THE ECONOMICS OF HEALTHCARE RATIONING

[in the *Oxford Handbook of U.S. Healthcare Law*]

Michael Frakes, Matthew B. Frank, and Kyle Rozema

Abstract: This article examines the economics of healthcare rationing. We begin with an overview of the various dimensions across which healthcare rationing operates, or at least has the potential to operate, in the first place. We then describe the types of economic analyses used in healthcare rationing decision-making, with particular reference to cost-benefit analysis and cost-effectiveness analysis. We also discuss healthcare rationing in practice, such as how economic analyses inform decisions regarding which services to cover, and conclude by discussing various practical and conceptual challenges that may arise with economic analyses and that span both economics and ethics.

Keywords: economics, healthcare rationing, healthcare, economic analyses, decision-making, cost-benefit analysis, cost-effectiveness analysis, ethics

Another article in this volume discussed various ethical considerations surrounding the rationing of healthcare. Any study of the rationing of goods and services, however, would be incomplete without the consideration of economic analysis. The need for rationing in the first instance arguably encapsulates the essence of economics itself. Though many definitions abound, economics is often described as the study of how individuals and agents of society allocate scarce resources among competing desires. Clearly, individuals

cannot each acquire and consume all of the products and services that they want without limitation. Many, if not all, goods and services must be "rationed" in some manner. In this article, we provide a brief overview of healthcare rationing through the lens of economic analysis.

At the outset, it is worth saying a bit about the term "rationing," which is often thrown around rather loosely by the media. In the broadest sense, rationing can be interpreted as *any* means of limiting individuals' consumption of products and services. This perspective implies that prices can ration. After all, those unwilling to pay the price for a particular good or service may be excluded from consuming or receiving it. Moreover, the first fundamental welfare theorem of economics holds that price mechanisms in competitive markets will, under a restrictive set of assumptions, perform this necessary allocation of resources in a Pareto-efficient manner—that is, it would be impossible to reallocate resources so as to make someone better off without making another worse off. Some economists, however, have uneasy feelings about referring to the free market's price mechanism as a means of rationing. Rather, "rationing" is often used as a term of art by many economists to explain any number of ways that policy-makers and industry participants influence who consumes what healthcare services, excluding from this definition the distribution of goods and services via market-clearing forces. For the purposes of this article, we largely refer to rationing in this more limited context.

If prices theoretically generate Pareto-efficient outcomes, one may wonder why nonprice means of rationing are even warranted in the first place. One reason lies in the failures of the assumptions of perfect completion inherent in the first fundamental welfare

theorem of economics. While it is beyond the scope of this article to discuss in depth the various market failures in the healthcare system, it is well known that certain features of this system cause it to deviate from the assumptions necessary to achieve Pareto efficiency. Such sources of failure include, among others, agency costs stemming from informational asymmetries between physicians and patients (sometimes manifesting in "physician induced demand" of excessive services) and moral hazard in healthcare utilization caused by third-party payment relationships. These features of healthcare markets may contribute to supra-optimal levels of healthcare utilization that may be crowding out spending on other nonhealth goods that likewise contribute to social welfare. Total spending aside, such market failures may also lead to inefficient utilization of some types of health services over others. These arguably undesirable outcomes have motivated policy-makers and industry participants to design nonprice means of rationing services, some of which are briefly summarized in this article.

A second reason for society's unwillingness to rely solely on price mechanisms to ration healthcare may arise from noneconomic considerations altogether. Even if prices were to generate Pareto-efficient outcomes, it may not necessarily be the case that such outcomes are ultimately desired. Economists evaluate policy issues by attempting to maximize specified social welfare functions. The maximization calculus inherent in this exercise need not collapse to one focused exclusively on economic efficiency, the kind of which price competition is meant to produce. Social welfare optimization criteria can, and often do, incorporate various distributional considerations. Metaphorically speaking, when structuring healthcare rationing programs, we do not just care about the size of the pie (efficiency considerations, generally); we also care about how it is sliced. This

analysis necessarily transcends economics. Key inputs into this optimization problem—that is, how exactly do we value different distributions of the pie—derive from ethical and philosophical considerations. Because another article in this volume deals exclusively with the ethical side of this coin, we largely abstract away from such matters in the discussion to follow.[1]

To set the stage for our review of the economics of healthcare rationing, Part I of this article summarizes the various dimensions across which healthcare rationing operates, or at least has the potential to operate, in the first place. We turn in Part II to a brief review the types of economic analyses used in healthcare rationing decision-making, with a focus on cost-effectiveness analysis. In Part III, we discuss healthcare rationing in practice. For instance, among other things, we examine how economic analyses inform decisions regarding which services to cover. In Part IV, we describe various practical and conceptual challenges that may arise with economic analyses, challenges that span both economics and ethics. Finally, in Part V we conclude.

I. OVERVIEW OF DIMENSIONS TO HEALTHCARE RATIONING

In this section, we attempt to lay the foundation for a basic understanding of healthcare rationing by delineating, in rather broad strokes, the types of healthcare resources that require rationing in the first place and the different levels at which rationing operates (e.g., macro- versus micro-level allocation decisions).

a. Types of Scarcity That May Require Rationing

Healthcare goods and services may be scarce due to financial or physical limitations. While we have separated these two concepts, they are naturally intertwined. Financial scarcity is perhaps self-evident. Individuals, entities, or states only have so much money

to devote to the purchase of goods and services, leading to inherent limitations on consumption. To provide an extreme example, robotic surgical systems currently cost upward of $2 million, rendering such devices prohibitively expensive for many healthcare institutions.[2] Physical scarcity arises in various contexts. As a paradigmatic example, organs used in transplantation are often scarce—the need for organs far outweighs the supply.[3] Physician services are also sometimes scarce. With only so many willing and able providers in existence (and only so many hours in the day), demand for labor-intensive procedures is not always fully satisfied. There are concerns, for instance, that the demand for primary care physicians cannot be met given the increasing insurance coverage under the Affordable Care Act.[4]

These sources of scarcity may be absolute or relative in nature. Absolute scarcity refers to shortages that cannot be addressed in the short term by reallocating resources, whereas relative scarcity refers to shortages that may be addressed by shifting resources from one use to another. For example, the shortage of organs for transplantation is a form of absolute scarcity in that the limited supply is presently fixed, although policies could be pursued to increase supply over time.[5] In contrast, a shortage of robotic surgical systems would be a form of relative scarcity to the extent that healthcare institutions may shift funding priorities to purchase the equipment within the short term.[6]

In the balance of this article, we are primarily concerned with relative financial scarcity: how should scarce financial resources be allocated in the short term between competing healthcare goods and services?

b. Levels of Rationing

At the outset we would like to distinguish between efficiency considerations between healthcare and other non-healthcare services and efficiency considerations within healthcare.[7] Indeed, the necessity of healthcare rationing manifests itself on both macro and micro levels. With respect to the former, one must determine how many resources should be allocated to healthcare delivery in the first instance as opposed to other services that may likewise contribute to social welfare—for example, education. Analysts often label this question one of allocative efficiency. At the national level, the United States currently spends approximately 18% of its total income on healthcare, up from 7% in 1970.[8] Some believe that social welfare could indeed be better served through a reallocation of these resources to other means. Others believe that much of the healthcare spending growth is justified by associated gains in health. The answer to this macro-level question is undeniably complicated, and an array of policy initiatives (both implemented and theorized) are aimed at curbing overspending in general.[9]

For the sake of tractability, we largely focus in this article on the more internal, micro-level counterpart: taking the national amount of healthcare spending as given, how do we allocate scarce healthcare resources among alternative uses—for example, deciding which clinical treatments for which diseases should or should not be covered. For instance, in treating patients suffering from prostate cancer, how many resources should be dedicated to the use of surgery (prostatectomy), proton-beam therapy, or watchful waiting by physicians (expectant management)? Analysts label this inquiry as one bearing on technical efficiency. We now turn in Part II to an overview of the various tools employed to guide this technical efficiency analysis.

