Health Services and Outcomes Research Methodology (2022) 22:59–78
https://doi.org/10.1007/s10742-021-00253-9

Detecting bad actors in value-based payment models

Brett Lissenden[1] · **Rebecca S. Lewis**[1] · **Kristen C. Giombi**[1] · **Pamela C. Spain**[1]

Received: 12 October 2020 / Revised: 30 April 2021 / Accepted: 25 May 2021 /
Published online: 28 June 2021

Abstract
The U.S. federal government is spending billions of dollars to test a multitude of new approaches to pay for healthcare. Unintended consequences are a major consideration in the testing of these value-based payment (VBP) models. Since participation is generally voluntary, any unintended consequences may be magnified as VBP models move beyond the early testing phase. In this paper, we propose a straightforward unsupervised outlier detection approach based on ranked percentage changes to identify participants (e.g., healthcare providers) whose behavior may represent an unintended consequence of a VBP model. The only data requirements are repeated measurements of at least one relevant variable over time. The approach is generalizable to all types of VBP models and participants and can be used to address undesired behavior early in the model and ultimately help avoid undesired behavior in scaled-up programs. We describe our approach, demonstrate how it can be applied with hypothetical data, and simulate how efficiently it detects participants who are truly bad actors. In our hypothetical case study, the approach correctly identifies a bad actor in the first period in 86% of simulations and by the second period in 96% of simulations. The trade-off is that 9% of honest participants are mistakenly identified as bad actors by the second period. We suggest several ways for researchers to mitigate the rate or consequences of these false positives. Researchers and policymakers can customize and use our approach to appropriately guard VBP models against undesired behavior, even if only by one participant.

Keywords Value-based care · Unintended consequences · Outlier detection · Medicare

1 Introduction

The Patient Protection and Affordable Care Act (ACA) of 2010 allocated billions of dollars, $10 billion each decade, for the Center for Medicare and Medicaid Innovation (CMMI) to test effective and efficient ways to reduce healthcare costs and maintain or improve quality of care. Many of the CMMI initiatives are value-based payment (VBP)

✉ Brett Lissenden
blissenden@rti.org

[1] RTI International, 3040 E. Cornwallis Road, P.O. Box 12194, Research Triangle Park, NC 27709, USA

models. VBP models are performance-based payment strategies linking financial incentives to performance on a defined set of measures. The goal of VBP models is to move the U.S. toward paying providers based on quality rather than quantity of care, ideally reducing Medicare costs. Cutler and Ghosh (2012) estimated that Medicare could save $4.7–$29.0 billion annually if they paid a bundled rate for certain medical conditions. While the U.S. Department of Health and Human Services (HHS) and Centers for Medicare & Medicaid Services (CMS) have been testing several VBP models since 2005, CMMI tremendously expanded the VBP initiative. As of November 2019, CMMI identified 50 VBP models, or, equivalently, "alternative payment models", which it operates or has announced (Quality Payment Program n.d.). The models generally last five years or less with the goal of expanding successful models into permanent programs. Of the 50 VBP models, only two (the Pioneer ACO model and the Diabetes Prevention Program) have thus far been made permanent. Other examples of CMMI's VBP models include the Home Health Value Based Program, Hospital Readmission Reduction Program, and End-Stage Renal Disease Quality Incentive Program.

Evaluating the efficacy of VBP models in terms of their intended effects on both healthcare costs and quality is important, but it is also important to monitor whether the models may lead to any unintended consequences. CMMI produces evaluation reports for each model. These reports focus on a difference-in-differences methodology to measure any intended or unintended effects of the model. For example, Eibner et al. (2020) discuss unintended effects of the Medicare Advantage Value-Based Insurance Design model to increase ambulatory care sensitive inpatient utilization and emergency department utilization. Outside of the evaluation reports, in efforts that are generally not made public, CMMI also monitors model participants for compliance with model requirements and any risks to program integrity. For example, the request for applications for the Direct Contracting model (Department of Health and Human Services 2019) states that CMMI will use "[a]udits of charts, medical records, Implementation Plans, and other data … and claims analyses to identify fraudulent behavior or program integrity risks such as inappropriate reductions in care, manipulation of organizational or corporate structures to participate as one entity type versus another, efforts to manipulate risk scores for aligned populations, overutilization, and cost-shifting to other payers or populations". A general risk for most VBP models is that participants may attempt to lower costs by actively seeking lower-risk beneficiaries or by sacrificing quality of care. While CMMI's VBP models are generally designed to not reward strategies that lower the quality of care, history has shown that unintended consequences are possible, and even likely, with changing financial incentives (e.g., Alexander 2017; Alexander 2020; O'Neil et al. 2015; Damberg et al. 2014).

Even if unintended consequences are not widespread within an early model, which typically consists of voluntary participants who know they are being closely monitored, they may become a larger issue if the program is scaled up. Undesired behavior by even a small number of atypical "bad actor" participants is thus important to detect. We use the bad actor terminology to distinguish participants who engage in undesired behavior, *as a result of the model*, that would constitute an unintended consequence of the model. The unintended consequence would be detectable by an atypical trend that begins after the start of the model. For example, for a bad actor that sacrifices quality of care to achieve costs savings, we may observe a decrease in quality of care metrics once the model begins. We use the term "honest" to refer to all other participants, who are not bad actors, even if they have outlier trends driven by random change or factors unrelated to the model.

We propose a novel approach, based on ranked percentage changes within participants over time, to identify potential bad actors. After we describe the approach in detail, we

provide a hypothetical case study to demonstrate how the approach is applied. Like all evaluation approaches, there will be uncertainty involved. Some honest participants may be flagged only due to random variation—a false positive. Similarly, a bad actor may fall through the cracks and not be flagged—a false negative. We use our hypothetical case study to simulate the likelihood of false positives and false negatives resulting our approach. Particularly because the relative costs of false negative and false positives may vary by application, we also discuss how researchers might adjust the methodology to either reduce the likelihood of false positives or reduce the likelihood of false negatives depending on the costs they perceive for false positives relative to false negatives.

Unintended consequences and undesired behaviors are notoriously difficult to identify. The contribution of this paper is to propose and validate a new outlier detection methodology specifically designed to identify bad actors in VBP models *in real time*, thus allowing policymakers to improve the model before it rolls out on a larger scale. Our methodology, shown to be highly effective in a hypothetical case study, has several key advantages relative to other methodologies. Our methodology does not require a rich set of claims or other data or any advanced calculations. It only requires at least one measure, repeated over time, that is tied to an a priori model risk. Another advantage is that our approach is a rank-based approach. No assumptions or calibrations of underlying distributions are required, including how different a participant must be from its peers to trigger concern. Our approach specifically structures detection around changes before and after a model, at repeated intervals to refine the pool of candidate bad actors, but does not require the use of a control group. The methodology thus only detects participants who may have undesired behavior *related to the VBP model*, and it detects those participants soon after their undesired behavior begins. The methodology works well in coordination with other monitoring and auditing efforts, such as targeted site visits, where both qualitative and quantitative data together can be used to determine where strategic and undesired behavior is most likely to be present.

