INFORMATION TECHNOLOGY AND MEDICAL MISSTEPS:
EVIDENCE FROM A RANDOMIZED TRIAL

Jonathan C. Javitt
James B. Rebitzer
Lonny Reisman

Information Technology and Medical Missteps: Evidence from a Randomized Trial
Jonathan C. Javitt, James B. Rebitzer, and Lonny Reisman
NBER Working Paper No. 13493
October 2007
JEL No. I12,M15

ABSTRACT

We analyze the effect of a decision support tool designed to help physicians detect and correct medical "missteps". The data comes from a randomized trial of the technology on a population of commercial HMO patients. The key findings are that the new information technology lowers average charges by 6% relative to the control group. This reduction in resource utilization was the result of reduced in-patient charges (and associated professional charges) for the most costly patients. The rate at which identified issues were resolved was generally higher in the study group than in the control group, suggesting the possibility of improvements in care quality along measured dimensions and enhanced diffusion of new protocols based on new clinical evidence.

Jonathan C. Javitt
Chairman and CEO, Health Directions
1700 Pennsylvania Ave.
Washington D.C.
jjavitt@healthdirections.net

James B. Rebitzer
Mannix Medical Mutual Professor
and Chair of Economics Dept.
Weatherhead School of Management
Case Western Reserve University
11119 Bellflower Rd
Cleveland, OH 44106-7235
and NBER
james.rebitzer@case.edu

Lonny Reisman
Chief Executive Officer
Active Health Management, Inc.
New York, NY
lreisman@activehealth.net

Introduction

In 1987, Nobel Laureate Robert Solow famously remarked, "you can see the computer age everywhere but in the productivity statistics."(Solow, 1987), p. 36). Solow's aphorism neatly summarized the state of knowledge in the late 1980's and early 1990s. Since that time, however, economists have been able to identify measurable economic effects of the revolution in information technology (IT). The emerging consensus from this research is that the effect of IT varies depending on the design of organizations and the nature of production processes. IT *complements* the work of people engaged in non-routine problem solving and communication while it *substitutes* for lower-skill tasks involving the sorts of explicit rules that are relatively easy to program into computers.[1]

Studying the effect of IT on work processes involving non-routine problem solving and communication is hard -- in large part because the inherent complexity of these processes make it difficult to identify meaningful performance measures that are also directly related to specific IT innovations. The search for good performance indicators and cleanly demarcated innovations has moved economists away from the analysis of aggregate productivity and technology data towards more narrowly focused studies. [2] The added institutional knowledge made possible by the limited scope of these studies also helps analysts address the selection problems created by the non-random distribution of new innovations across organizations and work places.[3]

---

[1]   For discussions of this perspective see (Autor, Levy and Murnane, 2003); (Brynjolfsson and Hitt, 2000), (Bresnahan, Brynjolfsson and Hitt, 2002); and (Levy and Murnane, 2004).

[2]   See for example (Athey and Stern, 2002) on IT and the delivery of emergency medical services; (Autor, Levy and Murnane, 2002) on banking; (Bartel, Ichniowski and Shaw, 2005) on computer controlled machines in manufacturing and (Hubbard, 2003) on capacity utilization in the trucking industry.

[3]   The econometric challenges involved in studying the effect of IT innovations closely parallel the issues

In this paper, we also analyze the effects of an IT enabled innovation in a narrowly defined production process characterized by non-routine problem solving and communication. The information technology we study is a decision support tool designed to notify physicians about potential medical "errors" as well as deviations from evidence based clinical practice guidelines. Our approach is closest in spirit to Athey and Stern's (Athey and Stern, 2002) study of emergency medical services. Like Athey and Stern, we focus on the introduction of a discrete innovation that altered the handling of information in a health care setting and we assess the efficacy of the innovation by tracking health related outcomes. Our econometric approach, however, differs from theirs in that we use a randomized controlled trial to identify the effect of the new technology.[4]

Although we focus on a specific production process, the results we report have broad implications for management and economic issues in health care. A large and influential body of research suggests that preventable medical errors have a substantial effect on the cost and quality of medical care.[5] In response to these findings, a number of high-profile public and private initiatives

---

involved in the study of innovations in human resource practices. For an illuminating discussion and review see (Ichniowski and Shaw, 2003) and for an application to the health care setting see (Gaynor, Rebitzer and Taylor, 2004).

[4]   (Athey and Stern, 2002) identify the effect of the technology in their study by comparing early and late adopters in a differences-in-differences framework. An obvious issue with this approach is that participants who choose to adopt early might be those for whom the benefits of the innovation are especially large. A randomized controlled trial eliminates this source of bias because the participants receiving the treatment are a random sample of the subject pool. Randomized trials have other limitations, however. The subject pools are often small and may not be representative of the underlying population. This can bias estimates of the effect of the intervention on a population. For a practical example of this sort of bias in a health care setting see (Duggan, 2005).

[5]   Evidence on the incidence of medical errors was recently reviewed by the Institute of Medicine of the National Academy of Sciences (Institute Of Medicine Committee On Quality Of Health Care In America, 2000). This report concluded that tens of thousands of Americans die each year as a result of medical errors during hospitalization. To date, very little is known about the incidence of errors in outpatient settings, but the incidence may be high (Lapetina and Armstrong, 2002). A recent study of errors in intensive care units is (Landrigan, Rothschild, Cronin, Kaushal, Burdick, Katz, Lilly, Stone,

have called for major new investments in information technology and decision support tools to reduce

the incidence of errors and increase compliance with evidence-based treatment guidelines (President's

Information Technology Advisory Committee, 2004), and (Institute Of Medicine Committee On

Quality Of Health Care In America, 2001).[6] Economists who have examined these issues generally

agree that new information technologies and decision support tools – perhaps combined with novel

incentive arrangements -will likely have a substantial influence on both errors and efficiency in the

delivery of health care, yet economic studies concerning the efficacy these interventions have been

scarce ((Newhouse, 2002)). [7]

The data in this study comes from a randomized trial of a physician decision support

technology introduced to a population of commercial HMO patients. We find that the intervention

reduced resource utilization: average charges were 6% *lower* in the study group than in the control

---

Lockley, Bates and Czeisler, 2004). A discussion of the literature on the use of IT to reduce errors and increase compliance with evidence-based guidelines can be found in (Institute Of Medicine Committee On Quality Of Health Care In America, 2001).

[6] As an indicator of the high level of public policy interest in these issues it is worth noting that reference to the use of IT to reduce errors appeared in President Bush's *Economic Report of the President* in 2004 (President's Council of Economic Advisors, 2004) and in his State of the Union Address. "By computerizing health records, we can avoid dangerous medical mistakes, reduce costs, and improve care."(cited in President's Information Technology Advisory Committee, 2004 p. 3). Private sector initiatives concerned with preventing medical errors have also been formed. The most prominent of these may be the Leapfrog Group, a coalition of more than 150 large public and private organizations that provide health care benefits. The purpose of this organization is to use employer purchasing power to speed the adoption of processes that improve patient safety ((The Leap Frog Group For Patient Safety, 2004)).

[7] There is a growing literature on disease management programs that often rely on information technology similar to that which we evaluate in this study. Disease management programs typically analyze billing records and other clinical information to identify patients whose care deviates from accepted clinical practice guidelines. Although disease management programs have become a ubiquitous part of health care and there is some evidence that they can be effective in reducing costs and improving quality (Shojania and Grimshaw, 2005; Gertler and Simcoe, 2004; Beaulieu et. al. 2007), many of the studies are poorly designed and few of them use evidence from randomized controlled trials of interventions (Shojania and Grimshaw, 2005).

group. These savings were the result of reduced in-patient charges (and associated professional charges) for the most costly patients.

The importance of IT based decision support systems for physicians extends beyond resource utilization: patients, providers, payers and policy-makers want to know whether this type of technology improves care quality. Decision support might improve quality if the system reminded physicians to do something beneficial for their patients that they had already intended to do but somehow forgot. Alternatively, decision support might improve quality if it provided useful new information to physicians in a form that was easy to incorporate into their daily practice and routines. This latter avenue of action is especially important because of the problem of physician information overload. In medicine, the number and variety of diseases and treatments and the rapid growth of new knowledge threaten to overwhelm the information processing capacities of individual doctors. Failure to keep abreast of this flood of information can cause physicians to overlook important new treatments or protocols that may improve care quality (Frank, 2004; Phelps, 2000; and Institute of Medicine 2000 and 2001).