II. ECONOMIC EVALUATIONS USED IN RATIONING HEALTHCARE

Analysts have derived several different approaches to assessing technical efficiency—that is, to determining how competing uses of healthcare resources should be compared with one another—the most well-known of which are called cost-benefit analysis (CBA) and cost-effectiveness analysis (CEA). These two approaches are distinct, though the lines between them are at times blurred, as we suggest in a stylized example below.[10] In this section, we provide an overview of these tools, focusing on the latter.

a. Overview of CBA and CEA

Cost-benefit analysis (CBA) is a rather intuitive concept—analysts quantify in monetary terms the costs of a particular healthcare service and compare these costs with a monetary calculation of the benefits. The healthcare service may be deemed appropriate if the calculated benefits exceed the costs. Generally, CBA utilizes a simple alternative decision rule known as the Kaldor-Hicks criterion: "a policy should be adopted if and only if those who will gain could fully compensate those who will lose and still be better off."[11] To be sure, the benefits of CBA actually go beyond a technical efficiency context—that is, beyond the question of how to allocate resources among competing healthcare options. To the extent that it places all benefits—health and nonhealth—on a common monetary framework (assuming that can be successfully done),[12] it can also be used to assess allocative efficiency—that is, to evaluate whether resources (e.g., land, labor, capital) should be directed toward healthcare or non-healthcare goods and services. A major problem with CBA relates to the difficulty associated with achieving this monetization of benefits.[13] Largely for this reason, CBAs are not commonly utilized in healthcare settings. As such, we will place less emphasis on CBA in this article.

Cost-effectiveness analysis (CEA) represents the primary approach to economic analysis of healthcare rationing. One can view CEA as effectively asking: taking as given a particular level of healthcare spending, how do we select among two or more alternative healthcare options to ensure that we receive the most "bang for the buck" from the selected approach? Not immune to the difficulties associated with valuing health benefits, CEA requires the construction of a common set of health outcomes in order to facilitate a comparison across different types of health interventions—for example, the number of life years gained as a result of given treatments. Comparing the ratio of the costs associated with a particular treatment to the value of the outcome metric provides a relative measure of the cost-effectiveness of each treatment. Lower ratios naturally entail greater efficiency. Depending on certain choice parameters (e.g., must only one treatment be selected or may some combination of treatments be selected) and using cost-effectiveness ratios as inputs, CEA then applies certain selection criteria to identify those health interventions that merit priority.

In perhaps the most general terms, CEA involves maximizing population health given a fixed healthcare budget. With perfect information and foresight, the economic framework of CEA collapses to a straightforward constrained optimization problem where policy analysts take as given various treatments i with exogenous health production functions $h_i()$, prices of those treatments p_i, and a healthcare budget W, and maximize aggregate health production by choosing to employ a set j of the various health technologies subject to the production and budgetary constraints:

$$\max_{j \in i} \sum_j h_j$$

$$\text{subject to} \begin{cases} W \ge \sum_j h_j p_j \\ h_i \le h_i(p,W) \, \forall i \end{cases}$$

CEAs are typically applied in a more simplistic fashion than this general optimization problem entails. For instance, when multiple treatments may be selected among a set of noncompeting alternatives, CEA often entails calculating the cost-effectiveness ratio for each treatment relative to the no-treatment option (often known as the average cost-effectiveness ratio for that treatment) and then ranking treatments from the lowest to higher ratios (from most efficient on average to least efficient on average). Society may then select treatments from among this rank ordering, moving down the list until the budget is exhausted. It is important to note that the analysis may take on a form different from this generalized optimization problem and from this stated simplification of that problem when analysts are required to select a *single* treatment among a set of mutually exclusive alternatives, or when analysts are determining whether to adopt a new treatment on the margin to the exclusion of the status quo treatment previously adopted. A "competing" CEA of this latter nature typically follows a particular selection process that we demonstrate in a highly stylized example below.

Before turning to that example, however, we note that much discussion in CEA settings bears on the selection of the necessary common measure of health benefits. Measures employed include the number of illnesses cured, the number of lives saved, and the number of life years gained.[14] In general, this exercise faces real difficulties when comparing the benefits of different treatments. For instance, how should one assess patient A's cataract surgery (to restore sight) against patient B's anterior cruciate ligament surgery (to restore mobility)?[15] Moreover, an obvious limitation with CEA is

that the analyst must decide how to define, measure, and combine the quality dimension and the quantity dimension.[16] Economists and other health researchers have attempted to create methods to standardize benefits—that is, methods that seek to make commensurate quantity and quality. CEAs often employ health benefit measures that account for a treatment's impact on length of life (mortality) adjusted for any decrements in quality of life due to functional limitations resulting from the treatment (morbidity), such as the number of quality-adjusted life years (QALYs) or disability-adjusted life years (DALYs).[17] Even the use of measures such as QALYs, however, comes with substantial challenges (discussed further in Part IV), not the least of which is the lack of data required to construct QALYs for a particular treatment.

In the stylized example below (and in the Appendix), we demonstrate how CEA and CBA might be applied in practice.

b. Stylized Example

Consider a sixty-year-old patient who suffered from a heart attack and now faces a hospital mortality rate of 10%.[18] Suppose there are three treatment options that differ in effectiveness and cost: (1) streptokinase, which reduces the mortality rate by 5 percentage points to 5% and costs $200 per patient; (2) tissue plasminogen activator (t-PA), which reduces the mortality rate by 7.5 percentage points to 2.5% and costs $2,000 per patient; or (3) do nothing, which does not change the mortality rate and has no cost. For the purposes of the calculations to follow, assume a relevant heart attack patient population of 1,000.

To determine what treatment option should be adopted—*assuming in this case that only one can be adopted*—a policy analyst seeking to ration costs might conduct a

simple CEA. CEAs begin by ruling out from consideration strictly dominated treatments, that is, treatments that cost the same or more as another treatment but yielded lower benefits. None of the three treatments in our example are strictly dominated by another. Thereafter, the analysis orders the remaining alternatives starting from the one that generates the lowest benefit but also costs the least to the one that costs the next highest amount and generates the next highest benefit level, and so on and so forth. A CEA in this "competing choice" setting essentially involves incrementally moving up this chain and determining whether the next alternative should be chosen over the preceding alternative in that chain. While the selection framework may be tailored to achieve different sets of objectives, each such determination often rests on whether the incremental cost-effectiveness ratio (ICER) associated with the relevant treatment in comparison with the status quo treatment falls below some specified threshold ratio. To be more precise, the ICER here equals the incremental cost of the treatment in question relative to the next-least effective treatment divided by the incremental benefit of the treatment at question relative to the next-least effective treatment. If the ICER indeed falls below the stated threshold—which often captures our priors regarding the value of a unit of the indicated health outcome—that marginal treatment should be preferred over the prior treatment and we should proceed to evaluating whether the next most effective treatment should be adopted (using the same criteria).