2 Related works

Existing empirical literature on the prevalence and severity of bad actors in CMMI VBP models is limited. However, there are enough studies to suggest bad actors are a relevant concern for VBP models. A large body of literature documents that physicians respond to financial incentives. Specifically, research has examined how reimbursement levels influence medical procedure choice (Clemens and Gottlieb 2014; Coey 2015; Alexander 2017; Gruber and Owings 1996; Gruber et al. 1999; Yip 1998). Alexander (2017) provides evidence that, at least in some cases, responses to financial incentives may have unintended consequences for quality of care. In particular, financial incentives to reduce the use of cesarean sections may have increased infant mortality. Alexander (2020) shows another type of unintended consequence that arose from an actual CMMI VBP model, the New Jersey Gainsharing Demonstration. Physicians earned bonuses by changing which patients were admitted to participating hospitals but did not reduce costs or change treatments conditional on patient health. The unintended shift in hospital admission patterns occurred despite having a risk adjustment mechanism within the VBP model.

Unintended consequences have been documented in other studies as well. After a 2005 increase to CMS physician payments for office-based bladder cancer care, O'Neil et al. (2015) found an increase in clinic-based procedures (the intended effect) but no decline in procedures at higher cost facilities and an increase in the rate of redundant procedures

(unintended effects). Damberg et al. (2014) reviewed early VBP literature and identified several unintended consequences of pay-for-performance programs including disregarding other clinically important areas that are not measured or incentivized by the program, avoiding the treatment of sicker patients, providing care that is not clinically recommended, and overtreating patients. Finally, Weeks et al. (2013) and Tsai and Miller (2015) discussed preliminary concerns for unintended consequences in bundled payment models, including an increase in overall episode volume, underuse of necessary services during an episode, and a reduction in access to care for sicker patients.

While identifying any unintended consequences of VBP models in aggregate is clearly important, it is also important, particularly in the early stages of the model, to identify any undesired behaviors concentrated to a small number of bad actor participants. We are not aware of any literature specifically focused on the detection of bad actors in VBP models. There are standard statistical approaches to identify outliers by examining the distance of observations from the distribution's interquartile range. A practical concern with these approaches, as discussed by Shahian et al. (2001), is that they fail to account for random variation across participants. Our approach specifically addresses random variation across participants by focusing on changes within participants over time.

More sophisticated outlier detection algorithms have been developed to examine multivariate outliers in big datasets. For example, van Capelleveen et al. (2016) provide outlier techniques to detect providers with potentially fraudulent patterns of submitted insurance claims. Their method identifies relevant metrics (e.g., reimbursement per beneficiary, number and amount of reimbursed claims), defines a distribution for each metric, and then assigns outlier scores based on standard deviations away from the mean. Their method is called an unsupervised method because it is not calibrated using data records previously identified as fraudulent or non-fraudulent. The approach we propose in this paper is also an unsupervised method. Joudaki et al. (2015) reviewed the various data mining techniques to identify health care fraud, including supervised and unsupervised methods. Other studies to use unsupervised fraud detection techniques include Lin et al. (2008) and Shin et al. (2012). Our approach differs from previously used unsupervised methods for several key reasons. First, our basic approach focuses on only one measure rather than a rich set of measures in claims data. Second, our approach focuses on changes after the start of a VBP model to detect undesired behavior that is specifically related to the model. Third, our approach is repeated over time to allow real-time identification of potential bad actors as well as refinement of the bad actor candidates over time.

The remainder of this article is organized as follows. We first discuss the setup required for our approach and describe the proposed methodology. We then illustrate the methodology with a case study using hypothetical data. We also simulate the properties of our methodology to determine how frequently it correctly identifies bad actors and how frequently it mistakes honest participants as bad actors. The discussion describes how the methodology can be tailored or extended to meet researcher needs, and then we conclude.

3 Research approach

3.1 Setup and requirements

Our approach is a rank-based approach identifying participants with the largest percentage changes since the start of the model. The approach requires a panel of data

over time for participants in the payment model. The panel data setup is a standard setup for CMMI VBP models. CMMI contracts with outside organizations to implement and monitor their models, and ensures those contracts have access to such a data panel. The data must include periods (e.g., quarters or years, depending on how frequently data measurements are updated) before and after the start of the VBP model. The approach works best if there are multiple baseline periods (i.e., periods of data before the start of the VBP model) and multiple intervention periods (i.e., periods after the start of the VBP model). Unlike a standard evaluation approach (e.g., differences-in-differences), our approach does not require data for a control group of non-participants. The goal of our approach is to detect outlier model participants rather than estimate overall model impacts. Standard evaluation approaches are useful to identify unintended consequences that are sufficiently widespread across model participants. Our approach, however, is useful regardless of how widespread unintended consequences may be. By identifying participants who are most likely to be the most egregious bad actors, our approach facilitates specialized attention to participants whose behavior may present the most risk to a scaled-up program.

The researcher must identify at least one measure in the data which is suitably related to an a priori risk of unintended consequences. For example, suppose there is concern that some participants may sacrifice quality of care to maximize their payment. Though CMMI's VBP models typically account for quality in some way, there may be important dimensions of quality that are not accounted for and could pose a risk to the model. Example measures may include process measures (e.g., depression screening, advance care planning) or outcome measures (e.g., medication adherence, patient satisfaction scores or complaints, hospitalizations, emergency department visits, end-of-life utilization). Note that measures tied to rare events may not be ideal if random variation over time tends to overwhelm changes that could be driven by undesired behavior. The most appropriate measure will depend on the context of the model. Advance care planning and end-of-life utilization metrics may be useful if the model includes a significant fraction of patients near the end of life. Patient satisfaction scores or complaints may be useful in models that push patients to use new technologies. If the researcher can identify a relevant quality measure, then our approach can be used to identify any participants that may be sacrificing quality as represented by the chosen measure. The same considerations apply to other potential risks to the model, such as a risk of participants targeting their services to lower-risk patients.

3.2 Proposed methodology

Suppose the researcher selects one measure of interest for an outlier analysis. The researcher is concerned about a bad actor who might change behavior after the start of the VBP model, in response to the VBP model, in a way that would result in a decrease in the measure (an unintended consequence). The five steps outlined in Fig. 1, and described in detail below, allow the researcher to detect potential bad actors in real-time, starting with the period in which their undesired behavior begins. Targeted actions to investigate and deter the undesired behavior before it compromises the VBP model are then feasible. Steps 1 through 4 are applied for every intervention period (or "calculation period"), and results are combined for further inference in Step 5.

