The Impact of a Merit-Based Incentive Payment System on Quality of Healthcare: An Artefactual 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. Our approach allows us to implement exogenous changes to provider’s incentives and directly measure the unintended consequences of this payment system. Within our sample, we find that merit-based incentive payment systems increase the number of the incentivized outcome measures met, however, lower overall quality 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). The failure to find significant and consistent associations is likely due to the fact that policy changes to financial models are difficult to study in real-world contexts. Natural experiments or field studies typically do not allow the researcher to distinguish between the effects of implicit and explicit incentives, to adapt for measurement error, and to implement exogenous changes to the models (Cox, Green et al. 2016). Further, it is difficult to measure unintended consequences in the field, and therefore, find an estimate for the true cost of a policy change. For these reasons, novel research methods are necessary to identify the potential impact of proposed finance reform in the healthcare industry.

To study the potential impact of merit-based incentive payment schemes on clinician behavior, we used artefactual experiments that leveraged healthcare simulations with standardized patients, actors with training on how to portray a specific patient case with standardized responses. We recruited practicing primary care physician assistants and nurse practitioners to participate in our healthcare simulation. The clinicians were asked to evaluate standardized patients (SP) and develop diagnostic and treatment plans for each. The treatment in the experiment varied the clinician’s incentive scheme. Clinicians were assigned to either a control group or 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 self-reported incentivized outcome measures that they satisfied. This approach allowed us to implement exogenous changes to clinicians’ incentives within a controlled environment. Measurements of patient satisfaction, standards of care, and adherence to the incentivized outcome measures were collected and compared across treatment groups.

---

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; however, lowered overall quality of care. Specifically, clinicians paid under MIPS had lower standards of care, which included measures of obtaining a patient history, conducting a physical exam, and providing 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 described in 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.

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). Unlike previous studies, our artefactual 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).

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 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,

---

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.
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. 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). 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 Artefactual Field Experiment

Our artefactual experimental framework replicates the primary care environment through the use of
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 artefactual field 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 form, 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, and a completed patient-provided information form reflecting the reason for the current
visit. Clinicians were given a total of 20-minutes to review the patient information, and evaluate the
patient. The clinicians could elect how to allocate their 20 minutes. After the 20 minutes expired,

---

3 Patients establishing new care did not have medical records.
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. 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 the treatment. In the following section, we describe the payment treatment in detail.

### Table 1: Treatments

<table>
<thead>
<tr>
<th>Name</th>
<th>Description of Treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>Clinicians were paid a flat rate of $200 for evaluating the three standardized patients.</td>
</tr>
<tr>
<td>MIPS</td>
<td>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)</td>
</tr>
</tbody>
</table>

#### 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

<table>
<thead>
<tr>
<th>Metric</th>
<th>Description of required actions to meet metric</th>
<th>Payment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Breast Cancer Screening</td>
<td>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.</td>
<td>$10/per patient</td>
</tr>
</tbody>
</table>
| 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

<table>
<thead>
<tr>
<th>Case</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial Visit</td>
<td>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.</td>
</tr>
<tr>
<td>Costochondritis</td>
<td>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.</td>
</tr>
<tr>
<td>Diverticulitis</td>
<td>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.</td>
</tr>
</tbody>
</table>

3.5 Standardized Patients (SPs)

The SP cases were systematically 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 or ‘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). (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 assess demographic and other relevant information to 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.
Results

Thirty-five nurse practitioners and physician assistants participated in the experiments. One participant’s data was dropped as they were 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 to spend 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, clinicians spent an average of 14.49 minutes and 15 minutes with their patients in initial visits, respectively (Acheson, Wiesner 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 reportedly the most difficult followed by the Costochondritis case and then the Diverticulitis case. However, there was no significant difference across treatments in self-reported perceived difficulty (Wilcoxon rank sum p-value 0.1473).

Table 4: Summary of Clinician Experience by Treatment & Case

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

Notes: In rows 1-3 and 4-7, data is disaggregated by self-reported time in minutes spent with patients in typical practice and observed time spent with patients in the experiment. Rows 8-11, report the subject's perceived level of difficulty of each case on a scale from 1-10, 10 being most difficult. Standard deviations reported in parentheses.

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

Clinicians were provided 20 minutes with the patients or reviewing charts and an additional 5 minutes specifically for reviewing charts.
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). Finally, 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 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

<table>
<thead>
<tr>
<th>Case</th>
<th>MIPS</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incentivized</td>
<td>3.82 (0.95)</td>
<td>2.94 (1.14)</td>
</tr>
<tr>
<td>Standards</td>
<td>31.88 (4.91)</td>
<td>37.06 (6.40)</td>
</tr>
<tr>
<td>Patient</td>
<td>3.44 (0.63)</td>
<td>3.58 (0.86)</td>
</tr>
</tbody>
</table>

| Costochondritis | Incentivized | 1.82 (1.59) | 0.41 (0.62) |
| Patient        | 3.82 (0.47)  | 3.80 (0.66) |

| Diverticulitis  | Incentivized | 2.29 (1.49) * | 0.88 (0.93) |
| Standards       | 29.47 (5.08) | 29.06 (3.58) |
| Patient         | 3.54 (0.65)  | 3.87 (0.65)  |

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 if provided, and 0 if not recommended by the clinician.