Let us return to our example to demonstrate. Start from the baseline in which we do no intervention. This costs nothing and generates no benefits. Let us then evaluate whether we should adopt the next intervention in this chain—in this case, streptokinase. Is it worth it to adopt this marginal intervention? To assess this, the analyst will calculate

the incremental costs and effectiveness of streptokinase versus no treatment. Using the assumed numbers, the incremental cost for the population is $200,000, and the incremental effectiveness is 50 individuals saved.[19] Thus, the so-called ICER becomes $4,000 per life saved ($200,000/50). If $4,000 falls below the specified threshold ICER level—stated differently, if society is willing to accept $4,000 per avoided fatality—then we would determine that providing streptokinase is preferred to no treatment. Assume that we do. Next, we move on to evaluating the marginal adoption of t-PA in place of streptokinase, which we have just established as our new status quo. Using the same analysis we can obtain the ICER of $72,000 per life saved.[20] Whether we adopt t-PA as the sole treatment over streptokinase depends on whether $72,000 falls below the threshold ICER level specified by the analyst.

[Insert Table 1]

To the extent that this competing choice model requires the selection of a threshold ICER level—often meant to capture the value that society places on a unit of the health outcome—the analysis will run into the same problems identified above for CBA—that is, placing a monetary value on health. For this reason, we note that the distinction between CBA and CEA may, in part, be blurred. We note, of course, that the simple, noncompeting model described briefly in Part II(a) above—in which we exhaust those treatments with the lowest *average* cost-effectiveness ratio before moving on to the next lowest average cost-effectiveness treatment—does not always require a comparison with a value threshold. Thus, as a more general matter, CEA faces at least fewer concerns regarding the need to place a monetary value on healthcare outcomes.

Though our focus in this article is largely on CEA, we extend this stylized example in the Appendix in order to demonstrate how a common formulation of CBA would approach the choice as between streptokinase and t-PA.

III. HEALTHCARE RATIONING IN PRACTICE

As suggested in the introduction above, we define healthcare rationing in rather broad terms to capture initiatives aimed at limiting utilization of healthcare services, other than via price mechanisms in competitive marketplaces. To discuss the full array of ways in which rationing takes place is beyond the scope of this modest article. Given our decision already to focus on technical (within health) rather than allocative (between health and nonhealth) efficiency, we focus our discussion on forms of rationing that more explicitly incorporate CEA into decisions regarding what healthcare resources society should prioritize. Arguably, it is precisely such types of initiatives that first come to mind when hearing the word "rationing" anyway. We begin this discussion by describing programs in which payers (private or public) use CEA to guide coverage decisions.

a. Covered Services

Excluding particular services from insurance coverage is a common supply-side rationing mechanism. While denying insurance coverage does not necessarily prevent patients from receiving desired services, treatments that are not covered—that is, reimbursed—are perhaps substantially less likely to be utilized because patients financially internalize the full cost of care. In some countries, determinations of services to cover are explicitly guided by CEA. In the United Kingdom, for example, the National Institute for Health and Care Excellence (NICE) is responsible for making recommendations to the British National Health Service (NHS) on the coverage of selected medical treatments. These

recommendations are based on the cost per quality-adjusted life year (QALY) gained for patients and on the overall impact of the proposed covered treatment on NHS costs. It is believed that NICE employs an implicit cost-effectiveness threshold and recommends coverage of treatments with costs per QALY gained within or below the $30,000 to $50,000 range.[21] For example, if a treatment extends life by 0.5 QALYs, the NHS might pay as much as $25,000 for the treatment, implying a cost-effectiveness of $50,000 per QALY gained.

We briefly digress to note the findings of an empirical literature that estimates the value people place on QALYs. Such valuations can provide a useful benchmark for payers to use in guiding coverage decisions; if the cost per QALY for a particular treatment is lower than the amount by which individuals value a QALY, such services may be deemed worthwhile. A comparison of this nature is evidently helpful for allocative as well as technical efficiency analyses. Depending on the method used,[22] estimated dollar values of a QALY range from $74,000 to $450,000, with practitioners of CEA tending toward the low end of the range.[23]

In the United States, private payers often use CEA to justify coverage decisions. However, CEA is less well incorporated in the publicly financed portion of the U.S. market. While some public insurers use CEA to guide coverage decisions (e.g., the Department of Defense and the Department of Veterans Affairs), the Medicaid and Medicare programs, two of the largest health insurers in the United States, may not, by law, consider the cost of a treatment in determining coverage.[24] Yet some evidence suggests that cost-effectiveness has played an informal role in some Medicare coverage decisions by guiding the level of scrutiny Medicare uses when assessing new

treatments.[25] With the rising burden of healthcare expenses as a percent of national income, there is a growing movement toward formally incorporating CEA into more public insurers coverage decisions; the outcome of this debate, however, remains uncertain.

b. Demand-Side Initiatives

Health insurers employ an array of cost-sharing requirements—copays, coinsurance, deductibles, and so forth—to discourage utilization of inappropriate care by patients who do not pay the full cost of the care they receive (i.e., to alleviate moral hazard and encourage patients to pursue cost-justified services). Such arrangements have undergone a dramatic growth in recent years with the proliferation of high-deductible health plans (HDHPs) that impose large deductibles on most medical services (averaging over $2,000 for individual plans and $4,000 for family plans).[26] The vast literature on this topic reveals that increases in cost-sharing reduce healthcare utilization.[27] While cost-sharing does not deny coverage of medical services (see Part III.A), its intent to discourage utilization of healthcare in some circumstances is arguably in the same spirit. Rather than give a full accounting of such mechanisms, we offer a brief discussion of those cost-sharing arrangements that have incorporated CEA-style components similar to that summarized in Part II above, thereby falling within the narrower theme of this article.

Cost-sharing mechanisms have historically been somewhat blunt in nature—for example, particular copayment amounts or percentages were—and still often are— applied indiscriminately to all covered services. Over time, insurers developed more tailored structures that attempt to place more cost-sharing pressure on some procedures over others. A particular concern with blunt cost-sharing approaches was that they would

discourage utilization of both their intended target—clinically inappropriate care (i.e., where the costs of care surpass its benefits)—and an unintended target—clinically appropriate care (i.e., where the benefits of care surpass its costs). Informational limitations and behavioral irrationalities (e.g., hyperbolic discounting) may cause patients to nonetheless forego valuable and cost-justified care in the face of cost-sharing obligations.[28] Common targets of concern here include chronic-care management (e.g., routine diabetes care), particularly within vulnerable populations.[29]

To address the concern that cost-sharing results in individuals foregoing clinically beneficial care, insurers have begun to modify this blunt, traditional approach to cost-sharing. One increasingly popular idea is called value-based insurance design (VBID). VBID programs promote utilization of clinically cost-effective care by lowering or even eliminating cost-sharing for "high value" medical treatments (such as metformin for diabetes management) and simultaneously discouraging utilization of unnecessary care by raising cost-sharing for "low-value" treatments (e.g., lung volume reduction surgery).[30] VBID programs are more common for prescription drugs where, for instance, generic drugs (e.g., metformin) are covered at higher rates than their more expensive branded equivalent (Glucophage) and more effective drugs (e.g., Flovent for asthma) are covered more generously than their less-effective counterparts (Singulair) within the relevant drug class.[31] Some VBID programs even target patients based on clinical criteria, as the value derived from certain drugs may differ based on patient condition and severity.[32] In addition, VBID has been incorporated into HDHPs, many of which now waive the deductible for selected types of preventive care (e.g., breast and cervical cancer

screenings, and well-child visits)[33] and other treatments perceived to be of high value (e.g., medications for chronic conditions).[34]