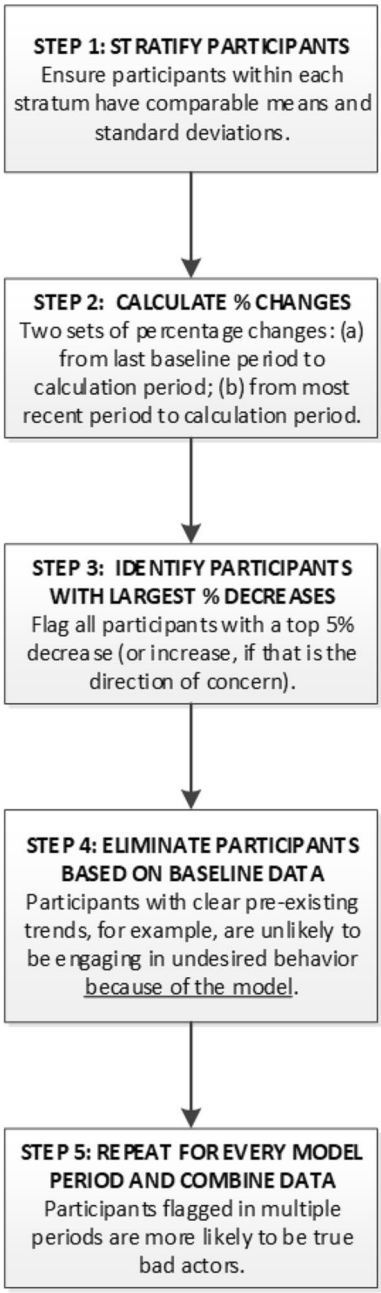

Fig. 1 Steps to identify potential bad actors

3.2.1 Step 1: Stratify participants

Stratify participants into groups where the value and variation over time in the measure is roughly uniform across all participants in each stratum. For example, participants can be stratified according to the average or standard deviation of the measure over the four

baseline periods. The size (e.g., number of patients) or geographic region of the participant may also be useful stratifiers. Stratifying in this way allows a convenient simplifying assumption that random deviations over time follow the same distribution for all participants and thus allows percentage changes (see step #2), within a participant between two periods, to be a comparable metric across participants. Note that step #4 includes suggestions to relax this assumption. The researcher can choose to keep the strata fixed over time or can re-stratify each intervention period to address relevant shifts such as participant mergers and acquisitions.

3.2.2 Step 2: Calculate percentage changes

For each participant, calculate two percentage changes. The first is the change in the measure from the last baseline period to the calculation period. The second is the change from the immediately preceding intervention period to the calculation period. When the calculation period is the first intervention period (I1), there is only one interval over which to compute a percentage change. For example, if baseline period 4 (B4) is the last baseline period, the percentage change calculation would be $B4/I1-1$.When the calculation is the second intervention period (I2), the first percentage change compares the second intervention period to the *last baseline period* (e.g., $I2/B4-1$) while the second percentage change compares the second intervention period (I2) to the *first intervention period* (e.g., $I2/I1-1$). The two sets of percentage changes are useful to detect different types of bad actors. The first percentage change, comparing to the last baseline period, allows bad actors with immediate implementation of unintended behavior to be identified. The second percentage change, comparing to the most recent intervention period, allows bad actors with delayed implementation of unintended behavior to be identified.

3.2.3 Step 3: Identify participants with largest percentage decreases

For each percentage change (step #2), within each stratum (step #1), sort the participants in ascending order from biggest decrease to smallest decrease to biggest increase. Identify the participants whose percentage decrease is in the top 5% (i.e., at the top of the sorted list). If the measure is defined in such a way that an increase, rather than a decrease, would be concerning, the researcher would then focus on the bottom 5% of the sorted list. The result is two lists of participants—(1) the participants with top 5% percentage decreases from the last baseline period to the calculation period and (2) participants with top 5% percentage decreases from the preceding intervention period to the calculation period. There may be overlap between the two lists. The union of the two lists, which contains between 5 and 10% of model participants, is the preliminary list of practices most likely to be bad actors. A key advantage of this participant identification approach is that it is robust to secular trends (e.g., seasonality, technology, policies outside of the model) affecting all participants. Regardless of whether the measure exhibited an overall decrease or increase across participants on average, this step still identifies the participants with the largest decrease or, if needed, the smallest increase.

3.2.4 Step 4: Eliminate participants based on baseline data

For each bad actor candidate, compare baseline and intervention data to refine the list of candidate bad actors. For example, consider eliminating any participants whose value of

the measure in the calculation period is higher than the value of the measure in at least one of the observed baseline periods. Participants whose pattern suggests they are more prone to large fluctuations over time, as opposed to engaging in undesired behavior because of the VBP model, would then discarded. Participants with a decreasing trend in the measure that originated before the VBP model and continues (with roughly the same slope) after the VBP model begins may also be eliminated. A trend that began before the VBP model must be unrelated to the VBP model. The latter is analogous to requiring "parallel trends" in traditional difference-in-differences analysis. The refinements in this step are not necessary, and applying them may slightly increase the risk of incorrectly eliminating bad actors. However, in many contexts these refinements may significantly reduce the risk of incorrectly flagging honest participants as bad actors.

3.2.5 Step 5: Repeat for every intervention period and combine results

The intent is to repeat the first four steps for each intervention period. Participants that were flagged as potential bad actors in multiple intervention periods are most likely to be actual bad actors. Consider requiring a participant be flagged in at least two intervention periods, or possibly two consecutive intervention periods, for it to trigger any high-cost action related to concerns of an unintended consequence.

The five steps above can easily be adjusted to fit a measure for which an increase could be a sign of unintended consequences. Similarly, the steps can be adjusted to fit combinations of measures. For example, one might require both an atypical decrease in a quality measure and an atypical decrease in a cost measure for a participant to trigger concern with respect to sacrificing quality of care. In this case the steps above can be followed independently for each measure. One may wish to relax the parameters when multiple measures are involved. For example, participants with percentage decreases in the top 10%, rather than the top 5%, of both measures may be appropriate candidates.

4 Case study

4.1 Case study design

To illustrate how the methodology is applied, and how well the methodology identifies actual bad actors, we constructed a hypothetical case study with simulated data. Our case study involves 100 participants, a reasonably representative sample size for CMMI VBP models, comprised of two bad actors and 98 honest participants. With this composition, having unintended consequences limited to 2% of participants, it would be difficult to detect unintended consequences with standard impact analyses. Our outlier approach is more promising. The researcher does not know which, if any, of the 100 participants are bad actors and aims to correctly identify any bad actors as quickly as possible. There are eight periods of data, four prior to the start of the model (baseline periods) and four after the model is in place (intervention periods). The two bad actors differ in terms of when their undesired behavior begins—one begins immediately in the first intervention period (I1), the other is delayed and begins in I3. The 98 honest participants do not engage in any undesired behavior, but their trend for any measure may look atypical by random chance. We simulated most (85) of the 98 honest participants according to the same parameters. To fully vet our approach, we also rigged some of the 98 honest participants to have different

parameters which might make it more or less likely for those participants to mistakenly appear as bad actors. The parameters for each participant are described in detail below.

