The Impact of a Merit-Based Incentive Payment System on Quality of Healthcare: A Framed Field Experiment

Ellen P. Green, Katherine S. Peterson, Katherine Markiewicz, Janet O'Brien, Noël M. Arring

Highlights:

- This is the first study to use health care simulations with patient actors to investigate a merit-based incentive payment system.
- The merit-based incentive payment system increased the number of incentivized measures met.
- The merit-based incentive payment systems resulted in lower patient satisfaction, lower standards of care, and unnecessary diagnostic testing.

Keywords: Medicare Access and Chip Reauthorization Act, pay-for-performance, Merit-Based Incentive Payment System, Framed Field Experiments

JEL Codes: I1, I18, C9, L2

Abstract:

We study the impact of a merit-based incentive payment system on provider behavior in the primary care setting using new experimental methods that leverage healthcare simulations with patient actors. Our approach allows us to exogenously change a provider's incentives and to directly measure the consequences of two alternative payment systems: flat rate and merit-based incentive payment systems. Within our sample, we find that merit-based incentive payment systems increase the number of the incentivized measures met, but also lower quality of care through unintended effects on adherence to standards of care and patient satisfaction.

1. Introduction

Despite the popularity of merit-based incentive payment systems[1] studies have not demonstrated that they are associated with a consistent improvement in performance or decline in costs of care (Emmert et al. 2012, Gillam, Siriwardena, and Steel 2012, Rosenthal, Landon, and Epstein 2007, Werner et al. 2011, Iezzi, Bruni, and Ugolini 2014, Rosenthal et al. 2006, Sherry 2016, Mullen, Frank, and Rosenthal 2010). The failure to find significant and consistent associations is likely explained by the fact that changes in payment systems are difficult to study in real-world settings or that the outcomes of merit-based incentive payment systems (MIPS) may be conceptually ambiguous.

The most common argument against MIPS is that clinicians must carry out actions that are both measurable (e.g., mammogram screening) and unmeasurable (e.g., thorough physician examinations). Incentives that place higher value on measurable tasks are likely to cause clinicians to divert resources from unmeasurable tasks (Holmstrom and Milgrom 1991). However, analysis from Mullen, Frank, and Rosenthal (2010) suggests that the outcomes of MIPS may be ambiguous. Their paper provides a theoretical model and empirical evidence that clinician's tasks may be jointly produced. Specifically, actions taken to meet rewarded tasks may lower the costs of producing unrewarded tasks, suggesting that the production of unrewarded tasks may not drop significantly under MIPS. Therefore, the inconsistency in results across studies of MIPS may be driven by the ambiguity of the potential outcomes caused by MIPS.

Another potential explanation for the inconsistencies is that field studies rarely occur in the environment necessary to identify the true effects of explicit incentives. For example, experiments are typically cluster randomized by health care facility so that differences in unobserved variables across facilities may interact with changes in payment structure and impact the estimated effect of different payment systems. In addition, changes in payment systems often coincide with changes to health care

[1] Merit-based incentive payment systems have also been termed pay-for-performance, performance-based financing, and outcome-based payment in literature and policy reform.

delivery processes, which makes it difficult to draw a causal link between the change in payment and patient outcomes. For example, the 2015 Medicare Access and Chip Reauthorization Act (MACRA), which employs MIPS, requires health care providers to invest in their electronic health record system.

Another concern is that it is difficult to fully measure unintended consequences in the field. For example, field experiments typically rely on claims data. However, claims data does not report all tasks (e.g., thoroughness of physical exams). This implies that the clinicians' action sets are not fully measured in field studies, which is markedly important in evaluating the influence of MIPS on clinician behavior. These problems necessitate novel research methods to explore payment systems and framed-field experiments offer the necessary tools and conditions to study clinician behavior under true *ceteris paribus* conditions.

To study the potential impact of MIPS on clinician behavior, we used health care simulations with actors trained to portray a specific patient case with standardized responses. Healthcare simulations mimic a clinician's interaction with patients in a controlled true-to-practice environment and are typically used to assess clinical skills in medical schools, board licensure exams, and continuing medical education programs. Importantly, clinician behavior in health care simulations has been shown to correlate with clininical performance in the workplace (see Brydges et al. 2015 for a review).

The clinicians recruited to our study were asked to evaluate and recommend diagnostic and treatment plans for a panel of patients and were assigned to one of two groups: a control group or a MIPS group. In the control group, clinicians were paid a flat rate for participating in the experiment. Under MIPS, clinicians were paid a lower flat rate plus a bonus for each of the incentivized measures that they satisfied. All interactions were evaluated for performance through a video review, which allowed us to measure actions typically unobserved (e.g., thoroughness of physical examinations, thoroughness of patient history, and misrepresented self-reported data). Hence, the use of health care simulations to study MIPS allowed us to both provide a fully exogenous change to compensation methods and a more complete measure of the consequences of such a change.

We found several results pertinent to the implementation of MIPS. Most importantly, we found that although MIPS increased the number of incentivized measures met that other critical, but unincentivized, tasks were less likely to be conducted. Specifically, clinicians paid under MIPS were less likely to obtain a complete patient history, conduct a physical exam, or provide a thorough summary of the patient encounter. We also saw an increase in unnecessary services under MIPS. Specifically, clinicians paid under MIPS were more likely to inappropriately order the screening tests rewarded by the incentivized measures.

These results provide insight into the potential impact of MIPS and contributes to the growing literature in support of Holmstrom and Milgrom's (1991) multi-task agent problem (see for Jacob (2005) an example in education). Specifically, the experiment provides evidence that clinicians respond to MIPS by diverting resources from unrewarded actions (e.g., thorough physical exams, patient satisfaction) to rewarded actions (incentivized measures). Notably, the reduction in unrewarded actions were largely to services that are not reported in claims data. This suggests that field experiments that fail to find a disruption in care may be hindered by their dataset.

The paper also responds to the call for economists to strengthen their skills as economic-engineers (Roth 2002) and economic-plumbers (Duflo 2017) by introducing simulations as a tool for studying policy change. In each of their papers, Roth and Duflo discuss the importance of detail in implementing policy change. They provide examples of how well-vetted interventions and theories fail due to missing details that were seemingly of second-order importance. The control offered by healthcare simulations provides researchers with a novel setting to manipulate and evaluate these details at a lower cost (time, money, and unintended consequences) than randomized control trails and under *ceteris paribus* conditions.

The paper proceeds as follows. Section 2 introduces the Medicare Access and Chip Reauthorization Act and the policy we simulated in our experiment. Section 3 introduces and analyzes our model for the multi-task agent problem. Section 4 discusses how our experiment contributes to the pertinent literature and provides an overview of our experimental design. Section 5 provides a summary of our results. Section 6 provides a

discussion of our results, including estimates of the monetary impact of the MIPS incentive scheme. Section 7 concludes.

2. Policy Background

On April 16th, 2015 the Medicare Access and CHIP Reauthorization Act of 2015 (MACRA) was signed into law. MACRA requires that Medicare pay health care providers through one of two Quality Payment Programs (QPP), the Merit-based Incentive Payment System (MIPS) or the Advanced Alternative Payment Model (APMs), by 2019. The focus of this study is on the measures incentivized under the MIPS QPP.

MIPS will alter the reimbursement mechanism for clinicians who bill for Medicare Part B, which accounted for approximately $167.8 billion in gross fee-for-service spending in 2015 (HHS 2016). Medicare Part B pays for all medically necessary services (e.g., lab tests, surgeries, and health care provider visits) and preventive services (e.g., medical screenings, flu shots, annual wellness visits). Under MIPS, reimbursements for these services will remain volume-based (i.e., fee-for-service); however, reimbursement rates will be adjusted from year to year based on the healthcare provider's performance score. Performance scores are calculated by a weighted composite score based on adherence to outcome metrics in quality, improvement activities, and advancing care information. Of these reported measures, healthcare providers will be required to report their achievements on 6 *quality measures*. The impact of incentivizing these *quality measures* on performance is the focus of our study. However, to reduce confusion, we will refer to the *quality measures* defined under MIPS as *incentivized measures* and reserve the term *quality* as an adjective to describe the clinician's ability to satisfy more inclusive measurements for standards of care throughout the study (for more detail, see Section: Incentivized Measures (Table 2)).

3. Theoretical Prediction of Outcome-Based Incentives

A substantial literature examines the potential for negative consequences of outcome-based incentives for multi-task agents such as Gibbons 1998 and Baker 2000. In this section, we use a simplified version of the Holmstrom and Milgrom (1990) multi-task principal-agent framework to demonstrate that outcome-based incentives may result in

a diversion of clinician effort from unrewarded actions to rewarded actions and to formulate hypotheses for our framed field experiment.

Consider the clinician's optimization problem. During a typical day, he or she will serve a number of patients and allocate some time, T, to each patient according to anticipated medical requirements of the patient. Our interest is the allocation of the time allotted to a patient. We assume that patient outcomes are determined by an easily measurable action (e.g., breast cancer screening) and a more difficult to measure action (e.g., collection of medical history and complete physical exam). The time spent on the easily-measured action is denoted t_1 and that on the less measurable one t_2, with sum being the allotted patient time, *T*.