While VBID programs may consider CEAs only informally when identifying high-value care, they focus on clinical areas with the greatest potential to improve health outcomes and/or constrain costs, including the treatment of chronic illnesses (e.g., diabetes and antihypertensive medications) and secondary disease prevention (e.g., statins that lower cholesterol).[35] Early evidence suggests that such programs have been successful at meeting this potential. For example, Chernew and colleagues evaluated the effect of a VBID program that eliminated copays for generic drugs and reduced copays by 50% for branded drugs in five chronic disease classes.[36] They found that adherence improved in four of the five classes by 2.6 to 4.0 percentage points. Subsequent VBID studies have confirmed this result, reporting statistically significant though quantitatively modest impacts that vary in magnitude across drug classes.[37] The current evidence base regarding the extent to which these improvements will translate into lower costs, however, remains inconclusive.[38]

c. Provider Networks

In addition to denying full coverage or imposing differential cost-sharing amounts for certain services, health insurers have also sought to regulate patient behavior by directing patients to certain providers over others. Insurers often construct provider networks associated with their plans, extending more favorable coverage to insureds who visit providers within the network (the most common model of this approach is known as a preferred provider organization, or PPO). To see out-of-network providers, patients

typically face higher cost-sharing. As a general matter, these plans may achieve cost savings by negotiating favorable prices with participating in-network providers.

Falling more within the spirit of the rationing approaches emphasized in this article (aimed at attaining technical efficiency), some provider network programs take matters one step further (i.e., beyond mere price concessions through bargaining power) and try to encourage patients to seek care from providers who offer greater productive efficiency—that is, those who have proven to deliver quality care at low costs. Such alternative approaches are known as "limited network" and "tiered network" plans.

Limited networks typically rank physicians and hospitals based on some cost and quality measures and exclude nonpreferred providers from the network altogether.[39] Tiered networks, in turn, are the network analog to VBID plans (discussed in Part III(b) above) in that they encourage patients to seek care from preferred providers by structuring cost-sharing differences among tiers.[40] Over the past five years, limited and tiered provider networks have become increasingly popular as the percentage of employers whose largest plan included a limited or tiered network increased from 16% in 2010 to 23% in 2013.[41]

Even if provider networks can effectively steer patients toward preferred providers, provider networks face a number of potential drawbacks. While one of the objectives of networks is to guide patient decision-making with respect to provider selection, such networks may disrupt long-standing existing relationships (and thus the benefits of continuity-of-care associated with such relationships). Another issue is that patients seeking specialty care from in-network providers with long waitlists may have few low-cost immediate alternatives, an issue that is largely mitigated in tiered networks.

Finally, questions abound regarding the availability of sound data on provider quality. On this point, physician groups in several states have sued health plans with provider networks on various grounds, including defamation, interference with advantageous relations, and violation of procedural due process, alleging that health plans are misclassifying providers through opaque processes that employ fallible ranking methods and rely on limited (and sometimes inaccurate) data on provider quality.[42] While these lawsuits have failed in the courts, the concerns expressed reveal some key challenges with provider tiering. Ultimately, the evidence of the impact of limited and tiered provider networks on utilization and cost containment is mixed though admittedly sparse.[43]

IV. CHALLENGES WITH ECONOMIC EVALUATIONS

In this section, we expand our discussion of CEA to highlight several measurement, analytic, and application challenges confronting analyses of this nature. While we discuss these challenges in the context of CEAs that use outcome metrics that incorporate quality of life adjustments—for example, QALYs—most of the issues generalize to other forms of economic analysis as well.

a. Measurement Challenges

By providing a common framework to assess the benefits of competing healthcare resources, measures such as QALYs are useful for cross-treatment comparisons, as noted in Part II. However, there are various conceptual issues with how QALYs are constructed. Without setting forth the full details of a QALY calculation, we note a few such issues in this section. One concern is the fact that the quality-of-life adjustment weights used in QALY calculations are derived from elicited utility preferences.[44]

Preferences, however, will likely differ across various types of groups due to various social, cultural, or economic differences, and there is a lack of consensus over whose preferences should be used. For example, individuals in manual labor settings may assign greater weight to physical functional limitations than individuals in knowledge-based occupational settings.[45]

Consider also the case of disabled individuals. Interestingly, disabled patients with a given condition often report less limitation on quality of life than community members without the condition. On the one hand, the ex ante preferences of a community member may less accurately capture the effect of a treatment on quality of life than the ex post informed preferences of a disabled person. But on the other hand, the ex post preferences of disabled persons may reflect adaptation to their functional limitations, a lowering of expectations that may provide a less acceptable basis for developing quality-of-life adjustment weights.[46] Relatedly, if community member preferences (assigning greater functional limitation to disabilities) are used in CEAs, then treatments extending the lives of disabled individuals will receive less weight. But if CEAs rely on disabled individual preferences (assigning less functional limitation to disabilities), then less weight will be given to treatments preventing disability.[47]

Another concern relating to QALY construction regards what life expectancy measures should be used.[48] QALYs gained by an intervention effectively capture the amount of time spent in a particular state of health that is made possible by the intervention itself. A calculation of this sort naturally requires some estimate of the general life expectancy facing the patient—after all, an intervention that avoids death and returns a patient to a healthy state will not do so indefinitely. In evaluating a treatment for

individuals with a particular condition, it may seem reasonable to use the average life expectancy for individuals with that condition. However, defining QALY gains in this way would result in fewer QALYs gained for individuals with life-shortening disabilities, making treatments targeting these individuals appear less cost-effective. In turn, life expectancies also vary by socioeconomic status, race, ethnicity, and gender. Some of these differences appear largely biological in nature, while others may have been exacerbated by unjust social conditions (e.g., between the poor and wealthy). How should such differences factor into QALY calculations?

A final and related debate frequently invoked bears on whether QALYs should be age-weighted. While CEAs count QALYs the same regardless of the age of the patient, some argue that QALYs should be age-weighted to give priority to the young. One justification, often referred to as the "fair innings" argument,[49] posits that as a matter of fairness, priority should be given to help individuals achieve a normal lifespan; the young, who have enjoyed less of their life share, should therefore receive greater priority.[50] While some forms of preferencing may be deemed discriminatory in certain contexts (e.g., based on race or gender), some argue that preferencing the young is not discriminatory because everyone is young at some point, during which time they would have received priority.[51] Another rationale argues that priority should be given to QALYs gained by adults during their working years, given that they are a primary source of financial, emotional, and physical support for children and the elderly.[52] However, one concern with this argument is that it values the health of adults based on their instrumental value to others.

b. Analytical Challenges

In addition to the measurement challenges described in Part IV(a), there are a variety of issues relating to how CEAs are conducted. In this section, we discuss several such issues. The first operational obstacle relates to the perspective from which CEAs are conducted. Often CEAs take the viewpoint of the agent making the decision—for example, the payer—but CEAs may also be conducted from the societal perspective. While CEAs conducted from the payer perspective use the prices charged for a treatment as the cost, the societal perspective includes all health benefits and costs regardless of who receives the benefits or bears the costs. The choice of a perspective, therefore, may have a large effect on the results of the analysis.

