Hospital Upcoding Decisions under Medicare Audits

Jianjing Lin
Rensselaer Polytechnic Institute
Department of Economics

Juan Pantano*
University of Arizona
Department of Economics

March 29, 2022

Abstract

We propose a general model to explain hospitals' coding practices, that encompasses some of the major mechanisms advanced in the literature and propose a new channel that moderates the role of financial incentives in driving upcoding practices. We then derive testable implications and leverage aggregate Medicare Part A claims data from past decade (2012-2019) to explore which of these mechanisms best explains hospitals' admissions and billing behavior in this recent period. We also develop a novel empirical strategy based on a random sample of audits that further helps us distinguish among the alternative mechanisms. This idea, of using audits data to sort out competing hypotheses in settings with potentially inappropriate but hidden actions has broader appeal, and could be applied in other contexts. Using the evidence from both aggregate claims and audit microdata we conclude that hospital decision-makers do engage in upcoding practices, but also seem to enjoy some utility from engaging in partially-compliant behavior. In particular, their demand for compliance is subject to substitution and income effects arising from changes in reimbursement, with the income effect likely being dominant during our study period.

*Corresponding Author: juanpantano@gmail.com

1 Introduction

Medicare, on average, processes 4.5 million claims and makes $1 billion payments every day.[1] Given the increasing healthcare expenditure and a growing beneficiary population, it is important to protect the integrity of the Medicare program. Hospitals inpatient care is the largest contributor to Medicare expenditure and improper payments. A strand of literature has been devoted to exploring to what extent hospitals engage in inappropriate billing practices and what are the mechanisms underlying their behavior. Prior studies found mixed evidence of improper billing and provided different explanations. We revisit this topic by proposing a general model that characterizes hospital coding decisions and seek to identify the most prevalent mechanism by combining the analysis from the aggregate claims and more novel data from a random sample of audits.

The primary source for hospital revenue comes from the provision of inpatient care, where hospitals consider what services are provided to treat the patient. The practice by which a hospital seeks to enhance its revenue by coding a service and/or submitting a bill without adequate medical justification, is known as *upcoding*, a common type of improper billing. Hospitals that are found to engage in these activities may undergo financial and other consequences. In this paper, we explore the prevalence of these activities and the various incentives behind them.

We construct a model where hospital decision makers must decide on admission and coding assignment, following the theoretical framework by McGuire and Pauly (1991); Gruber and Owings (1996). In particular, the hospital's key decision-makers aim to maximize a utility function that depends positively on consumption and negatively on the level of non-compliance associated with their improper admissions and billing practices. Consumption is funded by (legitimate and improper) profit generated from treating patients, and the utility from compliance arises from the satisfaction of an "internal conscience" as in the work of McGuire and Pauly (1991). In this model compliance can arise through two channels. First, compliance could be an optimal profit-maximizing decision that generates higher profits, after accounting for the financial expected costs associated with the inappropriate behavior being detected. But it could also arise simply from decision maker's utility from doing the "right thing" even when at the expense of extra ill-begotten profits.

Starting from this general model, we then develop four special cases that feature each of the mechanisms that could underlie hospital coding behavior and derive testable implications regarding the level of inappropriate admissions and coding as financial incentives

[1] See https://www.aarp.org/money/scams-fraud/info-2016/federal-strategy-to-fight-medicare-fraud.html.

vary. First, inspired by the seminal work by Dafny (2005), we introduce a model where the financial incentive is the predominant factor in the utility function (*without* considering compliance in the utility function). We allow hospitals to a) admit patients unnecessarily and b) upcode rightly admitted patients into a (higher) billing tier than warranted by their medical condition. This model predicts, as no effect on total admissions and more top bill coding as the reimbursement rate for the top bill code increases. Then, we incorporate the stricter coding requirements into this model and discuss how admission and coding intensity respond to the increasing reimbursements for sicker patients that also come with the more stringent documentation requirements for top tier billing justification.

Next, we present a model specialization where both financial incentives and compliance play a role in the payoff. Similar to the model in Gruber and Owings (1996), there is a substitution effect and an income effect. The substitution effect captures the finding in prior studies that hospitals tend to upcode more when the revenue from it gets higher, because compliance becomes more "expensive." In contrast, if the income effect is the primary force, hospitals tend to engage in fewer inappropriate activities such as admitting unnecessary patients or upcoding patients, because they can "afford" to be more compliant. The ultimate effect of a change in reimbursement rate depends on the interplay between the substitution and income effects.

