opment. Other protocol-specified prospective data collection programs address similar questions.

Examples of new or emerging treatments worthy of such review are easy to identify. They include accelerated partial breast irradiation to accompany lumpectomy in the treatment of localized breast cancer, bevacizumab to treat age-related wet macular degeneration, endovascular repair of thoracic aortic aneurysm, intracerebral stenting for the prevention of stroke, endobronchial values as an alternative to lung volume reduction surgery to treat emphysema, bronchial thermoplasty to treat moderate-to-severe asthma, and injectable bulking agents to treat vesicoureteral reflux in children.

Final Comments

Given the amount spent on health care in the United States, consumers of health services, professionals who provide those services, and purchasers who pay for them are entitled to know what works and what does not. They are entitled to know which health services are definitely beneficial, which are likely to be beneficial, which have insufficient evidence supporting their use to know if they are beneficial, and which services in common use today are known to be of no benefit or, worse, that are actively harmful. Persons making choices on which treatments to use should understand the range of treatments available to them, including advantages, harms, and alternatives. However, despite the plethora of information available today, such a "single source of truth" does not exist. The foregoing comments represent one attempt at defining the knowledge synthesis necessary to answer these vital questions.

METHODS THAT NEED TO BE DEVELOPED

Eugene H. Blackstone, M.D., Douglas B. Lenat, Ph.D., and Hemant Ishwaran, Ph.D.
Cleveland Clinic

Overview

RCTs and their meta-analyses are generally agreed to provide the highest-level evidence for comparative clinical effectiveness of clinical interventions and care. However, today cost and complexity impede nimble, simple, inexpensive designs to test the numerous therapies for which a randomized trial is well justified. Further, it is impossible, unethical, and prohibitively expensive to randomize "everything."

To fill this gap, balancing-score methods coupled with rigorous study design can approximate randomized trials. They are less controlled but

use real-world observational clinical data. They may provide the only way to test therapies when it is impossible to conceive of or conduct RCTs. Although a number of their important features remain to be understood and refined, they are comparatively inexpensive and use readily available electronically stored data. Interestingly, although the intent of EBM is to reduce practice variance, this methodology draws its power from heterogeneity of care.

Unfortunately, a longitudinal birth-to-death patient-centric health record, populated largely with discrete values for variables that would be useful for both streamlined randomized and balancing-score-based clinical trials, has not been brought to fruition. Instead, clinical information remains locked in narrative, mostly within segregated institutional silos. But a new methodology is emerging both to elegantly link these silos and to provide a population-centric view of clinical data for analysis: semantic representation of data. Meaning is emphasized rather than lexical syntax. This has the promise of transforming EBM into information-based medicine. Its elements include storage of patient data as nodes and arcs of graphs that can seamlessly link disparate types of data across medical silos, from genomics to outcomes, and, in theory, across venues of care to create a virtual longitudinal health record, to say nothing of the completely longitudinal personal health record. What is required are (1) a rich ontology of medicine, the taxonomy component of which is enough to enable semantic searching, and the formal knowledge base component, which is enough to permit—even today—natural language query of complex patient data (that is, separating logical understanding of query from the need to understand underlying data schemata); (2) a worldwide effort to assemble this ontology and the assertions that make it useful; and (3) intelligent agents to assist discovery of unsuspected relationships and unintended adverse (or surprisingly beneficial) outcomes.

But if such clinically rich data were available, especially a massive amount, could they be put to effective use? Computer-learning methods such as bootstrap aggregation (bagging), boosting, and random forests are algorithmic, as opposed to the traditional model-based methods that are computationally fast and can reveal complex patterns in patient genomic and phenotypic data. These methods refocus attention from "goodness of fit" to a given set of data to prediction error for new data. Methods like this are needed to propel the country yet another step toward personalized medicine.

Thus the results of trials, approximate trials, and automated discovery need to be transformed from static publications to dynamic, patient-specific medical decision support tools (simulation). Although such methodologies are widely used for institutional assessment and ranking, they need to lead

to clinically rich, easily used, real-time tools that integrate seamlessly with the computer-based patient record.

This article highlights five foundational methodologies that need to be refined or further developed to provide an infrastructure to learn which therapy is best for which patient. They are representative of those needed for progression from current siloed EBM to semantically integrated information-based medicine and on to predictive personalized medicine. The five methodologies can be grouped into three categories:

1. Evidence-based
   - Reengineering RCTs
   - Approximate RCTs
2. Information-based
   - Semantically interpreting, querying, and exploring disparate clinical data
   - Computer learning methods
3. Personalized
   - Patient-specific strategic decision support

Reengineering Randomized Controlled Trials

Following intense preliminary work, several cardiac surgical centers began designing a randomized trial to answer a simple question: Is surgical ablation of nonparoxysmal atrial fibrillation accompanying mitral valve disease effective at preventing the return of the arrhythmia? It took a short time—weeks—to design this study, but then it had to be vetted through committees, review boards, and the FDA, leading to multiple revisions, additions, and mounting complexity. The case report form became extensive and required considerable human abstraction of information from clinical records to complete. Two core laboratories were needed and competitively bid. After more than 2 years, the trial was launched. From inception to completion, the trial is likely to take 5 years at a minimum. The cost of what was intended to be a simple, easily deployed trial will be about $2 million; large multi-institutional, multinational trials may cost upwards of 10 times this figure.

Designing and executing RCTs like this has become one of the most demanding of human feats. It may not compare with climbing Mt. Everest, but it is close. A major reason to climb this mountain is that RCTs remain the gold standard for EBM. They are purpose designed, have endured ethical scrutiny, ensure concurrent treatment, capture highest-quality data, and have adjudicated end points. Their data meet the statistical assumptions of the methods used to analyze them.

Yet, like the trek up Everest, the design and conduct of an RCT is filled with pitfalls that need to be bridged. The following six areas are among those that must be addressed if RCTs are to achieve the kind of cost effectiveness that evidence-based medical practice requires in the future: complexity, data capture, generalizability, equipoise, appropriateness, and funding.