In our case study, there is one hypothetical measure monitored for decreases. The measure can be thought of either as a participant-level quality measure, with higher values indicating better quality, or a participant-level risk measure, with higher values indicating a more expensive case-mix of patients. The measure is standardized, so that 1.0 represents an average value. We simulate data, for all eight periods, for all 100 participants. The 100 participants have values of the measure randomly selected from the same relatively narrow distribution, which implies that stratifying the participants (step #1) is not necessary in our case. If stratification were necessary, the remaining steps would be applied independently to each stratum.

Each of the 100 participants is assigned a simulated baseline value of the measure in the first baseline period (B1). The value is drawn from a normal distribution with mean 1 and standard deviation 0.1. The simulated value in the first period then impacts the simulated value for the remaining seven periods. The values in the seven remaining periods are drawn from normal distributions with a mean that depends on the baseline value and standard deviation 0.04. To avoid the rare outcome of negative or very small values in any period, we do not allow values below 0.5. We found 0.5 to be a sufficient threshold, and immaterial to the results of the simulation, because the occurrences of the 0.5 threshold binding (i.e., over-riding a smaller value) were extremely few. Less than 0.003% of simulated values are replaced with 0.5. The parameters for our simulation, with standard deviation across participants equal to 10% of the mean and standard deviation within participants over time equal to 4% of the mean, were (approximately) calibrated using a standardized measure from an actual CMMI VBP. In applications where these standard deviations are substantially larger, the researcher may need to stratify participants to ensure a reasonably comparable baseline for percentage change calculations across participants (see step #1 of the methodology).

The trajectory of the simulations for the last seven periods varies by 5 groups of "honest" participants and two bad actor participants. These trajectories are summarized in Table 1 and described below in detail.

For most participants, the value for the seven remaining periods is a drawn from a normal distribution with mean equal to their initial value (B1) and standard deviation 0.04. This describes the complete simulation for 85 of our 100 participants (Group 1). To highlight how the methodology functions with alternative trends, we simulate values differently for the remaining 15 participants. For 10 of the participants, we simulate a "random walk" pattern in the values over time (Group 2). Their random draw in each period is from a truncated normal distribution centered on the mean of the immediately preceding period, rather than the initial period. For example, the value for the first intervention period (I1) is drawn from a distribution centered around the value from B4, which was drawn from a distribution centered around the value for B3, and so on. This implies that large random shocks in any given period will persist over time, which may be a more realistic pattern for how some participant-specific factors (e.g., a geographic expansion) impact the measure.

For the remaining 5 participants, we use their B1 value as a key component of the mean in each period and additionally force a shock to the mean in selected periods. The shock shifts the mean by a value of 0.15, 1.5 times the initial standard deviation across practices and 3.75 times the standard deviation of changes within a practice over time.

Three of the 5 practices with shocks have placebo shocks beginning in the baseline periods: one participant has a downward shock in B2 and all future periods (Group 3), one participant has a downward shock in B4 and all future periods (Group 4), and one participant has an upward shock in B4 and all future periods (Group 5). These shocks are meant

Table 1 How values were simulated for 100 participants

Group	Honest or bad actor	Number of participants	Simulation description
Group 1	Honest	85	Random B1 value, random mean-zero deviations from B1 value in all future periods
Group 2	Honest	10	Random B1 value, random mean-zero deviations from *immediately preceding period* in all future periods
Group 3	Honest	1	Random B1 value, random mean-zero deviation from B1 value *minus a shock* in all future periods
Group 4	Honest	1	Same as Group 1 for B1–B3, random mean-zero deviation from B1 value *minus a shock in B4* and all future periods
Group 5	Honest	1	Same as Group 1 for B1–B3, random mean-zero deviation from B1 value *plus a shock in B4* and all future periods
Immediate bad actor	Bad actor	1	Same as Group 1 for B1–B4, random mean-zero deviation from B1 value *minus a shock in I1* and all future periods
Delayed bad actor	Honest (I1, I2), and bad actor (I3, I4)	1	Same as Group 1 for B1–I2, random mean-zero deviation from B1 value *minus a shock in I3* and I4

to represent events unrelated to the payment model that can affect the measure, such as a merger or acquisition.

The last two participants are given shocks that represent an unintended model consequence from the perspective of CMMI. These participants are thus considered to be true bad actors. One participant (Immediate Bad Actor) alters their behavior immediately once the model begins and thus has a downward shock in the measure beginning in the first intervention period, I1. The other participant (Delayed Bad Actor) is delayed in altering their behavior and only has a downward shock in the measure in I3 and I4.

4.2 Case study results

Using our simulated values, the methodology outlined in the previous section is applied four different times—once for each of the four intervention periods. Per the case study design, no stratification (step #1 of our methodology) was required. Per step #4 of our methodology, we eliminate practices whose value in the calculation period is lower than the value in at least one baseline period. This refinement criterion generally only eliminates a small number of participants with our simulated data. We do not eliminate practices based on a continuation of a baseline decreasing trend.

Figure 2 illustrates the results for each intervention period. The participants identified by the methodology as potential bad actors are flagged with red markers. The Immediate Bad Actor and Delayed Bad Actor participants are labeled, regardless of whether they were flagged. In I1, there is one set of percentage changes (from B4 to I1), which is plotted on the vertical axis. The horizontal axis represents participant identifiers that range from 1 to

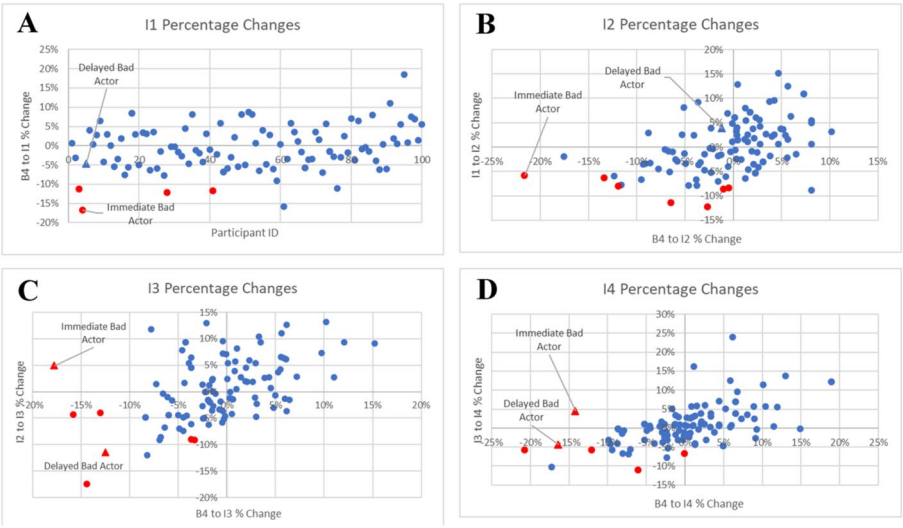

Fig. 2 Simulated percentage changes by participant for each intervention period. Source file: graphs R2.xlsx. (Supplementary information) Caption: There is one panel per calculation period—I1, I2, I3, and I4. The markers show the percentage change in the simulated measure from the last baseline period (B4) or the most recent intervention period to the calculation period. The participants flagged as potential bad actors by the methodology have red markers. The actual bad actors, the Immediate Bad Actor and the Delayed Bad Actor, have triangular markers and are labeled (Color figure online)