Finally, we consider a class of models where changes in treatment technology could drive both changes in admissions and reimbursement rates even if hospitals do not consider financial incentives in their decisions. For example, if treatment for a certain medical condition becomes less invasive leading to less costly, shorter hospital stays, it would be possible to observe more admissions (as patients who would otherwise manage their condition with medications now take up the hospital treatment) and lower reimbursements as the new treatment takes up less hospital input. To fix ideas, we in particular, describe a mechanism where hospitals are fully compliant with all the protocols but have to adjust admissions and top bill coding according to a capacity constraint. While, on average, the occupancy rate among U.S. hospitals lies between 60% and 70%, hospitals face a limit in accommodating patients due to hospital overcrowding or clinician shortage. However, with the emergence of more advanced treatment technology and better approaches for care management, the capacity constraint could be expanded if hospitals become more efficient in treating patients by applying modern technology and improving the strategies in resource allocation. As fewer resources are required to care for patients with the same condition, the reimbursement rates set by the Centers for Medicare and Medicaid Services (CMS) could also get lower.[2] In this specialized

[2] An important reason for the annual review of the weights of diagnostic related group (DRG) is to ensure the reimbursement rate reflecting the right amount of resource consumption.

model, we consider a mechanism that incorporates expansions in capacity that are tied in with decreasing reimbursement rates. The model predicts that hospitals admit and top-code fewer patients as the weight for the top bill code increases, when the constraint is binding.

We test these mechanisms by examining the testable implications in both the aggregate Medicare Part A inpatient prospective payment system (IPPS) claims data and a random sample of audit data during 2012 – 2019. Understanding which is the prevailing mechanism provides potentially important implications for hospital payment policy and other initiatives designed to curb down improper billing. For instance, if revenue enhancing is the predominant incentive that leads hospitals to engaging in upcoding, greater fraud enforcement might be needed to deter such activities. However, if coding behavior responds to the increasing coding requirements to a greater extent, policymakers might consider offering healthcare providers more instructive guidance, or even rewarding providers' efforts in producing accurate documentation. Alternatively, if the dominating income effect from the compliance utility model is the primary mechanism, policymakers might consider reinforcing the legal consequences from proper/improper billing (that could make compliance play a more important role in hospitals' decision-making) and ensure adequate compensation for treating patients (so hospitals are less likely to substitute improper billing practices for legitimate behavior). Finally, if capacity constraints constitute the main driver in hospitals' admission and coding practices, minimum government intervention is likely to be sufficient.

In the empirical analysis, we first exploit the longitudinal nature of the data to estimate fixed effect models using aggregate claims from Medicare Part A. By using fixed effects we rely on more convincing variation in reimbursements within a given diagnostic group (DRG) over time, relative to what could be done in a purely cross-sectional approach. We find that *fewer* admissions are reported with the top bill code as the reimbursement rate for this code *goes up*. While this could be interpreted as inconsistent with a model where financial incentives lead to improper admissions, our general model shows that the financial incentives model can still be valid, when augmented with either compliance utility or an assumption that increases in reimbursement are tied-in with stricter coding and documentation requirements. Then we examine the effect on another outcome—total admissions within a base DRG—and also find a negative effect on this outcome in response to increases in the corresponding top DRG weight. This finding seems less consistent with the predictions from the financial incentive model augmented by a data generating process assumption that increases in reimbursement tend to be tied in with stricter coding requirements. On the other hand, we show that a financial incentives model that is supplemented with a compliant utility feature remains consistent with the two lines of evidence. The evidence up to this point is also consistent with a model with capacity constraints and without financial incentives and with other models

of fully compliant behavior where the negative effects from reimbursement on admissions within diagnostic groups are driven by technological changes that simultaneously affect both the treatment cost (and thus reimbursement) and the number of patients that wish to treat their condition using hospital stays.

