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

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

Highlights:

- Merit-based incentive payment system can increase the number of incentivized outcome measures met.
- Merit-based incentive payment systems can lower standards of care in collecting health history, conducting a physical exam, providing a summary of patient encounter, and patient satisfaction.

Abstract:
We study the impact of a merit-based incentive payment system on provider behavior in the primary care setting using 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 alternative payment systems. Within our sample, we find that merit-based incentive payment systems increase the number of the incentivized outcome measures met, but lower overall quality of care through unintended effects on other indicators of care.

1. Introduction

Despite their popularity, studies of merit-based incentive payment schemes have not demonstrated that outcome-based payment is associated with consistent improvement in performance or decline in costs of care (Rosenthal, Landon et al. 2007, Werner, Kolstad et al. 2011, Emmert, Eijkenaar et al. 2012, Gillam, Siriwardena et al. 2012, Iezzi, Bruni et al. 2014). The failure to find significant and consistent associations is likely due to the fact that changes in compensation methods are difficult to study in real-world settings. Natural experiments and field studies rarely allow researchers to distinguish between the effects of implicit and explicit incentives, or correct for measurement errors, at least in part because the policies adopted are not truly exogenous (Cox, Green et al. 2016). It is also difficult to measure unintended consequences in the field, which implies that the effects of a change in compensation methods are rarely completely measured or assessed. In contrast, experimental methods permit both fully exogenous changes in compensation methods and a more complete measure of the consequences of such changes.

To study the potential impact of merit-based incentive payment schemes[1] on clinician behavior, we recruited practicing primary care physician assistants and nurse practitioners to participate in a healthcare simulation. The clinicians were asked to evaluate standardized patients (SP), actors with training on how to portray a specific patient case with standardized responses, and suggest diagnostic and treatment plans for each. Clinicians were assigned to two groups: a control group or a merit-based incentive payment system (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 outcome measures that they satisfied. This allows us to simulate the effects of an exogenous change in clinician incentives within a controlled environment. Measurements of patient satisfaction, standards of care, and adherence to the incentivized outcome measures were collected and compared across the two groups of clinicians.

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

Within our sample, we found that the MIPS increased the number of incentivized outcome measures met, although other unincentivized outcome measures fell. Specifically, clinicians paid under MIPS were less likely to obtain a complete patient history, conduct a physical exam, or provide a summary of the patient encounter. The MIPS group also had lower patient satisfaction scores. Further, clinicians paid under MIPS were more likely to inappropriately order the screening tests rewarded by the incentivized measures. For example, clinicians paid under the MIPS were more likely to prescribe a mammogram for a patient under the age of 50, (i.e., outside the recommended age group) than clinicians in the control group.

In our study, we confirm economic predictions and find that clinicians respond to outcome-based payment by diverting resources from unrewarded actions (e.g., thorough physical exams) to rewarded actions (incentivized outcome measures) (Holmstrom and Milgrom 1991, Prendergast 1999, Gravelle, Sutton et al. 2010, Maynard 2012, Papanicolas and McGuire 2015). Unlike previous studies, our framed field experiment allowed us to directly identify and measure unintended consequences of the MIPS incentive scheme in a controlled environment (Mullen, Frank et al. 2010, List 2011).

2. Policy Background & Approximations

On April 16th, 2015 the Medicare Access and CHIP Reauthorization Act of 2015 (MACRA) was signed into law.[2] 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 quality measures incentivized under MIPS.

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 these quality measures on performance is the focus of our study.

Quality 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).

[2] MACRA was passed by an overwhelming bipartisan majority in the House (392 Ayes and 37 Nays) and in the Senate (92 Ayes and 8 Nays). Therefore, even with the uncertain future of the Patient Protection and Affordable Care Act, the finance reform laid out in MACRA is unlikely to change without contrary evidence.