Consider, for example, a drug and a medical device of equal effectiveness. The drug's cost of production is low, but its price is high (under patent protection, it sells at a substantial markup), whereas the medical device's cost of production is moderate, and its price markup is small. On these facts, the medical device may be more cost-effective than the drug from the payer perspective because it offers equal effectiveness at a lower price to the payer. In turn, the drug may offer greater value from the societal perspective, given that it is less resource intensive to produce and offers equal effectiveness. Although the payer faces a higher price for the drug, from the societal perspective this constitutes a welfare transfer from payer to producer, not a welfare loss.[53]

When CEAs are used to inform coverage decisions, employing the payer perspective may also discourage the offering of preventive care. Given the high rate at which individuals switch health plans ("insurance churn"),[54] payers have less incentive to promote preventative care that provides future health benefits since many of the payer's

current members will not be members in the mid/long term, even if the preventive care would be supported from the societal perspective.

A second issue involves which costs and benefits should be taken into account in CEA calculations. Direct health benefits and costs are always included, but should indirect nonhealth benefits and costs be included, as well?[55] An efficient allocation of resources should indeed base its analysis on all costs and benefits stemming from the treatments under investigation, whether indirect versus direct or health versus nonhealth in nature. However, some object on grounds other than efficiency to the inclusion of some items. A prime example here is smoking-cessation treatment. Such programs may extend lives directly, but they may also tend to increase Social Security costs. Including these indirect nonhealth costs in a CEA would result in a lower smoking prevention cost-effectiveness ratio. Some view it unfair to allow a nonhealth benefit of this nature tip the scale in favor of allocation to one patient over another, as this treats patients instrumentally as a means to benefit others.[56] Including nonhealth benefits in other contexts raises fewer eyebrows—for example, the inclusion of enhanced worker productivity and reduced burdens on friends and family resulting from substance abuse treatments.

A final analytical/measurement issue related to QALYs concerns whether health benefits that accrue to individuals in the future should be discounted.[57] While future costs are discounted to reflect the time value of money, which is appropriate, it is not clear that health benefits should be discounted as well. Different interventions will take varying lengths of time to yield benefits, and the discounting of health benefits could result in greater priority for interventions that provide small benefits in the short term over

interventions that provide larger benefits in the future, including vaccine programs and other preventive interventions. While paradoxes may arise when health benefits are not discounted or discounted at a different rate than costs,[58] an adequate ethical justification for discounting health benefits has not been offered.

c. Application Challenges

Even assuming that the measurement and analytical challenges discussed in Parts IV(a) and IV(b) have been adequately addressed, issues with how the results of CEAs are applied remain. In this section, we review two issues sometimes called the "fair chances/best outcomes problem" and the "aggregation problem."[59]

The fair chances/best outcomes problem concerns rationing decisions between two or more treatments in which allocating resources for one treatment would provide better outcomes for a given group, but those in need of other treatments may nevertheless claim that they should still have a fair chance to receive the resources.[60] Consider the following stylized example: Two treatments yield different levels of value, say treatment A for condition X provides 1.00 QALY per unit of cost and treatment B for condition Y provides 0.99 QALYs at the same cost. A strict utilitarian applying these CEA results would conclude that the resources should be allocated to treatment A. However, most people would conclude that the 0.01 QALY difference between treatments A and B does not, ceteris paribus, sufficiently justify this decision. But what if treatment B provided only 0.90 QALYs (or 0.80, 0.50, 0.20)? In several of these cases, many people may believe that patients with condition Y (who need treatment B) have a reasonable complaint: since the treatments are "close enough" in value, everyone deserves a fair chance to receive the available resources.[61] However, most people believe there is some

point at which the value of treatment A so outweighs the value of treatment B that appeals to fair chances should no longer be deemed relevant. In the abstract, one proposal to address the fair chance/best outcomes problem is the use of a weighted lottery.[62] Yet even if this approach could be operationalized, it remains unclear how far intuitions about fair chances extend.

Second, the aggregation problem arises when modest benefits to a large number of individuals are more cost-effective than large benefits to a small number of individuals.[63] Philosophers disagree over the propriety of aggregating benefits across individuals in order to assess how limited resources should be allocated across different groups. An often cited example of this issue occurred in the state of Oregon Medicaid experiment.[64] In 1989, Oregon decided to expand its Medicaid program, financing the expansion by offering Medicaid beneficiaries a more limited set of healthcare services chosen and ranked based on cost-effectiveness. But when Oregon published its schedule of covered and uncovered services, dental caps for pulp exposure—providing modest benefits to many people—were determined to be cost-effective and therefore covered, while life-saving surgical procedures for ectopic pregnancy and appendicitis—providing substantial benefits to fewer people—were less cost-effective and given less preferential coverage treatment. Comparing the total benefits across these groups proved worrisome for many observers. Controversy over these and other coverage decisions led to the disbanding of the program.

Aside from the aggregation component per se, part of the discomfort in this example comes in comparing treatments with modest health benefits but high cost-effectiveness ratios to those with substantial health benefits but lower cost-effectiveness

ratios. One attempt to address the uneasiness appeals to the so-called "rule of rescue," which affords priority to less cost-effective treatments in life-or-death situations, assuming the cost-effectiveness is not unreasonably low.[65] The rule of rescue relates to the notion that the maximand of CEAs should be saving the most lives in a more immediate, "identifiable" sense, as opposed to extending life years in a more predicted and "statistical" sense. Some commentators have been critical of this distinction between identifiable lives and statistical lives, considering that any treatment that extends a patient's life also in a sense "rescues" that patient at some future time.[66] Yet some research suggests that individuals are willing to pay more per QALY for life-saving treatments than for life-extending or life-enhancing treatments,[67] suggesting that perhaps some (though not absolute) weight should be given based on the type of treatment being prioritized.

V. CONCLUDING REMARKS

A comprehensive discussion of the economics surrounding healthcare utilization and the means by which governments, private insurers, and market dynamics themselves may constrain such utilization (and thus "ration" limited resources) would require far more attention than a single article could hope to achieve. Our goals here have indeed been more modest. We have focused on how certain tools of economics, mainly cost-effectiveness analysis, have shaped distinct types of non-price-rationing practices. Indeed, much has been omitted in our discussion regarding the numerous other ways in which governments and private parties may regulate healthcare utilization. For instance, while we have addressed more categorical approaches to regulating which services patients may obtain, we have not addressed more case-by-case mechanisms that insurers

may employ—for example, utilization review. Nor have we discussed certain modern approaches to rationing care such as global payment systems that retreat from fee-for-service reimbursement models and instead reimburse providers (either individually or in a group) a flat fee for treating specified patients over a given period of time or during particular episodes of care.

Structures of this latter variety—that is, alternative means of reimbursing physicians—are worth emphasizing briefly in these concluding remarks. Our discussion in this article largely focused on policies and programs aimed at curbing or shaping patient decision-making. Excessive use of medical care, however, may also arise from another source: physicians. Perhaps one of the most important features of the U.S. healthcare system that may contribute to excessive healthcare utilization is the fee-for-service environment largely characterizing the U.S. healthcare delivery system.[68] Under a fee-for-service approach, physicians are effectively paid more for doing more, creating incentives on the part of physicians to perform an unnecessarily high degree of services, a phenomenon often labeled "physician induced demand." Whether fee-for-service reimbursement structures, in fact, cause physicians to provide an unnecessarily large number of procedures is the subject of an extensive literature.[69] While establishing causation in such studies is empirically challenging, the most convincing evidence put forth to date suggests that at least some physicians may indeed be inducing demand in this manner.[70] With these considerations in mind, more complete discussions regarding healthcare cost containment and healthcare rationing cannot ignore the decision-making role of physicians in this complex interaction between medical providers, insurers, and patients.