The results based on the aggregate claims data then leave us two specialized models: the one accounting for compliance utility and the one emphasizing changes in treatment technology and capacity constraints. A common route at this point would be to look for instruments or other exogenous sources of variation in reimbursement to rule out these models that assume full compliance. An alternative that we pursue is to rely on the different implications these models have when examining a random sample of audits. We separate these two models using a random sample of audits coming from the Comprehensive Error Rate Testing program (CERT), one of the auditing strategies adopted by CMS. The CERT program is unique and distinct from other national audit programs, as its review process is based on a random sample of Medicare fee-for-service (FFS) claims and, to the extent that even during an audit detection is not perfect, its results can provide a lower bound for the prevalence of Medicare improper billing. Moreover, we base our analysis on audit outcomes that clearly distinguish improper from proper billing by audit reviewers, enabling us to identify the key mechanism driving our results.

We define upcoding-related errors from the CERT data and estimate the effect of financial incentives on the probability of observing these errors using a similar fixed effects specification. We find that the probability of claims being flagged with errors associated with upcoding is lower when DRGs are subject to higher reimbursement rates, suggesting that the specialized model that focuses on technology changes and capacity constraints is not the primary mechanism. Instead, it conforms to the mechanism where financial incentives matter but they are modulated by a demand for compliant behavior which is itself driven by a strong income effect in the compliance utility model. Taken together, the findings in the aggregate claims data and the random sample of audit data suggest that the the financial incentives model augmented with compliance utility and a strong income effect in a hospital decision maker's demand for compliant practices seems to explain billing coding patterns during the period covered in this study.

Related literature. Upcoding and related practices that seek to improperly increase revenue for health care providers and/or private health insurers have been analyzed for different segments of the U.S health care system. For example Geruso and Layton (2020) focus on upcoding in the Medicare Advantage program, Fang and Gong (2017) explore physician overbilling in Medicare Part B, and Finkelstein et al. (2017) explore potential issues with the risk adjustment system in the Accountable Care Act exchanges where there

might be differential coding intensity across providers in different regions. We focus on the subset of this literature that examines the appropriateness of hospitals' coding practices in the Medicare Part A program following the seminal work of Dafny (2005).

Following Dafny (2005), a growing literature in economics explores how hospitals respond to price changes. Dafny (2005) found substantial upcoding but no increase in medical admissions for conditions that underwent large increases in reimbursement as a result of a 1988 payment reform. Silverman and Skinner (2004) further examined hospital upcoding behavior by ownership and found that for-profit hospitals experienced the largest increase in patients assigned to the most lucrative DRG for pneumonia and respiratory infections. A more recent paper by Cook and Averett (2020) found that approximately 3% of reimbursement arose from upcoding after the 2007 IPPS payment reform. We contribute to this literature by exploring data from audits, and identifying whether cases of improper payment due to upcoding are more likely to be identified in the audits when reimbursement for certain conditions increases. A recent paper by Shi (2021) also used audit data (from a different Medicare audit program—Recovery Audit Program) analyzing the costs and benefits of monitoring healthcare provider behavior.

Finally, our paper contributes to the literature on healthcare providers' decision-making about the provision of care. A strand of this literature specifically studied the physicians' decision to induce demand for their services (Dranove, 1988; McGuire and Pauly, 1991; Gruber and Owings, 1996). We build on these "inducement" models by introducing a preference for compliance into a hospital decision-maker's payoff. Key players in hospitals' decision-making, whether clinicians or administrators, are likely to care about this "internal conscience" to some extent. Starting from this general model, we are able to develop specialized models that feature specific mechanisms that have been advanced in the literature and motivate the empirical analysis.

The rest of the paper proceeds as follows. Section 2 introduces the institutional background. Section 3 discusses the general model and the various specializations that feature specific mechanisms. In this section we also derive testable implications that can be taken to the aggregate claims data. Section 4 describes the datasets and reports summary statistics. Section 5 presents the empirical strategy that is used to test the implications from the various models along with results based on the aggregate data for Medicare Part A IPPS claims. Section 6 derives the testable implications for different mechanisms based on the CERT data. Section **??** discusses the empirical analysis based on the CERT data. The last section concludes.

2 Background

2.1 Medicapre Part A IPPS

Medicare pays each inpatient admission based on a flat rate payment system, called the inpatient prospective payment system (IPPS). The payment for each inpatient admission depends on the assigned diagnostic related group (DRG), which is determined according to the patient's primary diagnosis/procedure, additional diagnoses/procedures, and discharge status. Each DRG is associated with a weight that is set by CMS, representing the average resources required to care for Medicare patients in that particular DRG. The flat rate paid for each admission is equal to the DRG weight multiplied by a base rate that depends on an area cost factor to account for geographic heterogeneity in the cost of hospital inputs.

