BLUE PILL OR RED PILL: THE LIMITS OF COMPARATIVE EFFECTIVENESS RESEARCH

Tomas J. Philipson University of Chicago

Eric Sun Stanford University

Comparative Effectiveness Research (CER) measures the effects of different drugs or other treatments on a population, with the goal of finding out which ones produce the greatest benefits for the most patients. Used properly, CER gives the patient, doctor, and payer hard information from thousands, or even millions, of cases, saving them time and money that otherwise would be spent on a trial-and-error quest for the right treatment.

Public and private payers for health care hope to use CER to cut costs without reducing quality of care. Great expectations have been placed on this approach. "If there's broad agreement ... [that] the blue pill works better than the red pill," President Obama has said, "and it turns out the blue pills are half as expensive as the red pill, then we want to make sure that doctors and patients have that information available to them."

The potential short-term savings are significant. For example, antipsychotic drugs represent one of the largest and fastest-growing expenses for Medicaid. In 2005, a CER analysis of antipsychotic drugs found little difference between the effectiveness of older, cheaper antipsychotics and that of more expensive "second-generation" drugs. We determined that if reimbursement policies had been changed in response and Medicaid had stopped paying for the more costly drugs, it would have saved \$1.2 billion out of the \$5.5 billion that it spent on these medications in 2005. However, the consequences of this policy shift would have been worse mental health for many thousands of people, resulting in higher costs to society that would equal or outweigh any savings in Medicaid costs.

This result seems counterintuitive: How can it be that, when a CER study shows no difference between two drugs, limiting coverage for the more expensive drug could actually increase costs? The answer is that in most CER studies, it is the drug or treatment with the larger average effect on an entire population that "wins." In the president's hypothetical, the blue pills are "just as effective" as the red ones because, on average, they do as much good for patients. But the average patient is not the same as any particular individual patient. Declaring a treatment most effective based on an average is a medical and an economic error, for two reasons.

First, individuals differ from one another and from population averages. Therefore, what may be on average a "winning" therapy may simply not work for a large number of patients. Conversely, a drug that is less effective on average may still be the best, or only, choice for a sizable proportion of patients.

The second reason is the variance in dependence in patient responses across therapies. Dependence, for any individual patient, is the degree to which response to one treatment predicts response to another. Dependence varies from illness to illness and from drug to drug but is often an important aspect of finding treatments that work. One cannot know in advance, as a general rule, that Drug A's failure guarantees the failure of Drug B. Yet a reimbursement policy based on CER could well make this error: by refusing to reimburse Drug B on the grounds that Drug A is "more effective," such a policy assumes that failure with Drug A will predict failure with Drug B.

To understand the effect of these points on costs, we looked at the real-world consequences of applying CER results to the antipsychotics we mentioned. These drugs are one of the largest classes of medication for Medicaid patients, and the program's expenditures on antipsychotics are among its fastest-growing: they rose from \$1 billion in 1995 to over \$5.5 billion in 2005.

In 2005, a national CER study, the Clinical Antipsychotic Trials in Intervention Effectiveness (CATIE), compared the effects of first-generation, cheaper antipsychotics with drugs discovered later. The CATIE study found that second-generation antipsychotics were no more effective at treating schizophrenia symptoms than are first-generation drugs. Naturally, this led to calls for Medicaid to limit reimbursement for second-generation antipsychotics.

As this debate continues, we set out to answer a simple empirical question: Would potential reimbursement policies based on the CATIE actually save money on health-care costs? Or would the effects of difference and dependence undo the cost savings?

We found that the latter is the case. Our analysis focused on antipsychotic coverage for roughly 250,000 non-elderly adult Medicaid enrollees with schizophrenia. First, we considered an extreme case: denial of all coverage for second-generation antipsychotics, on the grounds that the cheaper first-generation drugs are just as effective. We found that that this hypothetical policy would save Medicare \$1.2 billion, compared with full coverage. However, we estimate that it would reduce patient health by 13,138 quality-adjusted life years (QALYs) because of reduced health among the 75 percent of patients who were not responsive to first-generation antipsychotics and who, because of the restrictive policy, received no other drug therapy. Given that QALYs are typically valued at \$100,000, this suggests that the savings from denying coverage for second-generation antipsychotics (\$1.2 billion) would be outweighed by the costs of reduced health for patients (\$1.3 billion).

The second hypothetical policy we considered would cover perphenazine and risperidone (which are available in less costly generic forms) but exclude olanzapine (which is not). This policy would save Medicaid \$500 million annually but reduce health by 10,146 QALYs, mainly because of reduced health among patients who are unresponsive to either risperidone or perphenazine and who receive no therapy for six months or longer because of the restrictive policy. At a value of \$100,000 per QALY—again, the typical value assumed in the scholarly literature and by many payers—the health loss is nearly double the savings to Medicaid. Even at a value of \$50,000 per QALY, such a policy would only "break even." Therefore, using the CATIE findings to support restrictive coverage policies would not be cost-effective. It would limit freedom of choice for doctors and patients and yield no real compensation in savings.