We assume that the clinician cares about the welfare of his or her patients, and also about his or her own benefits, which is proxied by income *Y*. Two reward schedules are of interest, one in which the clinician is paid a fixed salary and one in which the clinician receives a salary plus some remuneration for the easily measured activity as under MIPS. In the second case, $Y = B^c + at_1$. In the former case, the clinician benefit is simply, $Y = \beta^c$, the clinician's base salary ($a = 0$). The expected benefit to the patient is increased by the time spent on both activities, $B^p(t_1, t_2)$ diagnostic activities.

As in Holmstrom and Milgrom's 1990 framework for teacher incentives, we assume that clinicians are intrinsically motivated to provide services to each patient, because they care to some extent about the welfare or medical outcomes of their patients. This implies that even under a fixed wage, clinicians will produce benefits for their patients. The clinician's overall benefit is a weighted sum of their income and the benefits that their diagnostic efforts provide for their patient, $B^c + at_1 + wB^p(t_1, t_2)$, with $0 < w < 0$. The opportunity cost of the time spent with the patient reflects benefits potentially available from spending more time with other patients or on personal leisure, which is represented as $C(T)$; this is assumed to be constant for a given allocation of time across patients or allocation to the average patient.

The clinician's expected payoff or net benefit (N) with respect to a particular patient is:

$$N = B^c + at_1 + wB^p(t_1, t_2) - C(T) \qquad \text{or}$$

$$N = B^c + at_1 + wB^p(t_1, T - t_1) - C(T) \qquad (1)$$

Under the fixed salary system ($a = 0$), the time the clinician spent on the easily observable diagnostic service will satisfy:

$$w(B^P_{t_1} - B^P_{t_2}) = 0, \qquad (2)$$

which for a given allocation of time to the patient, approximately the optimal division of time *T* between the two diagnostic methods, although the total amount of time allocated to the patient may be smaller than optimal. Time is divided between the two diagnostic methods so that the marginal benefit from each method to the patient are equal.

Under the MIPS system under which the clinician is rewarded for the easily observed diagnostic service, the allocation of time is:

$$a + w(B^P_{t_1} - B^P_{t_2}) = 0 \qquad (3)$$

which implies a somewhat larger amount of time spent on the observable service, since there is an additional marginal benefit from providing this service to patients ($a > 0$).

Ignoring any effect on the average allocation of time to the patients that might occur if, for example, the additional payment induces clinicians to reduce leisure or see more patients, the first implication of MIPS leads us to our first hypothesis:

Hypothesis 1: Clinicians incentivized under MIPS will meet a higher number of the MIPS-incentivized measures relative to the control group (salary reward system).

However, the impact of MIPS on unrewarded tasks is unclear and depends on whether or not the tasks are substitutes or jointly produced. The two tasks may be substitutes for the agent's time. This may be substantiated by clinical practice guidelines, which establish standards of care used by clinicians to guide clinical decision making (Institute of Medicine, 2011). Clinical guidelines can be visualized as decision trees which lead the clinician to diagnose (or not diagnose) a condition based on patient responses and results. To move along the decisions tree and closer to the correct diagnosis, the clinician must conduct (or ask) a series of unique and informative actions or questions. At each point along the process, the clinician becomes closer to the correct diagnosis and, therefore, increases patient benefit. Given the unique nature of each action or

question required to move along the decision tree, the activities would be substitutes. For a given allocation of time across patients, the implies that the time spent on the other diagnostic service necessarily falls, since $t_2^* = T - t_1^*$.

Alternatively, Sherry (2016) and Mullen et al. (2010) demonstrate that the effects of pay-for-performance are theoretically ambiguous if tasks are jointly produced. That is, the completion of the rewarded task either leads to the completion of an unrewarded task along the way or reduces the cost of production of the other task; this reduces the burden of the multi-task principal agent problem and changes the time constraint from $t_2^* = T - t_1^*$ to $t_2^* = T - t_1^*(t_2^*)$, where t_1^* is dependent on the production of t_2^*. Therefore, in this case, the impact of the incentives on the production of t_1^* depends on the joint production function of the two activities.

The results of the framed field experiment will also address whether or not the actions are substitutes or complementary, which leads us to our second hypothesis.

Hypothesis 2: If clinician tasks are substitutes and therefore, clinicians incentivized under MIPS will decrease production of unrewarded standards of care relative to the control (salary reward system).

4. Framed Field Experiment

Our experimental framework uses healthcare simulations to study the potential impact of MIPS on clinician quality of care. Healthcare simulation is used in the healthcare industry to assess performance for board licensure, graduate and undergraduate medical training, certification, and performance review. For example, in the second stage of the United States Medical Licensing Examination (USMLE), medical students' clinical skills are evaluated through simulation. Instead of a classroom, the examination center "simulates" a healthcare clinic. During the USMLE, each medical student rotates through 12 mock patient examination rooms encountering a different Standardized Patient (SP) portraying a different case in each room. Medical students are evaluated on their ability to gather a relevant medical history, conduct a thorough physical examination, communicate effectively with their patient, document findings, and order appropriate diagnostic exams. Without successful completion of the simulation portion

of the USMLE, a student will not become a licensed physician. Our experiment borrows from the USMLE design to assess clinicians' behavioral responses to MIPS.

Our framed field experiment contributes to the young and growing field of behavioral experiments in health (see Cox, Green et al. 2016 and Galizzi and Wiesen 2018 for an overview). Formative experimental studies set in experimental economics laboratories have demonstrated differences across agent behavior under traditional reimbursement schemes in health such as fee-for-service, capitation, and salary (Green 2014, Hennig-Schmidt, Selten, and Wiesen 2011, Brosig-Koch et al. 2016b, Bejarano, Green, and Rassenti 2017). Pay-for-performance incentives have also been investigated using artefactual field experiments and laboratory settings and demonstrate both an increase in productivity (Keser, Peterle et al. 2014) and a misallocation of resources (Brosig-Koch et al. 2016a, Keser, Peterle, and Schnitzler 2014, Green 2014). The discrepancy in results was dependent on the asymmetry of information in the experimental relationship. As in the field study of the Safelite Glass Corporation conducted by (Lazear 2000), Keser, Peterle et al. (2014) study a performance-pay model that anchored bonuses to an optimal outcome. To utilize this pay-for-performance model in the health care industry would require knowledge of the optimal patient outcome, which is a key obstacle in implementing a pay-for-performance model the health care industry (Holmstrom and Milgrom 1991, Emons 1997, Bejarano, Green, and Rassenti 2017).

While many of the aforementioned studies properly imitated the asymmetric information, problems present for agents in a health care environment, the decisions of subjects were abstracted from a physician's decision in the health care industry. For instance, Green 2014 used a real-effort task where student subjects (physicians) were hired to proofread essays and where correctness of the proofreading benefited or harmed an individual in another pool of student subjects (patients). The "physician's" pay varied based on his/her own behavior, but was independent of quality of proofreading the "patient" received. While the experimental design provided the appropriate asymmetic infromation and incentives to mimic the health care environment, it did not fully mimic the complexities in decision-making between a doctor and their patient. Cox, Sadiraj et al. (2016) improved on the laboratory experimental design by increasing complexity of the subject's choice. Cox, Sadiraj et al. (2016) created virtual patients and recruited

medical students as subjects in their investigation of how incentivized readmission rates impacted emergency room discharge. Our study contributes to this growing methodology by introducing patient actors trained to represent standardized patient cases in a true-to-practice laboratory environment or a framed field experiment (Harrison and List 2004). The patient cases convey the complexity of a clinician's patient evaluation, which allows us to more fully measure the unintended consequences of a change in policy like MIPS in a controlled environment.

While the introduction of SPs and health care simulations complement laboratory and field studies, there are limitations in the methodology. For instance, patient impact may be undervalued in a health care simulation. While the behavior of the clinician in our experiment was linked to the standardized patient payment, the potential for negative outcomes is capped to a monetary outcome and cannot meassure the true potential for a clinician's performance to negatively impact a patient's health. While clinicians may perform differently in a simulated environment knowing patient safety is not directly impacted by decisions, Hennig-Schmidt, Selten, & Wiesen (2011) note that simulation is a valuable research tool to complement other empirical methods.

4.1 Experimental Design

In our experiment, primary care NPs and PAs were recruited and asked to create a treatment plan for three unique SPs. Clinicians were recruited electronically using a list serv for the state's coalition of nurses in advanced practice, the alumni networks of five local universities, and professional e-mail addresses for advanced practice nurses in private and public healthcare organizations, including the state's primary care association. A standard recruitment letter was used (See Appendix 1-recruitment email). Flyers were also posted on university alumni websites and the state's professional organization for physician assistants (PA). The requirement for participation in the study was being an advanced practice provider currently in practice.

Each experimental session had a maximum of three clinicians. At the start of each session, the clinicians were consented and briefed on their task using a standard script (Appendix 2). During the instructions, clinicians were provided blank copies of patient medical records, patient-provided information forms, and documentation forms for

review. In the incentivized or MIPS treatment group, the clinicians were asked to complete a 3-question quiz to ensure understanding of how their decisions impacted their earnings. At the end of the instructions, the clinicians were given a virtual tour of the "simulated" patient examination rooms.

After completing the virtual tour, the clinicians were escorted to the examination rooms to start their evaluations. At the start of each patient evaluation, the clinicians were provided with the case specific medical record,[2] and a completed patient-provided information form reflecting the reason for the current visit. Clinicians were given a total of 20-minutes to review patient information and evaluate the patient.[3] The clinicians decided how to allocate their 20 minutes. After the 20 minutes expired, clinicians were escorted out of the examination room and given an additional 5 minutes to document the patient encounter and proposed treatment plan. Each clinician rotated through the three mock examination rooms encountering a different SP in each room.[4] To avoid ordering effects, each clinician started at a different patient case and subsequently rotated through the patient panel. After the third patient evaluation, the clinicians were asked to complete a brief survey (Appendix 3), were paid, and were then free to leave.