Although the experiment was not designed to analyze the source of missteps or the mechanisms by which the technology influenced physicians, we can learn something about quality by comparing the rate at which identified issues were resolved in study and control groups.[8] Under plausible assumptions, a higher rate of resolution in the study relative to the control group can be interpreted as an improvement in care quality – at least along measured dimensions. Our findings generally point towards higher resolution rates in the study group, although measurement issues

---

[8] In the control group physicians did not receive messages about identified issues, but these issues were tracked and often successfully resolved without outside intervention.

discussed below require that we present this conclusion cautiously.   The increase in resolution rates was especially large for a new treatment protocol that emerged from the results of a widely publicized clinical trial in the year 2000, the year before our study began. Computer generated messages suggesting that a patient appeared to be a good candidate for the new protocol were triggered quite frequently in our study, suggesting that it takes some time for physicians to incorporate even widely promoted new protocols into their treatment of patients. More importantly we found that the resolution rates in the study group were double those in the control group. This result suggests that the IT system may have been more effective than conventional communication channels in disseminating new knowledge to physicians. We discuss the economic and behavioral implications of these results below.

The plan of the paper is as follows. Section two describes the setting of the trial and the decision support technology. Section three presents the data analysis. Section four concludes and discusses new research questions raised by the study.

The Setting

Physician Mistakes:

Physicians make mistakes – and these mistakes are increasingly believed to have a substantial effect on the cost and quality of medical care. The causes of errors are not entirely clear, but a leading suspect is the volume and complexity of the information that physicians must process about their patients' medical conditions and the rapidly changing state of medical knowledge (Bohmer, 1998); (Institute Of Medicine Committee On Quality Of Health Care In America, 2001); (Landrigan, Rothschild et al., 2004)).

If errors come from keeping abreast of "too much information", then it makes sense to look to information technology to help manage this burden. Ideally one would like to use IT to collect and

5

analyze patient information and to communicate likely missteps to physicians. In order for these

messages to be influential it is important that they be delivered in a timely fashion, be targeted to

specific patients, and that they reliably inform physicians of overlooked issues or issues about which

he or she lacked adequate knowledge. Generating these timely, targeted and informative messages is

hard -- especially because most physician practices are not linked by a common IT system. In the

absence of such linkages, it is difficult for most managed care organizations to construct usable

electronic medical records for patients treated within their physician networks. [9] Since managed care

organizations are the predominant form of private sector health insurance, the problems posed by

balkanized IT systems can be a significant barrier to bringing computer assisted decision making to

medical care.

The decision support software evaluated in this trial was designed to overcome the problems

posed by fragmented IT systems.[10] It collects information about patients from billing records, lab

feeds and pharmacies to assemble a virtual electronic medical record. It then passes this information

through a set of decision rules culled from the medical literature. When the software uncovers an

---

[9]   "…to be effective, CDSS [clinical decision support systems] diagnostic systems require detailed,
         patient-specific clinical information (history, physical results, medications, laboratory test results),
         which in most health care settings resides in a variety of paper and automated datasets that cannot easily
         be integrated. Past efforts to develop automated medical record systems have not been very successful
         because of the lack of common standards for coding data, the absence of a data network connecting the
         many health care organizations and clinicians involved in patient care, and a number of other factors."
         (Institute Of Medicine Committee On Quality Of Health Care In America, 2001)

[10]   The technology we analyze is the property of Active Health Management, Inc., which was acquired by
         Aetna in 2005 but continues to operate as a standalone business. It is important to note that two of the
         authors had a proprietary interest in the company at the time of the study. Dr. Reisman was, and
         continues to be, the CEO of Active. Dr. Javitt was a shareholder and had a consulting relationship with
         the company. Rebitzer, who conducted the econometric analysis upon which this paper is based, has
         never worked for Active Health and has no financial relationship with or proprietary interest in the
         company.

issue it produces a message, called a care consideration, that includes the patient's name, the issue discovered, a suggested course of corrective action and a cite to the relevant medical literature. The care considerations (CCs) fall into three distinct, but not entirely mutually exclusive, categories: stop a drug; do a test; and add a drug. The CCs are also coded into three severity levels. A level one message (the most severe) includes potentially life-threatening situations: for example that a patient's blood potassium levels are dangerously off. A level two (moderate) CC refers to issues that might have an important effect on clinical outcomes. An example might be that the patient is a good candidate for an ACE inhibitor, a drug that is useful in treating many cardiovascular conditions. A level three (least severe) CC refers to preventative care issues such as being sure that diabetics have regular eye exams.

Contrary to most other studies of medical errors, issues tracked by the software are not limited to events occurring during hospitalizations. This is important because, most of the evidence regarding medical errors comes from studying the treatment of hospitalized patients, but in-patient treatments are a declining share of medical treatment.

Study Design

The participants selected for the study came from an HMO located in a Midwestern city. All were all under age 65 and all had some medical charges in the year prior to the experiment. Once selected, the participants were randomly allocated into study and control groups. The software was turned on for patients in the study group. This means that their physicians received CC's during the year-long course of the experiment. The software was *not* turned on for patients in the control group until the experiment was over, but the billing, pharmacy and lab data of these patients was collected and saved. At the end of the year, the control group's medical data was analyzed to find CC's that *would* have appeared if the software had been running. An important feature of the study design was

7

that randomization occurred at the level of the patient. This means that some physicians had patients in both the study and the control group.[11] Thus lessons that physicians learned from receiving a CC for a study group patient might spillover to their control group patients. These spillovers could therefore have the effect of biasing the estimated impact of the decision support software downwards.

Computer Generated Messages:

When the software generated a CC, doctors employed by the software company manually checked it. If the CC passed this scrutiny and was coded a level one (most severe) the HMO's medical director was called and he, in turn, called the appropriate physician. If the CC was at level two (moderate) or level three (least severe), a nurse employed by the HMO received the error message and decided whether to fax the CC on to the enrollee's physician. The HMO's nurse passed on most (but not all) level two CC's and some of the level three CC's. Unfortunately there are no records documenting the nurse's decisions concerning which messages to pass on. From informal discussions, however, it appears that some types of CCs were not sent because they duplicated advice in disease management programs already in place at the HMO. These were mostly level three CC's focused on preventative care. In addition, the nurse decided that some other CCs didn't make clinical sense.

---

[11] The initial randomization included 49988 members with at least 1378 distinct primary care providers (for 567 individuals, no PCP was identified at the time of the randomization). The vast majority of members were treated by physicians having patients in both the study and control groups. Of the patients with identifiable PCPs, only 424 patients (0.8 percent of the total) were treated by PCPs having only study or control group members.

<u>Outcomes:</u>

The data collected in this study yields two natural performance measures: the rate at which problems identified by CC's are resolved; and the average costs of medical care. We discuss each of these in turn.

Resolution Rates: Physicians often have better information about their patients than does the error detecting software. Actions that look like a misstep to the computer may in fact be the result of informed physician choice, informed patient choice and/or patient non-compliance. For this reason, the HMO and the software company viewed CCs as recommendations that physicians were free to ignore if they disagreed.[12] In addition, some issues identified by the software would have been resolved even if no messages had been sent to physicians. Given these ambiguities, how should we interpret differences in resolution rates between study and control groups?

In answering this question it is helpful to consider a simple example. Imagine there is a clinical action recommended by the computer that a physician can either take or not take and that this action can have an effect on the patient that is either positive or not. The probability that a physician in the study group takes an action that benefits a patient is $P(A=1|B=1,S=1)P(B=1)$ where: $A = 1$ if the action is taken and zero otherwise; $B=1$ if the recommendation in the CC is indeed beneficial to the patient and 0 otherwise; $S = 1$ if the individual is in the study group and 0 if in the control; and $P(\ )$ represents probabilities. With this notation we write the probability that the physician will take the action and the patient will *not* benefit as $P(A=1|B=0,S=1)(1-P(B=1))$. The corresponding probabilities

---

[12] Informal conversations with physicians who attended a discussion group about the software indicated that they also viewed the care considerations as suggestions that they could disregard.

for the control group are P(A=1|B=1,S=0)P(B=1) and P(A=1|B=0,S=0)(1-P(B=1)) respectively.