Denominator exclusions, such as patient specific medical conditions, are not included in the cases created for this study. More specifically, 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 et al. 2010, Sutton, Elder et al. 2010). In our study, if the metric was applicable to the patient case, and the clinician performed the appropriate screening, then the performance measure was met. This allows us to focus on the diversion of resources from unrewarded actions (e.g., thorough physical exams) to rewarded actions (incentivized outcome measures), by creating a clear monetary incentive for healthcare providers to meet as many outcome measures as possible without the possibility of denominator exclusions.

3. Experimental Design and Data

3.1 Framed Field Experiment

Our experimental framework uses healthcare simulations to study the potential impact of MIPS on clinician quality of care. Healthcare simulation is currently 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 SP (actor) portraying a different case in each room. Medical students are evaluated on their ability to gather 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 outcome-based payment schemes.

3.2 Experimental Design

In our experiment, primary care nurse practitioners and physician assistants were recruited and asked to create a treatment plan for three unique SPs. 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. During the instructions, clinicians were provided blank copies of patient medical records, patient-provided information forms, and documentations 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,[3] 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. 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

[3] Patients establishing new care did not have medical records.

encountering a different SP in each room.[4] After the third patient evaluation, the clinicians were asked to complete a brief survey, were paid, and were then free to leave. 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 based on their recommended treatments and compensation method. In the following section, we describe the payment treatment in detail.

Table 1: Treatments

Name	Description of Treatment
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 have the opportunity to earn a bonus of $10.00 for satisfying an outcome-based metric (for metrics see Table 5)

3.3 Treatments: Control & 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 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 necessary.

The 5 incentivized outcome-based metrics 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 measures list (PQRS) (CMS, 2017) and Healthcare Effective Data and Information Set (HEDIS). PQRS and HEDIS measures are U.S. based tool 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. The metrics were systematically selected based on two criteria, the appropriateness for the primary care environment, and that the measures could be addressed or easily modified to be addressed in a one-time patient encounter.

The one shot nature of the experiment required two of the incentivized outcome measures be evaluated differently than in actual clinical practice. 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. Furthermore, the pneumococcal vaccine 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.

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

Table 2: Incentivized Outcome Metrics

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

To create a link between the clinician's performance and their care for the patient, 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 (for more detail about each of these measures, see section 3.6). 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.

3.5 Standardized Patients (SPs)

The SP cases were created to reflect three distinctive 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 outcome metrics 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, both potentially distracted from appropriate care based upon the chief complaint.

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. In this case, all incentivized outcome metrics 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.

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

Incentivized Outcome Measures: Regardless of treatment, the success at completing the five incentivized outcome 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 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 et al. 2013). Additionally, the evaluation rubrics captured if 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 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, Bluhm 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 (Feingold, Steele et al. 2014, Stollman, Smalley et al. 2015). Finally, the standards of care rubrics were evaluated by five primary care nurse practitioners and one primary care physician assistant 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 access'. 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 access 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 of three hired Nurse Practitioners 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 medical student 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; O'Brien, 2015) by the SPs served as a proxy for patient satisfaction. The rubrics expand on current ICS evaluation tools used in clinical education (Cohen, Colliver et al. 1996, Hassett, Zinnerstrom 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 asked to complete a survey to obtain demographic and other relevant information for the study (e.g., experience, typical practice environment, etc.).

3.7 Subjects & Settings

Individuals participating in the study were required to be either a practicing primary care nurse practitioner or physician assistant. Both nurse practitioners and physician assistants bill Medicare for services rendered using their own unique National Provider Identifier number in the U.S. and have the same reporting responsibilities as physicians in primary care practices. Hence, the MIPS program will impact nurse practitioners and physician assistants.

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. On average, subjects earned $220.