Table 1: Description of Treatments in Framed Field Experiment

Name	Description
Control	Clinicians were paid a flat rate of $200 for evaluating the three standardized patients.
MIPS	Clinicians were paid a flat rate of $150 and had the opportunity to earn a bonus of $10.00 for satisfying each merit-based metric

[2] Patients establishing new care did not have medical records. The medical record and patient provided information form were created for each case based upon standard templates and were reviewed by clinical experts in primary care.

[3] The 20-minute timeframe was determined as the average appointment length by a panel of clinical experts in primary care, is also supported by the results of the National Ambulatory Medical Care Survey (Centers for Disease Control and Prevention, 2015).

[4] To remove ordering effects, practitioners rotated through the three patients in different sequences.

If desired, the clinicians were given the opportunity to spend an additional 15 minutes adding to their patient documentation after they received their payment.

In the experiment, the clinicians' final payment varied by treatment. In the following section, we describe the payment treatment in detail.

4.2 Treatments: Control and MIPS

In the *control*, clinicians were paid a flat rate of $200. In the *MIPS incentives* treatment, clinicians were paid a flat rate of $150 for evaluating 3 SPs and had the opportunity to earn a bonus of $10 for satisfying each of 5 outcome-based metrics for each patient. The $10 bonus was selected such that if a clinician in the MIPS incentive treatment provided the appropriate services for the age, gender, and purpose for the patient visit, he or she would receive the same payment as those in the control (i.e., $200). The success at fulfilling an outcome-based metric was self-reported for each patient

encounter. The clinicians were informed that they were not expected to fill out the entire outcome measure checklist, but only those sections that they felt were appropriate for the patient.

Table 2: Incentivized Measures

Metric	Description of required actions to meet metric	Payment
1. Breast Cancer Screening	Each female patient between the ages of 50-74 having had or been scheduled for a mammogram screen for breast cancer within the last 2 years.	$10/per patient
2. Colorectal Cancer Screening	Each patient between the ages of 50-75 having had or been scheduled for one or more screening for colorectal cancer. These include: · Fecal Occult Blood Test in the past year · Flexible Sigmoidoscopy during the past 4 years · Colonoscopy over the past 9 years	$10/per patient
3. Pneumococcal Screening	Each patient between the ages of 50-75 having had or been scheduled for a Pneumococcal Vaccination.	$10/per patient
4. Medical Assistance with Smoking and Tobacco Cessation	Each patient consulted on smoking and if appropriate, smoking cessation.	$10/per patient
5. Screening for Depression	Each patient over the age of 12 having been screened for clinical depression.	$10/per patient

The 5 incentivized measures fall under the U.S. Centers for Medicare & Medicaid Services' (CMS) definition of quality and can be found in the 2016 Physician Quality Reporting System (PQRS) measures list (CMS, 2017) and Healthcare Effective Data and Information Set (HEDIS). PQRS and HEDIS measures are U.S.-based tools used to measure performance on specific criteria related to care and service that is being largely adopted under MIPS (NCQA, 2012). The outcome measures used in the MIPS treatment included 1. Breast Cancer Screening, 2. Colorectal Cancer Screening, 3. Pneumococcal Screening, 4. Medical Assistance with Smoking and Tobacco Cessation, and 5. Screening for Depression (Table 2). The metrics were selected based on two criteria: the appropriateness for the primary care environment and that the measures could be addressed or could be easily modified to be addressed in a one-time patient encounter.

The one-shot nature of the experiment required that two of the incentivized measures be evaluated differently than in actual clinical practice. First, screening for breast and colorectal cancer consist of diagnostic exams ordered by a primary care clinician to be completed outside of the actual patient-provider appointment time and often even outside of the primary care office environment. Second, the pneumococcal vaccine screening is deemed met when it has been administered. In this study, clinician's verbalization to the patient of their intent to order the studies or vaccination was the primary means of determining whether or not these metrics were met in the simulated environment.

Additionally, under MIPS incentivized measures are expressed as a fraction with the numerator and denominator defined by specific International Classification of Diseases codes used for provider billing in healthcare (CMS 2015). The denominator describes the number of patients eligible for the performance measure within an individual clinician's patient caseload, while the numerator describes a clinical action that fulfills the performance measure. For example, for breast cancer screening the numerator describes the number of women who had a mammogram within the last two years to meet the measure of breast cancer screening and the denominator is the total number of women between the ages of 50 and 74 in a clinician's practice. The goal of the

incentive is to increase the number of women receiving a breast cancer screening; however, the format of the incentive allows clinicians to improve their ratio by excluding patients from the denominator (i.e., reducing the number of patients that are eligible for the screening). Denominator exclusions, such as patient specific medical conditions, are not included in the cases created for this study. The structure of our experiment eliminates the provider's ability to decrease the denominator, thus the potential to falsely increase the percentage of eligible patients who received the appropriate care as reported by the measure, as has been shown to occur in studies of the United Kingdom's pay for performance payment structure (Gravelle, Sutton, and Ma 2010, Sutton et al. 2010).

To mimic the impact clinician's performance has on their patient in the simulated environment, the SPs were given a bonus based on the quality of care provided. The bonus was calculated as the proportion of standards of care met in the documentation. Specifically, if the clinician met 75% or 30 out of 40 standards of care, they would be reimbursed 75% of $5 or $3.75. The clinicians were informed that a proportion of the patient's payment would be determined by the quality of documentation at the start of the experiment. The SPs were unaware of the bonus.

Table 3: Standardized Patient Case Descriptions

Case	Description
Initial Visit	Female patient age 66. Patient is establishing care at a new clinic with a new provider due to changes in insurance coverage. She has no specific complaints at this time.
Costochondritis	Established female patient age 45. Patient's chief complaint is a new onset of pain in the chest area for the past 3 days. Patient is otherwise healthy with no history of cardiac disease.
Diverticulitis	Established male patient age 73. Patient's chief complaint is a new onset of lower abdominal pain for the past 2 days. The abdominal pain is constant with occasional cramping and is worse in the lower left quadrant.

4.3 Standardized Patients (SPs) and Cases[5]

The SP cases were created to reflect three distinct sets of patient characteristics: a) patients that qualify for the measures, and standards of care are aligned with actions required by the measures; b) patients who do not qualify for the measures; and c) patients who qualify for the measures, but clinicians should focus on the acute problem.

Initial Patient Visit

Establishing care was the primary purpose of the Initial Visit case. A 66-year-old female patient is establishing care at a new clinic with a new provider due to changes in insurance coverage. She has no specific or acute complaints at this time. All incentivized measures were appropriate for this case, potentially improving care.

Costochondritis

Established female patient age 45. Patient's chief complaint is a new onset of pain in the chest area for the past 3 days. Patient is otherwise healthy with no history of cardiac disease. In this case the outcome-based metrics of breast cancer, colorectal cancer, and pneumococcal screening were not applicable given the patient's age. Screening for tobacco use with cessation counseling and screening for depression were applicable, but both potentially distracted from appropriate care based upon the chief complaint.

Although under the age cutoff for receiving Medicare, this patient was included to determine how clinicians would react to patients outside of the Medicare reimbursement scheme, but still in their practice, i.e., to provide a more complete measure of un-intended consequences.

Diverticulitis

Established male patient age 73. Patient's chief complaint is a new onset of lower abdominal pain for the past 2 days. The abdominal pain is constant with occasional

[5] The case tools are still in use for educational purposes today, thus it is imperative to maintain the confidentiality and fidelity of the cases and associated training tools. The academic institution that partnered in this study is concerned that publishing the complete case studies could place students at risk for violating the university's academic integrity policy. However, the case tools are available upon request.

cramping and is worse in the lower left quadrant. In this case, all incentivized measures aside from the breast cancer screening were applicable given the patient's age, but potentially distracted from appropriate care based upon the chief complaint.

The Costochondritis and Diverticulitis cases were complaint-based requiring accurate identification of probable or actual medical diagnosis and medical management of the acute condition. Nuances specific to each case, embedded within the patient's past medical history and history of present illness, were used to evaluate the clinician's misuse, overuse and/or underuse of screening tools and diagnostic studies in primary care.

The actors for each case in the experiment all had at least one year of experience as an SP. In preparation for the experiment, all SPs participated in a five-hour training session which consisted of three hours of case-specific training and two hours of training on the Interpersonal Communication Skills (ICS) tool. Case-specific material included background information on each patient such as medical/surgical and social history, as well as a series of potential questions and answers specific to the patients' cases, e.g., "When did symptoms begin?" SPs memorized all case material and practiced responding to questions during mock clinical exams. ICS tool training included SPs watching a series of videos of actual SP/participant interactions to practice evaluating each participant using the ICS tool. Benchmarks were provided for ratings on the ICS tool with descriptions of participant behaviors for each item and for different ratings. Afterwards, participants discussed reasons for selecting their ratings, scores were compared, and consensus was reached on final scores.

4.4 Outcome Measures

Quality of care was evaluated in three categories: (1) incentivized measures, (2) interpersonal communication skills, and (3) standards of care in the patient encounter.

