Since 1970, the annual growth in U.S. health care spending per capita has been more than double the real growth in GDP per capita: 4.3 percent versus 2 percent. Over that same time period countries belonging to the Organization for Economic Cooperation and Development (OECD) averaged an annual growth rate of 3.8 percent in health care spending per capita compared to only a 2.1 percent annual growth in GDP per capita. Eight of 20 countries had higher average annual growth rates in health care spending per capita than the United States (White 2007). In light of the pronounced institutional differences among these countries in medical financing arrangements, the similarity in the rate of health care spending growth is striking. Therefore, any explanation that seeks to account for the tremendous cost growth in health care over the last several decades must hold true across all OECD countries.

This article describes a construct for health care cost growth associated with social welfare loss that we refer to as the medical care cost
ratchet (MCCR). In this model, health care spending increases over time as new technologies are incorporated into the traditional standard of care that confer only modest clinical benefits. We explain how the current medical insurance model perpetuates the MCCR. We then explain how medical cost analyses are performed before presenting several clinical vignettes that validate our model. The article concludes by arguing that market-based approaches to health care reform would be effective at bending the cost curve over time by encouraging individuals to economize nonemergent health care decisions—doing so would upset the MCCR and reduce spending growth.

Technological Change and the MCCR

In the 1980s, the conventional explanation of health care cost growth emphasized the moral hazard from health insurance and particularly the tax treatment of health insurance (Newhouse 1992). According to this view, traditional health insurance reimburses as a function of expenditure or use. Because insurance drives the marginal price of medical care at the point of use to near zero, consumers—or physicians acting as their agents—demand care until the marginal product of additional care is nearly zero. Empirical evidence exists in support of the conventional view. Studies have found that a fully insured population spends about 40–50 percent more than a population with a large deductible and their status is not measurably improved by the additional services (Manning et al. 1987). This has been referred to by Enthoven (1980) as “flat-of-the-curve medicine,” where spending on medical care increases even though additional gains from such spending are very low or nonexistent. This idea has been recently reaffirmed by a landmark analysis of the Oregon Experiment where Baicker et al. (2013) found that Medicaid coverage generated no significant improvements in health outcomes in the first two years, despite increased use of prescription drugs, office visits, preventive care services including mammograms, and annual spending per individual (by insurance plan) in excess of $1,100.

The conventional view was challenged by Newhouse (1992: 11) who argued that the bulk of health care cost growth “is attributable to technological change, or what might loosely be called the march of science and the increased capabilities of medicine.” According to this view, increased medical spending is welfare enhancing. To support
this view, he offered that “patients are not going to the hospital more frequently . . . nor are patients staying longer. But the real cost of a day in the hospital rose by nearly a factor of 4 from 1965 to 1986. Thus, what is being done to and for people who are in the hospital is affecting hospital costs.”

While Newhouse acknowledged that there was some validity to the conventional view when looked at over a single period, he argued that it was insufficient to explain health care cost growth over time: “To explain increasing expenditure, one needs to point to something that is changing, indeed to factors that have been changing for 50 years” (Newhouse 1992: 5). He reasoned that the factor-of-five increase in real expenditure per person over the period 1950 to 1980 was more than eight times as large as one could predict from the effect of increased insurance on demand in the context of the one-period model.

Newhouse (1992) dismissed the idea that increased insurance could lead to too much technological change that was not welfare enhancing. If technological change diminished welfare, he reasoned, then countries that make centralized decisions about how much to spend on medical care would not adopt certain changes. Hence, their health care cost growth would be less than the United States. But since cost growth is similar, technological change must enhance welfare.

In our opinion, Newhouse was incorrect to dismiss the argument that too much insurance could lead to too much technological change that does not enhance welfare. In this article, we demonstrate that technological change has indeed increased health care costs in many cases without significantly improving health outcomes. This has occurred because medical insurance in its current state discourages individuals from economizing health care decisions and incentivizes the adoption and overconsumption of services with progressively diminishing returns on investment. Through this process health care costs increase as each technological advance is added to the medical repertoire. We refer to this construct as the “medical care cost ratchet” because a ratchet is a mechanical device that allows continuous linear or rotary motion in only one direction while preventing motion in the opposite direction.