Table 4: Summary of Clinician Experience by Treatment & Case

Variable	MIPS	Control	Overall
Self-Reported Typical Practice Experience (minutes)			
Minutes in Initial Visit	26.63 (13.39)	30.63 (9.67)	28.6 (11.99)
Minutes with Complaint Visit	19.11 (7.720)	19 (4.79)	19.06 (6.33)
Observed Time spent with patient (minutes)			
Initial Visit	14.49 (3.76)	12.28 (3.27)	13.39 (3.65)
Costochondritis	12.47 (3.14)	11.5 (2.8)	11.99 (2.96)
Diverticulitis	12 (3.33)	11.99 (3.04)	11.99 (3.14)
Difficulty of Cases (Scale 1-10, 10 most difficult)			
Initial Visit	5.27 (1.87)	4.75 (1.87)	5.09 (1.78)
Costochondritis	3.8 (1.57)	4.13 (2.10)	3.91 (1.73)
Diverticulitis	2.87 (1.56)	4 (1.85)	3.26 (1.71)

Notes: In rows 1-3 and 4-7, 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 8-11, 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

4. Results

Thirty-five nurse practitioners and physician assistants participated in the experiments. One participant's data was dropped as he/she was not trained as either a nurse practitioner or physician assistant. Table 4 reports the summary statistics of the participant's typical practice experience and experiment experience. 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) (ttest p-value 0.003). However, both in this study and studies in the field, clinicians spent an average of 14.49 minutes and 15 minutes with their patients in initial visits, respectively (Acheson, Wiesner et al. 2000).[5] 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 (ttest 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).

4.1 Incentivized Outcomes

Our first result is not surprising, clinicians in the MIPS treatment met significantly more of the incentivized outcome measures than those in the control for all three cases. Table 5 shows this through a comparison of the average number of incentivized measures met across treatments by case. 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 (Wilcoxon rank sum p-value=0.0258). 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 (Wilcoxon rank sum p-value=0.001). In the Costochondritis case, clinicians in the MIPS treatment met 1.82 of the 5 incentivized outcome measures on average, whereas in the control they

[5]Clinicians were provided 20 minutes with the patients or reviewing charts and an additional 5 minutes specifically for reviewing charts.

met 0.41 (Wilcoxon rank sum p-value=0.005). However, in the Costochondritis, the patient only needed 2 of the 5 incentivized outcome measures, whereas 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.

Table 5: Effect of MIPS on Quality of Care

Initial Visit	MIPS		Control
Incentivized Outcomes	3.82 (0.95)	**	2.94 (1.14)
Standards of Care	31.88 (4.91)	**	37.06 (6.40)
Patient Satisfaction	3.44 (0.63)		3.58 (0.86)
Costochondritis			
Incentivized Outcomes	1.82 (1.59)	***	0.41 (0.62)
Standards of Care	23.12 (4.81)		24.41 (5.83)
Patient Satisfaction	3.82 (0.47)		3.80 (0.66)
Diverticulitis			
Incentivized Outcomes	2.29 (1.49)	***	0.88 (0.93)
Standards of Care	29.47 (5.08)		29.06 (3.58)
Patient Satisfaction	3.54 (0.65)	*	**3.87** (0.65)
Observations	17		17

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

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, the patient in the Costochondritis case needed depression and smoking cessation screening, if the SP did not receive the screening, this was counted as 1, and 0 if the clinician recommended 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 counted 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 over-prescribe screenings for the breast cancer (47%) and 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 6: Effect of MIPS on Miuse of Medical Screening in Costochondritis Case

	MIPS		Control
Costochondritis			
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)
Tabacco 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%

Table 7 below summarizes the cost of over prescribing screening and pneumococcal vaccines. 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 et al. 1997, Frazier, Colditz et al. 2000, CMS 2017).

Table 7: Effect of MIPS on Unnecessary Costs and Inappropriate Screenings

	Inappropriate Screenings	Unnecessary Cost
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 the average cost per subject for screenings that were conducted that were not recommended for the Costochondritis Case. Standard deviations are reported in parentheses. Column 2 reports the average cost per subject paid for an incentivized metric that was not met to the standards of the experiment.

Self-Reported Screening vs. Observed Screening

Another potential drawback of the MIPS payment model is the self-reported nature of the incentivized outcome measures. Under the MIPS finance model, clinicians have an incentive to report screenings regardless of completion (Bejarano, Green et al. 2016). Table 7 also reports 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 outcomes 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 medical guidelines. Finally, clinicians under the MIPS treatment were paid

$28.00 on average for screenings that were self-reportedly prescribed, but after video review our evaluators did not find the screenings or recommendation for screenings were actually made.