Complexity

A deep pitfall of the current practice of RCTs is what John Kirklin, pioneer heart surgeon, called "the Christmas Tree Effect": ornamenting trials with unnecessary variables rather than keeping them elegantly simple and focused. Every additional variable increases the cost and difficulty of the trials, which reduces available resources, limiting the number of trials that can be performed. Nonessential complexity constructs a barrier to progress when instead a bridge is needed. In reengineering RCTs, data collection should be focused on the small number of variables that directly answer the question posed. A series of elegant, scientifically sound, clinically relevant, simple, focused trials will provide more answers more quickly than bloated multimillion-dollar trials that are justified as providing enormous riches of high-quality data for later (observational) data exploration.

Second, rapid development of simple pilot trials on clinically important questions should be encouraged, to be followed with simple, definitive trials. The National Heart, Lung, and Blood Institute has put into place a number of disease- and discipline-specific networks of centers devoted to simple RCTs. This is an important step forward. Two observations: (1) The trials being designed are simple only in the number of patients enrolled, not in design; funding would be better spent on highly focused, extremely simple RCTs. (2) There is no plan for funding pivotal trials based on clinical outcomes rather than surrogate and composite end points that stem from these pilot trials (Fleming and DeMets, 1996). Perhaps the focus should, therefore, shift to funding a mix of simple, inexpensive pilot trials and simple but definitive trials.

Third, adding administrative and bureaucratic complexity to many RCTs is needed for investigational device exemptions and new drug exemptions from the FDA. This introduces considerable delay by an organization that should itself promote efficient study designs focused on safety and efficacy. The heterogeneity of institutional review board requirements adds further administrative burden.

Fourth, to "survive," design and conduct of RCTs has become a "business" that is increasingly specialized and complex and distanced from the practice of medicine. Physicians with good questions believe they cannot attempt to scale the mountain. It was not always this way, and patient

recruitment suffers from it because patients' personal physicians are often no longer advocates for clinical trials. Again, simplification is key to bridging this chasm.

All four of these complexities argue for applying a kind of *symbolic sensitivity analysis* when an RCT is designed, eliminating variables that are more decorative than functional.

Data Capture

RCT technology as practiced today makes little use of discrete data elements acquired as part of clinical practice. Available computer-based clinical data could and should be used for patient screening, recruiting, and data gathering. With electronic patient records composed of "values for variables" (discrete data elements), one could electronically identify patients meeting eligibility criteria for trials, generating alerts so that healthcare providers could be on the front line of informing patients about a trial germane to their treatment. Insofar as possible, patient data, including end points, should be retrieved directly from the electronic patient record. Instead, study coordinators today laboriously fill out case report forms, translating from medical records. Reducing the data-gathering burden would not only reduce complexity and cost but also bring trials more into the sometimes messy reality of clinical practice—the very environment for which inferences about clinical effectiveness from the trial are to be made. Admittedly, redundant data abstraction, end point adjudication, and core laboratories all contribute to incrementally improving the quality of trial data, but it is questionable whether the improvements justify the accompanying costs, their impeding the climb, and permitting more climbs.

Generalizability

RCTs often focus on patient subgroups (usually the lowest-risk patients, ostensibly to reduce potential confounding and for which equipoise is unquestioned, rather than the spectrum of disease observed in the community (Beck et al., 2004). Yet results of these restrictive trials typically are extrapolated to the entire spectrum, a practice that may be treacherous no matter what the trial shows. One of the first large, costly trials sponsored by the NIH was the Coronary Artery Surgery Study of the late 1970s and early 1980s (NHLBI, 1981). About 25,000 patients were entered into a registry of patients with coronary artery disease, but only 780 were randomized (Blackstone, 2006). Yet treatment inferences from the study were applied to a broad spectrum of patients with coronary artery disease (Braunwald, 1983). Although it can be argued that pilot studies should be conducted in the patient subgroups most likely to demonstrate a treatment difference

(so-called enriched trials), these studies should be used to aid developing inclusive trials of adequate power. Just as the data acquired from clinical practice is often taken too lightly today, the data acquired from these restricted RCTs is often taken too seriously, when in truth both of these turn out in hindsight to be no more—and no less—than valuable heuristics (Ioannidis, 2005).

Equipoise

Among physicians' areas of expertise and responsibility is the task of selecting the right treatment for the right patient at the right time. Surgeons call this "indications for operation." This is the antithesis of equipoise. Thus, a number of important trials have been stopped or considerably protracted for lack of enrollment. Across time periods, nationalities, and schools of thought, each physician will follow his or her own generally consistent but somewhat idiosyncratic set of rules for deciding appropriate treatment. Thus, whenever one examines clinical practice, considerable variance is seen. This gives hope that equipoise on important medical dilemmas might be found at times. However, it also suggests the possibility of capitalizing on practice heterogeneity to conduct studies that approximate RCTs, as described later in this text, rather than seeking artificial, unnatural equipoise.

Appropriateness

Most investigators developing RCTs concentrate on efficacy. Studies are powered for anticipated (often overly optimistic) efficacy, but rarely focus on short- or long-term safety. This is even true of trials conducted for FDA approval. Indeed, for cardiovascular devices, the track record of mandated FDA safety surveillance is dismal. It usually involves a small cohort of patients for whom there is little power to detect increased occurrence of adverse events, and it generally employs a follow-up time too short to detect untoward effects of long-term device implantation. Rare adverse effects caused by long-term exposure to devices (or pharmaceuticals) may go undetected for a long time, but when they are finally detected they incite public anger, recalls, and withdrawals of effective drugs and devices (Nissen, 2006). This reaction might be avoided if a proper surveillance program were in place with impartial analysis of data, possibly assisted by the computer learning technology discussed later in this paper. The factual reporting of findings and a measured response could convince industry, the public, regulators, and even skeptics that the process is transparent and timely (Blackstone, 2005).

Are all the clinical trials that are being performed actually necessary?

Just because a trial can be mounted is no reason to initiate inappropriate trials. At the end of the appropriateness scale is the proverbial parachute trial. Not only will there not be a randomized trial of efficacy of parachutes, there is no compelling reason to do such a trial; magnitude of the effect is too large and logically obvious, although we concede that logic can trip us up. Many trials are expected at the outset to show no difference in efficacy, and yet futile trials are done, often because a regulatory body has required it. Many equivalency, nonsuperiority, and noninferiority trials could be replaced by objective performance criteria and an intense surveillance program (Grunkemeier et al., 2006).