We do not suggest that CER be dropped from the tool kit of private and public payers who want to cut costs while maintaining quality. On the contrary: we know that CER will become only more important to policymakers in the future. The 2009 federal stimulus law allocated \$1 billion for CER programs, and the 2010 health-care overhaul created an institute to promote CER and disseminate the results of this research to doctors and payers. The 2010 law also rescinds a prohibition on the use of CER for coverage decisions by Medicare. In the meantime, insurance companies and other private payers are also on the bandwagon. A recent survey found 85 percent of such organizations expecting that CER will soon be used to justify changes in reimbursement policies.

Our results suggest that CER will not fulfill its promise unless it is implemented differently by researchers and understood differently by policymakers. Simply put, seeking the treatment that is most effective on average will not improve health or save money. However, CER can be conducted in a way that takes difference and dependence into account and measures their effect. If CER is applied in this way—as a tool for matching individual patients to the best treatments for those individuals—it will realize its potential to reduce costs without inhibiting freedom of choice for doctors and patients.

CONTENTS

- 1 How CER Works
- 3 An Illustration: The CATIE Trial
- 5 Improving CER Evidence Metrics and Reimbursement Strategies
- 6 Conclusion
- 7 Appendix
 - Figure 1. Example of Sequencing of Treatments
 - Figure 2. Differences and Dependence in Treatment Effects
- 10 Endnotes

Blue Pill or Red Pill: The Limits of Comparative Effectiveness Research

About the Authors

TOMAS J. PHILIPSON is chairman of the Manhattan Institute's Project FDA. A managing director at Precision Health Economics, Philipson is also the Daniel Levin Chair in Public Policy at the Irving B. Harris Graduate School of Public Policy Studies and a member of the Department of Economics at the University of Chicago. In 2003-2004, Philipson served in the U.S. government as senior economic adviser to the commissioner of the Food and Drug Administration (FDA), and from 2004-2005 he was senior economic adviser to the administrator of the Centers for Medicare and Medicaid Services. He is the recipient of several international and national awards including the Kenneth Arrow Award of the International Health Economics Association in 2000 and 2006 (for best paper in health economics). Philipson is a co-editor of the journal *Forums for Health Economics & Policy* of Berkeley Electronic Press and is on the editorial boards of the journal *Health Economics* and the *European Journal of Health Economics*. Philipson earned his undergraduate degree in mathematics at Uppsala University, in Sweden, and his M.A. and Ph.D. in economics from the University of Pennsylvania.

ERIC SUN is a resident in the department of anesthesiology at Stanford University and a visiting fellow at the Bing Center for Health Economics at the RAND Corporation. His research has examined the costs and benefits of medical research and development, the role of the FDA and product liability in ensuring drug safety, and the economics of global public health. Sun's work has been published in the *Journal of Health Economics, American Journal of Managed Care, Journal of Public Economics, Health Affairs, Health Economics, Health Services Research, and BE Press Forum for Health Economics*. He holds an A.B. in molecular biology from Princeton University, an M.D. from the University of Chicago, and a Ph.D. in business economics, also from Chicago.

BLUE PILL OR RED PILL: THE LIMITS OF COMPARATIVE EFFECTIVENESS RESEARCH

Tomas J. Philipson & Eric Sun

n their quest to rein in costs without compromising quality, public and private payers for health care lately have placed hope in Comparative Effectiveness Research (CER): studies that compare alternative treatments for a given condition, with the aim of finding those that provide the most benefit to the most patients. In its report to the president and Congress, the Federal Coordinating Council for Comparative Effectiveness Research explains: "The purpose of this research is to improve health outcomes by developing and disseminating evidence-based information to patients, clinicians, and other decision-makers, responding to their expressed needs, about which interventions are most effective for which patients under specific circumstances." The intuition of these comparisons (often randomized clinical trials, though they are sometimes observational studies) is simple: doctors and patients can save time and money, as they go through the trial-and-error process of finding the right treatment, by knowing what is most effective for the whole population. CER is expected to raise the total health benefit and lower total spending by increasing the amount of useful information about each potential treatment for a given disease, be it diabetes, heart and lung ailments, schizophrenia, or a host of other conditions.