4.2 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 outcome 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) (Wilcoxon rank sum p-value=0.0207).

These findings suggest the clinicians in the MIPS group may have been more focused (i.e. distracted) on meeting the incentivized outcomes, than providing care based upon the current standards for each case. Alternatively, clinicians in the MIPS group may have been so distracted by incentivized outcomes, 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 medical guidelines; yet only 23.12 standards of care were met related to the patient's acute complaint of chest pain.

4.3 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 Interpersonal Communication Skills (ICS) instrument. 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 3.54, which was significantly less that those in the control, 3.87 (Wilcoxon rank sum p-value=0.081). 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.

5. Discussion

In this section, we discuss tradeoffs between the increased number of incentivized outcome measure 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, Chumley et al. 2009). Approximately 9% of diagnostic errors lead to major errors that would have been treatable, but instead resulted in death (Shojania, Burton 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, Weiner et al. 2012, AHRQ 2016, CDC 2017). The distractions introduced by the incentivized outcomes 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 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, Kraft-Todd et al. 2014).

Nonetheless, under MIPS clinicians met more of the incentivized outcome measures, which included screenings for cancer and chronic diseases. For example, 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 et al. 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, Unützer 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, cancers as well as health risk behaviors such as obesity, and tobacco and alcohol use (Chapman, Perry et al. 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, Petitti et al. 2009).

While some screening is relatively inexpensive, such as patient questionnaire type 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 by prescribing through prescribing mammograms 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, Tyne et al. 2009) revealed mammography screening reduces mortality by up to 15% for women 39 to 49 years of age, while data is lacking for women 70 years and older, 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 from 2001-2013, 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.

Further, in this study MIPS increased high-cost screenings regardless of recommended guidelines (mammograms and colonoscopies). In this study, 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 $6.9 billion additional cost to the healthcare system using the cost rate of

$704 for a mammogram and if only half the population of women between the ages of 45-65 visited their primary care physician (Salzmann, Kerlikowske et al. 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 $9.9 billion and $2.1 billion of unnecessary costs to the healthcare system if half of the population between the ages of 45-65 visited their primary care physician (Frazier, Colditz et al. 2000, Howden and Meyer 2010, CMS 2017). This study supports that MIPS does increase incentivized 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 $18.9 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 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.

6. Conclusion

While past empirical studies and theoretical models suggested the negative consequences of pay-for-performance (Holmstrom and Milgrom 1991, Prendergast 1999, Gravelle, Sutton et al. 2010), this is the first study to use health care simulations. 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 (Boodman 2014, Asif, Mohiuddin et al. 2017). 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 cost 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 determine exactly which aspects of MIPS 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 et al. 2016), and/or providing too many options (Ariely and Wertenbroch 2002), policies become inefficient or backfire (Kamenica 2012). Additional research is thus warranted on these issues.

Further 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 individuals 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, L. S., G. L. Wiesner, S. J. Zyzanski, M. A. Goodwin and K. C. Stange (2000). "Family history-taking in community family practice: implications for genetic screening." Genetics in Medicine **2**(3): 180-185.

Adamson, K. A., S. Kardong-Edgren and J. 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, I. (2006). Constructing a theory of planned behavior questionnaire, Amherst, MA.

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

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

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

Bejarano, H. D., E. P. Green and S. 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.

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

Calonge, N., D. B. Petitti, T. G. DeWitt, A. J. Dietrich, L. Gordis, K. D. Gregory, R. Harris, G. Isham, M. L. LeFevre and R. 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, D. P., G. S. Perry and T. 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." from https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Part-B-Drugs/McrPartBDrugAvgSalesPrice/2017ASPFiles.html.

Cohen, D. S., J. A. Colliver, M. S. Marcy, E. D. Fried and M. 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): S87-89.

Cox, J. C., E. P. Green and H. Hennig-Schmidt (2016). "Experimental and behavioral economics of healthcare." Journal of Economic Behavior and Organization **131**: A1-A4.