Funding

Typically, the costs of new pharmaceutical and device trials are borne by industry sponsors, with their attendant actual and potential conflicts of interest. Relative to this, only a small number of trials are sponsored by the NIH. Yet in an evidence-based medical system, the obvious benefactors are health insurers, and to a lesser extent the pharmaceutical and device manufacturers. Shouldn't insurers be interested in sponsoring clinically relevant RCTs, including making data available from the trials to the scientific community or at least bearing the patient costs of RCTs?

Approximate Randomized Clinical Trials

What effect does chronic exposure to urban pollution have on the risk of developing pulmonary disease or cancer? What is the effect of socioeconomic status on response to therapy? What is the effect on long-term outcomes of complete versus incomplete coronary revascularization? What is the effect of chronic atrial fibrillation on stroke? Can severe aortic stenosis be managed medically rather than surgically? Is the radial artery a good substitute for the right internal thoracic artery for bypass of the circumflex coronary system? These are but a few questions for which an evidence basis is needed. Some may be answerable with cluster randomized trials (Donner and Klar, 2000). Others require epidemiologic studies, and none seem readily amenable to randomized trials. It is not possible to randomize gender, disease states, environmental conditions, choice of ancestry, or healthcare organizations in local communities. It would be unethical to randomize patients to placebo or to incomplete or sham surgery when at least knowledge at the present time, if not solid data, indicates that to do so is unsafe. Thus, there is no knowledge in the modern era about the untreated natural history of certain diseases, such as critical aortic stenosis, hypoplastic left heart syndrome, transposition of the great arteries, untreated renal failure, unset fractures, untreated acute appendicitis, or jumping out of an airplane

without a parachute. Yet clinical decisions are made on incomplete evidence or flawed logic every day. Is it possible to do better than guessing? Is there an alternative to "randomizing everything?"

When literature comparing nonrandomized treatment groups is scrutinized, the natural response is to think, "They are comparing apples and oranges" (Blackstone, 2002). This is because in real-life clinical practice there remains wide variance in practice (that is, selection bias), and this results in noncomparable groups. If it is impossible to randomize patients or impractical or unethical, or if it can be demonstrated that one cannot draw a clean, causal inference even from a randomized trial (such as a trial that inextricably confounds treatment with the skill of the person implementing the treatment), is there a way to exploit the heterogeneity of clinical practice to make better comparisons that are closer to apples to apples? Basically, the goal would be to discover within the heterogeneity of practice the elements of selection bias and account for these to approximate a randomized trial.

A quarter century ago, Rosenbaum and Rubin (1983) introduced the improbable notion that observational data can be thought of as a broken randomized trial (Rubin, 2007), with an unknown key to the treatment allocation process. They proposed that the propensity of a patient to receive treatment A versus B be estimated statistically (for example, by logistic regression) to find that key. In its simplest form, a quantitative estimate of propensity for one versus the other treatment is calculated for each patient (propensity score) from the resulting statistical analysis and used for apples-to-apples comparisons (Blackstone, 2002; Gum et al., 2001; Sabik et al., 2002).

How does a single number, the propensity score, seemingly magically achieve a balance of patient characteristics that makes it appear as if an RCT had been performed (for that is exactly—and surprisingly—what it does)? It does so by matching patients with similar propensity to receive treatment A. A given pair of propensity-matched patients may have quite dissimilar characteristics but similar propensity scores. A set of such pairs, however, is well matched (Figure 2-3). What distinguishes these patients from those in an RCT is that at one end of the spectrum of propensity scores, only a few who actually received treatment *A* match those who actually received treatment *B,* and at the other end of the spectrum, only a few patients who actually received treatment *B* match those who received treatment *A*. Thus, balance in patient characteristics is achieved by unbalancing *n* along the spectrum of propensity scores (Figure 2-4). The generic idea is called *balancing score* technology, which can be extended from two treatments to multiple treatments, or even to balance continuous variables, such as socioeconomic status or age (Rosenbaum and Rubin, 1983).

Unlike an RCT in which the allocation mechanism (randomization)

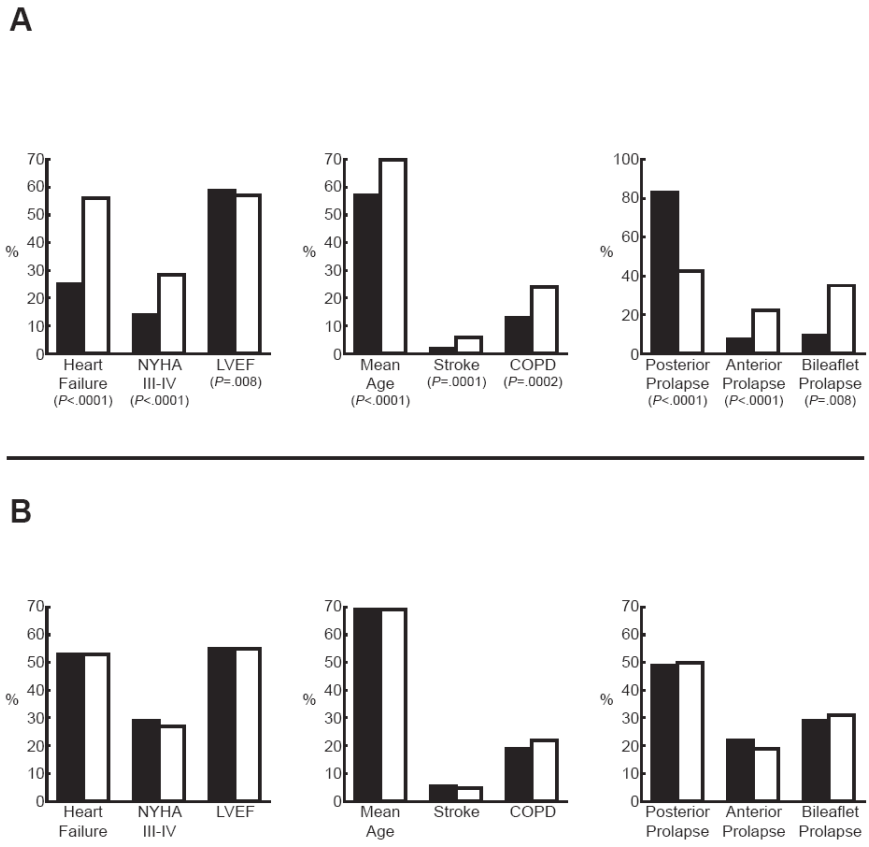


FIGURE 2-3 Comparison of patient characteristics before mitral valve repair (black bars) or replacement (unshaded bars). Unadjusted values are depicted in ***A*** and propensity-matched patients in ***B.***
NOTE: COPD = chronic obstructive pulmonary disease; LVEF = left ventricular ejection fraction; NYHA = New York Heart Association.