Incentivized Measures: Regardless of treatment, the success at completing the five incentivized measures described in Table 2 were evaluated by video review. The breast cancer screening measure was met if clinicians inquired about SPs having had a mammogram within the past 2 years or scheduled the study during the patient encounter. The colorectal cancer screening measure was met if clinicians verbally

inquired about SPs having had or been scheduled for at least one screening test during the encounter: (1) fecal occult blood test in the past year, (2) sigmoidoscopy in the past 4 years, or (3) colonoscopy in the past 9 years. The pneumococcal screening measure was met if clinicians verbally inquired about SPs having ever received or been scheduled to receive the vaccine during the encounter. Screening for smoking and tobacco use was satisfied if clinicians verbally inquired about tobacco use and provided consultation on cessation, if applicable, during the encounter. The measure for depression was met if the clinician screened the SP for depression during the encounter or was scheduled for a depression screening.

Standards of Care: Standards of care rubrics to evaluate the process of medication reconciliation, history of present illness, past medical/surgical history, physical exam, and summary to include report, impression and plan were created to assess the quality of care of (1) the patient evaluation and (2) documentation (Adamson, Kardong-Edgren, and Willhaus 2013). Additionally, the evaluation rubrics captured whether the clinician performed a complete patient evaluation (e.g., physical examination, family history, social history, health history, etc.) and addressed all identified health concerns in their plan of care.

To ensure that rubrics were inclusive, clinical practice guidelines served as the foundation of the case-specific evaluation recommendations. Clinical practice guidelines from the American Association of Clinical Endocrinologists, American Thyroid Association, and American Congress of Obstetricians and Gynecologists served as the foundation for the evaluation rubric for the Initial Visit case. Clinical practice guidelines from the American Academy of Family Physicians on the diagnosis and management of acute chest pain in adults, and National Guidelines Clearinghouse on the diagnosis and treatment of chest pain guided the development of the evaluation rubric for the Costochondritis case (Davis et al. 2012). The Diverticulitis case rubric was created based upon recommendations from the American Gastroenterological Association Institute and American Society of Colon and Rectal Surgeons, in addition to the National Guidelines Clearinghouse practice parameters for the treatment of sigmoid diverticulitis and management of acute diverticulitis (Stollman et al. 2015, Feingold et al. 2014).

Finally, the standards of care rubrics were evaluated by five primary care NPs and one primary care PA to ensure accuracy in inclusiveness of standards of care and differential diagnoses.

All rubric questionnaires were reported as either 'Yes' the clinician did meet the standard of care, 'No' they did not, or 'Could not assess.' For example, evaluators were asked in the physical evaluation of the SP, "Did the clinician palpate the thyroid?" Evaluators could report yes, no, or could not assess if they did not have the appropriate camera angle to evaluate the question. Rubric questionnaires were completed through review of the video recording of the patient-clinician interaction. Two NPs from the Doctor of Nursing Practice program at a large public research university evaluated each patient-clinician encounter. If there were discrepancies across the two nursing faculty's' responses, a third evaluator reviewed the video to break the tie. The evaluators were blind to the experimental treatments and goal of the study.

Patient Satisfaction: Evaluation of the Interpersonal Communication Skills (ICS) by the SPs served as a proxy for patient satisfaction. The rubrics expand on current ICS evaluation tools used in clinical education (Cohen et al. 1996, Hassett et al. 2006). We identified eight constructs which fall within ICS care domains for the SPs to evaluate: (1) Non-verbal Communication, (2) Information Gathering, (3) Listening Skills, (4) Empathy, (5) Information Giving, (6) Respectfulness, (7) Safety, and (8) Professionalism. To improve inter-rater reliability, SPs were trained on these 8 criteria prior to the experiment. The training included how to evaluate the subject on each construct on a scale from 1 to 5, with 1 indicating 'poor performance,' 2 indicating 'needs improvement,' 3 indicating 'meets expectations,' 4 indicating 'good performance,' and 5 indicating 'exceptional performance.'

Survey: Subjects in the study were also asked to complete a survey (Appendix 3-post survey) to obtain demographic and other relevant information for the study (e.g., experience, typical practice environment, etc.).

4.5 Subjects and Settings

Participants in the study are among those professions whose Medicare Part B reimbursement will be impacted by the introduction of MIPS. Our sample was limited to NPs and PAs. In the state of Arizona, where the study was conducted, NPs and PAs are equivalent in the clinical setting and collectively referred to as Advanced Practice Providers under CMS. That is, clinicians with either an NP or PA have been trained to and can obtain license permits to provide the same set of services in primary-care clinics (azpa.gov). As it pertains to the implementation of MIPS, NPs and PAs both bill Medicare for services rendered using their own unique National Provider Identifier number and have the same reporting responsibilities as physicians in primary-care practices.

The experiments were conducted in a Simulation & Learning Resources (SLR) facility at large public research university. The experiment took approximately 1 hour and 45 minutes to complete per subject. On average, subjects earned $214 ($228 in MIPS). A total of 18 experimental sessions were conducted on eight separate days over 10 months (August 6, 7, 13, 14, 2016; October 29, 30, 2016; and May 5, 6, 2017). Dates were selected based upon availability of the Simulation & Learning Resources (SLR) facility. Researchers predetermined each experimental session as either treatment or control. Clinicians self-randomized themselves into a treatment or control group based upon their selected date and time for participation in the study. Study team members were stationed in the corridor to facilitate participant transitions between examination rooms and to ensure participants didn't interact with each other.

We determine effect size of our primary outcome through the theory model outlined in Section 3. Assuming clinicians meet at least one more of the incentivized measures under MIPS than in the control with a standard deviation in behavior of .75, our effect size is d=1.33. The discussion section below demonstrates that this is a conservative effect size, as a single unit increase in the number of incentivized measures met would have a large impact on the cost of health care both from clinician payment and cost of services. A power analysis suggest that with these assumptions, we need 14 subjects in each cell to obtain 95% power (A Priori Power (two-sided Wilcoxon rank sum) using G*Power 3.1 and alpha=0.05) (Faul et al. 2007, Faul et al. 2009).

Table 4: Summary of Clinician Characteristics and Behavior by Treatment & Case

Variable	MIPS	Control	Overall	N
Subject Characteristics				
Percentage NPs	47%	76%	62%	34
Age	34 (9.37)	42 (12.45)	38 (11.44)	31
Gender (% Female)	75%	86%	80%	31
Subject Behavior				
Earnings	$228 (39.50)	$200 (-)	$214 (31.16)	34
Self-Reported Typical Practice Experience (minutes)				
Minutes in Initial Visit	26.63 (13.39)	30.63 (9.67)	28.6 (11.99)	34
Minutes with Complaint Visit	19.11 (7.720)	19 (4.79)	19.06 (6.33)	34
Observed Time spent with patient (minutes)				
Initial Visit	14.49 (3.76)	12.28 (3.27)	13.39 (3.65)	34
Costochondritis	12.47 (3.14)	11.5 (2.8)	11.99 (2.96)	34
Diverticulitis	12 (3.33)	11.99 (3.04)	11.99 (3.14)	34
Difficulty of Cases (Scale 1-10, 10 most difficult)				
Initial Visit	5.27 (1.87)	4.75 (1.87)	5.09 (1.78)	34
Costochondritis	3.8 (1.57)	4.13 (2.10)	3.91 (1.73)	34
Diverticulitis	2.87 (1.56)	4 (1.85)	3.26 (1.71)	34

Notes: Standard deviations are reported in parenthesis next to mean. Row 1 reports the percentage of advanced practice providers that were Nurse Practitioners (NPs). In rows 5-7 and 8-11, data is disaggregated by self-reported time in minutes spent with patients in typical practice and oberved time spent with patients in the experiment. Rows 12-15, report the subject's percieved level of difficultyof each case on a scale from 1-10, 10 being most difficult. Standard deviations reported in paretheses.

*35 subjects were recruited, but one participant's data was dropped as they were not trained as either a Nurse Practitioner or Physician assistant

*In the state of Arizona where the study was conducted, NPs and PAs are equivalent in the clinical setting and collectively referred to as Advanced Practice Providers under CMS.

5. Results

Thirty-five NPs and PAs participated in the experiments.[6] Table 4 reports the summary statistics of the clinicians' demographics, typical practice environment, and experiment experience disaggregated by treatment. The majority of the clinician participants had an NP degree (62%). NPs accounted for a larger proportion of the control group than the MIPS treatment. The control also had significantly older clinician participants on average than the MIPS treatment (42 and 34 years old respectively, Wilcoxon rank sum p-value 0.0783). To ensure that the distribution of provider type and age was not driving our results, we conducted a series of robustness checks for each of our outcome

[6] One participant's data was dropped as he/she was not trained as either an NP or PA.

measures, which can be found in Appendix 4 and are discussed in footnotes throughout sections 5.1-5.3 of the paper.[7] The robustness checks indicate that neither the age distribution nor the training of the clinicians were driving the results. As is typical of the professions recruited, the majority (80%) of the clinicians were female (Jones 2007, Waugaman and Lohrer 2000).

