 were performed in 2009--2017, for whom the one-year outcome was recorded.

21,310 were performed in 2009--2013, for whom the five-year outcome was recorded.

Methods

We use a bivariate kernel regression procedure to estimate the conditional probability of survival at specified values of patient age and KDPI.

The SRTR data list all transplants performed in the United States. It is not obvious how to measure the statistical precision of the estimates.

Previous research has viewed the SRTR data as a random sample from a population of potential transplants. In the absence of a ready alternative, we follow this practice.

We use the bootstrap to compute approximate 95% confidence intervals.

Although our analysis is nonparametric, it maintains the possibly unrealistic assumption that the *ex-post* distribution of outcomes for realized transplants equals the *ex-ante* distribution of outcomes for potential transplants.

A clinician making a transplant decision wants to predict outcomes before choosing whether to accept an offered organ.

Equality of ex ante and ex post outcome distributions has been assumed throughout the research literature that uses SRTR or other data to predict transplant outcomes.

Yet the accuracy of the assumption is unknown.

The assumption would be easy to motivate if transplant decisions were made randomly.

It may not hold when clinicians make purposeful transplant decisions. Then there is so-called “selection bias.”

The econometric literature on partial identification of treatment response weakens the assumption of equal ex ante and ex post outcome distributions.

It shows that observational data may still be informative to some degree, yielding bounds on treatment response.

Results

A kernel estimate is easy to compute for any value of the conditioning covariates (z, x).

Hence, nonparametric prediction is amenable to development of an online prediction tool.

A clinician could input the (z, x) value for a case of interest and obtain survival estimates and confidence intervals.

Presentation in a research article of estimates for all values of (z, x) is not realistic due to space constraints.

We compute estimates for the 27 values of the HLA mismatch indicator and for a grid of nine values of (recipient age, organ KDPI): {30, 50, 70} x {25, 50, 75}. This yields 27 x 9 = 243 estimates.

Sparsity of SRTR data near certain values of (z, x) occasionally prevents computation of a precise kernel estimate.

When computing estimates of one and five-year survival, we restrict attention to 25 and 22 values of HLA mismatch for which there exist at least fifty observations.

We exclude (age, KDPI) combinations where there exist no observations within the relevant local bandwidth.

Table 1 presenting nine estimates and their confidence intervals. Considering one-year survival, the table fixes values for (A mismatch, B, mismatch, age) and shows the variation in estimates with the values of DR mismatch and KDPI.

Table 1: Kernel Estimates and Confidence Intervals for Probability of One-year Survival by DR Mismatch (mm) and KDPI, (A mm = 1, B mm = 1, age = 30 years)

	KDPI = 25	KDPI = 50	KDPI = 75
DR mm = 0	0.97	0.96	0.89
	[0.91, 1]	[0.85, 1]	[0.80, 0.96]
DR mm = 1	0.97	1.00	1.00
	[0.90, 1]	[1, 1]	[1, 1]
DR mm = 2	0.92	0.96	0.94
	[0.84, 0.98]	[0.87, 1]	[0.90, 0.98]

Strong and interesting patterns emerge when we summarize the findings.

We exclude estimates for bandwidths with no variation in the survival outcome. With this exclusion, we have 207 one-year estimates and 194 five-year estimates.

We perform linear least-squares fits of the estimates of survival probability to recipient age, organ KDPI, total number of HLA mismatches, and a constant.

Table 2: Least Squares Fits of Kernel Estimates of Survival Probability to Covariates
(Robust standard errors in parentheses)

Covariates	One-Year Survival	Five-Year Survival
Age = (30, 50, 70 years)	-0.000622	-0.00276
	(0.000170)	(0.000304)
KDPI = (25, 50, 75)	-0.000634	-0.00149
	(0.000133)	(0.000274)
# (A, B, DR) mm = (0 to 6)	-0.00124	-0.00894
	(0.00169)	(0.00309)
Constant	0.990	1.007
	(0.0113)	(0.0238)
Observations	207	194

Transplant Decisions with Haplotype Imputations

The SRTR data on HLA types are incomplete in two respects.