Resolution rates will appear higher in the study than the control group when

(1)   P(A=1|B=1,S=1)P(B=1) + P(A=1|B=0,S=1)(1-P(B=1)) >

        P(A=1|B=1,S=0)P(B=1) + P(A=1|B=0,S=0)(1-P(B=1))

Regrouping we can rewrite (1) as

(2)   [P(A=1|B=1,S=1) - P(A=1|B=1,S=0)]P(B=1) >

        [P(A=1|B=0,S=0)- P(A=1|B=0,S=1)](1-P(B=1))

From expression (2) it is clear that if the care considerations are always right so that P(B=1) = 1, then

if more resolutions to care considerations are observed in the study than the control group the quality

of care in the study group must be higher than the control group. Conversely if the care considerations

do not benefit patients at all, P(B=0) = 1, then we can make no inferences about quality from

differences in rates of resolution. Given the state of the software (and the state of medical science) it

is reasonable to assume that 0 < P(B=1) < 1. In this case it is possible, but not certain, that the

messages delivered contain useful medical information. It follows that quality improves in the study

group when:

(3)   [P(A=1|B=1,S=1) - P(A=1|B=1,S=0)]P(B=1) >

        [P(A=0|B=0,S=0)- P(A=0|B=0,S=1)](1-P(B=1))

    Subtracting (3) from (2) it is easy to derive sufficient conditions under which resolution rates and

quality can both increase in the study group when 0 < P(B=1) < 1:

(4) [P(A=1|B=1,S=1) - P(A=1|B=1,S=0)]P(B=1) > 0; and P(A=0|B=0,S=1)- P(A=0|B=0,S=0) > 0.

The two conditions in (4) are intuitive. The first states that CCs contain enough useful information to

persuade physicians to take actions that offer benefit to their patients. The second states that physician

judgment is sufficiently good that computer messages will not persuade doctors to take actions when no action is in the best interest of their patient.[13] If these conditions hold, then higher resolution rates in the study group are an indicator of improved care quality – at least along the dimensions measured by the computer system.

Average Charges: The CC's issued by the system recommended roughly three types of actions: "add a drug", "stop a drug" and "do a test". The first and the third of these entail a direct increase in the utilization of medical resources. If, however, these actions prevent subsequent costly complications, the net effect might be to reduce charges relative to the control group.

Administrative data from the HMO was collected on average charges per member per month in the year prior to the study and also during the year of the study. Charges are "list prices". They are the prices that providers bill for services, but that are rarely used in actual transactions because most purchasers negotiate discounts. For this reason charges should not be interpreted as a price in the usual economic sense of the word. Rather we use average charges to track movements in aggregate resource utilization. Assuming that similar charges are applied to similar services in the study and control group, charges can be a useful dollar index of resource utilization. In the context of our study, the assumption of a common charge schedule in study and control groups is reasonable. The participants in our study were taken from a single HMO and more than 99% of them have primary care providers with patients in both the study and control group. While individual PCPs and medical

---

[13] Of course these conditions are not relevant if physicians simply viewed the CCs as "orders" that they had to follow or as messages with an inherent legitimacy. As noted above, conversations with physicians involved in the study indicate that participating doctors viewed CCs as suggestions that they were free to ignore. In the long-run, if the decision support technology proved to be sufficiently reliable, physicians may substitute it's advice for their own reading or discussion with peers. This will have interesting effects on the diffusion of new knowledge. See Rebitzer, Rege and Shepherd (2007) for an analysis.

practices may have a different schedule of charges for different HMOs, they use the same schedule

when charging a single HMO. Thus differences in physician charges between study and control

groups should be largely eliminated by the random assignment of patients to study and control groups.

This is especially true in fixed effects specifications because, if patients remained with their PCP over

the course of the study, the effect of cross PCP's variation in the charge schedule will be largely

absorbed in the "fixed effect". Most of the hospital admissions in this HMO were to a single hospital,

so it is also unlikely that differentials in average charges for in-patient services and for hospital related

professional services were the result of differences in the schedule of charges between the study and

the control group.

In addition to charges, we also have data on reimbursements. Reimbursements are the amount

insurers actually pay to providers for specific services and they typically are below charges. Data

from the *Medical Expenditures Survey* indicates that reimbursements differ from charges and that

these differences vary by service and also have been increasing over time (Slesnick and Wendling,

2006). This would suggest that charges may be a poor proxy for reimbursements – especially if

studying variation over long periods of time and diverse geographic settings. Reimbursements might

be preferred to charges as a measure of resource utilization because they are used to settle actual

transactions. In our setting, however, reimbursements are less informative about resource use than

charges. The reason for this has to do with the accounting practices used in the HMO we study.

Specifically the hospital that accounted for the bulk of in-patient activity *charged* the HMO for

resources used during a stay, but the HMO *reimbursed* the hospital a set amount charge per-diem.

This means that a high daily use of in-patient resources would be reflected in high in-patient charges

but not in in-patient reimbursements. This per-diem arrangement did not apply to in-patient related professional services. [14]

It is also worth noting that the study is designed to assess only the short-run effects of the intervention on resource utilization. Many of the benefits of avoiding missteps may appear years after the error occurred. Given the high rate of turnover among HMO patients, however, much of the benefits to individual insurers of reduced resource utilization may be best captured by short-run savings identified in this study. [15]

Data and Results

Descriptive Statistics

Table 1 presents descriptive statistics. The analysis excludes enrollees younger than age 11 because the decision support software had very few pediatric treatment guidelines in place at the time of the study.   The number of individuals in the study and control groups older than age 12 in the year 2001 was 19716 and 19792 respectively. There is some attrition from the study in the year 2001,

---

[14]   More information about the design as well as a preliminary analysis of results can be found in (Javitt, Steinberg, Couch, Locke, Jacques, Juster and Reisman, 2005). This study differs from the earlier one in a number of important respects. First, we use an additional two years of data to analyze the effect of the intervention on resource utilization. This allows us to run a "reverse experiment" that corroborates and extends the findings of the initial experiment. Our analysis of resource utilization also includes a discussion of the ways that the billing practices of the HMO influence measures of resource utilization based on average charges and average reimbursements. Second, we investigate the effect of the intervention on different quantiles of the cost distribution as well as on sub-populations whose pre-experiment characteristics give them a high-propensity to generate care considerations. Third, we offer a much more extensive and detailed analysis of the effect of the intervention on resolution rates including: a discussion of potential measurement error in the identification of CCs in the control group; a clear statement of the relationship of resolution rates to care quality; and a consideration of the behavioral implications of the observed pattern of resolution rates.

[15]   Cebul et. al. (2007) find insurance turnover rates in excess of 30% per year. In their study, Gertler and Simco (2004) report that a disease management program for diabetes reduced costs by about 8 percent in the first twelve months and larger cost savings were realized in subsequent quarters.

mostly because of the change of insurers that takes place at the beginning of each calendar/contract year. In both the study and control groups, roughly 72% of respondents stayed in the sample for all 12 months.

Who Gets CCs?

Table 2 examines who in the study group gets CCs, how many they get, and of what severity. Column (1) is a probit with a dependent variable equal to one if the member received any CC's at all. It is estimated for members of the study group because actual CC's were generated only in the study group. The first set of right hand side variables code for age. The omitted group consists of those between age 12 and 20 and the coefficients are expressed as derivatives. Thus a participant aged 20 to 30 is 1.1 percentage points more likely to receive a CC than a participant in the omitted group. The likelihood of generating a CC increases with each subsequent decade and peaks for the oldest group. HMO members aged 60-65 are 31 percentage points more likely to generate a CC than those in the omitted group. Women are 0.6 percentage points less likely to generate a CC than are men, a difference of roughly 12% from the mean. This gender differential probably is due to two factors: first, the software program did not have codes for many obstetric and gynecological issues and secondly, cardiac and other issues that were well represented in the software often manifest themselves a decade later in life for women than for men. Since this study focused on a commercial insurance population less than 65 years old, this decade delay in onset would reduce the number of CCs generated for women. Finally, participants with higher levels of charges in the year 2000 are more likely to generate CCs in the year 2001. The effect, while statistically significant, is also small. Moving from zero charges in 2000 to the mean level of $280 increases the odds of generating a CC by 0.003, an increase of 6% above the mean incidence of CCs.

Taken together, the results in column (1) of Table 2 indicate that older, male patients with high medical charges in the previous year are more likely to have errors. The findings are consistent with the notion that care complexity is an important determinant of physician missteps. As bodies age, more things are likely to go wrong -- leading to more treatment and also more opportunities for lapses. Similarly, the more charges a patient generates, the greater the medical activity undertaken on their behalf. Managing these activities creates additional opportunities for missteps.

The models estimated in columns (2) and (3) of Table 2 redo the analysis focusing respectively on the number of CCs received (a negative binomial model) and the severity of the most severe CC received (an ordered probit model). In both cases we find that older patients, male patients and patients with more charges prior to the experiment are likely to generate more, and more severe, CCs. These results are both statistically significant and large in magnitude. They further underscore the likely role that complexity of care plays in generating errors.