100. In the other intervention periods, there are two sets of percentage changes. The percentage change from B4 is plotted on the horizontal axis, and the percentage change from the preceding intervention period is plotted on the vertical axis. Note that the values and markers in the graphs are independent for each intervention period; for example, a participant flagged with a red marker in I1 is not flagged in I2 unless it meets the criteria specifically for I2.

Note that the random variation present for all participants, including the two bad actors, implies that the percentage changes for the two bad actors do not necessarily stand out dramatically from the percentage changes for the other participants. Also note that the participants with the largest percentage decreases are not always flagged. In I1, a participant with a percentage decrease of 16% is not flagged because this participant had a value of the measure in one of the baseline periods which was higher than the value in I1. In this particular simulation, the bad actors were correctly flagged in all of the periods reflecting their undesired behavior—all four intervention periods for the Immediate Bad Actor and the last two intervention periods for the Delayed Bad Actor. However, the bad actors do not always have the largest percentage decreases compared to honest participants. In each intervention period, there are several honest participants flagged.

In our example, there are 18 unique participants flagged in at least one of the four intervention periods. This includes the two bad actors and 16 honest participants. Based on this information alone, it could prove difficult for the researcher to identify which of the 18 flagged participants was actually a bad actor. One useful way to refine the list of candidate bad actors is to require participants to be flagged in at least two periods. In our example, the list of candidate bad actors would be reduced from 18 to just 3—the two bad actors and one honest participant. In other cases, considering only participants flagged in at least two periods could result in an actual bad actor (particularly a delayed bad actor) not being identified. We quantify the trade-off in our simulation section.

In practical applications it may be possible for researchers to further refine or prioritize the list of candidate bad actors by examining the full trend over time for each one. See Fig. 3 for several examples. Figure 3 includes the two bad actors (Fig. 3A, B), as well as the honest participant which was flagged in at least two intervention periods (Fig. 3C). Figure 3 also includes one other participant, which was flagged in I4 only (Fig. 3D). Visual inspection of Fig. 3 could let the researcher de-prioritize the participant shown in Fig. 3D since its atypical decrease flagged in I4 was driven more by an unusually high simulated value in I3 than an unusually low simulated value in I4. Subjective determinations based on visual inspection are not included in our simulations in the next section.

4.3 Simulated properties of case study results

In the single simulation above, the methodology correctly identified the two bad actors for a 100% true positive rate and 0% false negative rate. Further, it correctly identified the bad actors right away and in all periods they were engaged in undesired behavior—4 out of 4 periods for the Immediate Bad Actor and 2 out of 2 periods for the Delayed Bad Actor. However, it also incorrectly identified 16 out of 98 participants who were not bad actors for a 16.3% false positive rate. The identification of some honest participants in every intervention period is a mechanical feature of our setup, which includes at most two bad actors in an intervention period but generally (depending on the step #4 refinement criteria) flags more than two participants each intervention period. The properties of the simulation are more impressive with the criterion that participants must be flagged at least twice to be

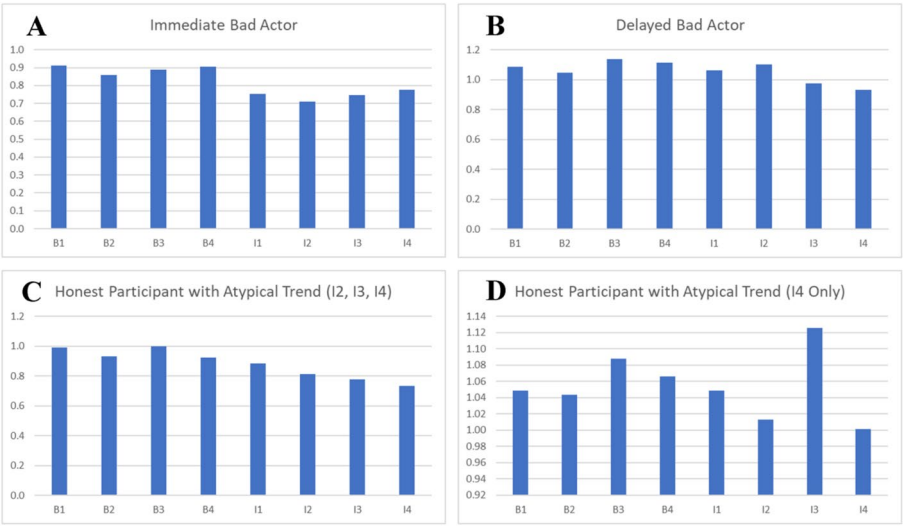

Fig. 3 Simulated values by period for selected participants. Source file: graphs R2.xlsx. (Supplementary information) Caption: There is one panel for each of four selected participants. The participants include the two actual bad actors as well as two honest participants flagged as being potential bad actors in at least one intervention period. The bars indicate the value of the simulated measure in each baseline period (B1–B4) and intervention period (I1–I4)

considered a likely bad actor. With this criterion the methodology missed 0 out of 2 bad actors (0% false negative rate) and incorrectly flagged only 1 out of 98 (1% false positive rate) honest participants. These results are dependent on the values that were simulated, however, and could vary importantly under different simulations.

To better understand the properties of our methodology, we repeated the simulation 1000 times. The key results are presented below in Tables 2 and 3. The focus is on false positive rates (incorrectly flagging honest participants) and false negative rates (incorrectly not flagging the two bad actors). These rates are presented for each intervention period separately for the Immediate Bad Actor, the Delayed Bad Actor, and the five groups of honest participants. In addition, we present statistics for the true positive rate, precision, and accuracy combined for all participants. The true positive rate is the complement of the false negative rate. Precision measures the proportion of all positives which are true bad actors. Accuracy measures the rate at which participants are correctly classified either as honest or as bad actor. Table 2 presents cumulative results for participants being flagged in at least one period up through the current period. Participants flagged in only one period may, for example, be considered low priority for further monitoring or corrective action. Table 3 presents cumulative results for participants being flagged in at least two periods up through the current period. Participants flagged in at least two periods may be considered high priority for further monitoring or corrective action.