Despite differences in health care financing across OECD countries, medical insurance is basically the same; its goal is to make medical care free at the point of delivery. This is achieved in several
ways. The *Bismarck system*, which originated in Germany, is dominated by private insurers that are created by or connected with employers, financed by employees and employers, and heavily regulated by the government. In the *Beveridge system*, conceived in Great Britain, the state provides comprehensive health care insurance and services to all citizens with no intermediaries (Colombatto 2012). In many countries today, including the United States, health care delivery reflects a hybrid of these two systems. Over the years, the main principle underlying them has not been challenged—that health care is a social right rather than a service to be purchased. The consequence of this attitude is that people demand the standard of medical care be made available despite the cost. Unfortunately, the general public has no conception of how valuable (or beneficial to them) the standard of care actually is and how it is determined. In many cases, providing the standard of care to a patient at the expense of a third party or the public could reasonably be considered a social welfare loss.

Before building our case we would like to clarify several points. We do not believe that our concept of the medical-care-cost-ratchet and Newhouse’s concept of technological change are mutually exclusive. Instead, they should be viewed as a continuum. For example, it is common for a new drug, diagnostic test, or procedure that significantly benefits a small subgroup of patients to be used in larger subgroups of patients who benefit from it much less or not at all. We will demonstrate an example of this later in the article. Also, we do not believe that insurance per se is the problem but rather, insurance that covers all medical expenses is. If the public desires comprehensive coverage, then its consequences should at least be clearly understood. However, public perception that health care costs are too high and insurance premiums are rising too fast suggests that many are seeking alternatives. While it is reasonable and socially desirable for individuals to insure against medical emergencies, the vast majority of health care decisions are not made in an emergent setting—they are not a matter of life or death. In an elective or nonemergent setting, personal economization should be encouraged. As F. A. Hayek (1960: 422) noted,

There is no objective standard for judging how much care and effort are required in a particular case; also, as medicine advances, it becomes more and more clear that there is no
limit to the amount that might profitably be spent in order to do all that is objectively possible. . . . As in all other decisions in which we have to deal not with certainties but with probabilities and chances, we constantly take risks and decide on the basis of economic considerations whether a particular precaution is worthwhile, i.e., by balancing the risk against other needs.

The current medical insurance model discourages individuals to economize nonemergent health care decisions and instead encourages them to do or accept whatever the physician advises. And, in most cases, whatever the physician recommends is the standard of medical care that insurers are obliged to cover. What is the standard of medical care? It is defined as the best available combination of benefit and risk; cost is a secondary consideration that is not formally recognized in most OECD countries. Where it is formally recognized, such as the United Kingdom, the acceptable cost is quite high. The standard of care for most major medical conditions is ubiquitous across OECD countries. From the physician’s standpoint, anything that provides the prospect of benefit—regardless of how small—must be offered so long as the perceived benefit is larger than the perceived risk. For example, suppose a new drug offers the prospect of reducing the risk of a heart attack in one year by 1 percent. That is a relatively small risk reduction. However, suppose the known risk of a major side effect or complication of the drug is 0.5 percent. In this situation the potential benefit outweighs the potential risk by 0.5 percent per year, so the new drug should be offered. Whether the drug costs $10, $100, $1,000, or $10,000 dollars per year does not factor into the physician’s decision.

To demonstrate the MCCR, we will present several prominent medical interventions. Each vignette will present an intervention that has been added to the standard of care, describe how it has affected clinical outcome measures, and discuss the cost of its inclusion. However, we first must explain how medical cost analyses are performed and address their limitations.

Evaluating the Cost Effectiveness of Medicine

A range of approaches exist to perform economic analyses in medicine. The two most widely used are cost-effectiveness and cost-utility analyses. Cost-effectiveness estimates are expressed in terms of “years
of life saved” (YLS) and cost-utility evaluations as “quality-adjusted life years” (QALY) gained (Meltzer 2001). These approaches aim to assess the cost, both direct and indirect, of any therapeutic intervention with respect to its predictable benefits—with the effectiveness or utility being measured as the mean YLS or QALY gained as a result of the intervention (Boriani et al. 2009).

First and most important, cost effectiveness and cost utility do not mean cost saving. Cost saving means that the costs of an intervention are less to the payer than the costs would be if the intervention was not performed. For example, if a payer did not cover drug X, which reduces the annual risk of a heart attack in patients with condition Y, the payer would end up paying more down the line to cover the costs of treating heart attacks that could have been prevented if drug X were used by all patients with condition Y. Little is done in modern medicine that is actually cost saving. It is also important to note that cost effectiveness and cost utility do not correspond to the price of the service in question. It is often presumed that if a service is cheap, then it is cost effective and vice versa. However, a very cheap medicine can have a very poor cost ratio if many people must be treated to stop one event from occurring and if survival or quality of life is not significantly altered by its use.