The data provide types for donor HLA loci (A, B, C, DP, DQ, DR), but only for recipient loci (A, B, DR). Hence, the data do not reveal mismatch at the (C, DP, DQ) loci.

The data code only low-resolution two-digit types, each representing multiple alleles.

Modern HLA typing codes high-resolution four-digit types, each being a unique allele.

Donor and recipient HLA types that appear matched with two-digit coding may be mismatched with four-digit coding.

Hoping to overcome the problem of incomplete data on HLA types, some transplant researchers have proposed imputation of complete high-resolution genotypes conditional on available partial data.

Imputation studies use data on the frequency distribution of distinct HLA types within various ethnic/nationality populations.

Such a database is embedded in HaploStats, a web application provided by the National Marrow Donor Program Bioinformatics group.

A clinician can input the partial HLA type data available for a donor or recipient.

HaploStats outputs the frequency distribution of complete types conditional on the data provided.

The frequency distribution output by HaploStats is formatted in order of prevalence of specific *haplotypes:* groups of HLA alleles that are inherited in conjunction.

Some have suggested use of one or a few most prevalent haplotypes to impute an unobserved donor or recipient HLA genotype.

Haplotype imputation may seem an appealing way to overcome incompleteness of clinically available HLA data.

The appeal does not survive under scrutiny. The obvious issue is that a donor or recipient may not actually have the imputed haplotype, even if the imputation is the most prevalent haplotype within the relevant population.

When an imputed haplotype is not accurate, computed HLA mismatch may be incorrect, with consequent misprediction of transplant outcomes.

Accurate probabilistic prediction of outcomes requires consideration of the entire frequency distribution of haplotypes conditional on the HLA data the clinician possesses.

Using More Complete Knowledge of HLA Mismatch

Now consider transplant outcome prediction when a clinician has more complete knowledge of HLA mismatch than is available in the SRTR data.

Current risk assessments condition at most on two-digit (A, B, DR) mismatch.

More refined partial prediction is possible combining SRTR and HaploStats data.

The methodology has been developed in research on the *ecological inference* problem, with application to medical risk assessment in Manski (*QE*, 2018).

The Ecological Inference Problem

Let each member of a population be characterized by outcome y and covariates (z, x, w).

In the transplant setting, y is the length of graft survival. (z, x) are the (donor, recipient) attributes considered earlier, including knowledge of (A, B, DR) two-digit mismatch.

The symbol w expresses further information on HLA mismatch that is observed by a clinician but not coded in the SRTR data.

In our application w represents mismatches for HLA antigens C and DQ.
(HaploStats does not provide information on the DP antigen).

Suppose that data are available from two sampling processes.

One sampling process yields observable realizations of (y, z, x), but realizations of w are not recorded. This is the situation with the SRTR data.

The other sampling process yields observable realizations of (w, x, z), but realizations of y are not recorded. This is the situation with the HaploStats data, when z denotes ethnicity/nationality.

The two sampling processes reveal P(y|z, x) and P(w|z, x).

The term *ecological inference*, describes the problem of inference on P(y|z, x, w) given knowledge of P(y|z, x) and P(w|z, x).

The Law of Total Probability (LTP) relates these distributions. Hold (z, x) fixed at specified values. For any value of w, say w = j,

$$P(y|z, x) \quad = \quad P(w = j|z, x)\cdot P(y|z, x, w = j) \quad + \quad P(w \neq j|z, x)\cdot P(y|z, x, w \neq j).$$

Knowledge of P(y|z, x) alone reveals nothing about P(y|z, x, w = j).

Partial conclusions may be drawn if one has evidence revealing P(w = j|z, x).

Analysis is simple when the objective is to learn the probability $P(y > T|z, x, w = j)$ that a graft will survive for at least T years and one makes no assumptions about $P(y|z, x, w)$.