The probability of detecting the immediate bad actor right away is 86% (Table 2), and that probability increases each period. By the second period, the probability of detecting the immediate bad actor is 96%. The 96% probability of at least one detection by I2 includes a 75% (Table 3) probability that the immediate bad actor is detected in both I1 and I2, providing a strong signal that the participant is truly a bad actor. The probability that an honest participant is flagged in both I1 and I2 is 0.5%. The pattern is similar for the delayed

Table 2 Rate of false negatives and false positives in 1000 simulations when participants flagged in any period are potential bad actors

	I1	I2	I3	I4
False positive rates (honest participants)				
Group 1 (85 participants)	3.6%	9.1%	13.2%	17.3%
Group 2 (10 participants)	1.4%	5.9%	10.3%	14.5%
Group 3 (1 participant)	5.5%	13.5%	19.5%	26.0%
Group 4 (1 participant)	1.9%	11.0%	17.7%	24.9%
Group 5 (1 participant)	0.2%	0.3%	0.3%	0.4%
Delayed bad actor	3.7%	9.4%	N/A	N/A
Total	*3.3%*	*8.7%*	*12.9%*	*17.0%*
False negative rates (bad actors)				
Immediate bad actor	14.0%	3.6%	4.8%	2.0%
Delayed bad actor	N/A	N/A	7.3%	2.3%
Total	*14.0%*	*3.6%*	*7.3%*	*2.3%*
Overall rates (all participants)				
True positive rate	*86.0%*	*96.4%*	*95.2%*	*98.1%*
Precision	*34.5%*	*18.4%*	*13.1%*	*10.5%*
Accuracy	*96.5%*	*91.4%*	*87.3%*	*83.3%*

The table indicates the results of the methodology, tracking participants flagged in at least one intervention period, after each intervention period (I1–I4)

Precision proportion of all positives which are true bad actors

Accuracy rate at which participants are correctly classified either as honest or as bad actor

Table 3 Rate of false negatives and false positives in 1000 simulations when participants flagged in multiple periods are potential bad actors

	I2	I3	I4
False positive rates (honest participants)			
Group 1 (85 participants)	0.5%	1.2%	2.1%
Group 2 (10 participants)	0.8%	3.4%	6.7%
Group 3 (1 participant)	1.3%	2.3%	3.9%
Group 4 (1 participant)	0.0%	0.2%	1.5%
Group 5 (1 participant)	0.0%	0.0%	0.0%
Delayed bad actor	0.7%	N/A	N/A
Total	*0.5%*	*1.4%*	*2.5%*
False negative rates (bad actors)			
Immediate bad actor	24.6%	9.3%	5.1%
Delayed bad actor	N/A	91.0%	25.0%
Total	*24.6%*	*50.2%*	*15.1%*
Overall rates (all participants)			
True positive rate	*75.4%*	*49.9%*	*85.0%*
Precision	*73.9%*	*41.6%*	*40.6%*
Accuracy	*99.0%*	*97.6%*	*97.2%*

The table indicates the results of the methodology, tracking participants flagged in at least two intervention periods, after each intervention period with at least one prior intervention period (I2–I4)

Precision proportion of all positives which are true bad actors

Accuracy rate at which participants are correctly classified either as honest or as bad actor

bad actor, which has a 93% chance of being detected by I3 (including the possibility of a false positive in I1 and I2). By I4, the delayed bad actor has a 98% chance of being detected at least once and a 75% chance of being detected at least twice.

Correctly identifying bad actors with more extreme versions of undesired behavior is easier, all else equal, and correctly identifying bad actors with less extreme versions of undesired behavior is harder. Similarly, for a given intensity of undesired behavior, it is harder to correctly identify bad actors with a noisier measure (i.e., one that varies more over time for reasons unrelated to the VBP model). To quantify how the intensity of undesired behavior, relative to the noise present in the measure, impacts the performance of our methodology, we repeated our 1000 simulations using two alternative shock values and holding constant the standard deviation parameters. Recall that the baseline shock was 0.15 (1.5 times the initial standard deviation across practices and 3.75 times the standard deviation of changes within a practice over time). For our sensitivity simulations, we used a lower shock of 0.10 (1.0 times the initial standard deviation across practices and 2.5 times the standard deviation of changes within a practice over time) and a higher shock of 0.20 (2.0 times the initial standard deviation across practices and 5.0 times the standard deviation of changes within a practice over time). The results are summarized in Table 4.

The false positive rates are generally consistent across the three shock values. However, the false negative rates vary notably according to the shock value. With undesired behavior that affects the measure by 10%, an immediate bad actor is missed 31.6% of the time when requiring a flag in at least two intervention periods. With undesired behavior that affects the measure by 20%, an immediate bad actor is missed only 0.5% of the time when requiring a flag in at least two intervention periods.

In addition to being easier to detect bad actors with higher shock values, it is also easier to detect bad actors who increase their undesired behavior each period. In unreported results, we found that assigning the bad actors a random walk pattern (similar to Group 2, except that a shock occurs starting in I1) led to the Immediate Bad Actor always being

Table 4 Rate of false negatives and false positives by participant type in 1000 simulations, by shock value

	Flagged by I4 in ≥ 1 intervention period			Flagged by I4 in ≥ 2 intervention periods		
Shock value:	0.10	0.15	0.20	0.10	0.15	0.20
False positive rates (honest participants)						
Group 1 (85 participants)	18.4%	17.3%	17.0%	2.3%	2.1%	1.9%
Group 2 (10 participants)	15.8%	14.5%	13.5%	7.2%	6.7%	5.6%
Group 3 (1 participant)	23.5%	26.0%	26.5%	3.9%	3.9%	4.3%
Group 4 (1 participant)	24.6%	24.9%	28.9%	2.0%	1.5%	2.0%
Group 5 (1 participant)	2.6%	0.4%	0.0%	0.0%	0.0%	0.0%
Total	*18.1%*	*17.0%*	*16.7%*	*2.8%*	*2.5%*	*2.3%*
False negative rates (bad actors)						
Immediate bad actor	13.5%	1.6%	0.0%	31.6%	5.1%	0.5%
Delayed bad actor	19.6%	2.4%	0.1%	63.6%	25.1%	4.5%
Total	*16.6%*	*2.0%*	*0.1%*	*47.6%*	*15.1%*	*2.5%*

The table indicates the results of the methodology by the fourth intervention period (I4) under alterative intensities (i.e., shocks) of unintended consequences by the bad actors

flagged in at least two intervention periods and the Delayed Bad Actor being flagged in at least one (of the possible two) intervention periods over 99% of the time.