Ultimately, cost analyses are modeling exercises rather than scientific experiments. This makes them dependent on the input variables such as the cost of treatment or the cost of a hospitalization for an adverse event. They are especially sensitive to the validity of data obtained in clinical trials and are therefore likely to overestimate the cost effectiveness or cost utility of any drug or intervention. Positive outcome bias, also known as “publication bias,” is the well-established tendency of investigators, reviewers, and editors to submit or accept manuscripts for publication that have positive findings and to ignore or reject negative studies (Dickersin 1990, Hasenboehler et al. 2007). Cost-effectiveness and cost-utility analyses rely on the validity of unbiased, balanced, and objective data from published studies, independent of the reported outcome. This is corrupted by the positive outcome bias of individual studies, rendering clinical recommendations and cost analyses flawed toward a positive effect of specific treatment strategies. In the era of evidence-based medicine, this prevalent, often unrecognized, positive outcome bias poses a severe challenge to cost analyses by promoting unjustified therapeutic concepts.
Other limitations of cost analyses include the narrowness of inclusion criteria used in clinical trials, which often exclude elderly patients and patients with more serious co-morbidities. This can bias the cost analysis in favor of an intervention when in real life it would be much less favorable due to its use in populations who were not represented in the clinical trial(s). Cost analyses may also be plagued by overly optimistic assumptions. For example, a recent cost analysis published in a major cardiovascular journal observed that the cost effectiveness of a new endovascular technique for lowering blood pressure was very favorable; however, the authors assumed that its proven benefit in lowering blood pressure directly translated to reductions in cardiac events, though the procedure had never been proven to reduce events (Geisler et al. 2012).

Despite these significant limitations, the numbers of cost-effectiveness and cost-utility analyses have increased steadily over the last several decades. These studies have covered a range of interventions, with drugs, surgical procedures, and various diagnostic procedures as well as a range of conditions, with particular emphasis on the cardiovascular system. A review of the literature up to 1997 found that the cost utility of medical interventions varied considerably with a median QALY gained ratio of $2,000 for vaccines, $6,000 for medical care delivery, $10,000 for surgical interventions, $11,000 for pharmaceuticals, $12,000 for screening, $20,000 for health education/counseling, $20,000 for diagnostics, and $40,000 for devices. Cost-utility analyses funded by industry had more favorable results than those that were non–industry sponsored (Neumann 2000).

Due to the limitations and complexity of formal medical cost analyses, some experts suggest thinking about cost effectiveness in another, simpler term—NNT, which stands for *numbers needed to treat* or the number of patients needed to apply a particular intervention to realize a benefit of that intervention. Simply stated, the higher the NNT is for a particular intervention, the lower the cost effectiveness is for that intervention.

**Illustrative Cases**

In this section, we refer to cost analyses as well as NNT derived from clinical trial data to demonstrate the MCCR.
Screening for Breast Cancer with Mammography

Screening mammography in women without any signs or symptoms of breast cancer has been studied in large randomized trials of nearly a half-million women. The theoretic basis for the intervention is sound. It is presumed that therapeutic intervention at a point when cancer is visible on mammogram but not yet palpable or visibly noticeable on the breast will result in earlier, ultimately life-saving therapy.

Gotzsche and Nielsen (2013) found that routine screening is likely to reduce the absolute rate of breast cancer mortality by only 0.05 percent over a 10-year period. However, screening led to an absolute rate of overdiagnosis and overtreatment by 0.5 percent over 10 years. This means that for every 2,000 women screened throughout 10 years, only 1 will avoid death from breast cancer, and 10 healthy women will be misdiagnosed and treated unnecessarily. Furthermore, more than 200 women will experience important psychological distress for months because of false positive findings.

Based on this data, a formal cost-effectiveness analysis of screening mammography is not required to appreciate the massive costs required to prevent a single death from breast cancer. Moreover, despite the miniscule reduction in breast cancer mortality, screening mammography does not reduce overall mortality (Baum 2013).

Coronary Stents

Coronary stents represent an excellent example of an intervention that is beneficial in subgroups of patients in emergent clinical settings but is much less so for the majority of patients who receive them electively. There are three broad categories of patients who receive a coronary stent. The first category is made up of patients who have a major heart attack that involves the full thickness of the heart muscle; this is termed an ST segment elevation myocardial infarction or STEMI. These patients are at significant risk of dying and being debilitated. In this setting, putting in a stent reduces the risk of death by 5–10 percent compared to the best available medical therapy alone (FTT Collaborative Group 1994, Huynh et al. 2009). Therefore, only 10 to 20 patients with an STEMI need to be treated to prevent one death and far fewer need to be treated to
prevent rehospitalization or significant disability. The overall number of STEMI patients is relatively small compared to the latter two categories.