is known explicitly and equally distributes both known and unknown factors, propensity score methods can at best account for only those selection factors that have been measured and recorded, not for those that are unknown. Thus claims of causality, which are strong with RCTs, are weaker with propensity-based methods. This considerable disadvantage is, however, offset in a number of ways: (1) innumerable treatments can be studied at low cost based on heterogeneity of practice and availability of clinically rich data and (2) treatments or characteristics that cannot be randomized (e.g., gender, place of birth, treating facility, presence of

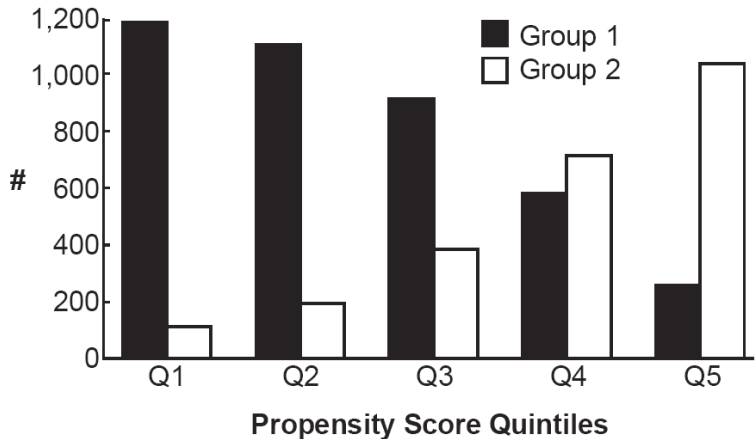


FIGURE 2-4 Achieving balance of clinical features by unbalancing *n*. Shown are two groups of patients that have been divided according to increasing quintile of propensity score. Notice at low propensity scores, the numbers of group 1 patients dominate over those of group 2, and at the other extreme, the numbers of group 2 patients dominate over those of group 1. Within each quintile, patient characteristics are well matched between groups, but these characteristics progressively change across quintiles (for example, low-risk profile in quintile 1 and high-risk profile in quintile 5).

disease) can be analyzed. Thus, there is broad applicability for a relatively inexpensive method.

It is important to say, however, that relying on clinical practice data alone is potentially irresponsible, biased, and dangerous, much like standing on untested terrain that may turn out to be thin ice, and patterns may turn out later to be artifactual "false peaks." However, these techniques may play a valuable role as a *heuristic* for helping to point RCTs in promising directions when that is possible and as better evidence than apples-to-oranges comparisons when it is not.

Taking yet another step backward, it has been claimed that traditional multivariable analysis is equally accurate in making risk-adjusted nonrandomized comparison (Sturmer et al., 2006). The problem, however, is that until now, there has been no independent support for this claim. It may be right more than 80 percent of the time, but what about the other 20 percent? Propensity-based methods provide this independent assessment. In addition, they also permit comparison when important clinical outcomes occur at a low frequency by supplying a single risk-adjustment variable: the propensity score (Cepeda et al., 2003).

Propensity methodology (and balancing scores in general) should be

elevated further. First, because propensity models are predictive ones (predicting which treatment was selected), the computer learning approach presented later in the text could be exploited to account for possibly complex interactions among selection factors. Second, comparisons based on clinically rich vs. administrative vs. electronically available laboratory databases should be tested for relative value. Third, the most appropriate method of comparing outcomes after propensity matching remains controversial and probably requires developing new statistical tests.

Semantically Interpreting, Querying, and Exploring Disparate Clinical Data

Computerized Patient Records

In 1991 the Institute of Medicine described what it called the computer-based patient record (CPR) (Barnett et al., 1993; IOM, 1991). Its creators envisioned a birth-to-death, comprehensive, longitudinal health record that contained not just narrative information but also values for variables (discrete data) to allow the record to be active, generating medical alerts, displaying trends, providing meaningful patient-level clinical decision support, and facilitating clinical research. It would not be simply an electronic embodiment of the paper-based medical record, which is what they believed the emerging electronic medical record (EMR) was.

The need for a CPR is, if anything, more acute today than it was in the early 1990s because of the increased complexity of care, the aging of the population with multiple chronic diseases, and the multiplicity of care venues from shopping malls to acute care facilities to clinics to large hospitals, to say nothing of OTC medications and a proliferation of alternative and complementary therapies, public awareness of clinical outcomes, the need to track unanticipated complications of therapy across time, and a cumbersome built-in redundancy of clinical documentation for reimbursement.

The originators recognized most of the same impediments to implementing such a system as still exist, not the least of which was that medical education would need to be altered to train a new generation of physicians how to use this new technology optimally.

What has not been clear to developers of EMRs is how discrete data might provide the underpinnings for a learning medical system. Nor did those who were willing to give up pen and ink and adopt the electronic record demand discrete data gathering as a by-product of patient care. Thus, before describing various methods that exploit discrete medical data, it is important to ask why discrete data is an asset and envision what could be done with this asset. For individual care, discrete data can be used to generate smart alerts based on the real-time assessment of data by algo-

rithms, care plans, or models developed on the basis of past experience. For informed patient consent, patient-specific predicted outcomes of therapy can be displayed based on models that are risk adjusted for individual patient comorbidities and intended therapeutic alternatives (see the later section, "Patient-Specific Decision Support"). From a population-centric vantage point, discrete data can provide outcomes and process measures for quality metrics and necessary feedback for improving patient care. This is in part because discrete data can make possible the automated compiling of quality outcomes and process measures along with variables needed for proper risk adjustment. Discrete data assist institutions in responding to clinical trials eligibility specifications to determine feasibility of studies and provide historical outcomes for estimating sample sizes. In addition, discreet data could alert physicians that a patient being seen satisfies all eligibility criteria for a clinical trial. Discrete data coupled with an intelligent query facility can be used to identify patient cohorts for observational clinical studies and approximate clinical trials. They provide the observational data for developing propensity scores, balancing scores, and conducting studies of comparative clinical effectiveness. If a true longitudinal record is created, then discrete data may identify adverse events and the substrate by which unsuspected correlated events may be identified, quite possibly with the use of artificial intelligence and computer learning techniques.