Although CER mainly compares treatments based on clinical utility, it is a small leap from clinical comparisons to economic ones. (In fact, the council report lists the ability to reduce costs as one criterion for judging the merit of potential CER studies.) Yet American policymakers were long reluctant to make this leap because of fears that CER-type evaluations would limit access to treatment, reduce doctors' autonomy, and even lead to rationed care.¹ In recent years, though, pressure to reduce costs has overcome this reluctance.² In an interview with ABC News, for example, President Obama signaled his support for a CER-based approach to cost control. "If there's broad agreement ... [that] the blue pill works better than the red pill," he said, "and it turns out the blue pills are half as expensive as the red pill, then we want to make sure that doctors and patients have that information available to them."

The change in attitude has expressed itself in recent legislation and regulation. For example, the 2009 stimulus bill (the American Recovery and Reinvestment Act) provided nearly \$1 billion in funding for CER. The act allocated \$400 million to the Office of the Secretary in HHS, \$400 million to the National Institutes of Health, and \$400 million to the Agency for Healthcare Research and Quality. The 2010 healthcare reform bill (the Patient Protection and Affordable Care Act) provides for the creation of a private, nonprofit organization, the Patient-Centered Outcomes Research Institute. This institute is tasked with identifying CER priorities, funding CER studies, and disseminating the results of CER to payers and physicians. Crucially, the bill also gives Medicare authority to incorporate CER into determining coverage decisions. Though the details remain to be worked out, this change in the law assures that CER will become an important factor in health-care expenditures.

Private payers have been no less interested in CER's potential to save money without hurting quality. A recent survey by the health-care consulting firm Xcenda found that 81 percent of payers believe that the importance of CER will increase in the next two years. And 85 percent foresee situations in which CER will be used to justify shifting the expense burden of cost-lier treatments onto patients.

In this broad-based acceptance of CER as a factor in payment decisions, its benefits—improved treatments for patients at less cost to payers—have been more often assumed than proved. The implications of CER and its strengths and weaknesses have only recently begun to attract scholarly attention.³ But it is already clear that CER approaches will have the effect of shift-ing demand from the therapeutic "losers"—drugs and treatments shown to be less effective, or equally effective but more costly—toward the "winners."

It seems intuitive that these "winners" will provide more health benefits to society and, when cost is taken into account, that shifting to these favored treatments will save society money on health-care costs. Unfortunately, we have found that this intuition is wrong. CER as it is usually implemented will not have this positive health impact and may even lead to greater costs to society. CER is a promising method for matching patients to the right treatments, but it will have to be applied differently by researchers and understood differently by policymakers, if it is to fulfill its promise.

HOW CER WORKS

In most CER studies, it is the drug or treatment with the larger average effect on an entire population that "wins." In the president's hypothetical, the blue pills are "just as effective" as the red ones because, on average, they do as much good for patients. But just as the average human being theoretically has one ovary and one testicle, so the average patient is not the same as any individual patient. And declaring a treatment most effective based on an average is a medical and economic error. There are two reasons that average effectiveness cannot be equated with "best."

First, individuals differ from one another and from population averages. Therefore, what may on average be a "winning" therapy may simply not work for a large number of patients. Conversely, a drug that is less effective on average may still be the best, or only, choice for a sizable proportion of patients. CER researchers have recently attempted to address this problem by breaking down populations by gender, age, ethnicity, or other relevant categories. But these divisions into subgroups do not address the fundamental difficulty: treatment is a matter of matching an individual to a therapy. And average performance is often non-informative about how to find the right therapy for an individual. A drug that is less effective on average may still be the best choice for a sizable proportion of patients (see Appendix).

The second reason that average effectiveness cannot be equated with "best" is rooted in the dependence in patient responses across therapies. Dependence,

2

for any individual patient, is the degree to which response to one treatment predicts response to another. Dependence varies from illness to illness and from drug to drug but is often an important aspect of finding treatments that work. One cannot know in advance, as a general rule, that Drug A's failure guarantees the failure of Drug B. Yet a reimbursement policy based on CER could well make this error: by refusing to reimburse Drug B on the grounds that Drug A is "more effective," such a policy assumes that failure with Drug A will predict failure with Drug B.

There are cases in which dependence is almost nonexistent-where the effectiveness of a vaccine, for example, perfectly predicts the effectiveness of another. However, dependence is usually more complex and will vary from illness to illness and from drug to drug. It cannot be ignored. On the contrary; in most cases, there is no hope of finding the optimal therapy for a given patient without knowing the differences in treatment effects across patients and the dependence in effects across treatments. Yet this is not the orientation of current CER studies, which identify and compare simple average treatment effects in a population. Thus, using average treatment effects to identify "winners" could actually worsen patient health by reducing freedom to choose the best therapies for an individual patient.