Typically, an admission is first assigned to a base DRG, based on the patient's primary diagnosis or primary procedure. Then the admission is categorized into an exact DRG, depending on the presence/absence of complicating conditions (CCs) or major CCs (MCCs). A base DRG could include one to three associated DRGs. For instance, DRGs 88 – 90 are "concussion w/ MCC," "concussion w/ CC," and "concussion w/o CC/MCC," respectively, all belonging to the same base DRG. We call each DRG within a base DRG a tier, a level, or a severity subclass interchangeably, and specifically refer to the most severe DRG as the top bill code. CMS publish a list of CCs and MCCs such that, whenever a patient is coded with a secondary diagnosis from that list, s/he will qualify for a more severe DRG. The one with more complications implies a higher severity level and results in greater reimbursements, but it also imposes stricter requirements on the justifying documentation that needs to be submitted with the claim.

The DRG-based IPPS was first established in 1983 and it undergoes an annual review of its DRG classification and adjustment to the DRG weights. The periodic recalculation of DRG weights aims to capture the changes (increases or decreases) in the necessary use of resources and inputs, which might arise from the presence of new treatments/technology, improved efficiency in care management, and any other factors that will affect the resources needed to care for the patients.[3] A significant payment reform to IPPS occurred in October 2007, in which CMS reclassified some of the DRGs, revised the CC/MCC list, and adjusted DRG weights, with the goal to better align payments with the resources used by hospitals. One of the most dramatic outcomes from this reform was that the percentage of patients who qualified for a DRG with (at least) CCs dropped from 77.7% (based on the pre-reform

[3]See https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/MS-DRG-Classifications-and-Software.

criteria) to 40.3% (based on the post-reform criteria).[4] Recently, CMS again suggested severity-level changes in the proposed rule for IPPS in Fiscal Year (FY) 2020, with the majority of adjustments downgrading MCCs to CCs or CCs to non-CCs.[5] All of these efforts reflect a consistent CMS goal to offer adequate compensation for treating ill patients while maintaining the accuracy of Medicare billing.

2.2 CERT Audits

CMS have adopted multiple audit strategies to ensure the accuracy of Medicare payments, including the Comprehensive Error Rate Testing (CERT) program, the Recovery Audit Program, Medicare Administrative Contractors, Supplemental Medical Review Contractors, and the Zone Program Integrity Contractor audits. The CERT program is unique and distinct from these other programs because it is the only one that reviews claims based on a random sample. The results provide a reasonable estimate of the overall prevalence of improper bills in Medicare, and thus, CMS rely on the results from CERT to identify program vulnerabilities and also use them as guidance to improve reimbursement mechanisms and related policies. To the extent that auditors fail to detect some of the upcoding that hospitals engage in (even when those cases are randomly selected for audit) the estimates from CERT provide a lower bound for the true prevalence of upcoding.

The CERT program randomly select approximately 50,000 Medicare fee-for-service (FFS) claims for review, following a stratification strategy.[6] For each claim selected for review, the CERT contractors send a request to providers for all related documents. If the documentation is available, providers have incentives to respond as failure to submit these documents could lead to CMS recouping the Medicare payments for that claim. CERT auditors evaluate all supporting documentation before making a determination. Claims that are found with improper billing will be categorized into one of the following payment error types: medical necessity, no documentation, insufficient documentation, incorrect coding and others.[7] Medical necessity refers to a situation where the appropriate medical service should have been rendered in an ambulatory setting instead of an inpatient hospital admission. The other four errors involve inaccurate coding for inpatient services or inadequate justification due to the

[4] See Office of the Federal Register and National Archives and Records Service (2007), p. 47,153 – 47,154.

[5] See https://www.federalregister.gov/documents/2019/05/03/2019-08330/medicare-program-hospital-inpatient-prospective-payment-systems-for-acute-care-hospitals-and-the.

[6] The selection is based on claims types, including Part A IPPS, Part A excluding IPPS, Part B, and Part B durable medical equipment, prosthetics, orthotics and supplies (DMEPOS). For each claim type, CMS further divide the claims into approximately 100 service levels except for Part A excluding IPPS, which contains fewer than 20 strata due to the difficulty in service level stratification.