APPENDIX

Return to the heart attack example in Part II, in which we evaluated the use of two treatment interventions: streptokinase and t-PA. We now evaluate the choice between these two interventions employing cost-benefit analysis (CBA), as opposed to cost-effectiveness analysis. Assume that the baseline treatment is streptokinase, and a policy-maker is deciding whether to adopt t-PA as a treatment for society, guided by CBA. Effectively, this analysis entails calculating the monetary benefits of t-PA relative to that of streptokinase and asking whether such relative benefits exceed the costs of t-PA relative to streptokinase. If such net benefits are indeed greater than 0, the CBA suggests that efficiency would be enhanced by selecting t-PA over streptokinase.

Assume that the use of t-PA results in a lower risk of death than the use of streptokinase. How should such reduced risk of death be valued? CBAs often approach this inquiry by calculating something known as compensating variation (CV). In this context, CV can be thought of as the amount of money that must be taken away from the individual in order to leave her just as well off as she was prior to the reduction in risk. Conversely, CV can be interpreted as the maximum amount of money the patient would be willing to pay to obtain the treatment reducing risk of death (t-PA) over the current treatment (streptokinase).

Continuing our example, suppose a representative heart attack patient has lifetime expected utility under treatment t of $E\big[U\big(Y\,|\,t\big)\big]=\big(1-p_t\big)u\big(Y\,|\,live\big)+p_t u\big(Y\,|\,dead\big)$, where p_t is the probability of death with treatment t. In words, conditional on obtaining treatment t, the patient lives with probability $\big(1-p_t\big)$ and obtains a utility level

$u(Y \mid live)$ and dies with probability p_t and obtains a utility level $u(Y \mid dead)$. Let's denote the probably of death under streptokinase treatment as p and probability of death under the t-PA treatment as q. Assume that $u(Y \mid dead) = 0$, that is, the patient gets no utility if he or she is dead. Further assume that the patient has $100,000 in income. Solving for CV means that we find the CV that leaves the consumer indifferent between an income of $100,000 with a probability of death of p and an income of $100,000–CV with a probability of death of q.

For concreteness, let's solve a simple example by specifying a functional form for the utility function, say, $u = \ln(Y)$, and assuming that $Y = \$100K$, $p = 0.05$, and $q = 0.025$.[71] We can find CV = $25,562.[72] In words, under the simplified example here, the representative heart attack patient's willingness to pay for the t-PA treatment (as opposed to streptokinase) is $25,562, which far exceeds the monetary cost of the procedure over streptokinase ($1,800). If the relative monetary cost of the treatment ($1,800) equals the social opportunity cost of the treatment, the decision rule from the CBA would be to adopt t-PA as a treatment because it has positive net benefits.

Table 1. Cost-Effectiveness Analysis Example

Strategy	Cost	Incremental Cost	Effectiveness	Incremental Effectiveness	Cost/Effectiveness	ICER
No treatment	$0	$0	900	900	$0	$0
Streptokinase	$200,000	$200,000	950	50	$211	$4,000
t-PA	$2,000,000	$1,800,000	975	25	$2,051	$72,000

---

[1] Commentators have identified a range of normative frameworks to guide this welfare analysis. Such guiding principles include: (1) "first-come-first-serve" (assigning

priority to individuals simply on the basis of time), *see* American Thoracic Society, *Fair Allocation of intensive Care Unit Resources*, 156 AM. J. RESPIR. CRIT. CARE MED. 1282 (1997); *but see* Norman Daniels, *Fair Process in Patient Selection for Antiretroviral Treatment in WHO's Goal of 3 by 5*, 366 LANCET 169 (2005); (2) "instrumental value" (assigning priority to patients who contribute more to society—*e.g.*, giving healthcare professionals priority access to scarce vaccines given that these workers are needed to treat others; or assigning priority based on one's responsibility for their medical need—*e.g.*, providing lower priority to alcoholics), *see* Ezekiel Emanuel & Alan Wertheimer, *Who Should Get Influenza Vaccine When Not All Can?*, 312 SCI. 854 (2006); (3) "best outcomes" (employing a utilitarian principle that aims to maximize social welfare, typically without regard to the distribution of benefits and burdens; this principle generally underlies the economic analyses used for healthcare rationing, *see* Part II); and (4) "equity" (allocating scarce resources based on a fair distribution of benefits and burdens, rather than merely maximizing net benefits), *see* Alan Williams & Richard Cookson, *Equity in Health*, *in* HANDBOOK OF HEALTH ECONOMICS 1863–1910 (Anthony Culyer & Joseph P. Newhouse eds., 2000). Selected equity principles include "prioritarianism" (giving priority to those who are worst off), *see* Derek Parfit, *Equality and Priority*, 10 RATIO 202 (1997), "sufficientarianism" (giving priority to individuals or groups below a predefined threshold level of well-being), *see* Roger Crisp, *Equality, Priority, and Compassion*, 113 ETHICS 745 (2003), and Rawls's "fair equality of opportunity principle" as applied to health (arguing that if there is a social obligation to

protect fair equality of opportunity, there is also a derivative obligation to promote health because health has special moral importance given its link with opportunity), *see* NORMAN DANIELS, JUST HEALTH (Cambridge University Press 2007); *see also* JOHN RAWLS, A THEORY OF JUSTICE (Belknap Press 2nd ed. 1999). The above list of guiding principles roughly follows the taxonomy set forth in Govind Persad et al., *Principles for Allocation of Scarce Medical Interventions*, 373 LANCET 423 (2009); *see also* I. Glenn Cohen, *Rationing Legal Services*, 5 J. LEGAL ANALYSIS 221 (2013).

[2] *See The Slow Rise of the Robot Surgeon*, MIT TECH. REV. (Mar. 24, 2010).

[3] *See* Judd Kessler & Alvin Roth, *Organ Allocation Policy and the Decision to Donate*, 102 AM. ECON. REV. 2018 (2012).

[4] *See* Adam Hofer et al., *Expansion of Coverage under the Patient Protection and Affordable Care Act and Primary Care Utilization*, 89 MILBANK QUARTERLY 69 (2011).

[5] For recent research on various policy levers that can potentially be used to combat the shortage of human organs for transplant, *see* Eric Johnson & Daniel Goldstein, *Do Defaults Save Lives?*, 302 SCIENCE 1338 (2003); *see also* Judd Kessler & Alvin Roth, *Loopholes Undermine Donation: An Experiment Motivated by an Organ Donation Priority Loophole in Israel*, 114 J. PUB. ECON. 19 (2014).

[6] While it is beyond the scope of this article, we note that dynamic allocation decisions of this nature implicate matters beyond just rationing and cost containment—mainly, innovation policy. There is an inevitable conflict between healthcare rationing policies that aim to constrain spending on the one hand, and healthcare innovation

policies that seek to promote research and development on the other. For instance,
healthcare utilization of new technologies may sometimes appear to be inefficient
and excessive from a static perspective, perhaps arising from moral hazard,
physician-induced demand, or related features of healthcare systems. However,
from a more dynamic perspective, the promise of such financial returns may
stimulate the desire to produce such innovations in the first place. In fact, a
leading explanation for growth in U.S. healthcare spending is technological
innovation. *See* Amitabh Chandra & Jonathan Skinner, *Technology Growth and
Expenditure Growth in Health Care*, 50 J. ECON. LIT. 645 (2012).