AN ILLUSTRATION: THE CATIE TRIAL

To illustrate these points, consider the real-world effect of a CER study of antipsychotic drugs. Used to alleviate symptoms of psychosis in schizophrenia, bipolar disorder, and other mental illnesses, these medications consist of a first generation of drugs discovered in the 1950s, including chlorpromazine and haloperidal (called "typical antipsychotics"); and drugs discovered in later decades (the "atypical antipsychotics"), including risperidone and olanzapine. Antipsychotics are one of the largest classes of drugs for Medicaid patients and a growing part of its expenses: Medicaid expenditures on antipsychotics increased from \$1 billion in 1995 to over \$5.5 billion in 2005.⁴

In 2005, a national CER study, the Clinical Antipsychotic Trials in Intervention Effectiveness (CATIE), compared

typical and atypical antipsychotics using the "gold standard" of medical studies, the randomized clinical trial (RCT).⁵ (In RCTs, patients are assigned at random to take a drug or placebo; bias effects are avoided because neither patients nor researchers know which drug is being taken by which patient.)

Typical as well as atypical antipsychotics are used to control symptoms of schizophrenia. While the typical antipsychotics (e.g., haloperidal and perphenazine) are cheaper than the atypical antipsychotics (e.g., olanzapine and quetiapine), many of which remain branded only, the typical antipsychotic drugs generally have more severe side effects, including diabetes, sexual dysfunction, and motion impairment. The CATIE study confirmed this side-effect difference but found that second-generation antipsychotics are no more effective at treating schizophrenia symptoms than traditional antipsychotics.⁶ Subsequent cost-effectiveness analysis using those results concluded, therefore, that first-generation antipsychotics were cost-saving: they delivered the "same" health benefit for less expense.7

Those results, unsurprisingly, led to calls for public payers to limit their coverage of second-generation antipsychotics.8 (Costs of antipsychotics were one of the fastest-growing pharmaceutical expenditures in Medicaid in the late 1990s and early 2000s.)9 This argument has been adopted by some influential media outlets and pharmacy benefit managers. In an editorial, for example, the New York Times held that "the nation is wasting billions of dollars on heavily marketed drugs that have never proved themselves in head-to-head competition against cheaper competitors."10 There has been considerable policy debate on whether the evidence generated by the CATIE should be used as a basis for limiting reimbursement or coverage for atypical antipsychotics-in other words, whether coverage and reimbursement should be responsive to the CER generated by the CATIE.¹¹

We recently set out to answer a fundamental question at the heart of this debate: Would using the CATIE to guide Medicaid reimbursement policy actually result in cost savings? Because the CATIE study permitted us to examine individual differences, we were able to assess not only average effects but also individual differences in drug response. That assessment has led us to conclude that the answer to our question is no. Using the CATIE to guide Medicaid reimbursement would not save American society money. Rather, what was gained in lower Medicaid payments would be lost in lost wages, tax payments, and other costs the consequences of a poorer level of mental health among Medicaid recipients.

A unique aspect of the CATIE was that it followed a novel approach in which patients who discontinued their first drug assignment were given an alternate drug. Therefore, unlike typical randomized trials, the CATIE provides data on how individual patients responded to alternate therapies. This gave us a way to reanalyze the individual-level CATIE data. We found that optimal therapy varies significantly across patients-for example, nearly 75 percent of patients who failed to respond to first-generation antipsychotics would respond to second-generation antipsychotics.12 Thus, while there may have been no significant difference between first- and secondgeneration antipsychotics on average, a substantial proportion of patients would benefit more from second-generation antipsychotics than from first-generation ones.

In light of these differences in patient responses, we analyzed how coverage policies would affect health and costs among Medicaid patients if they use the CATIE study to guide payment criteria. Our analysis focused on antipsychotic coverage for roughly 250,000 non-elderly adult Medicaid enrollees with schizophrenia.¹³ We considered coverage for three drugs: perphenazine, a first-generation antipsychotic; and risperidone and olanzapine, two second-generation antipsychotics. These three drugs were chosen because they account for 70 percent of antipsychotic prescriptions in the United States. If Medicaid were to provide coverage for all three drugs, we estimate annual costs to be \$4.5 billion.

We examined two potentially restrictive coverage policies that might be adopted in response to the CATIE findings. First, we considered an extreme case: denial of all coverage for second-generation antipsychotics, on the grounds that the cheaper first-genera-

tion drugs are just as effective. (Such denial is not legal under current law but, as we have noted, already has advocates; in the current climate of enthusiasm for CER as a cost-cutting measure, future changes to the law are certainly possible.) We found that this hypothetical policy would save Medicare \$1.2 billion, compared with full coverage. However, we estimate that it would reduce patient health by 13,138 quality-adjusted life years (QALYs) because of reduced health among the patients who were not responsive to first-generation antipsychotics and who, because of the restrictive policy, received no other drug therapy. Given that QALYs are typically valued at \$100,000, this suggests that the savings from denying coverage for second-generation antipsychotics (\$1.2 billion) would be outweighed by the costs of reduced health for patients (\$1.3 billion).