[7] See CERT Program (2019) for more detail on the definition of each error type.

lack of documentation, but the service considered was provided in the right setting.

In this paper, we group the four error types other than medical necessity together and call them upcoding-related errors in our analysis based on the CERT data, as each of them most likely reflects upcoding in one way or another.

3 Theoretical Framework

In this section, we first describe a model that characterizes hospital decisions on admission and tier coding assignment. We then discuss four specialized mechanisms that generate different predictions on the admission and coding behavior. Note that our main focus is on the implications on coding behavior, but we also analyze the impact on admissions, which help distinguish between different mechanisms.

3.1 A General Model

We develop the model by adapting the theoretical insights about physician behavior from McGuire and Pauly (1991) and Gruber and Owings (1996) to a hospital decision-making context. In our model hospital decision makers do not simply seek to maximize profits, but rather a utility function that depends positively on consumption and compliance with the billing and coding protocols. The level of non-compliance in admission and coding practices is defined as the extent of deviation from the appropriate levels.

For a given base DRG j, we let s index the severity of the patient's condition. s follows a uniform distribution between 0 and 1. Also, we let the capacity to accommodate patients be denoted by $\bar{u}$. Let $\underline{s}^a$ denote the level of severity above which a patient should be admitted following the protocol. Similarly, let $\underline{s}^u$ denote the condition severity above which a patient should be coded into the top billing tier following the proper coding guidelines. For each visit i, hospital decision-makers choose whether to admit the patient: $a_i = 1$ if they admit her and $a_i = 0$ otherwise. Similarly, they decide whether to top-code a rightly admitted patient: $u_i = 1$ if yes and $u_i = 0$ otherwise. Hospitals always admit every patient with $s \geq \underline{s}^a$ and, among admitted patients, always code into the upper tier those who, in addition, have $s \geq \underline{s}^u$. For the remaining cases, decisions must be made on admission for those medically unnecessary cases ($s < \underline{s}^a$) and tier coding assignment for those who are rightly admitted but should be coded into the lower tier ($\underline{s}^a \leq s < \underline{s}^u$). The model is decentralized at the DRG level, effectively assuming that a different decision maker with similar utility function is in charge of admissions and tier coding for each DRG.[8]

[8] We abstract from a more complex modeling where a single decision maker would decide across all DRGs

Hospitals get reimbursed for each admission belonging to the base DRG j by CMS with a baseline amount B multiplied by the DRG weight. ω^h and ω^ℓ, respectively, denote the DRG weight for the bottom and top tier within a base DRG. For simplicity, we model base DRGs with only two tiers, though our empirical work extends to base DRGs with three tiers.

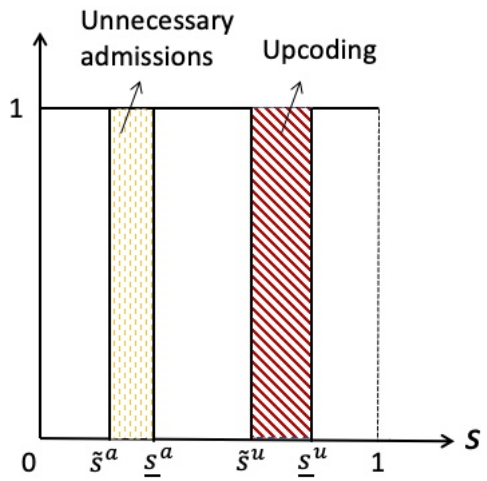
Decision-makers at the hospitals may decide to admit patients unnecessarily (inpatient admissions not justified by medical necessity) or "upcode" properly admitted patients, that is, improperly classify patients into the upper tier when they should not, as they stand to profit by doing so. They do however expose themselves to being audited and having to return the extra reimbursement and suffer additional future costs. Thus, hospital decision-makers must choose the extent to which they allow unnecessary admissions and inappropriately top-code patients. In doing do they trade-off profit motives with the expected costs from audits and the disutility from non-compliance.