As a result of the overprovision of incentivized outcomes, which we discuss in more detail below, earnings were greater in the MIPS treatment ($228) relative to the control ($200).[8]

Clinicians self-reported spending significantly more time (a total of 28.6 minutes) with their patients in initial visits in typical practice than was provided in the study (20 minutes) (t-test p-value 0.003). However, both in this study and studies in the field,

Table 5: Effect of MIPS on Quality of Care

Initial Visit	MIPS		Control	Difference	Difference 95% C.I.
Incentivized Measures	3.82 (0.95)	**	2.94 (1.14)	-0.88 (0.36)	[-1.62, -.15]
Standards of Care	31.88 (4.91)	**	37.06 (6.40)	5.18 (1.96)	[1.19, 9.16]
Patient Satisfaction	3.44 (0.63)		3.58 (0.86)	0.20 (0.29)	[-.40, .80]
Costochondritis					
Incentivized Measures	1.82 (1.59)	***	0.41 (0.62)	-1.41 (0.41)	[-2.25, -.57]
Standards of Care	23.12 (4.81)		24.41 (5.83)	1.29 (1.83)	[-2.44, 5.03]
Patient Satisfaction	3.82 (0.47)		3.80 (0.66)	-0.01 (0.20)	[-.42, .39]
Diverticulitis					
Incentivized Measures	2.29 (1.49)	***	0.88 (0.93)	-1.41 (0.43)	[-2.28, -0.55]
Standards of Care	29.47 (5.08)		29.06 (3.58)	-0.41 (1.51)	[-3.48, 2.66]
Patient Satisfaction	3.63 (0.57)	*	3.91 (0.54)	0.33 (0.22)	[-.13, .79]
Observations	17		17	34	34

Notes : Standard deviations reported in parentheses. Wilcoxon rank sum tests the hypothesis that the standards of care are the same across treatments disagregated by case. * indicates significant at 10%; ** indicates significant at 5%; *** indicates significant at 1%.

[7] We confirm that we have reported all measures, conditions, data exclusions, and the determination of sample size within the results section of the manuscript. Given that this was the first experiment of its kind, an a priori power analysis required estimates of effect size.

[8] The $10 bonus was selected such that if a clinician in the MIPS incentive treatment provided the appropriate services for the patient, they would receive the same payment as those in the control (i.e., $200).

clinicians spent an average of 14.49 minutes and 15 minutes with their patients in initial visits, respectively (Acheson et al. 2000). There was no significant difference in the amount of time spent with complaint-based visits in typical practice and the amount of time provided in the study (t-test p-value 0.392). The Initial Visit case was reported to be the most difficult followed by the Costochondritis case and then the Diverticulitis case. However, there was no significant difference across the two groups of clinicians in (self-reported) perceived difficulty (Wilcoxon rank sum p-value 0.1473).

In the following sections, we compare clinician performance across treatment disaggregated by SP case (i.e., Initial Visit, Costochondritis, and Diverticulitis). We analyze the data by SP Case as each case represents a distinct patient type found in primary care provider's patient panel and the three cases combined do not represent the full depth and breadth of a primary care providers patient panel (Ashman, Rui, and Okeyode 2018). A patient panel in a primary care setting often includes infants to the older elderly, as well as acute, chronic, preventative and post-surgical care for all age groups.

5.1 Incentivized Measures

Our first result is not surprising: clinicians in the MIPS treatment met significantly more of the incentivized measures than those in the control for all three cases. This result confirms **Hypothesis 1** of our theoretical model, namely, clinicians incentivized under MIPS will meet a higher number of the MIPS incentivized measures. Table 5 shows this result through a comparison of the average number of incentivized measures met across treatments by case. Statistical comparisons were made by the Wilcoxon rank sum test as well as the Fisher-Pitman permutation test (Kaiser 2007). In the Initial Visit case, clinicians in the MIPS treatment met 3.82 of the 5 outcome measures on average, whereas in the control clinicians met 2.94 (difference of -0.88 for incentivized measures, 95% C.I. -1.62, -.15) (Wilcoxon rank sum p-value=0.0258, Fisher-Pitman p-value=0.02). In the Diverticulitis case, clinicians in the MIPS treatment met 2.29 on average out of the 4 outcome measures needed, whereas in the control they met only .88 (difference of -1.41 for incentivized measures, 95% C.I. –2.28, -0.55) (Wilcoxon rank sum p-

value=0.001, Fisher-Pitman p-value=0.00).[9] In the Costochondritis case, clinicians in the MIPS treatment met 1.82 of the 5 incentivized measures on average, whereas in the control they met 0.41 (difference of -1.41for incentivized measures, 95% C.I.-2.25, -.57) (Wilcoxon rank sum p-value=0.005, Fisher-Pitman p-value=0.00).[10] However, in the Costochondritis, the patient only needed 2 of the 5 incentivized measures, whereas

Table 6: Percentage of Clinicians that Exhibited Misuse by Screening

	MIPS		Control
Overtreatment			
Breast Cancer Screening	47% (0.51)	***	0% (0.00)
Colorectal Cancer Screening	24% (0.44)	**	0% (0.00)
Pneumococcal Screening	18% (0.39)	*	0% (0.00)
Undertreatment			
Tobacco Use Screening	35% (0.49)	*	65% (0.49)
Depression Screening	71% (0.47)	*	94% (0.24)
Observations	17		17

Notes: Undertreatment is the failure to use proven treatments when appropriate. If the SP did not receive the depresseion screening or the tabacco use screening, this was counted as 1 and 0 if treated. Overtreatment is the use of treatments that were not needed. If the SP received the breast cancer, colorectal cancer screening, or phneumococcal screening, this was counted as 1, and 0 if not provided. Standard deviations reported in parentheses. Frequency of misuse (undertreatment or overtreatment) of each screening is reported next to the standard deviation.

The Proportion test was used to test the hypothesis that the proportion of over/under is the same across treatments by screening. *significant at 10%; ** significant at 5%; *** significant at 1%

[9] Incidentally, the diverticulitis case did not need breast cancer screening; however, two clinicians in the MIPS treatment recommended this screening. This overtreatment was not statistically different across treatments.

[10] As a robustness check of our results, we also investigated individual subject characteristics (See Appendix 4: Table A.1). In the ordered probit regression (clustered at the randomization level), when controlling for age, the likelihood of meeting incentivized measures remained significantly greater in the MIPS treatment than under the control. When controlling for age and practitioner type, the likelihood of meeting incentivized measures remained significantly greater than under the control was no longer significantly different in the Initial Visit; however, remained significantly greater in the Costochondritis case and the Diverticulitis Case.

some clinicians reportedly spent part of their examination time addressing all 5 of outcome measures. We further explore the actions in the Costochondritis case by looking into the misuse of medical services.

Costochondritis Case: Misuse

To further explore how the MIPS treatment impacted clinicians' performance we explore metrics of undertreatment and overtreatment in the Costochondritis case, where 0 indicates appropriate or non-use of a screening, and 1 indicates inappropriate or missed screening. Specifically, undertreatment is the failure to use proven treatments when appropriate. For example, a patient in the Costochondritis case needed depression and smoking cessation screening. If the SP did not receive the screening, this was coded as 1, and 0 if the clinician recommended or completed the screening. Overtreatment is the use of treatments that were not needed. For example, the patient in the Costochondritis case did not need pneumococcal vaccine, breast cancer screening or colorectal cancer screening. If the SP was recommended to receive any of these screenings it was coded as 1, and 0 if not recommended by the clinician.

Table 6 reports the frequency of recommended screenings across the two groups of clinicians for the Costochondritis case. The higher the frequency reported in Table 6, the worse the outcome for the patient in terms of cost, time, and inconvenience. Here we see that clinicians in the MIPS group were more likely to over-prescribe screenings for breast cancer (47%), colorectal cancer (24%), and the pneumococcal vaccine (18%). In fact, there was no overtreatment in the control. Conversely, clinicians in the control group were more likely to under-provide screening for tobacco use (65%) and depression screening (94%), in comparison with MIPS (35% and 71%, respectively).

Table 7 below summarizes the cost of healthcare waste and abuse and healthcare fraud. Healthcare waste and abuse is any practice that results in providing medical services that are not consistent with practice standards resulting in unnecessary cost to the healthcare system (HHS, 2017). Healthcare fraud is the intentional misrepresentation and/or false representation that medical treatment is needed and/or provided to attain benefit (HHS, 2017). In our study, we are hesitant to use the term

fraud since we did not assess providers' intention and have labeled this instead "potential fraud". Clinicians in the MIPS group were inappropriately paid an additional $5.29 on average, totaling $89.93 over the course of the experiment, for procedures that should not have been recommended. This figure does not include the additional fees a patient would incur for the additional services recommended, which in the United States cost $704 for mammograms, $1,012 for colonoscopies, and $291.49 for vaccines per patient (Salzmann, Kerlikowske, and Phillips 1997, Frazier et al. 2000, CMS 2017).

Table 7: Cost of Healthcare Fraud & Waste and Abuse

	Waste and Abuse	Potential Fraud
Initial Visit	- -	$8.24 (8.09)
Costochondritis	$5.29 (9.43)	$9.41 (13.45)
Diverticulitis	- -	$9.41 (11.44)
Total	$5.29 (9.43)	$27.06 (16.11)
Observations	17	17

Notes: Column 1 reports waste and abuse observed in the experiment as the average cost per subject for screenings that were conducted that were not recommended for the Costochondritis Case. Column 2 reports the potential fraud as the average cost per subject paid for an incentivized metric that was not observed upon video review. Standard deviations are reported in parentheses.