[7] *See* ANTHONY BOARDMAN ET AL., COST–BENEFIT ANALYSIS (4th ed. 2011).

[8] *See* Micah Hartman et al., *National Health Spending in 2013: Growth Slows, Remains
in Step with the Overall Economy*, 34 HEALTH AFF. 1 (2014).

[9] Determining the optimal amount of resources to healthcare, as any good or service,
depends on, among other things, societal preferences, national wealth, and the
distribution of that wealth. Contrary to public perception, increases in health
spending does not necessarily imply decreases in social welfare or even create a
(greater) need for rationing. Naturally, as we become richer, society's willingness
to pay for a longer and healthier life increases, which translates into greater
healthcare spending overall and perhaps also as a percentage of national income.
As a result, some preconceived amount of national wealth devoted to healthcare,
whether it be 18% or even 50%, does not serve as a sufficient condition for which
to justify the enactment of policies aimed at rationing healthcare. *See* Robert Hall

& Charles Jones, *The Value of Life and the Rise in Health Spending*, 122 Q. J. ECON. 39 (2007).

[10] For a recent literature review on the ways CBA and CEA have been converging, *see* Linda Ryen & Mikael Svensson, *The Willingness to Pay for a Quality Adjusted Life Year: A Review of the Empirical Literature*, HEALTH ECON. (2014) (forthcoming).

[11] Nicholas Kaldor, *Welfare Propositions of Economics and Interpersonal Comparisons of Utility*, 49 ECON. J. 549 (1939); John Hicks, *The Valuation of the Social Income*, 7 ECONOMICA 105 (1940).

[12] The necessity of monetizing the benefits of healthcare interventions in CBAs arises because of the need to place a value on opportunity costs used to decide where to employ resources. *See* Alan Garber & Charles Phelps, *Economic Foundations of Cost-Effectiveness Analysis*, 16 J. HEALTH ECON. 1 (1997). Any discussion of CBAs for health interventions naturally leads to the controversial methods of valuing health and life, which are far from similar. *See* W. Kip Viscusi*, The Value of Individual and Societal Risks to Life and Health*, *in* HANDBOOK OF THE ECONOMICS OF RISK AND UNCERTAINTY 385 (Mark Machina & Kip Viscusi, 2014). The most frequently used terminology for valuing probabilistic decreases to risks to life is called the "value of statistical life" (VSL), which is simply a convenient way of summarizing one's willingness to pay for small reductions in fatal risks. *See* W. Kip Viscusi & Joseph Aldy, *The Value of a Statistical Life: A Critical Review of Market Estimates Throughout the World*, 27 J. RISK & UNCERTAINTY 5 (2003). For clarity, VSL is neither the value of saving the life of

a specific identified person nor the value of reducing high probabilities of death, for instance, from 1 to 0. Because many of those health interventions that we associate with healthcare rationing deal with large reductions in risk to life, VSL does not lend itself well to a CBA and should be avoided.

[13] *See* Donald Kenkel, *Using Estimates of the Value of a Statistical Life in Evaluating Consumer Policy Regulations*, 26 J. CONSUMER POL'Y 1 (2003).

[14] *See* Jeremiah Hurley, *Chapter 2: An Overview of the Normative Economics of the Health Sector*, *in* HANDBOOK OF HEALTH ECONOMICS 55–118 (Anthony Culyer & Joseph Newhouse eds., 2000); Nien-he Hsieh et al., *The Numbers Problem*, 34 PHIL. & PUB. AFF. 352 (2006); Joseph Raz, *Numbers, With and Without Contractualism*, 16 RATIO 346 (2003).

[15] *See* Magnus Johannesson, *Should We Aggregate Relative or Absolute Changes in QALYs?*, 10 HEALTH ECON. 573 (2001).

[16] *See* BOARDMAN ET AL., COST–BENEFIT ANALYSIS, at 464.

[17] Cost-effectiveness analyses that use QALYs or other utility-based measures are sometimes referred to as "cost utility analyses" because the measures serve as a proxy for changes in utility levels. DALYs can be distinguished from QALYs in how "the mortality and morbidity burdens of various diseases" differ in different populations. Christopher Murray & Arnab Acharya, *Understanding DALYs*, 16 J. HEALTH ECON. 703 (1997).

[18] This example was modified from an example presented in BOARDMAN ET AL., COST–BENEFIT ANALYSIS.

[19] The incremental effectiveness is calculated by differencing the expected number of heart attack survivors with the streptokinase treatment [(1–0.05)1000 = 950] with the expected number of heart attack survivors with no treatment [(1–0.1)1000 = 900].

[20] The ICER is again calculated by dividing the incremental costs of $1.8 million [(1000)($2000)–(1000)($200)] by the incremental effectiveness of 25 individuals [(1–0.05)1000–(1–0.025)1000].

[21] *See* Christopher McCabe et al., *The NICE Cost-Effectiveness Threshold: What It Is and What That Means*, 26 PHARMACOECONOMICS 733 (2008).

[22] For a discussion of the various ways in which QALYs have been valued, s*ee* Donald Kenkel, *A WTP- and QALY-Based Approaches to Valuing Health for Policy: Common Ground and Disputed Territory*, 34 ENVTL. & RES. ECON. 419 (2006).

[23] *See id.*

[24] *See* Susan Bartlett Foote & Peter Neumann, *The Impact of Medicare Modernization on Coverage Policy: Recommendations for Reform*, 11 AM. J. MANAGED CARE 140 (2005). Medicare has made several attempts to incorporate some form of CEA into its coverage considerations, but it has been unable to do so due to political opposition. *See id.*

[25] *See* Sean Tunis, *Why Medicare Has Not Established Criteria for Coverage Decisions*, 350 NEW ENG. J. MED. 2196 (2004).

[26] For concerns with HDHPs, *see* J. Frank Wharam et al., *High-Deductible Insurance: Two-Year Emergency Department and Hospital Use*, 17 AM. J. MANAGED CARE 410 (2011). Regarding growth, 38% of covered workers were in HDHPs in 2013

and enrollment may accelerate under the Affordable Care Act. *See* Amelia Haviland et al., *Growth of Consumer-Directed Health Plans to One-Half of All Employer-Sponsored Insurance Could Save $57 Billion Annually*, 31 HEALTH AFF. 1009 (2012).

[27] *See* Dana Goldman et al., *Pharmacy Benefits and the Use of Drugs by the Chronically Ill*, 291 J. AM. MED. ASS'N 2344 (2004).

[28] *See* Kevin Frick & Michael Chernew, *Beneficial Moral Hazard and the Theory of the Second Best*, 46 INQUIRY: J. HEALTH CARE ORG., PROVISION, & FIN. 229 (2009).

[29] *See* Amelia Haviland et al., *How Do Consumer-Directed Health Plans Affect Vulnerable Populations?*, 14 F. FOR HEALTH ECON. & POL'Y 1 (2011).

[30] *See* Niteesh Choudhry et al., *Assessing the Evidence for Value-Based Insurance Design*, 29 HEALTH AFF. 1988 (2010).

[31] *See* A. Mark Fendrick et al., *A Benefit-Based Copay for Prescription Drugs: Patient Contribution Based on Total Benefits, Not Drug Acquisition Cost*, 7 AM. J. MANAGED CARE 861 (2001). For example, a VBID plan offered by the University of Michigan lowers copayments for medications to treat diabetes, high blood pressure, high cholesterol, and depression as follows: generics by 100%, tier 2 by 50%, and tier 3 by 25%. *See* Alicen Spaulding et al., *A Controlled Trial of Value-Based Insurance Design—The MHealthy: Focus on Diabetes (FOD) Trial*, 4 IMPLEMENT SCI. 19 (2009).