The next category is made up of patients who have a heart attack that does not involve the full thickness of the heart muscle; this is termed non-ST segment elevation myocardial infarction or NSTEMI. In the setting of an NSTEMI, 31 patients need to be treated to prevent one death (Fox et al. 2010). STEMI and NSTEMI are both acute situations where patients often show up to the hospital sick and unstable. However, the third category is composed of patients who are not having an acute heart attack but who have a stable blockage in one or more of their heart arteries, also known as “stable angina.” Patients with stable angina represent a significant percentage of those receiving stents in the United States and across OECD developed countries. Multiple large-scale clinical trials have demonstrated that in patients with stable angina, placing a stent does not reduce the risk of death or a heart attack. Therefore, an infinite number of patients with stable angina would need to be treated with a stent to save one life or to prevent one heart attack (Boden et al. 2006, De Bruyne et al. 2012). Weintraub et al. (2008) estimated that the cost per patient for a significant improvement in chest pain frequency, not prevention of death or a heart attack, was $154,580.

Figure 1 demonstrates how overall spending increases as coronary stenting is applied to subgroups of patients who derive progressively diminishing returns from it. The area of each box represents a hypothetical utility function that takes into account the number of eligible patients and the effectiveness of the intervention in that subgroup. A smaller area represents an intervention with higher utility because the cost would be low relative to the expected benefits. This is an excellent example of how the medical-care-cost-ratchet and Newhouse’s idea of technological change can be viewed as a continuum. In certain subgroups of patients (STEMI and to a lesser extent NSTEMI), coronary stenting increases costs but significantly improves cardiovascular outcomes, consistent with Newhouse’s technological change argument. However, in another large subgroup of patients with chronic blockages, coronary stenting has little effect if any on improving cardiovascular outcomes, consistent with our argument for the MCCR.
Response to Newhouse’s Criticisms

Newhouse (1992) specifically criticized the idea that too much insurance could facilitate technological change that was not welfare-enhancing. He argued that if consumers thought the cost of medicine did not justify its benefits, companies would provide policies offering coverage for outdated services at cheaper prices. But since insurance companies have not offered such policies, changes associated with increased costs must be welfare-enhancing. However, Newhouse himself recognized the flaw with that argument—companies would open themselves up to malpractice complaints for not providing the standard of care.

Newhouse also claimed that countries that make centralized decisions about how to allocate health care resources would not adopt changes to the standard of care that did not enhance welfare. But since they did adopt the same changes as the United States in many cases, and their cost growth was very similar, the changes must be welfare-enhancing. However, in cost-effectiveness analyses, services with...
incremental costs less than $50,000 per QALY are considered low-cost, $50,000 to $100,000 per QALY are considered intermediate-cost, and greater than $100,000 per QALY are considered high-cost (Fang, Minichiello, and Auerbach 2005). In most OECD countries, there are no formal rules to define the upper limits for what is acceptable. One recent study has estimated the upper bounds for a cost-effectiveness decision rule in the United States to be $297,000 per QALY (Braithwaite et al. 2008). In the United Kingdom, the National Institute for Clinical Effectiveness (NICE) has a stated range for suitable cost effectiveness between £20,000 and £30,000 per QALY. However, research on NICE decisions reveals the threshold to be considerably higher (Devlin and Parkin 2004). This criterion assures that services with progressively diminishing returns on investment will be adopted indefinitely. Therefore, the fact that countries with more socialized health care systems continue to experience cost growth similar to the United States does not invalidate the MCCR concept—but rather, affirms it.

Policy Implications

Today there seem to be two prevailing schools of thought on how to reduce health care costs and bend the cost curve. One school, represented by Emanuel et al. (2012) and the Center for American Progress, seeks a systematic approach to containing health care costs. This approach includes global targeting of payment rates, replacing the fee-for-service model of payment, and simplifying administrative systems for payers and providers. At its core, this school believes that it can reduce costs by reducing waste and inefficiencies. This school takes for granted that technological change is welfare enhancing and that the standard of care should be provided for free at the point of service. By seeking approaches that would make health delivery in the United States more like other countries with centrally planned delivery systems, this school fundamentally fails to address the underlying problem of health care cost growth. Their solutions would work to lower costs in the context of a one-period model, for example, by reducing regional disparity in health care service intensity that does not significantly improve patient-level outcomes. However, they would not address cost growth over time. Furthermore, because their approach strongly relies on government to regulate health care delivery, it is subject to the heavy influence of industry lobbyists.
Our theory of the MCCR recognizes that health care costs around the world have not been increasing at a sensational rate because of inefficiencies within the system. Rather, the rapid cost increase is the result of incorporating new technologies into the standard of care and disseminating them to the public. The MCCR provides a useful construct for understanding how health care spending increases with the incorporation of new technologies as the clinical benefits to patients and the public progressively decline. However, medical insurance that encourages individuals to economize on nonemergent health care decisions would lower costs over time without diminishing social welfare. Reform efforts should focus on rejuvenating market forces that have been systematically suppressed. The market-based school offers several approaches.