Self-Reported Screening vs. Observed Screening

Another potential drawback of the MIPS payment model is the self-reported nature of the incentivized measures. Under the MIPS finance model, clinicians have an incentive to report screenings regardless of completion (Bejarano, Green, and Rassenti 2016). Table 7 reports potential fraud as the average additional compensation received by clinicians who reported having recommended a screening that they did not verbally conduct or prescribe in their patient encounter. On average, clinicians under the MIPS treatment were overpaid by $27.06, totaling $460.02 over the 17 subjects, for services that they did not complete.

In summary, MIPS increased the number of incentivized measures met. However, in the Costochondritis case the MIPS incentives resulted in the over-prescription of expensive screening (i.e., mammograms and colonoscopies), when they were not warranted and resulted in bonuses for services that should not have been recommended based on practice guidelines. Finally, clinicians under the MIPS treatment were paid $27.06 on average for screenings that were self-reportedly prescribed; however, after video review

our evaluators did not find the screenings or recommendation for screenings were actually made.

5.2 Standards of Care

Table 5 (above) provided a second indication of the negative impact of MIPS on the clinician performance. Despite the increase of incentivized measure met, clinicians in the MIPS groups met significantly fewer of the Standards of Care (31.88) than those in the control group for the Initial Visit case (37.06) (difference of 5.18 for standards of care, 95% CI 1.19, 9.16) (Wilcoxon rank sum p-value=0.0207, Fisher-Pitman p-value=0.01).[11] This confirms our **Hypothesis 2**: If clinician tasks are substitutes and therefore, clinicians incentivized under MIPS will decrease production of unrewarded standards of care relative to the control (salary reward system).

Additionally, the findings confirm that the clinicians in the MIPS group were more focused on (i.e. distracted by) meeting the incentivized measures, than providing care based upon the current standards for each case. Alternatively, clinicians in the MIPS group may have been so distracted by incentivized measures that some attention to detail of case specific facts were overlooked. An example of this is found in the Costochondritis case in which breast cancer screening was completed on a 45-year-old female, which is not recommended by practice guidelines.

Further, we see that the rewarded and unrewarded tasks are substitutes for the clinician's time for some patient cases. Specifically, an increase in the completion of rewarded actions resulted in a decrease in unrewarded actions for the initial visit and

[11] As a robustness check of our results, we also investigated individual subject characteristics (See Appendix 4: Table A.2). In an OLS regression (clustered at the randomization level), when controlling for age, standards of care remained significantly lower in the MIPS treatment than under the control for the Initial Visit. Additionally, standards of care were significantly greater for the Costochondritis case as well. When controlling for age and practitioner type, patient satisfaction remained significantly greater in the MIPS treatment than under the control in both the Initial Visit and the Costochondritis case.

Costochondritis case as well as a decrease in standards of care. Conversely, the increase in rewarded actions in the diverticulitis case did not result in a decrease in unrewarded actions. However, the null result of the diverticulitis case may be due to our small sample size.

5.3 Patient Satisfaction (Interpersonal Communication Skills)

Table 5 also reported the negative impact that MIPS had on the clinician's patient satisfaction as evaluated by the ICS tool. Here we see that participants in the MIPS group received lower patient satisfaction ratings than those in the control group in the Initial Visit and Diverticulitis cases. However, the difference in patient satisfaction ratings was only statistically significant in the Diverticulitis case, the case in which the patient's chief complaint was not aligned with the outcome measures being incentivized. In the diverticulitis case, clinicians paid under MIPS received an average patient satisfaction score of 3.63, which was significantly less than those in the control, 3.91 (difference of 0.33 for patient satisfaction, 95% C.I. -.13, .79) (Wilcoxon rank sum p-value=0.081, Fisher Pitman p-value=0.09).[12] In a review of the video recording of the patient evaluations, clinicians appeared to be distracted by their incentives and rushed through their evaluations of the Diverticulitis case.

6. Discussion

In this section, we discuss tradeoffs between the increased number of incentivized measures met, and lower standards of care and patient satisfaction found in our study.

First, the use of MIPS resulted in poor information gathering (i.e., physical examination and patient health history). Poor and inadequate information gathering is the leading cause of diagnostic error in the primary care setting (Delzell et al. 2009). Approximately

[12] As a robustness check of our results, we also investigated individual subject characteristics (See Appendix 4: Table A.3). In an OLS regression (clustered at the randomization level), when controlling for age, patient satisfaction remained significantly lower in the MIPS treatment than under the control for Diverticulitis case. When controlling for age and practitioner type, patient satisfaction remained significantly greater in the MIPS treatment than under the control in the Diverticulitis case.

9% of diagnostic errors lead to major errors that would have been treatable, but instead resulted in death (Shojania et al. 2003). Other diagnostic errors lead to inappropriate care plans that result in either under or over -utilization of healthcare, both of which entail a cost to the system.

The estimated average cost of an inappropriate care plan is $436 per primary care visit, which would result in an estimated $36 billion in medical errors (Schwartz et al. 2012, CDC 2017, AHRQ 2016). The distractions introduced by the incentivized measures under MIPS have the potential to increase diagnostic errors, thereby decreasing the quality and increasing the cost of health care. In addition to the increased potential for diagnostic related errors, patient satisfaction was lower when clinicians were incentivized under MIPS. These results are similar other studies that found patient satisfaction was not improved or potentially decreased due to less patient-centered care that resulted from the provider's focus on the incentivized care (Gillam, Siriwardena, and Steel 2012, Maisey et al. 2008). These findings are significant because the quality of the patient-practitioner relationship has been linked to patient health outcomes such as blood pressure management and patient perceptions of pain control (Kelley et al. 2014).

Nonetheless, under MIPS clinicians met more of the incentivized measures, which included screenings for cancer and chronic diseases. These results are similar to field studies that found incentives increased performance for care that was incentivized (Doran et al. 2011, Gillam, Siriwardena, and Steel 2012, Iezzi, Bruni, and Ugolini 2014, Morland et al. 2017). For example, we found that clinicians assigned to the MIPS treatment group were more likely to screen for tobacco use and cessation. In the U.S. cigarette smoking accounts for more than 480,000 deaths per year, with an estimated cost of more than $289 billion per year reported in 2014 (General 2014). This expenditure includes an estimated $133 billion in direct medical care costs associated with tobacco use for adults and more than $156 billion in lost productivity from premature death. Therefore, the increased screening in the MIPS treatment would help relieve some of the burden of cost for tobacco use at a small cost.

Additionally, clinicians assigned to the MIPS were more likely to screen for depression. With depression being the estimated second leading cause of disability throughout the world by 2020 (Murray, Lopez, and Organization 1996), routine screening for depressive disorders is equally as impactful as screening for more traditionally considered chronic diseases such as heart disease, diabetes and cancers. Depression treatment is estimated to increase a patient's quality of life years by $15,331 and $36,467 (Schoenbaum et al. 2001). Evidence also suggests that depressive disorders are strongly correlated to the occurrence, success of treatment, and overall course of many chronic diseases such as cardiovascular disease, diabetes, and cancers as well as health risk behaviors such as obesity, and tobacco and alcohol use (Chapman, Perry, and Strine 2005). The results of our study support that incentivizing depression screening increases the likelihood of clinicians to screen for depression, which may lead to earlier diagnosis and treatment of depressive disorders resulting in improved healthcare outcomes and a decrease in healthcare associated costs (Calonge et al. 2009).

While some screening is relatively inexpensive, such as patient questionnaire tools used by clinicians to evaluate depression or lifestyle related behaviors such as tobacco and alcohol use, other screenings require expensive diagnostic studies like mammography. Clinicians assigned to the MIPS group were more likely to screen for breast and colon cancer for the Initial Visit and Costochondritis cases. There is some question about the cost-benefit of some of these high cost screenings. A 2009 systematic review performed by U.S. Preventive Services Task Force (Nelson et al. 2009) revealed mammography screening reduces mortality by up to 15% for women 39 to 49 years of age; however, data is lacking for the age group that is recommended to receive regular mammograms. On the other hand, a recent study revealed that 22% of tumors detected by mammography screening were slow growing which resulted in unnecessarily aggressive treatment (Lannin and Wang 2017). More specifically, the study suggests that these patients would have likely have died from something else without aggressive cancer treatment, additional cost, and burden of stress.

Contrary to past studies, in our experiments, MIPS increased high-cost screenings regardless of recommended guidelines (Morland et al. 2017). We found that in the Costochondritis case, where the patient did not meet all clinical indicators for breast cancer screening, 47% of clinicians ordered a mammogram. Extrapolating these behaviors to the larger healthcare system, these unnecessary screenings would result in an estimated $3.4 billion additional cost to the healthcare system using the cost rate of $704 for a mammogram if only half the U.S. population of women between the ages of 40-49 visited their primary care physician (Salzmann, Kerlikowske, and Phillips 1997, Howden and Meyer 2010). Both colonoscopies and pneumococcal vaccines were also over-ordered at a rate of 24% and 18% respectively, which would add an estimated $5 billion and $1.1 billion of unnecessary costs to the healthcare system if half of the U.S. population between the ages of 40-49 visited their primary care physician (CMS 2017, Frazier et al. 2000, Howden and Meyer 2010). Through our study, we show that MIPS can increase screenings; however, we found this comes at an additional cost to the healthcare system through the ordering of unnecessary high-cost screenings resulting in potentially $9.5 billion in overtreatment cost.