Table 6 reports the frequency undertreatment or overtreatment by each screening mechanism across treatments for the Costochondritis case. The higher the frequency reported in Table 6, the worse the outcome for the patient. Here we see that the outcome measures in the Costochondritis case were more likely to be over provided screenings for the breast cancer (47%) and colorectal cancer (24%), and the pneumococcal vaccine (18%) when the clinicians were incentivized under MIPS than in the control. In fact, there was no overtreatment in the control. Conversely, the clinicians were more likely to under provide screening for tobacco use (65%) and depression screening (94%) under the control in comparison with MIPS (35% and 71%, respectively). Table 7 below reports that the overprovision of breast screening, colorectal screening, and pneumococcal vaccine resulted in subjects being 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 $138.17 for mammograms, $170.11 for colonoscopies, and $192.63 for vaccines per patient.  

<table>
<thead>
<tr>
<th>Screening Mechanism</th>
<th>Unnecessary Cost</th>
<th>Inappropriate Screenings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial Visit</td>
<td>$8.24 (8.09)</td>
<td>-</td>
</tr>
<tr>
<td>Costochondritis</td>
<td>$9.41 (13.45)</td>
<td>$5.29 (9.43)</td>
</tr>
<tr>
<td>Diverticulitis</td>
<td>$9.41 (11.44)</td>
<td>-</td>
</tr>
<tr>
<td>Total</td>
<td>$27.06 (16.11)</td>
<td>$5.29 (9.43)</td>
</tr>
<tr>
<td>Observations</td>
<td>17</td>
<td>17</td>
</tr>
</tbody>
</table>

Notes: Column 1 reports the average cost per subject paid for an incentivized metric that was not met to the standards of the experiment. Column 2 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.

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 reports the average unnecessary cost of the 17 subjects in the MIPS treatment. Specifically, the additional cost incurred for the clinicians reporting to have completed a screening that they did not verbally conduct or prescribe in their patient encounter. On average, subjects under the MIPS treatment were overpaid by $27.06, totaling $460.02 over the 17 subjects, for services that they did not complete.

6 These figures do not include the diagnostic fees associated with review of the mammogram or colonoscopy.
In summary, MIPS increased the number of incentivized outcomes met. However, in the Costochondritis case the MIPS incentives resulted in the overprovision of expensive screening tools (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 conducted, but after video review our evaluators did not find the screenings or recommendation for screenings were actually present.

4.2 Standards of Care
Table 5 provides the second indication of the negative impact MIPS had on the clinician's performance. Table 5 reports the average number of standards of care met by treatment across each of the cases. Despite the increase of incentivized outcome measure met, clinicians in MIPS (31.88) met significantly fewer Standards of Care than under the control in 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. Additionally, 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 reports the negative impact outcome-based payment had on the clinician's patient satisfaction as evaluated by the Interpersonal Communication Skills (ICS) instrument. Here we see that participants in the outcome-based treatment received lower patient satisfaction ratings than those in the control treatment 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, participants were distracted by their incentives and rushed through the evaluation of the Diverticulitis case.

5. Discussion
Through our artefactual field experiment, we found that the number of incentivized outcome measure met was higher under MIPS than under the control; however, overall quality of care was lower. For example, clinicians under the MIPS treatment met fewer standards of care in collecting their patient history, conducting a physical exam, and providing a summary of the patient encounter in the initial visit. Further, in the acute case of diverticulitis, clinicians had lower patient satisfaction. In this section, we discuss the tradeoff between the increased number of incentivized outcome measure met, and lower standards of care and patient satisfaction.

First, in our study the addition 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. 2009).
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 healthcare. 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).

On the other hand, 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 this to the larger healthcare system, these screenings would result in an estimated X$ additional cost to the healthcare system using the Medicare cost rate of X for a mammogram. Both colonoscopies and pneumococcal vaccines were also over ordered at a rate of 24% and 18% respectively, which would add an estimated X$ and X$ of unnecessary costs to the healthcare system. 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 SX 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. However, 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 the 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.

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. We further explored these costs in the discussion section above. 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 unclear if the costs of poorly conducted patient examinations exceed the net benefits of increased screenings. Hence, we recommend that more research be conducted to measure the cost and benefits of specific health screenings prior to the implementation of policy changes brought on by MIPS.

Additionally, our study does not allow us to determine if the change in behavior is solely due to the changes in incentive structure or if it was caused by parameters used in the experiment and more research is necessary to determine appropriate parameters. 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).

Further studies in Psychology demonstrate that behavior change is influenced by normative beliefs and subjective norms (Ajzen 2006). Normative beliefs are the perceived expectations from people or groups with authority. Whereas, subjective norms consist of the social pressure to perform the behavior. The simulation study created an environment where the expectations for each study arm were clear and the social pressure supported following the study expectations. The monetary incentives helped to increase the attitudes toward the expected behaviors. With this in mind, our results do not imply that the changes in behavior are solely a result of the monetary incentives; hence, more research should be done to determine how social norms, such as monitoring or public report cards, impact behavior independently of monetary incentives in a controlled environment.

In conclusion, we hope that our study is a catalyst for research utilizing experimental economics and healthcare simulations to define the most precise a priori foundation prior to implementing policy to reduce the likelihood of unintended consequences.
References:


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