Let $\tilde{s}^a$ denote the *least* severe patient that does *not* qualify for admission but gets admitted, i.e., $\tilde{s}^a = \min\{s_i : \ s_i < \underline{s}^a;\ a_i = 1\}$. Similarly, let $\tilde{s}^u$ denote the *least* severe patient that is top-coded but does *not* meet the threshold for top-tier coding, i.e., $\tilde{s}^u = \min\{s_i : \underline{s}^a < s_i < \underline{s}^u;\ u_i = 1\}$. Note that implicit in the definition of $\tilde{s}^u$ is an assumption we view as a reasonable approximation to the set of opportunities the hospital faces to engage in improper practices. While mild cases can be unnecessarily admitted ($\tilde{s}^a \leq s < \underline{s}^a$), they cannot also be upcoded. In other words, some cases are at the margin of being admitted unnecessarily and some cases are at the margin of being coded inappropriately into the top tier, but no case can be both, unnecessarily admitted *and* upcoded. Similarly, we assume that hospital decision-makers only consider marginal cases—those close to $\underline{s}^a$—for unnecessary admission given that the cost of doing so is too high among the cases farther away from the cutoff.

Figure 1 displays graphically the key elements of the model. In particular, it is easy to see in the figure that in our model, for each base DRG, we have $0 < \tilde{s}^a < \underline{s}^a < \tilde{s}^u < \underline{s}^u < 1$. The dark-shaded area captures the extent of upcoding, whereas the light-shaded area captures the extent of medically unnecessary admissions.

within a hospital but the main insights we derive do carry over to that model.

Figure 1: Distribution of Potential Patients' Condition for a Base DRG
Medically-Appropriate Thresholds $(\underline{s}^a, \underline{s}^u)$ vs. Threshold Choices $(\tilde{s}^a, \tilde{s}^u)$



The decision-maker in charge of hospital admission and coding practices then solves:

$$
\begin{aligned}
&\max_{\{\tilde{s}^a,\, \tilde{s}^u\}} E\left[U(c,m)\right] \\
\text{s.t. } c = &\int_{\tilde{s}^a}^{\tilde{s}^u} \Big[B\omega^\ell - c^\ell - E[\zeta^\ell(s)]\Big] ds + \int_{\tilde{s}^u}^{1} \Big[B\omega^h - c^h - E[\zeta^h(s)]\Big] ds; \\
&\tilde{s}^a \in (0, \underline{s}^a); \quad \tilde{s}^u \in (\underline{s}^a, \underline{s}^u); \quad 1 - \tilde{s}^a \leq \bar{u}; \\
&m = \tilde{s}^a + \tilde{s}^u - \underline{s}^a; \\
&E[\zeta^\ell(s)] = \mathbb{1}\{s < \underline{s}^a\} \times \phi^{\text{AUDIT}} \times \phi^a(\underline{s}^a - s) \times \xi^a; \\
&E[\zeta^h(s)] = \mathbb{1}\{s < \underline{s}^u\} \times \phi^{\text{AUDIT}} \times \phi^u(\underline{s}^u - s) \times \xi^u;
\end{aligned}
\tag{1}
$$

where m denotes the level of compliance; c^h and c^l denote the cost of treating the patient at the corresponding tier the patient is assigned to. $E[\zeta^\ell(s)]$ ($E[\zeta^h(s)]$) denotes the expected cost from inappropriate admissions (upcoding) if the hospital is found engaging in these practices during an audit. ϕ^{AUDIT} refers to the probability of an admission being selected for an audit, whether by the CERT program or other audit programs more generally. ϕ^a and ϕ^u denote, respectively, the probabilities that, in the event of an audit, auditors detect medically unnecessary admissions and upcoding. Moreover, we assume that the probability of getting detected increases with the deviation from the appropriate level of coding, that is, $\phi^a(\cdot)$ ($\phi^u(\cdot)$) increases with $|\underline{s}^a - s|$ ($|\underline{s}^u - s|$). ξ^a and ξ^u denote, respectively, the costs associated with an audit determination of improper payment due to a medically unnecessary admission and upcoding. Note that our model incorporates both the cost and benefit to improper billing, with the former captured by $E[\zeta^\ell(s)]$ and $E[\zeta^h(s)]$ including the possibility

of being detected in an audit and the associated costs, and with the latter captured by the financial payoff, ω^ℓ and ω^h.

Note that the expected cost of either unnecessarily admitting patients or upcoding them is an increasing function of the deviation from the appropriate cutoff. Therefore, hospitals will always find it optimal to first improperly admit or upcode the marginal patients, that is, those whose severity level is below the threshold but relatively close to it.