The second hypothetical policy we considered would cover perphenazine and risperidone (which are available in less costly generic forms) but exclude olanzapine (which is not). This policy would save Medicaid \$500 million annually but reduce health by 10,146 QALYs, mainly because of reduced health among patients who are unresponsive to either risperidone or perphenazine and who receive no therapy for six months or longer because of the restrictive policy. At a value of \$100,000 per QALY—again, the typical value assumed in the scholarly literature and by many payers—the health loss is nearly double the savings to Medicaid. Even at a value of \$50,000 per QALY, such a policy would only "break even."

These results reveal the economic consequences of the facts that we have described about individual differences in treatment response and the inability of responses to a first treatment to predict response to a second or third. They follow from the fact that treatments labeled "losers" by a CER study may nonetheless benefit significant numbers of patients who would not be cured by the "winner" of the trial. The CATIE study found no differences between first- and second-generation antipsychotics on average, but a significant number of individual patients would benefit from second-generation drugs and not from firstgeneration medications. Therefore, using the CATIE findings to support restrictive coverage policies would not be cost-effective. It would limit freedom of choice for doctors and patients and yield no compensating savings to society.

IMPROVING CER EVIDENCE METRICS AND REIMBURSEMENT STRATEGIES

We can CER methods be improved to better serve the goals of cost control and quality? As we discussed, the traditional metric from a randomized clinical trial—the average response to treatment—is limited in its ability to answer the clinically relevant question of how best to match individual patients to available treatments. To do this, studies must provide insight into, first, individual differences in response and, second, dependence (the extent to which an individual's response to one treatment predicts response to another).

Therefore, we reject the simpleminded notion that CER can find the right, cost-effective "blue pill" for every patient and every condition and eliminate the costly, less effective "red pill." If, as seems very likely, health-coverage decisions will soon be influenced by CER, then CER must be implemented differently and used more insightfully by policymakers. Our recommendations are:

1. Coverage policies should reflect information about difference and dependence effects, not CER population-wide averages. Specifically, public and private payers should never deny coverage for the so-called losers of CER studies that were based on average effects. Instead, they should use information on differences (the variation in response to a given drug from patient to patient) as well as dependence (the variation in each individual's response to different drugs). Such studies should then be used to find the most cost-effective therapies for each patient—by informing the trial-and-error sequence through which a doctor tries first one treatment and then the next.

For example, "prior authorization" insurance policies now aim to provide this kind of guidance, by requiring failure on one therapy before they will authorize reimbursement for a second (usually more expensive) treatment. With better information about

dependence effects, this type of policy could be expanded to save costs. A policy could, for instance, use data on differences and dependencies to specify, for a given condition, precisely which initial treatments should be tried, and then map subsequent steps based on nonresponse (essentially adding an economic perspective to the sequence tree in the Appendix). As we have stated, dependence is not a major issue for some diseases and therapies: in heart disease, for example, patients who fail to respond to a first drug are unlikely to do better on a second. In those instances, an informed reimbursement policy could limit payments for second and third treatments and save costs without reducing the overall health of the population. Well-designed protocols could also be built in to clinical software (for example, e-prescribing programs that write prescriptions) in order to further extend the impact of the effectiveness research on cost.

2. Going forward, CER should be used and implemented differently from the way it has been used to date. Of course, effective policies of this sort do not just depend on policymakers making the right use of CER results. They will also require changes in the conventions of CER itself, so that more studies supply the information that policymakers need. Hence, we also recommend that funders promote and support the more useful form of CER trial: not the kind that seeks only "winners" and "losers" in average effects but rather the kind that tracks individual differences and dependence in treatment effects.

Examples of CER techniques that do this include "crossover designs," such as the CATIE, in which patients are switched from one treatment arm to another.¹⁴ Another approach incorporates "adaptive assignments,"¹⁵ in which patients are switched between arms based on their treatment responses. In both designs, the switching of individual patients between trial "arms" provides information on the way a single drug produces different responses in different individuals and how, for any given individual, reaction to one drug predicts reaction to another.

Another way in which well-designed CER can provide fine-grained information about difference and dependence is by taking in the consequences of side effects, unpleasant reactions, and patient preference. It is a fact of life that a drug may appear more costeffective after a clinical trial than it is in real-world conditions. This is because many randomized clinical trials involve measures to keep patients compliant. Outside the controlled conditions of the trial, though, a drug that is unfamiliar to patients-or that causes sexual dysfunction or provokes nausea-may well be less cost-effective because patients do not take it as frequently as they would a less troublesome alternative. These effects on compliance with a drug protocol are themselves prone to exhibit differences among individuals and dependence across treatments. Therefore, in order to provide information on what therapies are likely to be effective for an individual patient, future CER studies should also measure differences and dependence in adverse effects and compliance. As the FDA plays an important role in regulating randomized clinical trials, the agency should encourage the collection of these types of data for drugs that it must approve.