Consumer-directed health care (CDHC) as defined by Goodman (2006) is a potential solution to control health care cost growth. By carving out areas in which it is appropriate and desirable for individuals to self-insure or pay out of pocket, CDHC encourages the economization of nonemergent health care choices. It recognizes that scarce resources must be allocated among unlimited wants. As the MCCR demonstrates, the costs of providing the standard of care in many cases, far exceeds its social benefits. Empirical evidence exists that patients with CDHC plans would reduce health care spending without jeopardizing their health.

The RAND Health Insurance Experiment (HIE) is the gold standard when assessing the impact of level of insurance coverage on health. The study randomly assigned individuals to different levels of health insurance generosity and compared utilization and health outcomes across experimental groups. It found no significant effect of insurance generosity on various measures of health status for the average adult—despite increased use of medical services in patients with higher levels of coverage (Levy and Meltzer 2008). Buntin et al. (2006) reviewed studies assessing the effects of CDHC on cost and quality that have followed the HIE. Despite limited data, the authors conclude that “the early evidence, consistent with the HIE, suggests that higher deductibles reduce total health care use and spending” (Buntin et al. 2006: w523). In some studies, this reduction in spending was attributed to deferral of appropriate medical care like screening mammograms. Some opponents of CDHC have cited this as a weakness of such plans. However, based on the data presented
in this article, deferral of screening mammography could as easily be considered prudent as lamentable.

A properly constructed CDHC plan would, for example, cover the cost of stenting in a patient with an STEMI, but it would not cover the cost of elective stent placement for a patient with stable angina. The latter patient would have to pay for the stent with money from his own pocket or could draw from a health savings account. A CDHC plan could be structured in such a way that it tiered co-payments for elective services so that those services with lower NNT would have lower deductibles than those with higher NNT. If this were the case, the co-pay required of an elective coronary stent would be very high. Ulrich, Brock, and Ziskind (2003) reported that from 1987 to 2001 the rate of stenting increased 128 percent. Given the high cost of revascularization, the growth in procedure volume, especially for elective cases, was only possible because the cost to the patient was near zero.

Moving toward CDHC requires a reversal of the policies and programs that have suppressed market forces. Antos, Pauly, and Wilensky (2012) offer proposals for doing this in both the public and private insurance markets. A premium-support model would shift Medicare from a defined-benefit to a defined-contribution plan by providing a fixed subsidy for each beneficiary’s purchase of insurance. “Seniors would receive a uniform subsidy to purchase insurance from competing health plans (including traditional Medicare), with each offering at least a core set of benefits” (Antos, Pauly, and Wilensky 2012: 955). This is a reasonable approach, but it cannot work unless traditional Medicare is reformed in some way to encourage seniors to economize on health care. If traditional Medicare remains an option and is set as the low bid, there would be no incentive for seniors to pay for an alternative plan that approaches the CDHC model.

Antos, Pauly, and Wilensky (2012: 954) also propose that “the principle of defined contribution be applied to the currently unlimited tax subsidy for employer-sponsored insurance.” Under the current tax code, employers offer insurance to employees with pre-tax dollars that encourages the purchase of health insurance policies with minimal cost-sharing, which helps fuel cost growth. It is also structured in such a way that is particularly unfair to low- and medium-income employees because “shielding premium payments from income taxes
is worth more to employees in higher income-tax brackets” (Antos, Pauly, and Wilensky 2012: 957). The existing tax exclusion could be turned into a predetermined tax credit that is made available to anyone purchasing insurance, whether through an employer or on the individual market. A fixed subsidy would eliminate the bias of the tax exclusion toward more coverage and higher spending.

Other reform proposals exist, but they are beyond the scope of this discussion. Our goal in this article was to describe a model for health care cost growth that has not received adequate attention. We believe that an understanding of the medical care cost ratchet (MCCR) is important to guide reforms that will lower health care costs—that is, bend the health care cost curve—over time without diminishing social welfare.

References