Computer-Based Patient Record Efforts at the University of Alabama at Birmingham

Kirklin and Blackstone, then at the University of Alabama at Birmingham (UAB), recognized the formidable barriers to the CPR and in October 1993 embarked on a $23 million proof-of-concept CPR in partnership with IBM. Initially, they sought an object model of medicine. Two simultaneous efforts to accomplish this resulted in the same conclusion: There is no object model of medicine because "everything is related to everything." Requirements for complex relationships coupled with the extensibility needed to keep pace with rapid medical advancement, assimilation of disparate types of data, provisions for examining data from multiple vantage points (e.g., viewing diabetes from the vantage points of genetics, anatomy, endocrinology, laboratory medicine, pharmacology, and other medical perspectives), and the feeding of computer systems without slowing patient care were huge challenges. IBM brought to the table experts in a host of different types of databases and concluded that nothing existed that would satisfy the IOM's vision of an active CPR. However, a novel vision for a system emerged from the collaboration that would be infinitely extensible, self-defining, active, secure, and fast (response time less than 300 milliseconds to those using the system clinically). It required that the container hold-

ing the data know nothing of its content and thus be *schemaless*. Rather, values for variables themselves needed to be surrounded by their context (metadata) (Kirklin and Vicinanza, 1999). Such a system was built on the IBM-transaction processing facility platform, the same as used at that time by airlines and banking. Its major unsolved problem, however, was cross-patient (population centric vs. patient centric) queries: In theory, an infinitely extensible, comprehensive, centralized data store could take an infinite time to query.

Semantic Representation of Data and Knowledge

Meanwhile, computer scientists at Stanford University (Abiteboul et al., 1997) and the University of Pennsylvania (Buneman et al., 2000) were developing methods to query semistructured (schemaless) data stored as directed acyclic graphs (DAGs). We recognized that the storage format of our UAB data could also be considered DAGs and be queried by the techniques those investigators were developing. Blackstone's move to Cleveland Clinic in late 1997 provided the opportunity to pursue development of the CPR, but in the test bed of a highly productive cardiovascular clinical research environment. Clinical researchers know, of course, that discrete data are required for statistical analysis, and for the preceding 25 years, human abstractors at the clinic had laboriously extracted data elements from narratives for every patient undergoing a diagnostic or interventional cardiac procedure, resulting in the Cardiovascular Information Registry. We also found that other investigators at the clinic had developed more than 500 clinical data registries, often containing redundant, unadjudicated, non-quality-controlled data about various aspects of medicine—even of the same patient—stored in disparate clinical silos, such as orthopedics, cardiology, oncology, and ophthalmology. For the most part, these registries did not communicate with one another.

We therefore continued our work in developing what we then called a *semantic database* that, like any DAG representation, could be extended infinitely, was self-defining, and was also self-reporting by use of intelligent agents. Some 15 years and $50 million later, we at last have a technology that can underlie an extensible multidisciplinary CPR without the need for special integration, because it is natively integrated. Each data element in such a system is a node or an arc that connects nodes (databases) in a graph resource description framework), along with context and meaning (knowledge base). Additional nodes represent medical concepts and these are all linked. Each node has an address just like an Internet in a thimble. The Internet analogy is not an empty one. The infrastructure for the World Wide Web (Cleveland Clinic is 1 of some 400+ organizations worldwide that make up the World Wide Web Consortium) is the prime example of a

container that is ignorant of content, has all the properties of a DAG, and can easily be extended to assimilate new concepts that have never before entered the mind of humankind. Our test data set for cardiovascular surgery contains 23 million nodes (terms) and 93 million relationships (statements) representing 200,000 patients.

What are the advantages of such graph structures besides infinite extensibility? First, medical taxonomies, such as those of Systematized Nomenclature of Medicine (SNOMED) (Schulz et al., 2009) or the National Library of Medicine's metathesaurus (UMLS [unified medical language system]) (Thorn et al., 2007), underlie the data model and enable semantic searches. An investigator can search for patients and their data without knowing anything about underlying data structure. Specifically, this is achieved by separating semantics from the underlying syntax, in much the same spirit as the vision for the semantic web (Berners-Lee et al., 2001). Rather than being confined to lexical searches for information, a semantic web search is based on meaning. An example of this is the contrast between a dictionary based on meaning, such as the *American Heritage Dictionary* (Pickett, 2000), and one based on lexical definitions, such as *Merriam-Webster* (*Merriam-Webster Dictionary*, 2004). Thus, a heart attack, myocardial infarction, MI, acute myocardial infarction, AMI, and the variety of ways this medical concept may be expressed in both language and specific idiosyncratic syntax in a given database, are all recognized as a meaningful single semantic concept. Conversely, when the meaning of a term (such as *myocardial infarction*) *changes*, there is no semantic confusion because at the semantic level those are separate terms (Thygesen et al., 2007). There is a many-to-many mapping between lexical terms and their semantic denotations; the latter are the loci of medical knowledge.

Second, patients' graphs are connected by a data model to both general and medical *ontologies*, not just controlled term lists or taxonomies. These ontologies are built on a skeleton of taxonomically arranged concepts, but they contain as many—and as sophisticated—assertions *about* those concepts as are needed to compose an adequate model of an area of practice (Buchanan and Shortliffe, 1984). Think of an orthopedic ontology: It contains not only a taxonomy of all the bones in the body, but also assertions about them, such as "the knee bone's connected to the thigh bone, and the thigh bone's connected to the hip bone" (Weeks and Bagian, 2000), the type of joints between them, relative sizes, and so on.