Charge Differentials

Table 3 analyzes the effect of the intervention on average charges per member per month. We adopt an "intention to treat" approach and compare the average charges in the study group and the treatment group. There are many possible treatment mechanisms in this study and it is hard to appropriately identify them all. As discussed above, we observe CCs generated by the software and approved by MD's working for the software company, but the HMO's nurse passed only a subset of these along to the treating physician. Similarly it is hard to know if the effect of the intervention was due to the information content of the particular CC or simply the fact that a physician received a CC at all. The "intention to treat" approach allows us to be agnostic about the mechanisms of action.

Column (1) in Table 3 estimates a fixed effects model of the determinants of total charges (pmpm) between the study group and the control group. The variable *Year = 2001* is an indicator variable that is equal to one in 2001 and 0 in the year 2000. The coefficient 69.099 means that average charges rose from the pre-study year to the study year by $69.10 pmpm. This increase is driven by two factors: the growth in charges from one year to the next and also the increase in medical services delivered as individuals become one year older. The key variable that identifies the average treatment effect is *Study*Year = 2001*. The coefficient reported implies that the increase in average charges from the pre-study year to the study year was $21.92 less in the study group than in the control group. Thus the intervention reduced the average of total charges in the study group by 6.1% of the average $352 pmpm control group charges. This difference is both economically significant and statistically significant at the 5% level.

Columns (2)-(5) of Table 3 present estimates of the average treatment effect for the components of total charges. These are in-patient charges (charges incurred during hospitalization); out-patient charges, prescription (Rx) charges, and professional charges (charges resulting from professional services such as radiology). Focusing attention on the key variable, *Study*Year= 2001*, in-patient charges are reduced by $12.83 ($t = 1.8$) in the study group relative to the control. This accounts for 58% of the total cost differential. In contrast, out-patient and Rx charge differentials are quite small and statistically insignificant, -$1.82 ($t = 0.60$) and $0.70 ($t=0.90$) respectively. Professional charges, however, are $7.96 ($t= 2.21$) smaller in the study group. The final column of Table (3) estimates a linear probability model of the determinants of hospitalization. The dependent variable is equal to one if the participant had ever been hospitalized and is 0 otherwise. The coefficient on *Study*Year = 2001* is very small and statistically insignificant, -0.002 ($t=0.83$). Since hospitalization is expensive, this suggests that the reduction in in-patient costs observed in column (2)

16

is likely due to reduced resource use when hospitalized or due to a reduction in the number of

hospitalizations in the year. This makes sense, as the likely effect of some of the CCs is to prevent re-

hospitalization or to reduce resource utilization per hospital visit – perhaps by improving the health of

individuals who might become hospitalized.[16]

Further support for this interpretation of Table 3 emerges from an analysis of average

reimbursement differentials between study and control groups. In fixed effects estimates identical to

those presented in Table 3, we find that the coefficient on *Study*Year=1* is -$8.96 (t=1.60). This is a

six percent drop in charges. Fifty one percent of this difference is due to the $4.62 (t= 2.39)

differential in professional reimbursements while forty three percent is due to the $3.88 (t=1.15)

differential in average in-patient charges. Comparing these results to Table 3 we see that in-patient

savings account for less of the total treatment effect and are more imprecisely measured when we use

average reimbursements rather than average charges to measure resources used. The fact that the

effect on in-patient reimbursements was smaller and much more imprecisely measured is not

surprising given what we know about the HMO's accounting practices (discussed in section two

above). The hospital that the HMO relied on for the bulk of in-patient services *charged* the HMO for

all services provided but was *reimbursed* on a fixed per-diem basis. If patients in the study group used

on average fewer resources per day than the control group counterparts, they would appear to have

lower average charges but not lower average reimbursements. In contrast, professional charges and

reimbursements were both based on services provided and we find that the intervention reduced these

---

[16]   Examples of issues reported as CCs that were likely to prevent re-hospitalization include inadequate use of ACE inhibitors or beta-blockers for patients with myocardial infarctions or congestive heart failure.

significantly, regardless of whether reimbursements or charges were used as the measure of resource utilization.

The results in Table 3 imply that in-patient and professional charges account for 95% of the average charge differential between study and control groups. In-patient charges arise from the use of hospital resources, but professional charges include services that can be delivered in either an inpatient or outpatient setting. Our findings suggest that the experiment did not reduce all professional charges, but only those associated with hospitalization. Taking the coefficient on *Study*Year =2001* in column (2) and dividing it by the analogous coefficient in column (5) we get 1.62. Thus every dollar decrease in professional charges that is due to being in the study group is associated with a $1.66 reduction in in-patient charges. Similar calculations using results from column (3) of Table 3 suggest that a dollar reduction in professional charges resulting from the experimental intervention is associated with a drop in outpatient charges of only $0.23.

The results in Table 3 indicate that the reduction in total charges is largely driven by inpatient costs and associated professional charges. Since hospitalization is expensive, this suggests that savings generated in the experiment are the result of reduced resource use for high-cost participants. Table 4 examines this directly through the use of quantile regressions. Column (1) of Table 4 is a median regression. The variable of interest is once again *Study*Year=2001*. The coefficient of 0.541 (t=0.20) suggests that the median participants in the study and control groups had virtually identical total charges (pmpm). The corresponding coefficients in Column (2) suggest charge differentials are $26.51 ($t = 1.06$) at the 90th percentile and $658.61 ($t=1.85$) at the 99th percentile. Clearly the intervention is having its effect at the far right tail of the distribution of costs.

The finding that the "action" generated by the intervention primarily affects high-use patients raises concerns about the appropriate way to calculate the t-statistics reported in Table 3. Specifically

one might be concerned about the assumption that the error term in our cost equations is normally distributed. For this reason we recalculated the t-statistics in Table 3 using a bootstrap technique that makes no assumptions about the functional form of the errors. As reported in the notes to Table 3, we find that bootstrapped standard errors are quite close to the conventional standard errors generated by fixed effect estimators.

If the experimental intervention is the primary cause of the differences between study and control groups, one might expect the cost savings to be greatest for individuals with the highest propensity to generate CCs. Table 5 examines this issue by restricting our estimates to those individuals whose pre-study characteristics put them most at risk for producing an error message. Columns (1) through (5) of Table 5 examine charges for individuals who were over age 50 in the year prior to the experiment. This sample is of particular interest because the results in Table 2 indicate that this age group is most likely to generate error messages. In this table we use only data from the study year, so the key variable of interest is *Study*.[17] We find that the average total charges are $72.17 less in the study group than the control and this difference is rather precisely measured. As we observed in Table 3, much of this differential is the result of differences in in-patient charges (study group members had in-patient charges that are nearly $50.00 lower than the control group with a t-statistic of 2.12.). Out-patient charges and Rx charges are not significantly different between the study and control group although the point estimate of magnitude of the out-patient differential ($9.59) is sizeable. Professional charges are, as before, estimated to be the second leading contributor to the differential between the study and control group differential, but the size of the professional charges

---

[17]   As we observe in the notes to Table 3, point estimates of the average treatment effect were very close whether we used a fixed effects model or restricted our attention to the study year differentials.

differential was imprecisely measured and we cannot reject the hypothesis that the true effect was zero.

Restricting our sample to individuals over age 50 causes us to discard much of our data. An alternative approach, which we present in columns (6) through (10) is to use the "CC" equation in Table 2 to estimate each individual's propensity for receiving a CC based in pre-study characteristics. We can then weight each observation by its propensity score and, in this way, observe how our estimates change when greater weight is given to individuals likely – on the basis of pre-experiment characteristics -- to be exposed to a care consideration.[18]

In column (6) of Table 5 the coefficient on *Study* is -66.36, a cost differential in between study and control groups that is roughly three times the estimate in Table 3. Interestingly this number is not too far from the coefficient on *Study* in column (1). [19] This result confirms the impression from column (1) that cost savings from the study are greatest for individuals with a high propensity to receive a computer generated message. The weighted estimates for in-patient charges (column (7)) and professional charges (column (10)) follow a similar pattern. The coefficients on the variable *Study* in each of these cases is negative, statistically significant, and roughly three times the size of the effect observed in the un-weighted regressions in Table 3. The results on out-patient and Rx charges follow the now familiar pattern: the coefficient on variable *Study* is small in magnitude and imprecisely

---

[18]   We set the analytical weight in STATA's regression command equal to the observation's propensity score. This weighting scheme assumes that observations with high propensities to generate CCs are measured more accurately than observations with low propensities. Specifically, analytical weights are assumed to be inversely proportional to the variance of an observation; i.e., the variance of the j-th observation is assumed to be $\dfrac{\sigma^2}{w_j}$ where the weights, w_j, are rescaled to sum to the number of observations in the data.

[19]   The ninety five percent confidence interval is between -104.14 and -28.58.

measured. One cannot reject the hypothesis that the true effect of the intervention on outpatient and Rx charges is zero.