Finally, we tested how simulation results varied with more bad actors. Since our methodology relies on outlier patterns based on ranks, the implications are somewhat nuanced. With more bad actors, the likelihood of detecting any given bad actor decreases slightly. The likelihood of detecting at least one of the bad actors, however, increases substantially. The results are illustrated in Table 5, where we shifted one of the participants out of Group 1 to be a second Immediate Bad Actor. Recall that our methodology is intended to detect at most a small number of bad actors. Detecting even one, or a small number of bad actors, allows policymakers the opportunity to learn and better safeguard scaled-up versions of the VBP model against unintended consequences. If the researcher suspects more bad actors, and is interested in identifying each of them, the threshold in step 3 of the methodology can be adjusted to be higher than 5%.

5 Discussion

We have described, applied, and tested the properties of a methodology to identify a small number of bad actor participants in a VBP model. The methodology is simple (e.g., easily programmable in Microsoft Excel), yet quite effective at correctly identifying bad actors in the hypothetical case study we constructed. The case study was devised to illustrate a realistic scenario in which our methodology would be particularly appropriate. However, the effectiveness of the methodology for other practical applications may vary. For researchers implementing our approach, we suggest to first assess how different the configuration may be from our case study. Depending on the potential differences, the researcher may wish to

Table 5 Rate of false negatives and false positives by participant type in 1000 simulations, by number of bad actors

	Flagged by I4 in ≥ 1 intervention period		Flagged by I4 in ≥ 2 intervention periods	
# Immediate bad actors	1	2	1	2
False positive rates (honest participants)				
Group 1 (84 or 85 participants)	17.3%	16.1%	2.1%	2.7%
Group 2 (10 participants)	14.5%	12.9%	6.7%	5.4%
Group 3 (1 participant)	26.0%	22.8%	3.9%	3.2%
Group 4 (1 participant)	24.9%	25.3%	1.5%	2.3%
Group 5 (1 participant)	0.4%	0.1%	0.0%	0.0%
Total	*17.0%*	*15.8%*	*2.5%*	*3.0%*
False negative rates (bad actors)				
Immediate bad actors (1 or 2)	1.6%	2.2%	5.1%	8.6%
At least one immediate bad actor*	1.6%	0.0%	5.1%	0.5%
Delayed bad actor	2.4%	2.8%	25.1%	29.2%
Total	*2.0%*	*2.4%*	*15.1%*	*15.4%*

*The rate at which both of the immediate bad actors were incorrectly not flagged

first simulate our methodology's effectiveness under case studies which may more closely reflect their data configuration.

Beyond the unguaranteed generalizability to data configurations differing from our case study, our methodology has at least two more limitations. First, like any other methodology, our approach is not perfect. There is a non-zero chance the approach will miss a bad actor, and there is always a trade-off of flagging honest participants. Second, the methodology is dependent on the researcher being able to identify and obtain a relevant measure, or measures, which would reflect unintended consequences. Identifying a relevant measure may be challenging in real-world applications, as little is known about which measures are most often tied to unintended consequences. A fruitful area of future empirical research is to scrutinize any bad actors in VBP models, quantify which measures are most reflective of their undesired behavior, and share that information with the appropriate research community.

There are several key advantages of our methodology relative to other approaches. First, since participants are compared to each other, there is no need for a control group. Identifying an adequate control group, and obtaining data from the control group for relevant measures, is often a major challenge in evaluations of models with voluntary participants. Second, with repeated measurements over time, there is no need to focus only on extreme observations as potentially representing a bad actor. Percentage changes that consistently rank at the bottom or top across all participants over time are powerful indicators of an underlying unintended consequence versus a random or unrelated trend.

The methodology is also flexible, so that the researcher can select methodological parameters (e.g., the minimum number of periods in which a participant must be flagged, the threshold for selection of top percentage changes, any refinement criterion based on baseline trends) based on his or her willingness to accept false positives. As expected, our simulation results demonstrated a trade-off between false positive rates and false negative rates. We suggested some ways in which the researcher might further reduce the trade-off (e.g., participants with pre-trends may be eliminated in step #4 of the methodology), but some degree of trade-off between false negatives and false positives is likely unavoidable.

In general, we suggest that researchers consider participants flagged in at least one period for only low-cost follow-up actions. If higher-cost follow-up actions are needed, we suggest that researchers only consider participants flagged in at least two periods in order to limit costs. Considering all participants flagged in at least one period would correctly include bad actors 98% of the time by I4 in our case study. However, many participants who are not actually bad actors would also be included—17% of honest participants would be flagged by I4 in our case study. If severe penalties were imposed on participants based on these results alone, the cost of the false positives is likely too high in most applications. Intermediate, low-cost follow-up methods may be appropriate. Intermediate and low-cost follow-up options may include visual trend inspection (e.g., Fig. 3), further quantitative investigation of related measures, or conversations with the participant that may be able to determine an alternative explanation (such as an unrelated shift in patient composition, for example). The researcher may feel comfortable concluding that some of the flagged participants are not actually bad actors based on these types of follow-up.

Even after intermediate, low-cost follow-up, the policymaker may determine that it is too costly to follow up with all candidate participants, flagged only once as possible bad actors, to the extent required to stop undesired behavior. For example, some combination of comprehensive audits, financial penalties, or corrective action plans may be required. Some required actions may also lead to reputational costs for participants. The policymaker may instead wish to target resources towards a smaller set of participants who are most likely

to be bad actors. Participants that are flagged in at least two periods are useful towards this end. Focusing only on participants flagged in at least two periods reduces the fraction of honest participants who are flagged from 17 to 3% in our case study. Similarly, the fraction of flagged participants who are bad actors increases from 11 to 41%. The trade-off is that requiring a flag in at least two periods also increases the chance of missing a bad actor from 2 to 15% (from 2 to 5% for an immediate bad actor, from 2 to 25% for a delayed bad actor). Reducing the false positive rate below 5% at the expense of a high false negative rate for delayed bad actors may be an acceptable trade-off for the policymaker, particularly if the costs of necessary corrective follow-up are determined to be high.

Relatedly, though not necessarily critical, the methodology is intended to integrate with targeted primary data collection. When monitoring VBP models, CMMI and their contractors typically interview model participant staff (e.g., physicians, practice managers, data analysts) and can ask targeted questions of VBP model participants flagged as being potential bad actors. The outlier analysis is helpful to inform which participants to interview as well as what questions to ask. These questions do not necessarily have to rely on bad actor participants being forthcoming about unintended behavior. The goal of the questions, for example, can be to determine if there were any external factors that may be able to explain atypical trends. More broadly, qualitative data can be triangulated with the quantitative outlier analysis results to help determine if a practice may truly be a bad actor versus an honest participant.

Our methodology was designed to detect unintended consequences in VBP models, such as the ones being tested by CMMI. Another application of the methodology is to the VBP models that are being used by private and state payers (Chien and Rosenthal 2019). Additionally, the methodology applies more broadly to any setting where a researcher wishes to identify participants with repeated underlying atypical behavior. For example, the methodology could be used to identify participants with the most *desirable* impacts of an intervention. Identifying participants for whom the model had the most meaningful desirable impact could allow for useful refinement of the model, including participation guidelines. The methodology could also be used to identify the participants most affected by COVID-19 or other external factors that may have lasting impacts over the intervention periods.