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, J. E., H. Chumley, R. Webb, S. Chakrabarti and A. Relan (2009). "Information-gathering patterns associated with higher rates of diagnostic error." Advances in health sciences education **14**(5): 697.

Emmert, M., F. Eijkenaar, H. Kemter, A. S. Esslinger and O. Schöffski (2012). "Economic evaluation of pay-for-performance in health care: a systematic review." The European Journal of Health Economics **13**(6): 755-767.

Feingold, D., S. R. Steele, S. Lee, A. Kaiser, R. Boushey, W. D. Buie and J. F. Rafferty (2014). "Practice parameters for the treatment of sigmoid diverticulitis." Diseases of the Colon & Rectum **57**(3): 284-294.

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

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

Gillam, S. J., A. N. Siriwardena and N. 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, U. and A. Rustichini (2000). "Pay enough or don't pay at all." The Quarterly Journal of Economics **115**(3): 791-810.

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

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

Hassett, J. M., K. Zinnerstrom, R. H. Nawotniak, F. Schimpfhauser and M. T. Dayton (2006). "Utilization of standardized patients to evaluate clinical and interpersonal skills of surgical residents." Surgery **140**(4): 633-639.

HHS (2016). HHS FY2015 Budget in Brief.

Holmstrom, B. and P. Milgrom (1991). "Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design." Journal of Law, Economics, & Organization **7**: 24-52.

Howden, L. M. and J. A. Meyer (2010). Age and Sex Composition: 2010. U.S. Department of Commerce and E. a. S. Administration. https://www.census.gov/prod/cen2010/briefs/c2010br-03.pdf.

Iezzi, E., M. L. Bruni and C. Ugolini (2014). "The role of GP's compensation schemes in diabetes care: evidence from panel data." Journal of health economics **34**: 104-120.

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

Kelley, J. M., G. Kraft-Todd, L. Schapira, J. Kossowsky and H. 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.

Lannin, D. R. and S. Wang (2017). "Are Small Breast Cancers Good because They Are Small or Small because They Are Good?" New England Journal of Medicine **376**(23): 2286-2291.

List, J. A. (2011). "Why economists should conduct field experiments and 14 tips for pulling one off." The Journal of Economic Perspectives **25**(3): 3-15.

Maynard, A. (2012). "The powers and pitfalls of payment for performance." Health economics **21**(1): 3-12.

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

Mullen, K. J., R. G. Frank and M. 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, C. J., A. D. Lopez and W. H. 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, H. D., K. Tyne, A. Naik, C. Bougatsos, B. K. Chan and L. Humphrey (2009). "Screening for breast cancer: an update for the US Preventive Services Task Force." Annals of internal medicine **151**(10): 727-737.

Papanicolas, I. and A. McGuire (2015). "Do financial incentives trump clinical guidance? Hip Replacement in England and Scotland." Journal of health economics **44**: 25-36.

Prendergast, C. (1999). "The provision of incentives in firms." Journal of economic literature **37**(1): 7-63.

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

Salzmann, P., K. Kerlikowske and K. 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, M., J. Unützer, C. Sherbourne, N. Duan, L. V. Rubenstein, J. Miranda, L. S. Meredith, M. F. Carney and K. Wells (2001). "Cost-effectiveness of practice-initiated quality improvement for depression: results of a randomized controlled trial." Jama **286**(11): 1325-1330.

Schwartz, A., S. J. Weiner, F. Weaver, R. Yudkowsky, G. Sharma, A. Binns-Calvey, B. Preyss and N. Jordan (2012). "Uncharted territory: measuring costs of diagnostic errors outside the medical record." BMJ Qual Saf **21**(11): 918-924.

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

Stollman, N., W. Smalley, I. Hirano, M. A. Adams, S. D. Dorn, S. L. Dudley-Brown, S. L. Flamm, Z. F. Gellad, C. B. Gruss and L. R. Kosinski (2015). "American Gastroenterological Association Institute guideline on the management of acute diverticulitis." Gastroenterology **149**(7): 1944-1949.

Sutton, M., R. Elder, B. Guthrie and G. 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.

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