The overall impression from Table 5 is that the magnitude of the reductions in resource utilization in the study group are largest for those participants whose pre-study characteristics mark them as likely to be directly effected by the intervention.

Reverse Experiments:

The experiment was concluded at the end of December 2001, but the system was kept in place for study group members until the end of February 2002. At that time the entire software system was turned off. In June 2002 the software was started up again and CCs were sent to all HMO enrollees, including those in the original study and control groups. The general rollout of the system makes possible an additional test of the system's effects on costs. If the reduction in charges observed in the study group was indeed the result of the intervention, one should expect charges in the two groups to converge when the controls began receiving CCs.

Table 6 compares charges in the study and control groups in the two years following the end of the experiment. Panel A of Table 6 analyzes cost data from calendar 2002. The coefficient on *Study* in column (1) indicates that average total charges in the study group were about $8.58 lower in the study than the control group. This difference is about 40% of that observed in the year of the experiment and it is imprecisely measured ($t = 0.78$) and not statistically distinct from zero at conventional significance levels. The corresponding coefficients for inpatient, outpatient, prescriptions and professional charges are presented in columns (2)-(4) respectively. They are similarly small, imprecisely measured, and not different from zero at conventional significance levels. Column (6) is a probit where the dependent variable is 1 if a patient was ever hospitalized in the year

and 0 otherwise. The probability of any hospitalization in 2002 was 0.5 percentage points lower in the

study group than the control group (z=2.27). Column (7) is a quantile regression comparing the study

and control groups at the 99th percentile. The coefficient on *Study* is –238.91, slightly more than a

third of the analogous coefficient in Table 4 and imprecisely measured (*t*=1.07). Panel B compares

the remaining members of the study and control groups in the year 2003, two full years after the

experiment. We find no statistically significant difference in charges between the two groups in any

component of costs. Taken together, the absence of cost differentials in years when the intervention

was rolled out to both treatment and control groups supports the conclusion that the cost differentials

observed during the study year were the result of the intervention. These findings also suggest that the

effect that the intervention had on our dollar index of utilization was fast-acting (appearing in the first

year of the study) and also quickly dissipated.

Resolution Rates:

In the decision support system we study, messages are aimed at physicians. The charges data

we analyzed above are a useful index of utilization, but charges are far removed from the actions

physicians may take in response to messages. We can move closer to observing physician behavior by

examining the rate of resolution of issues identified by the computer in the study and the control

group.

The recommendations issued by the software fell into three, not-quite mutually exclusive,

categories: "add a drug", "do a test", and "stop a drug". To identify compliance with an "add a drug"

recommendation, the computer scanned pharmacy records following the recommendation. If a

prescription for the indicated drug was filled, the issue was declared resolved. Similarly, billing

records were scanned following a "do a test" recommendation. If bills for the suggested test were

sent, the recommendation was also declared to be resolved. Calculating resolution rates for the "stop a drug" recommendations was more problematic than for the other two categories of suggestions. Individuals might have months-long supplies of the drug at home and the records only tell us that no new prescriptions for the drugs were filled. To identify compliance with "stop a drug" recommendations, pharmacy records were scanned for 60-150 days after the CC was transmitted, and the issue was declared resolved if no new scripts for the indicated medication were filled in that time.

Table 7 presents descriptive statistics regarding care considerations. The number of CC's issued was 1299 in the study group and 1519 in the control group. The CC differential between the study and the control group is statistically significant and therefore unlikely to be a purely statistical artifact.[20] If random chance is unlikely to generate this differential, we need to consider what other causes might be and the consequences these might have for estimating resolution rate differentials. [21]

Why Were Extra CC's Triggered in the Control Group?

We have identified four potential causes for excess CC's in the control group. First, it is possible that the randomization didn't work in the sense that a larger proportion of expensive patients with characteristics that trigger CCs ended up in the control group than the study group. In Table 3, however, we report that the average cost differential between study and control groups is almost completely unchanged by the presence or absence of fixed, participant effects. This suggests that the

---

[20]   The fraction of individuals receiving a CC is 20% higher in the control than the study group. A probit regressing the variable *Any CC Generated* against the variable *Study* finds that the difference between the study and control group is statistically significant at the 1% level.

[21]   The estimates of the propensity to generate CCs that we used in Table 5 used only the actual CCs issued to the study group and hence were not influenced by the process of CC generation in the control group.

randomization "worked" in the sense that inherently expensive patients were evenly distributed between the study and control groups.

A second possible cause of excess CCs in the control group might be that the decision support tool was triggering extra scrutiny of patient files and therefore preventing subsequent CCs for patients who received any CCs. The evidence is, however, inconsistent with this hypothesis. In Table 7 we see that 80.96 percent of the control group CCs were issued to participants with only one CC compared to 76.87 percent in the study group.

A third explanation for excess CCs arises from the way that the CCs were identified in the control group after the end of the experiment. The software company saved all the data generated by the controls and ran this data through their software in early 2002. As was true in the study group, a committee of three physicians examined the resulting CCs and sent along those that made sense to them. The clinical thinking of the committee likely evolved over the year 2001, thus the real-time decisions they made over the course the experiment might have been different than the ones they made when evaluating the control group data after the experiment ended in early 2002. If the committee of physicians approved a larger number of CCs over time as they gained confidence in the computer system, then it would be reasonable to expect that more CC's would have been approved for the control group.[22]

The fourth possibility is one that we believe to be the most important. During the course of the study, the computer system generated messages based on information available at the time the

---

[22] We have some reason to believe, however, that increasing confidence in the system by the committee is not the cause of the differential we observe. In analysis of a subsequent randomized trial on Medicare patients, we found excess CCs in the study group rather than the controls – not the pattern one would expect if the differential was driven by a continued rise in confidence in the software.

program was run- roughly every week. In contrast, the control group CCs were generated using

information accumulated over the entire year of the experiment. This subtle difference in the handling

of information appears to be sufficient to generate significant differences in the number of CCs in the

study relative to the control group.[23]

Of the 101 distinct care considerations that the software could generate, we observe 90 types

issued in the study group and 83 issued in the control group. Our resolution rate estimates will be

biased if the mix of CCs in the control group was more or less likely to resolve spontaneously than the

mix of CCs in the study group. Unfortunately we have no way of assessing the spontaneous resolution

rates of different CCs and hence we have no way of assessing the sign or magnitude of potential

biases. One might suppose that very serious errors are more likely to be picked up by care providers

because of other pre-existing safety measures and this might provide some hint about the direction of

biases. The breakdown of CCs by severity level in Table 7 suggest that there are more severe CCs in

the study than in the control group, but the number of such serious issues are very small and are

therefore unlikely to be driving the results.

We can control for some of the potential differences in the mix of CCs between study and

control groups by examining resolution rate differentials by type of CC. The results of this

comparison can be found in the probits presented in Table 8. In equation (1) we observe that

resolution rate for "Add a Drug" CCs is 8.6 percentage points higher in the study group than the

---

[23]   In the second randomized trial mentioned in the preceding footnote, we were able to eliminate the CC
differential between study and control groups by forcing the computer system to conduct an evaluation
of patients using data generated at the same moment in time. It appears that the triggering of some CCs
is quite sensitive to small changes in the window of data available to the computer. For technical
reasons, it is not possible to construct such simulated CCs for this experiment. The analysis of this
second randomized trial is currently in process.

control group. This is a 48% improvement over the control group. In column (2) we observe that

resolution rates for "Do a Test" CC's are 5.8% higher in the study group, an improvement of 19% over

the control group. We do *not*, however, observe higher resolution rates in the study group for "Stop a

Drug" CCs. The coefficient on *Study* in column (3) has the wrong sign, but is also imprecisely

measured.[24]

We do not know why there appears to be no effect of CCs on "stop a drug" resolution rates and

substantial effects on "add a drug" and "take a test" resolution rates. One plausible explanation is that

many of the drug-drug inter-actions that trigger "stop a drug" CCs are also caught by pharmaceutical

databases used by major pharmacies. If this were true, the computer system would not be providing

much additional information about "stop a drug" issues to the study group.

Looking more closely at Table 8, we observe that eighty three percent of excess CCs are "stop

a drug" and "do a test" CCs. These CCs also had relatively high resolution rates in the control group,

suggesting that excess CCs tended to appear where there was a high likelihood that issues would

resolve spontaneously and without intervention. This would tend to bias downwards our estimate of

resolution rate differentials between study and control groups.