Third, because natural language queries that seem clear to human investigators are fraught with ambiguous terms and grammatical constructions (e.g., attachment of prepositional phrases), pronouns, elisions, and metaphors, the knowledge represented in rich ontologies (vs. a taxonomy) suffices—barely—to permit investigators to ask database questions in natural language rather than in the language of a database expert. For the last

few years, Cleveland Clinic has collaborated with Douglas B. Lenat and his group in Austin, Texas, who, for the last 24 years, have built a top-down ontology of general concepts that starts with "thing" at the top and goes all the way down to such domain-specific concepts as "kidney" and "dialysis," and millions of general rules and facts that interrelate and, therefore, partially define those terms and model a portion of human knowledge (Lenat and Guha, 1990). Not surprisingly, to cope with divergence across humans' models of the world, that ontology—Cyc—required its knowledge base to be segmented into locally consistent (but *only* locally consistent) contexts. Since 2007 a group of us from Cleveland Clinic and Cycorp have worked together to tie low-level medical ontology concepts to the general Cyc ontology of things.

An investigator can now type into a Semantic Research Assistant™ a simple English sentence such as "Find patients with bacteremia after a pericardial window." Although complete automatic parsing of realistically large and complex investigator queries is still far beyond today's state-of-the-art artificial intelligence software (Lenat, 2008), one thing that is possible today, and which the current system does, is to successfully extract entities, concepts, and relations from the text as it understands the meaningful fragments of the query. These fragments are understood as logical clauses (in the system's formal representation), each of which is translated into a short, comprehensible English phrase and presented to the investigator. The investigator selects those fragments believed to be relevant, at which time an amazing thing happens almost every time: There is only a single semantically meaningful *combination* of those fragments, and only a single query that makes sense, given common sense constraints, domain knowledge constraints, and discourse pragma. *Combining* the fragments entails, for example, deciding which variables from each fragment unify with variables from other fragments, or whether they represent separate entities, and deciding whether each variable should be quantified existentially or universally, and in what order. The full query is then assembled, an English paraphrase of it is presented to the investigator, and a SPARQL translation of it is presented to the semantic database, which returns answers that are displayed to the investigator. Often, in the course of this process, some clauses that were not explicitly included by the investigator can be suggested; at other points in the process, the investigator may tweak the query by replacing a term with one of its generalizations or siblings or descendants in the ontology.

Fourth, a semantic-ontology approach also permits truly intelligent patient search of medical concepts. This is becoming increasingly important as patients seek out information about their medical conditions. A patient might type into a medical semantic search engine, "I have a racing heart." The semantic search engine produces a number of hits that don't include

NASCAR racing but rather tachycardias, such as atrial fibrillation, presenting the patient with definitions and treatment options.

What now needs to be developed to implement semantic databases and knowledge bases for intelligent search of all of medicine is a comprehensive formal ontology of medicine. This will require a worldwide effort. Already some of this is going on. For example, the Cardiovascular Gene Ontology provides full annotation for genes associated with cardiac disease processes.[13]

In the future such systems may actively ask relevant questions about correlations and trends within longitudinal records by means of true artificial intelligence. Automated intelligent agents could assist in discovering unsuspected relationships, unintended adverse outcomes, and surprising beneficial effects (AAAI, 2008). It could be central to realizing a learning medical system, a key component of what 21st-century medicine must become.

Computer Learning Methods

As much as one can dream of a longitudinal database that might permit innovative research for information-based medical care, it is important to ask, "If we actually had these data, would we know what to do with them?" One useful way to look at the issue is to use a "trees and woods" analogy in which individual patients, their data, and their genes are like the individual trees, and groups of patients or populations are the woods (Blackstone, 2007). The expression "Ye can't see the wood for the trees" (Heywood, 1546) implies that there may be patterns in the wood that can be discerned by overview that are not visible by attention paid only to individual trees. Here is an example: If one sits on the sidelines of an Ohio State University football game, one can only see individual band members playing at half-time and their feet moving around. But from an aerial view, one can see the band is in formation spelling the word *Ohio*. Patterns in medical data represent the general ways that patients react to their disease or treatment. They are the incremental risk factors, the modulators, or the surrogates for underlying disease and treatment mechanisms (Kirklin, 1979).

The rapidly developing science of computer learning promises methods far more robust than traditional statistical methods for discovering these patterns (Breiman, 2001). Many of them are based on multiple bootstrap samples (Diaconis and Efron, 1983; Efron, 1979, 1982; Efron and Tibshirani, 1986), each of them analyzed and aggregated (Breiman, 1996). This can be illustrated by analyzing 15 potential risk factors for death after

[13] See http://www.geneontology.org/GO.cardio.shtml (accessed September 8, 2010).

mitral valve surgery. These are designated A through O in panel A in Figure 2-5, which shows the first five bootstrap analyses. The tall vertical bars designate variables identified in each analysis. Note that no analysis yields identical risk factors. But now consider a running average of these results (Figure 2-5, panel B). Notice the running average of these unstable results progressively reveals a clear pattern: Variables A, C, D, I, and J are signal, and the rest are noise (Figure 2-5, panel C).

Imagine extending this concept. For example, at each iteration the algorithm could average the contribution of a predictor based on its appearance in previous iterations (boosting)—an adaptive weighted average (Bartlett et al., 2004; Freund and Schapire, 1996; Friedman, 2001, 2002). Bagging produces an average, but unlike boosting, it uses the same weight for each iteration.

Other computer learning techniques are being developed, such as Bayesian analysis of variants for Microarray methodology, which is being used to discover empiric gene expression profiles that are highly predictive for colorectal cancer recurrence (Ishwaran and Rao, 2003; Ishwaran et al., 2006). Unsupervised hierarchical bootstrap clustering almost completely separates patients experiencing cancer recurrence from those whose cancer has not recurred. What is important to recognize is that these methods solve the problem of having a large number of parameters (P) compared to number of individuals (n), a key factor in genomic analysis and research.