The identification region for $P(y > T|z, x, w = j)$ is the interval

$$P(y > T|z, x, w = j) \quad \in \quad [0, 1]$$

$$\cap \quad \left[\frac{P(y > T|z, x) - P(w \neq j|z, x)}{P(w = j|z, x)}, \quad \frac{P(y > T|z, x)}{P(w = j|z, x)}\right].$$

This result was sketched by Duncan and Davis (1953). A proof is given in Horowitz and Manski (1995), who extend the analysis to derive identification regions for conditional means $E(y|x, z, w)$ and α-quantiles $Q_\alpha(y|z, x, w)$.

Bounded-Variation Assumptions

Tighter bounds may be obtained by combining evidence on P(y|z, x) and P(w|z, x) with credible assumptions.

Manski (2018) characterizes the identifying power of *bounded-variation* assumptions. These are inequalities restricting how P(y > T|z, x, w) varies with (z, x, w).

The state of knowledge in transplant immunology makes various bounded-variation assumptions credible when considering kidney transplants.

Study of transplant immunology suggests that, considering any locus and holding all else equal, survival probability decreases as the number of mismatches at this locus grows.

This implies a set of bounded-variation assumptions.

Let (x, w) and (x', w') be alternative mismatch vectors, with (x', w') > (x, w) in the vector sense that (x', w') has at least as many mismatches as (x, w) in every locus and more mismatches in at least one locus.

Then it is credible to assume that, for any value of covariates z and survival length T,

$$P(y > T|z, x, w) \quad \geq \quad P(y > T|z, x', w').$$

Consider mismatches at different HLA loci.

There exist two main classes of loci, which operate biologically in different ways. Class I includes the (A, B, C) loci. Class II includes the (DP, DQ, DR) loci.

Study of transplant immunology suggests that DQ mismatches have the most severe consequences for graft rejection, followed by DR. Next in severity are mismatches at the A and B loci. It is thought that C mismatches have the least severe consequences.

These ideas about relative severity imply a set of bounded-variation assumptions.

Estimation of Bounds on $P(y > T|z, x, w)$

To illustrate, consider a (donor, recipient) case with (age = 50, KDPI = 50).

In the polar case with no mismatch at any of the five loci, the estimated bounds on one and five-year survival are [0.924, 1] and [0.823, 1].

In the other polar case with two mismatches at each of the five loci, the estimated bounds on one and five-year survival are [0.496, 0.927] and [0.544, 0.775].

These findings are reasonably representative in terms of the width of the bounds.

Conclusion

Previous studies have used SRTR data to predict kidney transplant outcomes conditional on donor and recipient covariates, including partial characterization of HLA mismatch.

Whereas previous studies assumed proportional hazards models, we use nonparametric regression methods.

These do not make the unrealistic assumption that relative risks are invariant as a function of time since transplant. Nor is relative risk assumed to be log-linear.

Clinicians and patients would like to refine the predictions possible with SRTR data.

It has been suggested that HaploStats data might be used to impute complete HLA types. We counsel against this.

Nevertheless, HaploStats data are useful when combined appropriately with SRTR data.

We explained the ecological inference problem and showed how to combine the two data sources, generating partial predictions of outcomes conditional on refined HLA typing.

Informative bounds are achieved when one brings to bear immunological knowledge of the effects of mismatch at different HLA loci.

While this analysis enables one to make the most of the available data, it would be better to collect more refined HLA data.

OPTN has not required transplant centers to report C and DQ types. It could do so.

It could also encourage reporting of four-digit rather than two-digit HLA types.

Book Conclusion

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

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

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

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

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

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

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

Then adaptive diversification of treatment can be valuable.

Separating the Information and Recommendation Aspects of Guidelines

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

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

Having guideline development groups characterize knowledge can improve practice.

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

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

Synthesis by expert panels seems essential.

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

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

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

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

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

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

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

Care recommendations may be contentious if perspectives vary.

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

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

Educating Clinicians in Care under Uncertainty

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

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

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

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