3.2 Specialized Models

In the following, we discuss four specialized models each of which features a specific mechanism from the general model we developed above. For each of them, we describe the specific setup and derive the implications on admission and coding practices as the reimbursement rate varies. We start with the two mechanisms that have been advanced in the literature and then introduce the mechanism we propose as well as an alternative model that could generate similar predictions.

3.2.1 Financial Incentive Mechanism

This specialized model is inspired by the seminal work by Dafny (2005), who studied hospital coding behavior in response to a fee change in the IPPS in 1988. In this specialized model, hospitals are profit-maximizing entities, without considering the disutility of non-compliance. In other words, the utility for a hospital decision-maker, which does not involve the level of compliance, depends on $\tilde{s}^a$ and $\tilde{s}^u$ only through their effects on profits and ultimately consumption.[9] Assuming $\bar{u}$ is large enough, we revise the model as follows:

$$\begin{aligned}
&\max_{\{\tilde{s}^a,\, \tilde{s}^u\}} E[U(c)] \\
\text{s.t. } c = &\int_{\tilde{s}^a}^{\underline{s}^a} \left(\pi^\ell - \phi^{\text{AUDIT}} \times \phi^a(\underline{s}^a - s) \times \xi^a\right) ds + \int_{\tilde{s}^u}^{\underline{s}^u} \left(\pi^h - \pi^\ell - \phi^{\text{AUDIT}} \times \phi^u(\underline{s}^u - s) \times \xi^u\right) ds \\
&+ (\underline{s}^u - \underline{s}^a)\pi^\ell + (1 - \underline{s}^u)\pi^h
\end{aligned} \tag{2}$$

where $\pi^\ell = B\omega^\ell - c^\ell$ and $\pi^h = B\omega^h - c^h$ denote the operating profits from coding a hospital admission into a low or high billing tier not taking yet into account the possible financial costs from a potential audit.

[9] For simplicity we are assuming this decision-maker is the owner in the sense that he or she consumes all of the profit. But similar implications arise if he or she is only entitled to a fraction of the hospital profits.

To condense notation let

$$
\begin{aligned}
f(x_1, x_2) &= \int_{x_1}^{x_2} \left[\phi^{\text{AUDIT}} \times \phi^a(\underline{s}^a - s) \times \xi^a \times \frac{1}{x_2 - x_1} ds \right] \\
g(y_1, y_2) &= \int_{y_1}^{y_2} \left[\phi^{\text{AUDIT}} \times \phi^u(\underline{s}^u - s) \times \xi^u \times \frac{1}{y_2 - y_1} ds \right]
\end{aligned} \tag{3}
$$

where $x_1,\ x_2 \in [0, \underline{s}^a]$ and $y_1,\ y_2 \in [\underline{s}^a, \underline{s}^u]$. $f(x_1, x_2)$ captures the average expected costs of detection per inappropriate hospital admission patient whose severity level falls into the range of $[x_1, x_2]$. Likewise, $g(y_1, y_2)$ captures the average expected costs of detection per inappropriate upcoding patient whose severity level falls into the range of $([y_1, y_2])$. For instance, $f(\tilde{s}^a, \underline{s}^a)$ and $g(\tilde{s}^u, \underline{s}^u)$ measure the average expected costs from engaging in improper admissions and coding practices, respectively. Thus, the maximization problem becomes

$$
\begin{aligned}
&\max_{\tilde{s}^a,\, \tilde{s}^u} E[U(c)] \\
\text{s.t. } & c = \Pi^I + \Pi^P \\
& \Pi^I = (\underline{s}^a - \tilde{s}^a)\left(\pi^\ell - f(\tilde{s}^a, \underline{s}^a)\right) + (\underline{s}^u - \tilde{s}^u)\left(\pi^h - \pi^\ell - g(\tilde{s}^u, \underline{s}^u)\right) \\
& \Pi^P = (\underline{s}^u - \underline{s}^a)\pi^\ell + (1 - \underline{s}^u)\pi^h
\end{aligned}
$$

where Π^P collects total profits from proper admission and billing practices and Π^I collects total profits from improper ones. Prior studies have found that hospitals tend to upcode more when the revenue from coding becomes higher (Silverman and Skinner, 2004; Dafny, 2005; Jürges and Köberlein, 2015). Thus, we consider an increase in ω^h, which leads to an increase in π^h. Note that in practice c^h may also change in the same direction as ω^h. For instance, an important reason for the adjustment of DRG weights arises from the change in the treatment cost. Throughout the paper, we assume that $\Delta B \omega^h > \Delta c^h$.[10] We present the first testable proposition as follows:[11]