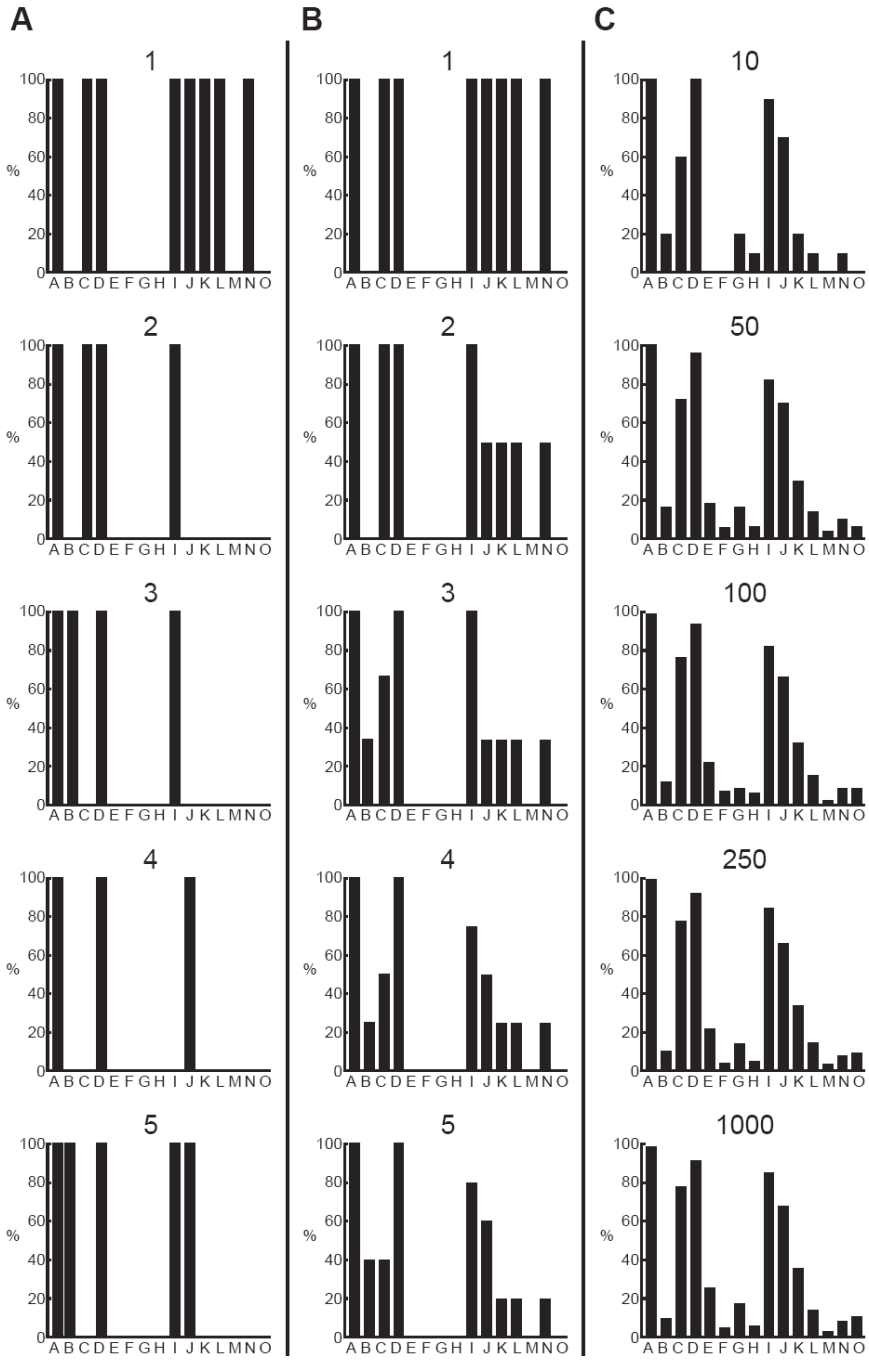
These methods are still in their infancy; many are based on computer-intensive methods such as bootstrap sampling or random forest technology. Variables may be selected by importance value (Breiman, 1996, 2001; Ishwaran, 2007) or by signal-to-noise ratios rather than by traditional P values, which become progressively less useful as n becomes large. Prediction error is minimized rather than maximizing goodness of fit.

An important feature of all ensemble learners is that they are computationally highly parallelizable—either for large-scale parallel computers or for grid computing. This may become important as researchers start looking at a huge number of patients, when speed of computation for clinical inferencing may be of the essence.

Patient-Specific Strategic Decision Support

Finally, to come full circle, consider personalized medicine. Joel Verter once said that RCTs are "sledgehammers, not fine dissecting instruments." Medicine needs to head toward fine dissecting instruments, toward personalized strategic decision support. With $n = 1$, a new paradigm of RCT needs to be developed for genomic-based personalized medicine (Balch, 2006).

Consider a 59-year-old man with ischemic cardiomyopathy and anterior MI resulting in left ventricular aneurysm. He has an ejection fraction



of 10 percent; 4+ mitral valve regurgitation; extensive coronary artery disease, including 90 percent left anterior descending coronary artery stenosis; and multiple comorbidities. Should the recommended therapy be continued medical treatment, coronary artery bypass graft (CABG), CABG plus mitral valve anuloplasty, a Dor operation, or cardiac transplantation? This complex information is too multidimensional for assimilation by the human mind. It calls for a cognitive prosthesis (Reason, 1999). Ideally, this patient's data would be entered automatically by a CPR into a strategic decision aid and the long-term expected survival would be depicted for multiple alternative therapies along with uncertainty limits, although not all therapies may be applicable.

Locked in the medical literature even today are static risk factor equations that could be used in dynamic mode for strategic decision support for a patient such as this (Levy et al., 2006). Random forest technology also can generate outcome risk estimates for individual patients by "dropping" their characteristics down a forest of trees, where they will land at a specific node in each tree with patients having similar characteristics and known outcome. Results of all patients at each node become the average ensemble predicted outcome for an individual patient. Thus, it is possible to imagine that in the future there will be methods by which patient-specific prediction of outcomes are generated and alternative therapies compared for patient decision support.

A library of modules must be developed for constructing strategic decision aids such as this. These in turn must be coupled to values for variables in a CPR so that no human intervention is required to depict comparable predictions of results. Then it must be prospectively verified that the simulated results match actual outcomes. The medical record thus becomes an active revealing and learning tool.