"Add a Drug" messages accounted for only seventeen percent of excess CCs and the most

common CC in this class appears to have been relatively unaffected by differences in the handling of

data in the study and control groups. This CC was sent out if a patient was a good candidate for taking

ACE inhibitors on the basis of protocols emerging from the Heart Outcomes Prevention Evaluation

---

[24]   In evaluating these resolution rate differentials, it is important to observe that it is hardly automatic that
physicians respond to interventions in a positive way. (Shojania and Grimshaw, 2005) report in their
assessment of quality improvement studies that interventions targeting provider behavior typically
produce only modest improvements in compliance with care guidelines and the variation in results
across studies is often large.

(HOPE) trials. The HOPE trial results were published in 2000, shortly before our experiment began in 2001 and the resulting treatment protocols were widely publicized – so much so that they were even included in the disease management guidelines of the HMO.[25] Of the 601 individuals with an "add a drug CC" in Table 1, 311 received the recommendation to use an ACE inhibitor on the basis of the HOPE trial. The distribution of these individuals was nearly perfectly balanced between study (155) and control (156) group members. We found that the resolution rate (the rate at which patients identified as good candidates on the basis of the HOPE trial began using ACE inhibitors) was 0.27 in the study group compared to 0.14 in the control group [26]. The resolution rates for the remaining individuals with "add a drug" CCs was 0.26 with no significant difference between the study and the control group. In short, the experimental treatment improved the rate of resolution of this issue roughly two-fold and brought resolution rates in line with those prevailing for other, "add a drug" CCs.

Models of physician learning in complex environments have typically assumed that physicians give disproportionate weight to easy-to-access "local" sources of information such as peers, teachers or even accumulated observations from the physician's own experience (Frank, 2004; Phelps, 2000; Rebitzer, Rege, Shepherd, 2007). This assumption implies that the diffusion of new information that is not "local" will be relatively slow and that inefficient, geographically specific, practice styles are likely to emerge (Phelps, 2000; Rebitzer, Rege and Shepherd, 2007). From the perspective of learning

---

[25] "The HOPE trial, published in 2000, looked specifically at high risk patients. These patients were defined as being either: diabetics with one additional risk factor (i.e. smoking, high blood pressure, high cholesterol) or patients with known CAD as demonstrated by remote MI, need for cardiac surgery or cardiac stenting, or angiographic evidence of narrowed coronary arteries. ACE inhibitors were shown to reduce the rate of cardiovascular death, MI, stroke, and new cases of diabetes. The results were so striking that the study was terminated early" (Shepherd, p. 6).

[26] This difference in resolution rates was statistically significant with a z score of 2.82.

27

models, it is perhaps not surprising that a CC based on a newly implemented protocol would be a commonly issued CC – it takes time for information distributed in medical journals and conferences to be incorporated into "local information" and hence practice styles. Our results, however, suggest something more: the messages in the CCs seem to have had a bigger effect on physicians than the conventional medical channels used to promote the HOPE trial findings. It seems plausible that the CCs were influential because they linked a general recommendation ("take ACE inhibitors") to a specific patient and a specific cite to the medical literature. If this explanation proves to be correct, then computer based decision support may be a way to make new knowledge as easy-to-access as "local knowledge" (Rebitzer, Rege and Shepherd, 2007).   The wide-spread implementation of computer based IT systems might plausibly enhance the diffusion of new knowledge and also break down the small area practice variations that stubbornly persist in the face of new evidence-based medical findings (Skinner and Staiger, 2005; Skinner, Fisher and Wennberg, 2001).[27]

Conclusions

This paper analyses the effect of a decision support tool designed to help physicians detect and correct medical "errors". Prior research suggests that physician missteps have a substantial effect on the cost and quality of medical care, and a number of high-profile public and private initiatives are premised on the notion that new information technologies can reduce the incidence of errors. Economic studies of the efficacy of these technological fixes have, however, been scarce.

---

[27]   While it is plausible that the program we analyze ameliorates some of the effects of information overload, our experiment was not designed to test this hypothesis.

The data in this study comes from a randomized trial of the new technology in a population of commercial HMO patients. We find that average charges were 6% *lower* in the study group than in the control group. These reductions in resource utilization were the result of reduced in-patient charges (and associated professional charges) for the most costly patients. Patients with the greatest propensity to trigger the computer messages were the most expensive and also the most likely to experience a reduction in charges over the course of the experiment. Consistent with these results, we also observed that when the experiment was ended and the computer system was rolled out to all HMO members, the cost differential between the study and control group rapidly disappeared. This reverse experiment also suggests that the effect of the intervention on resource utilization is quite rapid.

We also found that the rate at which identified issues were resolved was generally higher in the study group than in the control group – especially so for a set of messages promoting a new ACE inhibitor protocol resulting from a high profile clinical trial completed shortly before our study began. These resolution rate differentials must be interpreted cautiously because of a number of difficult measurement issues. They do offer suggestive evidence that the intervention may have improved care quality along measured dimensions. To the extent that the intervention stimulated physicians in the study group to adopt protocols that, like the ACE Inhibitor protocol, were also being widely promoted through more conventional channels, our results suggest that IT based decision support software may offer new and more effective ways of communicating new clinical knowledge to physicians.

We conclude by assessing the limitations of the experiment and suggesting avenues for future research. One important limitation is the study's short duration. To the extent that some of the benefits of correcting missteps spill over into future years, our analysis *understates* the saving due to the physician decision support system. A similar bias may result from physicians having patients in

both the treatment and control groups. If lessons learned from patients in the study group spill over to the treatment of patients in the control group, our estimates of the intervention's effect will be further understated.

A second limitation is that the study was conducted on a commercial insurance population where everyone was less than sixty-five years old. Since the likelihood of errors increases dramatically with age, much of the impact of this new technology will be found in older age groups not included in this study. In future work we will analyze a similar trial conducted for Medicare populations aged sixty-five and older.

A final issue left unresolved by this study is the mechanism by which care considerations influence outcomes. Specifically, the analysis does not identify the lessons physicians learned from the messages they received. It is possible that physicians learned only that the patient named in the care consideration required additional attention. Alternatively, it may be that the specificity of the messages provided by the system enabled physicians to more easily incorporate evidence based clinical protocols to the care of particular patients. This latter possibility suggests that investments in IT-based decision support in medicine may have quite a different long-term effect than IT investments in other settings. If IT systems can be used to increase the rate of diffusion of evidence based clinical knowledge, this may have the salutary effects of breaking down inefficient geographic variations in physician practice style and increasing the dynamic efficiency of the health care system. Understanding the effect of IT based decision support tools on the diffusion of new medical knowledge will be the subject of future investigations.

References

Athey, Susan and Stern, Scott (2002). "The Impact of Information Technology on Emergency Health Care Outcomes." RAND Journal of Economics **33**(3): 399-432.

Autor, David H., Levy, Frank, and Murnane, Richard .J. (2002). "Upstairs, Downstairs: Computers and Skills on Two Floors of a Large Bank April 2002." Industrial and Labor Relations Review **55**(3): 432-447.

Autor, David H., Levy, Frank, and Murnane, Richard J. (2003). "The Skill Content of Recent Technological Change: An Empirical Exploration." Quarterly Journal of Economics **118**(4): 1279-1333.

Bartel, A. P., C. Ichniowski, and K. L. Shaw (2005): "How Does Information Technology Really Affect Productivity? Plant-Level Comparisons of Product Innovation, Process Improvement and Worker Skills," National Bureau of Economic Research, Inc, NBER Working Papers: 11773.

Beaulieu, N., D. Cutler, K. Ho, G. Isham, T. Lindquist, A. Nelson, and P. O'connor (2007): "The Business Case for Diabetes Disease Management for Managed Care Organizations," *Forum for Health Economics & Policy*, 9, 1072.

Bohmer, Richard (1998). "Complexity and Error in Medicine." Harvard Business School Teaching Note (9-699-024): 16.

Bresnahan, Timothy F., Brynjolfsson, Erik, et al. (2002). "Information Technology, Workplace Organization, and the Demand for Skilled Labor: Firm-Level Evidence February 2002." Quarterly Journal of Economics **117**(1): 339-376.

Brynjolfsson, Erik and Hitt, Lorin M. (2000). "Beyond Computation: Information Technology, Organizational Transformation and Business Performance Fall 2000." Journal of Economic Perspectives, Fall **14**(4): 23-48.

Cebul, R., Herschman, R, Rebitzer, J.B., Taylor, L. J. and Votruba, M. (2007): "Employer-Based Insurance Markets and Investments in Health," Department of Economics, Weatherhead School, Case Western Reserve University.

Duggan, Mark (2005). "Do New Prescription Drugs Pay for Themselves? The Case of Second-Generation Antipsychotics." Journal of Health Economics **24**(1): 1-31.