Overall, we found that MIPS did increase adherence to incentivized measures; however, this came at a cost of lower quality of care, less-satisfied patients, and a great risk for overtreatment. MIPS potentially distract clinicians from conducting accurate health histories, physical assessments, and summaries, which increases the risk of diagnostic errors. Patient satisfaction, a key metric in quality of care, was lower in the MIPS arm. Finally, outcome measures were more likely to be met under MIPS. However, clinicians paid under MIPS were more likely to overtreat to meet the outcome measure and be paid more for doing so. Incentivizing measures seems to bring similar risks that have been seen in fee-for service models, resulting in additional cost to the healthcare system through overtreatment. In spite of this evidence, we are unable to fully evaluate if the diversion of resources will reduce overall patient health as few studies satisfactorily measure the multifaceted costs of poorly conducted patient examinations and overuse or misuse of screening tools. More research is needed to determine if the benefits of early detection and intervention (both quality of life and

direct costs) outweigh the costs of expensive diagnostic testing, the over prescription of diagnostic testing, lower quality patient examinations, and lower patient satisfaction.

7. Conclusion

While past empirical studies and theoretical models suggested the negative consequences of merit-based incentive scheme (Holmstrom and Milgrom 1991, Gravelle, Sutton, and Ma 2010), this is the first study to use health care simulations to study the payment method. This approach allows us to identify several unintended consequences of an outcome-based payment scheme such as MIPS. Specifically, MIPS resulted in lower patient satisfaction, lower standards of care, and unnecessary diagnostic testing.

Our study does not, however, shed direct light on whether patient health—as opposed to their costs—is worsened by MIPS-like compensation methods. It is clear that an incomplete assessment and/or poor health history leads to health care decisions being made based on inaccurate data which can result in misdiagnoses and poor patient outcomes, including death (Asif et al. 2017, Boodman 2014). However, it is unclear whether the costs of poorly conducted patient examinations exceed the net benefits of increased screenings. We recommend that more research be conducted to estimate the costs and benefits of specific health screenings prior to the implementation of changes in compensation methods such as MIPS.

Additionally, our study provides useful information about the impact that a MIPS-like compensation schedule has on clinicians. However, it does not allow us to isolate which aspects of the MIPS treatment generated the problems. For instance, experimental studies set outside of the healthcare industry have shown that if policy makers ignore the impacts of paying too much (Beilock 2010), paying too little (Gneezy and Rustichini 2000), prosocial behavior (Mellström and Johannesson 2008, Green 2014, Bejarano, Green, and Rassenti 2016), and/or providing too many options (Ariely and Wertenbroch 2002), policies become inefficient or backfire (Kamenica 2012).

Further, by incentivizing the 5 outcome measures we also direct the clinician's attention to the measures, which alone may be substantial motivation to change clinician behavior without monetary incentives (Meeker et al. 2014). Studies in psychology

demonstrate that behavior change is influenced by social norms, subjective norms, and social pressure (Ajzen 2006). Within a given organization, norms are transmitted by people or groups with authority, whereas subjective norms consist of an individual's own norms and sense of the social pressure to perform the recommended behavior. The simulation study created an environment where the expectations for each clinician were clear and the social pressure supported following the study expectations. The monetary incentives provided additional support for the expected behaviors. Our results do not imply that the changes in behavior observed were solely a result of the monetary incentives. Additional research on how social norms, such as monitoring or public report cards, impact behavior independently of monetary incentives in a controlled environment would be useful.

In conclusion, we expect our study to be a catalyst for research utilizing experimental economics and healthcare simulations that provide better empirical foundation for the development and implementation of compensations methods. Such studies would reduce the likelihood of unintended consequences.

References:

Acheson, Louise S, Georgia L Wiesner, Stephen J Zyzanski, Meredith A Goodwin, and Kurt C Stange. 2000. "Family history-taking in community family practice: implications for genetic screening." *Genetics in Medicine* 2 (3):180-185.

Adamson, Katie Anne, Suzan Kardong-Edgren, and Janet Willhaus. 2013. "An updated review of published simulation evaluation instruments." *Clinical Simulation in Nursing* 9 (9):e393-e400.

AHRQ. 2016. Patient Safety Primer. Diagnostic Error.

Ajzen, Icek. 2006. Constructing a theory of planned behavior questionnaire. Amherst, MA.

Ariely, Dan, and Klaus Wertenbroch. 2002. "Procrastination, deadlines, and performance: Self-control by precommitment." *Psychological science* 13 (3):219-224.

Ashman, Jill J, Pinyao Rui, and Titilayo Okeyode. 2018. "Characteristics of Office-based Physician Visits, 2015." *NCHS data brief* (310):1-8.

Asif, Talal, Amena Mohiuddin, Badar Hasan, and Rebecca R Pauly. 2017. "Importance Of Thorough Physical Examination: A Lost Art." *Cureus* 9 (5).

Baker, George. 2000. "The use of performance measures in incentive contracting." *American Economic Review* 90 (2):415-420.

Beilock, Sian. 2010. *Choke: What the secrets of the brain reveal about getting it right when you have to*: Simon and Schuster.

Bejarano, Hernan Daniel, Ellen P Green, and Stephen Rassenti. 2016. "Angels and Demons: How Individual Characteristics, Behavioral Types and Choices Influence Behavior in a Real-Effort Moral Dilemma Experiment." *Frontiers in Psychology* 7:1464.

Bejarano, Hernán, Ellen P Green, and Stephen Rassenti. 2017. "Payment scheme self-selection in the credence goods market: An experimental study." *Journal of Economic Behavior & Organization* 142:396-403.

Boodman, Sandra G. 2014. Patients Lose When Doctors Can't Do Good Physical Exams. Kaiser Health News.

Brosig-Koch, Jeannette, Heike Hennig-Schmidt, Nadja Kairies-Schwarz, and Daniel Wiesen. 2016a. *Physician performance pay: Evidence from a laboratory experiment*: Ruhr Economic Papers.

Brosig-Koch, Jeannette, Heike Hennig-Schmidt, Nadja Kairies-Schwarz, and Daniel Wiesen. 2016b. "Using artefactual field and lab experiments to investigate how fee-for-service and capitation affect medical service provision." *Journal of Economic Behavior & Organization* 131:17-23.

Brydges, Ryan, Rose Hatala, Benjamin Zendejas, Patricia J Erwin, and David A Cook. 2015. "Linking simulation-based educational assessments and patient-related outcomes: a systematic review and meta-analysis." *Academic Medicine* 90 (2):246-256.

Calonge, Ned, Diana B Petitti, Thomas G DeWitt, Allen J Dietrich, Leon Gordis, Kimberly D Gregory, Russell Harris, George Isham, Michael L LeFevre, and Rosanne M Leipzig. 2009. "Screening for depression in adults: US Preventive Services Task Force recommendation statement." *Annals of internal medicine* 151 (11):784-792.

CDC. 2017. National Center for Health Statistics: Ambulatory Care Use and Physician Office Visits.

Chapman, Daniel P, Geraldine S Perry, and Tara W Strine. 2005. "PEER REVIEWED: The vital link between chronic disease and depressive disorders." *Preventing chronic disease* 2 (1).

CMS. 2015. 2015 Physician Quality Reporting System (PQRS): Implementation Guide.

CMS. 2017. "2017 ASP Drug Pricing Files." https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Part-B-Drugs/McrPartBDrugAvgSalesPrice/2017ASPFiles.html.

Cohen, Devra S, Jerry A Colliver, Michelle S Marcy, Ethan D Fried, and Mark H Swartz. 1996. "Psychometric properties of a standardized-patient checklist and rating-scale form used to assess interpersonal and communication skills." *Academic Medicine* 71 (1):S87-9.

Cox, James C, Ellen P Green, and Heike Hennig-Schmidt. 2016. "Experimental and behavioral economics of healthcare." *Journal of Economic Behavior and Organization* 131:A1-A4.

Cox, James C, Vjollca Sadiraj, Kurt E Schnier, and John F Sweeney. 2016. "Incentivizing cost-effective reductions in hospital readmission rates." *Journal of economic behavior & organization* 131:24-35.

Davis, T, J Bluhm, R Burke, Q Iqbal, K Kim, M Kokoszka, T Larson, V Puppala, L Setterlund, and K Vuong. 2012. "Diagnosis and treatment of chest pain and acute coronary syndrome (ACS)." *Bloomington (MN): Institute for Clinical Systems Improvement (ICSI)*.

Delzell, John E, Heidi Chumley, Russell Webb, Swapan Chakrabarti, and Anju Relan. 2009. "Information-gathering patterns associated with higher rates of diagnostic error." *Advances in health sciences education* 14 (5):697.

Doran, Tim, Evangelos Kontopantelis, Jose M Valderas, Stephen Campbell, Martin Roland, Chris Salisbury, and David Reeves. 2011. "Effect of financial incentives on incentivised and non-incentivised clinical activities: longitudinal analysis of data from the UK Quality and Outcomes Framework." *Bmj* 342:d3590.

Duflo, Esther. 2017. "Richard T. Ely Lecture: The Economist as Plumber." *American Economic Review* 107 (5):1-26.

Emmert, Martin, Frank Eijkenaar, Heike Kemter, Adelheid Susanne Esslinger, and Oliver Schöffski. 2012. "Economic evaluation of pay-for-performance in health care: a systematic review." *The European Journal of Health Economics* 13 (6):755-767.