FIGURE 2-5 Example of automated variable selection by bootstrap aggregation (bagging). Fifteen variables labeled A through O are depicted as potential predictors of death after mitral valve surgery. In column A, analyses of five bootstrap samples are shown. Tall bars indicate the variable was selected at $P < .05$, and gaps represent variables not selected. Variables A and D were selected in all cases, but otherwise the analyses appear to be unique. Panel B shows a running average of these five analyses. Variables A, D, I, and J were selected more often than others. Panel C shows averages of 10, 50, 100, 250, and 1,000 bootstrap analyses. Notice that no variable was selected 100% of the time, and all 15 were selected at one time or another. But if variables appearing in 50% or more analyses are considered reliable risk factors, then variables A, C, D, I, and J fit that criterion of "signal," and the rest are "noise."

Summary

Moving beyond today's Mt. Everest level of difficulty, RCTs need to become more nimble and simple to better reflect the real world and to have their financing restructured. Heterogeneity in practice facilitates approximate randomized trials via propensity score methods that are inexpensive and widely accessible but which require patient-level clinical data stored as discrete values for variables. Emerging semantic technology can be exploited to integrate currently disparate, siloed medical data—responding to investigators' complex queries and patients' imprecise ones—and in the near future holds the promise to automate discovery of unsuspected relationships and unintended adverse or surprisingly beneficial outcomes. A next generation of analytic tools for revealing patterns in clinical data should build on successful methods developed in the discipline of machine learning. Both new knowledge learned and resulting algorithms should be transformed into strategic decision support tools. These are but a few concrete examples of methods that need to be developed to provide an infrastructure to determine the right treatment for the right patient at the right time.

Resources Needed

What resources are needed to develop this infrastructure?

Reengineering Randomized Controlled Trials

The cost of an NIH-sponsored simple trial appears to be in the range of $2 million, but multi-institutional, multinational large trials driven by clinical end points can consume 10 times that figure. If one uses $100 million as a metric, this means 5 to 50 such trials of therapy can be supported. Considering all the therapies of medicine for which the evidence base is weak, it is clear that demanding gold-standard RCTs for everything is unaffordable. The cost of RCTs that are highly focused, ethically unambiguous, and feasible could be brought down to a quarter, perhaps even a tenth, of this figure based on practical experience. This will require maximum use of electronic patient records, consisting of values for variables, and quite specifically longitudinal surveillance data to study the long-term side effects of therapies.

Approximate Randomized Controlled Trials

The NIH and National Science Foundation (NSF) should join forces and solicit 3-year methodology grants of approximately $250,000 per year,

10 per year. For this $7.5 million investment, a strong understanding of how best to use nonrandomized data would emerge. With this would come production of publicly available statistical software.

If rich discrete clinical data were available for analysis, a typical study using these methods for nonrandomized comparison would cost approximately $75,000. The cost would double if extensive integration of data was necessary, possibly over healthcare networks. For $100 million, it would be possible to conduct more than 1,000 such approximate randomized trials. This could have a major impact on acquiring what might be called "silver-level" evidence for practice.

Semantically Integrating, Querying, and Exploring Disparate Clinical Data

Based on several years of work, it seems that developing a comprehensive ontology of medicine—a new framework for analysis across disparate medical domains—will cost about 1 hour of time per term for an analyst, programmer, and clinical expert. One need not start from scratch, but can exploit SNOMED, UMLS, and other term lists and ontologies to start the process. Assuming that 100,000 terms would need to be defined in this fashion, that the wages would be $300 per hour, and that 25 ontologists would be needed, this work could be completed in 2 years at a cost of $36 million. This would include the software that must be programmed to implement a global effort in rallying medical experts to this task.

Computer Learning Methods

Knowledge discovery in medicine involves both methodologic development and applications. These should go hand in hand in this new field because it would accelerate the development of methods as they encounter problems requiring further methodologic work. The NSF has begun an initiative called Cyber-Enabled Discovery and Innovation (Jackson, 2007). This began with a $52 million first-year budget and is intended to ramp up $50 million per year and finish within 5 years for a total of $750 million. It would be useful to add $10 million per year for direct application to biomedicine, for a total sustained level for these activities within 5 years of $50 million.

Patient-Specific Strategic Decision Support

Costs in this area are largely for developing software, including the interfaces to EMR systems. This could be done for approximately $10 million. One could envision every study of clinical effectiveness having a

patient-specific prediction component built into it. Again, based on experience doing this, approximately $25,000 per study would be required to adapt and test the software and couple it with EMRs for decision support. It is also likely that at some point, the FDA may become involved with tools such as this and would introduce regulations that are more costly to meet than those of performing the studies.

COORDINATION AND TECHNICAL ASSISTANCE THAT NEED TO BE SUPPORTED

Jean R. Slutsky, Director, Center for Outcomes and Evidence, Agency for Healthcare Quality and Research

Overview

CER as a concept and reality has grown rapidly in the past 5 years. While it builds on an appreciation for the role of technology assessment, comparative study designs, and the increased role of health information technology to gather evidence and distribute it to the point of care, the capacity and infrastructure for this research has received less targeted attention. Understanding the landscape of organizations and health systems undertaking CER is challenging but essential. Without knowing what capacities and infrastructure currently exist, rational strategic planning for the future cannot be done. It is also important to address which functions can be most effective if they are centralized, which are most effective if they are local or decentralized, and how different activities relate to each other in a productive way. This paper will explore the practical realities of what exists now, what is needed for the future, and how the needs of the country's diverse healthcare system for CER can best be met.

The Agency for Healthcare Research and Quality Perspective

AHRQ plays a significant role in CER. Under a mandate included in Section 1013 of the Medicare Prescription Drug, Improvement, and Modernization Act of 2003, AHRQ is the lead agency for CER in the United States. AHRQ conducts health technology assessment at the request of CMS and analyzes data and suggests options for coverage with evidence development (CED) and post-CED data collection. AHRQ also provides translation of CER findings, promotes and funds comparative effectiveness methods research, and funds training grants focused on comparative effectiveness. AHRQ has an annual budget of over $300 million ($372 million for 2009), and received funds specifically for work on CER ($30