6 Conclusion

Unintended consequences by even a very small number of bad actors can compromise the integrity of VBP models. If there are a small number of bad actors, outlier techniques will be needed to identify them. Our proposed approach, which requires minimal data inputs and relies only on simple calculations such as percentage changes, allows researchers to identify in real-time the participants who are most likely to be bad actors. In a hypothetical case study with bad actors present, the approach is highly effective at correctly identifying the bad actors. There are trade-offs with respect to incorrectly identifying honest participants as bad actors, but the proposed approach is sufficiently flexible to allow researchers to adjust parameters based on their tolerance for false positives. More empirical work is needed to determine what strategies are commonly used by bad actors and which empirical measures those strategies impact.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10742-021-00253-9.

Author's contributions All authors contributed to the development of the methodology, the simulation of results, and the writing of the manuscript.

Funding None.

Code availability Generally not applicable; Excel formulas used can be provided.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Alexander, D.: Does physician pay affect procedure choice and patient health? Evidence from medicaid C-section use. Working Paper, No. 2017-07. Federal Reserve Bank of Chicago, Chicago, IL (2017)

Alexander, D.: How do doctors respond to incentives? Unintended consequences of paying doctors to reduce costs. J. Polit. Econ. **128**(11), 000–000 (2020)

Chien, A.T., Rosenthal, M.B.: A 3D model for value-based care: the next frontier in financial incentives and relationship support. UnitedHealthcare (2019)

Clemens, J., Gottlieb, J.D.: Do physicians' financial incentives affect medical treatment and patient health? Am. Econ. Rev. **104**(4), 1320–1349 (2014)

Coey, D.: Physician's financial incentives and treatment choices in heart attack management. Quant. Econ. J. Econ. Soc. **6**(3), 703–748 (2015). https://doi.org/10.3982/QE365

Cutler, D.M., Ghosh, K.: The potential for cost savings through bundled episode payments. N. Engl. J. Med. **366**(12), 1075–1077 (2012). https://doi.org/10.1056/NEJMp1113361

Damberg, C.L., Sorbero, M.E., Lovejoy, S.L., Martsolf, G.R., Raaen, L., Mandel, D.: Measuring success in health care value-based purchasing programs: findings from an environmental scan, literature review, and expert panel discussions. Rand Health Q. **4**(3), 9 (2014)

Department of Health & Human Services: Direct Contracting Model: Global and Professional Options Request for Applications. https://innovation.cms.gov/files/x/dc-rfa.pdf (2019)

Eibner, C., Khodyakov, D., Audrey Taylor, E., Buttorff, C., Armstrong, C., Booth, M., Bouskill, K.E., Cefalu, M., Dellva, S., Dworsky, M., Girosi, F., Haas, A.C., Martineau, M., Eshete-Roesler, B., Kim, A., Lai, J., Rastegar, A., Schwam, D., Sherry, T., Zhang, S.: Evaluation Report of the First Three Years (2017–2019) of the Medicare Advantage Value-Based Insurance Design Model Test RAND Health Care, Centers for Medicare & Medicaid Services, Santa Monica, CA https://innovation.cms.gov/data-and-reports/2020/vbid-yr1-3-evalrpt (2020)

Gruber, J., Owings, M.: Physician financial incentives and Cesarean section delivery. Rand J. Econ. **27**(1), 99–123 (1996)

Gruber, J., Kim, J., Mayzlin, D.: Physician fees and procedure intensity: the case of Cesarean delivery. J. Health Econ. **18**(4), 473–490 (1999)

Joudaki, H., Rashidian, A., Minaei-Bidgoli, B., Mahmoodi, M., Geraili, B., Nasiri, M., Arab, M.: Using data mining to detect health care fraud and abuse: a review of literature. Glob. J. Health Sci. **7**(1), 194–202 (2015). https://doi.org/10.5539/gjhs.v7n1p194

Lin, C., Lin, C.-M., Li, S.-T., Kuo, S.-C.: Intelligent physician segmentation and management based on KDD approach. Expert. Syst. Appl. **34**(3), 1963–1973 (2008). https://doi.org/10.1016/j.eswa.2007.02.038

O'Neil, B., Graves, A.J., Barocas, D.A., Chang, S.S., Penson, D.F., Resnick, M.J.: Doing more for more: unintended consequences of financial incentives for oncology specialty care. J. Natl. Cancer Inst. **108**(2), djv331 (2015). https://doi.org/10.1093/jnci/djv331

Quality Payment Program: Alternative Payment Models in the Quality Payment Program as of November 2019. Centers for Medicare & Medicaid Services, Baltimore, MD https://qpp-cm-dev-content.s3.amazonaws.com/uploads/733/2019%20Comprehensive%20List%20of%20APMs%20Nov%206.pdf (n.d.)

Shahian, D.M., Normand, S.-L., Torchiana, D.F., Lewis, S.M., Pastore, J.O., Kuntz, R.E., Dreyer, P.I.: Cardiac surgery report cards: comprehensive review and statistical critique11This review is an abridged version of a report submitted by the Massachusetts Cardiac Care Quality Commission to the Massachusetts Legislature, May 2001. Ann. Thorac. Surg. **72**(6), 2155–2168 (2001). https://doi.org/10.1016/s0003-4975(01)03222-2

Shin, H., Park, H., Lee, J., Jhee, W.C.: A scoring model to detect abusive billing patterns in health insurance claims. Expert Syst. Appl. **39**(8), 7441–7450 (2012). https://doi.org/10.1016/j.eswa.2012.01.105

Tsai, T.C., Miller, D.C.: Bundling payments for episodes of surgical care. JAMA Surg. **150**(9), 905–906 (2015). https://doi.org/10.1001/jamasurg.2015.1236

van Capelleveen, G., Poel, M., Mueller, R.M., Thornton, D., van Hillegersberg, J.: Outlier detection in healthcare fraud: a case study in the Medicaid dental domain. Int. J. Account. Inf. Syst. **21**, 18–31 (2016). https://doi.org/10.1016/j.accinf.2016.04.001

Weeks, W.B., Rauh, S.S., Wadsworth, E.B., Weinstein, J.N.: The unintended consequences of bundled payments. Ann. Intern. Med. **158**(1), 62–64 (2013). https://doi.org/10.7326/0003-4819-158-1-201301010-00012

Yip, W.C.: Physician response to Medicare fee reductions: changes in the volume of coronary artery bypass graft (CABG) surgeries in the Medicare and private sectors. J. Health Econ. **17**(6), 675–699 (1998). https://doi.org/10.1016/s0167-6296(98)00024-1