[32] Inhaled steroids, for example, are effective for treating asthma but not chronic obstructive pulmonary disease. *See* Choudhry et al., *Assessing the Evidence for Value-Based Insurance Design*.

[33] *See* John Rowe et al., *The Effect of Consumer-Directed Health Plans on the Use of Preventive and Chronic Illness Services*, 27 HEALTH AFF. 113 (2008).

[34] *See* Sheila Reiss et al., *Effect of Switching to a High-Deductible Health Plan on Use of Chronic Medications*, 46 HEALTH SERV. RES. 1382 (2011).

[35] *See* Choudhry et al., *Assessing the Evidence for Value-Based Insurance Design*.

[36] *Impact of Decreasing Copayments on Medication Adherence Within a Disease Management environment*, 27 HEALTH AFF. 103 (2008).

[37] *See* Matthew B. Frank et al., *The Effect of a Large Regional Health Plan's Value-Based Insurance Design Program on Statin Use*, 50 MED. CARE 934 (2012).

[38] *See* Teresa Gibson et al., *Value-Based Insurance Plus Disease Management Increased Medication Use and Produced Savings*, 30 HEALTH AFF. 100 (2011).

[39] *See* Jon Kingsdale, *After the False Start—What Can We Expect from the New Health Insurance Marketplaces?*, 370 NEW ENG. J. MED 393 (2014).

[40] *See* James Robinson, *Hospital Tiers in Health Insurance: Balancing Consumer Choice with Financial Incentives*, 22 HEALTH AFF. 135 (2003). Some insurers have started to experiment with stronger incentives to steer patients toward preferred providers, including larger cost-sharing amounts and greater inter-tier cost sharing differences. *See id.* To the extent that provider networks steer patients toward higher value providers, limited networks have the potential to be more effective than tiered networks because of their more severe incentive structure. On the other hand, relative to limited network plans, tiered network plans preserve patients' ability to choose from a wide range of providers (albeit with higher cost sharing for nonpreferred providers) and thus could be more palatable to patients.

[41] *See* Choudhry et al., *Assessing the Evidence for Value-Based Insurance Design*.

[42] Mass. Med. Soc'y et al. v. Group Ins. Comm'n et al., No. # 2008-cv-2124 (Suffolk Sup. Ct., filed May 21, 2008); Wash. St. Med. Ass'n et al. v. Regence Blue Shield, No. # 06-2-30655-ISEA (King County Sup. Ct. 2006); Fairfield County Med. Soc'y et al. v. CIGNA Corp. et al., No. # CV-075002943 (Conn. Sup. Ct. 2008).

[43] *See* Matthew Frank et al., *The Impact of a Tiered Network on Hospital Choice*, HEALTH SERVICES RESEARCH, in press (2015).

[44] These preferences are elicited using a range of valuation techniques, including standard gamble, time tradeoff, rating scale, multi-attribute utility, and person tradeoff techniques. *See* Milton Weinstein et al., *QALYs: The Basics*, 12 VALUE IN HEALTH S5 (2009).

[45] *See* Dan Brock, *Ethical Issues in the Use of Cost Effectiveness Analysis for the Prioritization of Health Resources*, *in* HANDBOOK OF BIOETHICS 294 (George Khushf, 2004).

[46] *See* MYRIAM HUNINK & PAUL GLASZIOU, DECISION MAKING IN HEALTH AND MEDICINE: INTEGRATING EVIDENCE AND VALUES 268–269 (2001).

[47] *See* Paul Menzel et al., *The Role of Adaptation to Disability and Disease in Health State Valuation: A Preliminary Normative Analysis*, 55 SOC. SCI. & MEDICINE 2149 (2002).

[48] *See* Brock, *Ethical Issues in the Use of Cost Effectiveness Analysis*.

[49] *See* Alan Williams, *Intergenerational Equity: An Exploration of the "Fair Innings" Argument*, 6 HEALTH ECON. 117 (1997).

[50] *See* JOHN HARRIS, THE VALUE OF LIFE: AN INTRODUCTION TO MEDICAL ETHICS 87–102 (1990).

[51] *See* NORMAL DANIELS, AM I MY PARENTS' KEEPER? AN ESSAY ON JUSTICE BETWEEN THE YOUNG AND THE OLD (1988).

[52] *See* Greg Bognar, *Age Weighting*, 24 ECON. & PHIL. 167 (2008).

[53] *See* Anupam Jena & Thomas Philipson, *Cost Effectiveness as a Price Control*, 26 HEALTH AFF. 696 (2007).

[54] *See* Bradley Herring, *Suboptimal Provision of Preventive Healthcare Due to Expected Enrollee Turnover Among Private Insurers*, 19 HEALTH ECON. 438 (2010).

[55] *See* Brock, *Ethical Issues in the Use of Cost Effectiveness Analysis*.

[56] *See* FRANCIS KAMM, MORALITY, MORTALITY: DEATH AND WHOM TO SAVE FROM IT 369 (1993).

[57] *See* Brock, *Ethical Issues in the Use of Cost Effectiveness Analysis*.

[58] *See* Emmett Keeler & Shan Cretin, *Discounting of Life-Saving and Other Nonmonetary Effects*, 29 MGMT. SCI. 300 (1983).

[59] *See* Norman Daniels, *Four Unsolved Rationing Problems A Challenge*, 24 HASTINGS CENTER REP. 27 (1994).

[60] This issue may also arise in other healthcare rationing contexts in determining the allocation of physically scarce resources such as solid organs and units of vaccine.

[61] Some may argue that patients with condition Y already had a fair chance and lost in the "natural lottery," while others (e.g., luck egalitarians) want to compensate for bad brute luck.

[62] *See* John Broome, *Selecting People Randomly*, 95 ETHICS 38 (1984).

[63] *See* Eric Rakowski, *The Aggregation Problem*, 24 HASTINGS CENTER REP. 33 (1994).

[64] *See* David Hadorn, *Setting Health Care Priorities in Oregon: Cost Effectiveness Meets the Rule of Rescue*, 265 J. AM. MED. ASS'N 2218 (1991).

[65] *See* Albert Jonsen, *Bentham in a Box: Technology Assessment and Health Care Allocation*, 14 J. L., MED. & ETHICS 172 (1986).

[66] *See* Cohen, *Rationing Legal Services*.

[67] *See* Cam Donaldson et al., *The Social Value of a QALY: Raising the Bar or Barring the Raise?*, 11 BMC HEALTH SERVICES RES. 8 (2011).

[68] *See* David Orentlicher, *Cost Containment and the Patient Protection and Affordable Care Act*, 6 FIU L. REV. 67 (2010).

[69] For an overview of the empirical literature (up to 2000) exploring this hypothesis, *see id*.

[70] *See* Jonathan Gruber & Maria Owings, *Physician Financial Incentives and Cesarean Section Delivery*, 27 RAND J. ECON. 99 (1996).

[71] Note that p and q will not likely be the same as the hospital mortality rates in the above example, but for simplicity let's assume they are the same.

[72] Note the expected *monetary* benefits of the procedure in terms of income to the representative individual are: $E[B] = (0.05–0.025) * 100,000 = \$2,500 < CV = \$25,562$. Thus, note that CV is a lot larger than the expected monetary benefit because of risk aversion.