3. CER approaches should vary with the characteristics of diseases, medications, and patient populations. As we have mentioned, individual differences as well as dependence will vary in their strength and importance, depending on the disease to be treated and the patient population. In the extreme case where all patients respond equally to a given treatment-such as the case of vaccines, where responses are likely to be similar across patients-there is great value in learning about treatment effects in a centralized manner by conducting studies aimed at identifying average effects in a population. However, if treatment responses vary widely across patients, there is little value in this centralized learning and in tailoring reimbursement policies to it. Accordingly, we recommend that CER approaches be tailored to circumstances. The population-average approach should be used when a more decentralized approach is expensive in time, money, and ill health. Electronic medical records could be used to determine cases where decentralized learning is particularly costly.

4. Observational studies should get more attention and support. The "gold standard" of CER remains the

randomized clinical trial. But observational studies, which use data collected from the health-care system (for example, insurance-claims records), should receive more attention. This kind of study provides a unique opportunity to inform CER efforts for two reasons. First, it can provide data on the effectiveness and utilization of treatments in real-world settings. Second, claims-based data allow researchers to identify and compare outcomes among patients who switched therapies. That permits them to estimate the proportion of patients who would benefit from specific therapies after failure on others.

Observational studies can have problems with selection bias, but these can be mitigated with properly designed studies and good practices—for instance, by utilizing quasi-randomization methods, such as valid instrumental variables, propensity score methods, or the GRACE (Good Research for Comparative Effectiveness) principles for observational studies.¹⁶ The growing use of electronic medical records provides a potential wealth of data for observational studies.

CONCLUSION

s currently implemented, CER uses population-based measures of response to identify **L**"winning" treatments and shifts demand toward these treatments by affecting clinical decision making and reimbursement policies. That approach falls short of the goal of finding the best treatment for an individual patient and could actually reduce patient health if the "winning" or optimal therapy varies significantly for each patient. Indeed, in the case of schizophrenia, we showed that while there may be no difference in average treatment effects across therapies, significant numbers of patients will benefit from one treatment over another. As a result, efforts to exclude therapies based on a lack of average differences could actually reduce patient health by denying access to patients who would specifically benefit from excluded therapies. CER has great potential as a means of cutting costs and improving overall health, but to realize its promise, it needs to change. So do the notions that policymakers have about it.

6

Figure 1. Example of Sequencing of Treatments

The two main difficulties with the standard CER paradigm are illustrated below by Figure 1, which depicts the typical sequencing of medical treatments.

Figure 1 outlines the questions facing a physician who wishes to treat a given patient. First, should the patient be started on Treatment A or Treatment B? Moreover, how is failure observed? And if it occurs, what should the next treatment be? The answers to these questions depend on more than simply the average effect in a population. Rather, they depend on the nature of differences in responses across patients—the physician wants to know the range of possible responses found across many different individuals. In addition, the answers depend on the dependence in responses across treatments. If an individual patient fails on Drug A, what does that information predict about what will work next—Drug B or Drug C? Knowing the differences in response across patients as well as the dependence across treatments is crucial in finding the optimal therapy for a given patient.

However, CER studies as currently performed, even randomized clinical trials, do not provide the kind of data necessary to make these judgments. Rather, as typically implemented, CER studies typically compare the clinical effectiveness of a set of treatments by comparing the average treatment effects among groups of patients receiving each treatment. CER is then justified as an improvement in health because it guides doctors and patients toward the "winning" treatments, which are those with higher average treatment effects. This centralized approach toward acquiring and implementing information about a treatment's effectiveness stands in contrast to current practice: a rather decentralized, trial-and-error approach that occurs between physicians and patients. While a centralized approach may seem to be a more efficient way to learn about a treatment's effectiveness in the population, its applicability for an individual patient has the following limitations:

• The use of centralized population-based averages overlooks differences in treatment response across patients. Fundamental to this problem is the use of population-based measures to infer treatment responses at the patient level. This approach is problematic when treatment responses vary significantly at the patient level, so that even if one treatment is better than the other on average, a significant portion of patients may still benefit from the "losing" treatment. A good example is the case of antipsychotics, where data from the CATIE suggest that even though there is little difference in average effects between the antipsychotics risperidone and olanzapine, a sizable fraction (nearly half) of patients experienced increased benefit with olanzapine compared with risperidone. • The use of centralized population-based averages overlooks dependence in treatment effects between therapies. When there is such dependence in treatment effects, response to one treatment may provide information about the likely response to other treatments. If patients who fail one treatment are likely to succeed on a second one, the second treatment has an important therapeutic role, even if it is less effective on average compared with the first. Antibiotics provide a good example of this dependence—if a patient fails a particular regimen, he is more likely to respond to subsequent regimens because failure on the first regimen provides clues to the susceptibility of the organism causing the patient's disease.