Frank, R. G. (2004): "Behavioral Economics and Health Economics," National Bureau of Economic Research, Inc, NBER Working Papers: 10881.

Gaynor, Martin, Rebitzer, James B., Taylor, Lowell J. (2004). "Physician Incentives in HMOs." Journal of Political Economy **112**(4).

Gertler, Paul and Simcoe, Tim (2004). "Disease Management: Using Standards and Information Technology to Improve Medical Care Productivity." WORKING PAPER.

Hubbard, Thomas N. (2003). "Information, Decisions, and Productivity: On-Board Computers and Capacity Utilization in Trucking September 2003." American Economic Review **93**(4): 1328-1353.

Ichniowski, Casey and Shaw, Kathryn (2003). "Beyond Incentive Pay: Insiders' Estimates of the Value of Complementary Human Resource Management Practices." Journal of Economic Perspectives **17**(1): 155.

Institute of Medicine Committee on Quality of Health Care in America (2000). To Err Is Human: Building a Safer Health System. Linda T. Kohn, Janet M. Corrigan and Molla S. Donaldson. WASHINGTON DC, NATIONAL ACADEMY PRESS.

Institute of Medicine Committee on Quality of Health Care in America (2001). WASHINGTON D.C, NATIONAL ACADEMY PRESS.

Javitt, Jonathan. C., Steinberg, Gregory., Locke, Todd, Couch, James B., Jacques, Jeffrey., Juster, Iver, and Reisman, Lonny (2005). "Use of a Claims Data-Based Sentinel System to Improve Compliance with Clinical Guidelines: Results of a Randomized Prospective Study." American Journal of Managed Care (11): 21-31.

Landrigan, Christopher P., Rothschild, Jeffrey M., Cronin, John W., Kaushal, Rainu, Burdick, Elisabeth, Katz, Joel T., Lilly, Craig M., Stone, Peter H., Lockley, Steven W., Bates, David W., Czeisler, Charles A. (2004). "Effect of Reducing Interns' Work Hours on Serious Medical Errors in Intensive Care Units." The New England Journal of Medicine **351**(18): 1838.

Lapetina, Elizabeth M. and Armstrong, Elizabeth M. (2002). "Preventing Errors in the Outpatient Setting: A Tale of Three States." Health Affairs **21**(4): 26.

Levy, Frank and Murnane, Richard J. (2004). The New Division of Labor. PRINCETON, NEW JERSEY, PRINCETON UNIVERSITY PRESS.

Newhouse, Joseph P. (2002). "Why Is There a Quality Chasm." Health Affairs **21**(4): 13

Phelps, C. E. (2000): "Information Diffusion and Best Practice Adoption " in *Handbook of Health Economics*, ed. by J. P. Newhouse, and A. J. Culyer. AMSTERDAM: ELSEVIER, 224-264.

President's Council of Economic Advisors (2004). Economic Report of the President. Washington, D.C., United States Government Printing Office.

President's Information Technology Advisory Committee (2004). Revolutionizing Health Care through Information Technology. Arlington, VA, Executive Office of the President of the United States.

Rebitzer, J. B., Rege, M., and Shepherd, C. (2007): "Information Technology and Information Overload in Health Care," Economics Department, Weatherhead School, Case Western Reserve University.

Shojania, Kaveh G. and Grimshaw, Jeremy M. (2005). "Evidence-Based Quality Improvement: The State of the Science." Health Affairs **24**(1): 138.

Skinner, J., and D. Staiger (2005): "Technology Adoption from Hybrid Corn to Beta Blockers," National Bureau of Economic Research, Inc, NBER Working Papers: 11251.

Skinner, J., E. Fisher, and J. E. Wennberg (2001): "The Efficiency of Medicare," National Bureau of Economic Research, Inc, NBER Working Papers: 8395.

Slesnick, D. T., and B. Wendling (2006): "Charges and Reimbursements: Evidence from the Medical Expenditure Panel Survey," University of Texas at Austin.

Solow, Robert S. (1987). We'd Better Watch Out. New York Times Book Review.

The Leap Frog Group for Patient Safety (2004). Fact Sheet.

Table 1
Descriptive Statistics

	Study	Control
Number of Members in 2001 > 12 years old	19716	19792
Fraction of Members in Study for all 12 Months 2001	0.730	0.724
Average Charges in 2001 (pmpm)		
Total Charges	327.54	352.31
In-patient Charges	58.15	72.06
Out-patient Charges	71.69	74.11
Rx Charges	65.21	65.27
Professional Charges	132.48	140.87
Average Charges in 2000 (pmpm)		
Total Charges	280.45	283.39
In-patient Charges	58.23	59.43
Out-patient Charges	57.39	57.97
Rx Charges	47.13	47.88
Professional Charges	117.71	118.11

Table 2
Who Recieved Care Considerations In the Study Group?

	Probit Receive any CC?	Negative Binomial Number CC's Received	Ordered Probit Severity of CCs
20<AGE≤ 30	0.011	1.446	-0.134
	(1.31)	(1.26)	(1.26)
30 ≤AGE < 40	0.041	3.742	-0.447
	(5.03)**	(5.26)**	(4.93)**
40 ≤ AGE < 50	0.092	8.825	-0.823
	(9.83)**	(9.14)**	(9.56)**
50 ≤ AGE < 60	0.175	16.251	-1.136
	(13.57)**	(11.76)**	(13.10)**
60 ≤ AGE	0.314	27.828	-1.4860
	(16.45)**	(13.75)**	(16.06)**
Female	-0.00600	0.83151	0.0620
	(2.29)*	(2.63)**	-(1.92)
Charges (pmpm) in 2000	0.000011	1.0005	-0.00014
	(5.67)**	(10.00)**	(6.25)**
Constant		-7.290	
		(31.42)**	
Observations	19693	19693	19693
log pseudolikelihood	-3389.5289	-4077.0834	-4077.0834

Column (1) is a probit with coefficients expressed as derivatives. For dummy variables this is a discrete change from 0 to 1.
For continuous variables the derivative is evaluated at the mean.
Column (2) is a negative binomial count model of the number of CCs received. Parameter $\alpha = 4.02$.
   The coefficients are expressed as incident rate ratios so that the number of cc's for those 20-30 is 3.742 times that of the omitted age group.
Colum (3) is an ordered probit of an indicator of cc severity. CC's were ranked from least (3) to most (1) dangerous. Those
   with no CCs were given a 4. Members were assigned the level of the most dangerous CC they recieved.

The omitted age category is teenagers between 12 and 20.

Numbers in ( ) are z scores

Table 3

The Average Effect of the Intervention on Utilization as Measured by: Dollar Charges per Member per Month and Hospitalization

	(1)	(2)	(3)	(4)	(5)	(6)
	Total Charges (pmpm)	In-patient Charges (pmpm)	Out-patient Charges (pmpm)	Rx Charges (pmpm)	Professional Charges (pmpm)	In Hospital
Year = 2001	69.099	12.692	16.172	17.399	22.837	0.002
	(8.88)**	(2.51)*	(7.47)**	(31.53)**	(8.92)**	(0.83)
*Study*Year = 2001*	-21.92	-12.833	-1.823	0.7	-7.963	0.000
	(1.99)*	(1.8)#	(0.60)	(0.90)	(2.20)*	(0.10)
Constant	281.856	58.829	57.661	47.499	117.867	0.05
	(72.37)**	(23.29)**	(53.23)**	(171.96)**	(92.01)**	(47.34)**
Individual Fixed Effects	yes	yes	yes	yes	yes	yes
Observations	78976	78976	78976	78976	78976	78976
Individuals	39508	39508	39508	39508	39508	39508

Absolute value of t statistics in parentheses
significant at 10% * significant at 5%; ** significant at 1%

The absolute value of bootstrapped z-scores for *Study*Year in* columns (1)-(6) are respectively:
(-1.85), (-1.83), (-0.58), (0.84), (-2.42), (-0.10). These were calculated by sampling, with replacement,
100 times for each equation.

Similar results were found if we estimated the model using only study year data.
For columns 1-5 the coefficients (t-statistics) for *Study*year* are: -24.777 (2.44), -13.906 (2.38), -2.423 (0.82), -.057 (0.04), -8.390 (-2.24)
Bootstrapped standard errors were very close to OLS in these models

F tests that individual fixed effects are jointly zero can be rejected at 1 percent significance levels.