Emons, Winand. 1997. "Credence goods and fraudulent experts." *The RAND Journal of Economics*:107-119.

Faul, Franz, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. 2009. "Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses." *Behavior research methods* 41 (4):1149-1160.

Faul, Franz, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. "G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences." *Behavior research methods* 39 (2):175-191.

Feingold, Daniel, Scott R Steele, Sang Lee, Andreas Kaiser, Robin Boushey, W Donald Buie, and Janice Frederick Rafferty. 2014. "Practice parameters for the treatment of sigmoid diverticulitis." *Diseases of the Colon & Rectum* 57 (3):284-294.

Frazier, A Lindsay, Graham A Colditz, Charles S Fuchs, and Karen M Kuntz. 2000. "Cost-effectiveness of screening for colorectal cancer in the general population." *Jama* 284 (15):1954-1961.

Galizzi, Matteo M, and Daniel Wiesen. 2018. "Behavioral experiments in health economics."

General, Surgeon. 2014. "The health consequences of smoking—50 years of progress: a report of the surgeon general." US Department of Health and Human Services.

Gibbons, Robert. 1998. "Incentives in organizations." *Journal of economic perspectives* 12 (4):115-132.

Gillam, Stephen J, A Niroshan Siriwardena, and Nicholas Steel. 2012. "Pay-for-performance in the United Kingdom: impact of the quality and outcomes framework—a systematic review." *The Annals of Family Medicine* 10 (5):461-468.

Gneezy, Uri, and Aldo Rustichini. 2000. "Pay enough or don't pay at all." *The Quarterly Journal of Economics* 115 (3):791-810.

Gravelle, Hugh, Matt Sutton, and Ada Ma. 2010. "Doctor behaviour under a pay for performance contract: treating, cheating and case finding?" *The Economic Journal* 120 (542).

Green, Ellen P. 2014. "Payment systems in the healthcare industry: an experimental study of physician incentives." *Journal of economic behavior & organization* 106:367-378.

Harrison, Glenn W, and John A List. 2004. "Field experiments." *Journal of Economic literature* 42 (4):1009-1055.

Hassett, James M, Karen Zinnerstrom, Ruth H Nawotniak, Frank Schimpfhauser, and Merril T Dayton. 2006. "Utilization of standardized patients to evaluate clinical and interpersonal skills of surgical residents." *Surgery* 140 (4):633-639.

Hennig-Schmidt, Heike, Reinhard Selten, and Daniel Wiesen. 2011. "How payment systems affect physicians' provision behaviour—an experimental investigation." *Journal of Health Economics* 30 (4):637-646.

HHS. 2016. HHS FY2015 Budget in Brief.

Holmstrom, Bengt, and Paul Milgrom. 1991. "Multi-task principal-agent analyses: Incentive contracts, asset ownership, and job design." *Journal of Law, Economics, & Organization* 7:24-52.

Howden, Lindsay M., and Julie A. Meyer 2010. Age and Sex Composition: 2010. edited by U.S. Department of Commerce and Economics and Statistics Administration. https://www.census.gov/prod/cen2010/briefs/c2010br-03.pdf.

Iezzi, Elisa, Matteo Lippi Bruni, and Cristina Ugolini. 2014. "The role of GP's compensation schemes in diabetes care: evidence from panel data." *Journal of health economics* 34:104-120.

Jacob, Brian A. 2005. "Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago Public Schools." *Journal of public Economics* 89 (5-6):761-796.

Jones, P Eugene. 2007. "Physician assistant education in the United States." *Academic Medicine* 82 (9):882-887.

Kaiser, Johannes. 2007. "An exact and a Monte Carlo proposal to the Fisher–Pitman permutation tests for paired replicates and for independent samples." *Stata Journal* 7 (3):402-412.

Kamenica, Emir. 2012. "Behavioral economics and psychology of incentives." *Annu. Rev. Econ.* 4 (1):427-452.

Kelley, John M, Gordon Kraft-Todd, Lidia Schapira, Joe Kossowsky, and Helen Riess. 2014. "The influence of the patient-clinician relationship on healthcare outcomes: a systematic review and meta-analysis of randomized controlled trials." *PloS one* 9 (4):e94207.

Keser, Claudia, Emmanuel Peterle, and Cornelius Schnitzler. 2014. "Money talks-Paying physicians for performance."

Lannin, Donald R, and Shiyi Wang. 2017. "Are Small Breast Cancers Good because They Are Small or Small because They Are Good?" *New England Journal of Medicine* 376 (23):2286-91.

Lazear, Edward P. 2000. "Performance pay and productivity." *American Economic Review* 90 (5):1346-1361.

Maisey, Susan, Nick Steel, Roy Marsh, Stephen Gillam, Robert Fleetcroft, and Amanda Howe. 2008. "Effects of payment for performance in primary care: qualitative interview study." *Journal of health services research & policy* 13 (3):133-139.

Meeker, Daniella, Tara K Knight, Mark W Friedberg, Jeffrey A Linder, Noah J Goldstein, Craig R Fox, Alan Rothfeld, Guillermo Diaz, and Jason N Doctor. 2014. "Nudging guideline-concordant antibiotic prescribing: a randomized clinical trial." *JAMA internal medicine* 174 (3):425-431.

Mellström, Carl, and Magnus Johannesson. 2008. "Crowding out in blood donation: was Titmuss right?" *Journal of the European Economic Association* 6 (4):845-863.

Morland, Thomas B, Marie Synnestvedt, Steven Honeywell Jr, Feifei Yang, Katrina Armstrong, and Carmen Guerra. 2017. "Effect of a Financial Incentive for Colorectal Cancer Screening Adherence on the Appropriateness of Colonoscopy Orders." *American Journal of Medical Quality* 32 (3):292-298.

Mullen, Kathleen J, Richard G Frank, and Meredith B Rosenthal. 2010. "Can you get what you pay for? Pay-for-performance and the quality of healthcare providers." *The Rand journal of economics* 41 (1):64-91.

Murray, Christopher JL, Alan D Lopez, and World Health Organization. 1996. "The global burden of disease: a comprehensive assessment of mortality and disability from diseases, injuries, and risk factors in 1990 and projected to 2020: summary."

Nelson, Heidi D, Kari Tyne, Arpana Naik, Christina Bougatsos, Benjamin K Chan, and Linda Humphrey. 2009. "Screening for breast cancer: an update for the US Preventive Services Task Force." *Annals of internal medicine* 151 (10):727-737.

Rosenthal, Meredith B, Bruce E Landon, and Arnold M Epstein. 2007. "Pay for Performance in Commercial Hmos." *The New England Journal of Medicine* 356 (8):873.

Rosenthal, Meredith B, Bruce E Landon, Sharon-Lise T Normand, Richard G Frank, and Arnold M Epstein. 2006. "Pay for performance in commercial HMOs." *New England Journal of Medicine* 355 (18):1895-1902.

Roth, Alvin E. 2002. "The economist as engineer: Game theory, experimentation, and computation as tools for design economics." *Econometrica* 70 (4):1341-1378.

Salzmann, Peter, Karla Kerlikowske, and Kathryn Phillips. 1997. "Cost-effectiveness of extending screening mammography guidelines to include women 40 to 49 years of age." *Annals of Internal Medicine* 127 (11):955-965.

Schoenbaum, Michael, Jürgen Unützer, Cathy Sherbourne, Naihua Duan, Lisa V Rubenstein, Jeanne Miranda, Lisa S Meredith, Maureen F Carney, and Kenneth Wells. 2001. "Cost-effectiveness of practice-initiated quality improvement for depression: results of a randomized controlled trial." *Jama* 286 (11):1325-1330.

Schwartz, Alan, Saul J Weiner, Frances Weaver, Rachel Yudkowsky, Gunjan Sharma, Amy Binns-Calvey, Ben Preyss, and Neil Jordan. 2012. "Uncharted territory: measuring costs of diagnostic errors outside the medical record." *BMJ Qual Saf* 21 (11):918-924.

Sherry, Tisamarie B. 2016. "A Note on the Comparative Statics of Pay-for-Performance in Health Care." *Health economics* 25 (5):637-644.

Shojania, Kaveh G, Elizabeth C Burton, Kathryn M McDonald, and Lee Goldman. 2003. "Changes in rates of autopsy-detected diagnostic errors over time: a systematic review." *Jama* 289 (21):2849-2856.

Stollman, Neil, Walter Smalley, Ikuo Hirano, Megan A Adams, Spencer D Dorn, Sharon L Dudley-Brown, Steven L Flamm, Ziad F Gellad, Claudia B Gruss, and Lawrence R Kosinski. 2015. "American Gastroenterological Association Institute guideline on the management of acute diverticulitis." *Gastroenterology* 149 (7):1944-1949.

Sutton, Matt, Ross Elder, Bruce Guthrie, and Graham Watt. 2010. "Record rewards: the effects of targeted quality incentives on the recording of risk factors by primary care providers." *Health economics* 19 (1):1-13.

Waugaman, Wynne R, and Donna J Lohrer. 2000. "From nurse to nurse anesthetist: The influence of age and gender on professional socialization and career commitment of advanced practice nurses." *Journal of Professional Nursing* 16 (1):47-56.

Werner, Rachel M, Jonathan T Kolstad, Elizabeth A Stuart, and Daniel Polsky. 2011. "The effect of pay-for-performance in hospitals: lessons for quality improvement." *Health Affairs* 30 (4):690-698.