Figure 2, which plots the hypothetical responses to two treatments for a population of patients, illustrates these two points. Patients lying below the 45-degree line respond better to Treatment A, while patients above the 45-degree line respond better to Treatment B. The "+" sign in the figure indicates the average responses for the two treatments and lies below the 45-degree line, showing that, on average, patients respond better to Treatment A. However, there are clear differences in treatment effects across patients, since a large proportion of patients lie above the 45-degree line. Thus, despite having a lower average effect, Treatment B still has an important therapeutic role.

Moreover, Treatment B has an important therapeutic role because of substantial dependence between treatment effects. A large proportion of patients lie in the circled areas in Figure 2, indicating that poor response to Treatment A predicts a large response to Treatment B and vice versa. As stated above, in the case of antibiotics, this dependence can arise because failure implicitly selects for patients who are more likely to succeed on alternative therapies. Thus, patients who fail one therapy are likely to succeed on the other.



Our analysis suggests that finding the optimal therapy for a given patient depends on knowing the differences in treatment effects across patients and the dependence in effects across treatments. As we have stated, this is not the orientation of most current CER studies.

Some CER studies attempt to address these limitations with analyses focused on subpopulations based on observable patient characteristics such as race, gender, and medical history. While this approach helps address differences in responses across patients, it is limited in value for two reasons. First, it fails to completely address the issue of differences across patients, since even within a subpopulation there are often large differences in treatment response. Second, and more important, estimating average treatment effects in a population or subpopulation still does not address the issue of dependence between treatment responses that is informative to optimal matching through sequencing treatments. Thus, using average treatment effects to identify "winners" could actually worsen patient health by reducing access to therapies with potentially important roles.

ENDNOTES

- 1. By contrast, CER plays an important role in clinical decision making in many other countries. For example, in the U.K., the National Institute for Health and Clinical Excellence (NICE) is a government agency tasked with integrating CER into technology appraisals and clinical guidelines that heavily influence clinical practice. For a review of the use of CER outside of the U.S., see K. Chalkidou and G. Anderson, *Comparative Effectiveness Research: International Experiences and Implications for the United States* (Washington, D.C.: Academy Health, 2009).
- 2. G. C. Alexander and R. S. Stafford, "Does Comparative Effectiveness Have a Comparative Edge?," *Journal of the American Medical Association* 302, no. 23 (2009).
- 3. For a more detailed discussion, see A. Basu and T. J. Philipson, "The Impact of Comparative Effectiveness Research on Health and Health Care Spending," NBER Working Paper no. 15633 (2010); A. Basu, "Individualization at the Heart of Comparative Effectiveness Research: The Time for I-Cer Has Come," *Medical Decision Making* 29, no. 6 (2009); and D. Meltzer, A. Basu, and R. Conti, "The Economics of Comparative Effectiveness Studies: Societal and Private Perspectives and Their Implications for Prioritizing Public Investments in Comparative Effectiveness Research," *PharmacoEconomics* 28, no. 10 (2010).
- 4. M. R. Law, D. Ross-Degnan, and S. B. Soumerai, "Effect of Prior Authorization of Second-Generation Antipsychotic Agents on Pharmacy Utilization and Reimbursements," *Psychiatric Services* 59, no. 5 (2008): 540.
- 5. See J. A. Lieberman et al., "Effectiveness of Antipsychotic Drugs in Patients with Chronic Schizophrenia," *New England Journal of Medicine* 353, no. 12 (2005).
- 6. See M. M. Desai et al., "Mental Disorders and Quality of Diabetes Care in the Veterans Health Administration," *American Journal of Psychiatry* 159, no. 9 (2002); and R. A. Rosenheck et al., "Cost-Effectiveness of Second-Generation Antipsychotics and Perphenazine in a Randomized Trial of Treatment for Chronic Schizophrenia," *American Journal of Psychiatry* 163, no. 12 (2006).
- 7. See Rosenheck et al., "Cost-Effectiveness of Second-Generation Antipsychotics."
- See W. T. Carpenter and R. W. Buchanan, "Lessons to Take Home from CATIE," Psychiatric Services 59, no. 5 (2008); H. A. Huskamp et al., "Coverage and Prior Authorization of Psychotropic Drugs under Medicare Part D," *Psychiatric Services* 58, no. 3 (2007); J. J. Parks, A. Q. Radke, and R. Tandon, "Impact of the CATIE Findings on State Mental Health Policy," *Psychiatric Services* 59, no. 5 (2008); and Centers for Medicare & Medicaid Services, *Psychotropic Medications: Addressing Cost without Restricting Access* (Baltimore, 2004).
- 9. See J. S. Banthin and G. E. Miller, "Trends in Prescription Drug Expenditures by Medicaid Enrollees," *Medical Care* 44, no. 5 supp. I (2006).
- See "Comparing Schizophrenia Drugs," New York Times editorial, September 21, 2005; G. Harris, "States Try to Limit Drugs in Medicaid, but Makers Resist," New York Times, December 18, 2003; and S. B. Soumerai and M. R. Law, "Cost-Effectiveness of Schizophrenia Pharmacotherapy," American Journal of Psychiatry 164, no. 4 (2007).