Table 4

Exposure to Technology Reduces Costs at the Far-Right Tail of the Distribution of Costs

	(1)	(2)	(13)
		Quantile regressions	
	Median Total Charges (pmpm)	90th %tile Total Charges (pmpm)	99th %tile Total Charges (pmpm)
Study	-0.541	-20.942	158.436
	(0.36)	(1.31)	(0.74)
Year = 2001	13.910	166.308	962.548
	(7.16)**	(9.07)**	(3.49)**
*Study*Year = 2001*	0.561	-26.512	-658.612
	(0.20)	(1.06)	(1.85)#
Constant	86.47083	636.9142	3189.613
	(77.92)	(54.00)	(19.50)
Observations	78976	78976	78976
Individuals	39508	39508	39508
R-squared	0.0005	0.0032	0.0037

Quantile regressions estimated for the median, 90th and 99th percentiles respectively.
 The standard errors for these regressions were bootstrapped with 1000 repetitions
 The R-squared for the quantile regressions are pseudo R-squared.

Table 5

The Effect of the Intervention on Individuals Likely to Recieve a CC

	Sample in Study Year over age 50 in Base Year					Weighting Observations Using the Propensity Score Entire Sample in Study Year				
	OLS	OLS	OLS	OLS	OLS					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Total Charges (pmpm)	In-patient Charges (pmpm)	Out-patient Charges (pmpm)	Rx Charges (pmpm)	Professional Charges (pmpm)	Total Charges (pmpm)	In-patient Charges (pmpm)	Out-patient Charges (pmpm)	Rx Charges (pmpm)	Professional Charges (pmpm)
Study	-72.171	-49.633	-9.59	0.447	-13.395	-66.363	-38.067	-2.913	-0.821	-24.562
	(2.04)*	(2.12)*	(1.10)	(0.13)	(1.18)	(3.44)**	(3.5)**	(-0.57)	(0.46)	(2.86)**
AGE $\geq$ 60	234.697	114.847	24.101	18.271	77.478					
	(5.82)**	(4.30)**	(2.42)*	(4.54)**	(5.97)**					
Female	-29.64	-56.477	9.918	13.444	3.475					
	(0.83)	(2.39)*	(1.12)	(3.78)**	(0.3)					
Charges (pmpm) in 2000	0.383	0.105	0.08	0.034	0.165					
	(26.60)**	(11.05)**	(22.35)**	(23.41)**	(35.54)**					
Constant	422.06	121.917	75.666	92.749	131.729	706.486	181.492	137.071	111.2238	276.7003
	(12.13)**	(5.29)**	(8.81)**	(26.76)**	(11.78)**	(51.76)**	(23.54)**	(37.73)**	(87.94)**	(45.55)**
Observations	8291	8291	8291	8291	8291	39468	39468	39468	39468	39468
R-squared	0.09	0.02	0.06	0.07	0.14	0.0003	0.0003	0.000	0.000	0.0002

Absolute value of t statistics in parentheses
* significant at 5%; ** significant at 1%

In columns (1)-(5), the absolute value of the bootstrapped z scores for variable *Study* are respectively (2.14), (2.17), (1.06), (0.13), (1.07)
The standard errors for all bootstrapped estimates involved sampling, with replacement, 100 times.

For the weighted regressions in (6)-(10), we set the analytical weight in STATA's regression command equal to the observation's propensity score, i.e. the predicted probability from equation (1) in Table 2 that an individual will generate a CC. This weighting scheme assumes that observations with high propensities to generate CCs are measured more accurately than observations with low propensities. Specifically, analytical weights are assumed to be inversely proportional to the variance of an observation; i.e., the variance of the j-th observation is assumed to be $\sigma/\omega j$ where the weights, ωj, are rescaled to sum to the number of observations in the data.

Table 6

Charges in the Study and Control Groups After the Care Engine was Rolled Out to Both Groups

	(1)	(2)	(3)	(4)	(5)	(6)	(7) Quantile Regression 99th percentile
	OLS	OLS	OLS	OLS	OLS	Probit	
	Total Charges (pmpm)	In-patient Charges (pmpm)	Out-patient Charges (pmpm)	Rx Charges (pmpm)	Professional Charges (pmpm)	In Hospital	Total Charges (pmpm)
Study	-24.777	-13.906	-2.423	-0.057	-8.39	-0.031	-500.176
	(2.44)*	(2.38)*	(0.82)	(0.04)	(2.24)*	(1.50)	(1.58)
Constant	352.313	72.057	74.112	65.27	140.874	-1.611	4,152.16
	(44.04)**	(14.63)**	(32.62)**	(71.12)**	(49.78)**	(109.67)**	(16.36)**
Observations	39508	39508	39508	39508	39508	39508	39508

Utilization Measures in 2002

	(1)	(2)	(3)	(4)	(5)	(6)	(7) Quantile Regression 99th percentile
	OLS	OLS	OLS	OLS	OLS	Probit	
	Total Charges (pmpm)	In-patient Charges (pmpm)	Out-patient Charges (pmpm)	Rx Charges (pmpm)	Professional Charges (pmpm)	In Hospital	Total Charges (pmpm)
Study	-8.528	-5.796	2.416	-1.783	-3.365	-0.005	-238.91
	(0.78)	(0.86)	(0.89)	(1.29)	(0.90)	(2.27)*	(1.07)
Constant	338.544	70.051	65.342	65.258	137.891		3948.585
	(41.94)**	(13.89)**	(33.61)**	(62.98)**	(49.50)**		(25.23)**
Observations	38056	38056	38056	38056	38056	38056	38056

Utilization Measures in 2003

	(1)	(2)	(3)	(4)	(5)	(6)	(7) Quantile Regression 99th percentile
	OLS	OLS	OLS	OLS	OLS	Probit	
	Total Charges (pmpm)	In-patient Charges (pmpm)	Out-patient Charges (pmpm)	Rx Charges (pmpm)	Professional Charges (pmpm)	In Hospital	Total Charges (pmpm)
Study	15.574	14.072	0.589	2.273	-1.36	0.001	-184.170
	(0.94)	(1.45)	(0.12)	(1.17)	(0.24)	(0.22)	(0.39)
Constant	379.225	78.716	88.1	50.632	161.776		5214.387
	(37.25)**	(14.55)**	(25.85)**	(39.57)**	(43.77)**		(16.03)**
Observations	27288	27288	27288	27288	27288	27288	27288

Robust t statistics in parentheses in columns (1)-(5) and (7). Column (6) presents z statstics.
The standard errors for the quantile regressions were calculated by bootstrapping 100 times with replacement.
* significant at 5%; ** significant at 1%

The experiment concluded at the end of December 2001, but the "care engine" remained on for two months
  At the end of February, the entire system was "turned off" and in June the system was rolled out to all members.
  Panel A compares charges in the study and control group during the study year, 2001.
Panels B and C compare study and control groups in 2002 and 2003, after the experiment was rolled out to all members.

Table 7
Descriptive Statics Regarding Care Considerations

	Study	Control
Number of CC's Issued	1299	1519
Fraction of Members with at least 1 CC in 2001	0.050	0.061
Distribution of CC's Among Members who Have Any CCs		
Percent of Members having any CC who have 1 CC	76.87	80.96
Percent of Members having any CC who have 2 CCs	16.46	14.56
Percent of Members having any CC who have 3 CCs	4.55	3.17
Percent of Members having any CC who have 4 CCs	2.12	1.14
Percent of Members having any CC who have 5 CCs	0	0.16
Mean CC's for Members with CC's	1.321	1.253
Number of Distinct Types of CC's Issued	90	83
Severity of CCs		
Fraction of Members with at least 1 Level 1 CC in 2001 (most serious)	0.001	0.002
Fraction of Members with at least 1 Level 2 CC in 2001 (less serious)	0.036	0.042
Fraction of Members with at least 1 Level 3 CC in 2001 (least serious)	0.019	0.024

Table 8

The Effect of the Experiment on Resolution Rates and the Probability of Recieving any Care Consideration

	(1) Probit *Successful Resolution Any "Add A Drug" CC* [0.18]	(2) Probit *Successful Resolution Any "Do a Test" CC* [0.31]	(3) Probit *Successful Resolution Any "Stop a Drug" CC* [0.34]
Study	0.086	0.058	-0.060
	(2.51)*	(2.24)*	(1.53)
Observations (only for 2001)	601	1354	592
Observations in Control Group	322	724	355

Robust z statistics in parentheses. [ ] is mean of dep. var. in the control group in 2001
* significant at 5%; ** significant at 1%

Probits in column (1) are estimated for all members who received at least one "add a drug" CC in 2001
Probits in columns (2) and (3) are for "Do a Test" CCs and "Receive Any CC" respectively.

 Coefficients are expressed as "derivatives.   Thus 0.18 of the control group who received an
   "add a drug" CC in 2001, had a successful resolution of an 'Add a Drug' CC. The resolution
   rate in the study group was 8.6 percentage points higher or 0.266.