- 11. See Carpenter and Buchanan, "Lessons to Take Home from CATIE"; Huskamp et al., "Coverage and Prior Authorization of Psychotropic Drugs under Medicare Part D"; Parks, Radke, and Tandon, "Impact of the CATIE Findings on State Mental Health Policy"; and Centers for Medicare & Medicaid Services, *Psychotropic Medications*.
- 12. Details of our methods can be found in Basu and Philipson, "The Impact of Comparative Effectiveness Research on Health and Health Care Spending." Other studies have also used the CATIE data to examine how therapy responses varied for an individual patient—for an example, see T. S. Stroup, "Heterogeneity of Treatment Effects in Schizophrenia," *American Journal of Medicine* 120, no. 4 supp. 1 (2007). However, our analysis is the first to examine how this heterogeneity affects health and spending based on potential coverage policies.
- 13. See E. Q. Wu et al., "Annual Prevalence of Diagnosed Schizophrenia in the USA: A Claims Data Analysis Approach," *Psychological Medicine* 36, no. 11 (2006), who estimated the prevalence of diagnosed schizophrenia among non-elderly Medicaid patients to be 1.66 percent. We apply this estimate to about 15 million non-elderly adult Medicaid enrollees (www.statehealthfacts.org) to obtain a total number of 250,000 non-elderly adult enrollees with schizophrenia.
- 14. For a discussion of crossover trials, see S. Senn, *Cross-over Trials in Clinical Research*, 2nd ed. (Hoboken, N.J.: Wiley, 2002).
- 15. See S. A. Murphy, "An Experimental Design for the Development of Adaptive Treatment Strategies," *Statistics in Medicine* 24, no. 10 (2005).
- 16. For an example of these methods, see C. C. Earle et al., "Effectiveness of Chemotherapy for Advanced Lung Cancer in the Elderly: Instrumental Variable and Propensity Analysis," *Journal of Clinical Oncology* 19, no. 4 (2001), who use these methods to examine the effectiveness of chemotherapy for elderly adults with stage IV non–small cell lung cancer, as well as www.graceprinciples.org.

Committee Members

Tomas J. Philipson, *Chairman* Dennis A. Ausiello Arthur Daemmrich Joseph DiMasi Henry G. Grabowski Paul Howard Daniel P. Petrylak Lance K. Stell Thomas P. Stossel

PROJECT FDA Sponsored by the Manhattan Institute

Utilizing 21st Century technologies to help develop better FDA regulations and a faster and safer drug and medical-device pipeline

The time and costs required to bring new medical products to market is growing ever larger. Today, it may take more than a decade, and hundreds of millions of dollars, to bring a single medical innovation to the public, from initial conception to Food and Drug Administration (FDA) approval. The slow pace and high cost of development contributes to the cost of health care and delays patient access to potentially lifesaving products.

At the same time, the FDA is facing a crisis in confidence among consumers, policymakers, and the media, with some critics declaring the agency "broken"—unable to ensure that medical products offered for sale in the United States are reasonably safe and effective. Doctors and academic medical centers, too, face growing concerns about allegedly harmful interactions with industry during the development and marketing of medical products. The result is a growing call for sweeping new regulation of the industry at both the state and federal levels.

Advances in the molecular and genetic understanding of disease have the potential to make health care more predictive and preventive than empirical and reactive—thus improving patient outcomes and reducing health-care costs. Unfortunately, in our zeal to reduce risks, regulate potential conflicts, and mandate transparency, we may be reducing incentives for companies to develop and market improved products; discouraging doctors from collaborating with companies in designing safer and more effective products; and slowing the FDA's efforts to incorporate into its oversight activities the latest scientific and technical advances.

The membership of Project FDA are scientists, economists, medical ethicists, policy experts, and practicing physicians. The committee's mission is to examine the current framework and direction of federal and state regulation and to promote the development of a robust medical innovation pipeline.

The Manhattan Institute is a 501(C)(3) nonprofit organization. Contributions are tax-deductible to the fullest extent of the law. EIN #13-2912529

http://www.manhattan-institute.org/projectfda