Proposition 1 As ω^h and thus π^h gets higher, we expect no changes in total admissions within the base DRG, i.e., $\partial \tilde{s}^a / \partial \pi^h = 0$, and more admissions to the high-severity level, i.e., $\partial \tilde{s}^u / \partial \pi^h < 0$.

[10] That is we assume that the absolute change in a DRG weight times the baseline reimbursement B is greater than the variations in the treatment cost. This assumption is not innocuous. For instance, the increase in reimbursement rate might not be able to cover the rising treatment cost for some conditions. However, we follow the related studies that focus on the effect of revenue on coding that rely on a similar assumption.

[11] Additional technical assumptions that we maintain include: $\frac{\partial f(\tilde{s}^a, \underline{s}^a)}{\partial \tilde{s}^a} = f_1 < 0$, $\frac{\partial f(\tilde{s}^a, \underline{s}^a)}{\partial \underline{s}^a}|_{d\tilde{s}^a = 0} = f_2 > 0$, $\frac{\partial^2 f(\tilde{s}^a, \underline{s}^a)}{(\partial \tilde{s}^a)^2} = f_{11} \geq 0$, $\frac{\partial g(\tilde{s}^u, \underline{s}^u)}{\partial \tilde{s}^u} = g_1 < 0$, $\frac{\partial g(\tilde{s}^u, \underline{s}^u)}{\partial \underline{s}^u}|_{d\tilde{s}^u = 0} = g_2 > 0$, and $\frac{\partial^2 g(\tilde{s}^u, \underline{s}^u)}{(\partial \tilde{s}^u)^2} = g_{11} \geq 0$

Proof. See Appendix I.

With the evolving changes in healthcare delivery, such as the shift to outpatient care and increasing availability of post-acute care facilities, inpatient admissions are on average more severe, and subject to more acute conditions. In response to that, CMS have been constantly revising DRG classification and payments, with the goal of better aligning reimbursements with treatment costs and ensuring the accuracy of Medicare billing.[12] Thus, an increased severity of illness adjustment payment could usually come with stricter requirements in coding and documentation, as can be seen in the 2008 DRG payment reform (Sacarny, 2018; Gowrisankaran et al., 2019) and the recent proposal of revising CC/MCC list by CMS.[13] In the following, we consider the effect of a change in ω^h with a simultaneous change in $\underline{s}^u$ to capture the presence of increasing reimbursement for severe illness tied with stricter coding requirements, that is, an increase in both the financial payoff from top tier billing and change in the relevant threshold for what it takes in terms of severity to be classified into the top tier. Specifically, we assume $cov(\omega^h, \underline{s}^u) > 0$.[14]

Since two parameters, ω^h and $\underline{s}^u$, vary at the same time, we first derive how $\tilde{s}^u$ and $\tilde{s}^a$ vary in response to the change in $\underline{s}^u$ and then discuss the combined effect from a simultaneous change in both ω^h and $\underline{s}^u$.[15] As is shown in Appendix I, a higher ω^h leads to more top-tier coding but a simultaneous increase in $\underline{s}^u$ produces the opposite effect. The ultimate effect depends on the force that dominates. Thus, we propose the second testable implication:

Proposition 2 If the effect from greater $\underline{s}^u$, i.e., stricter coding requirements, dominates the effect from ω^h, we expect no changes in total admissions within the base DRG and fewer admissions to the high-severity level.

Proof. See Appendix I.

[12] See https://www.cms.gov/icd10m/version37-fullcode-cms/fullcode_cms/Design_and_development_of_the_Diagnosis_Related_Group_(DRGs).pdf.

[13] For the proposed change in the CC/MCC list, see https://www.federalregister.gov/documents/2019/08/16/2019-16762/medicare-program-hospital-inpatient-prospective-payment-systems-for-acute-care-hospitals-and-the.

[14] We use the increasing $\underline{s}^u$ as a proxy for the stricter requirements for documentation. The idea is that a stricter documentation requirement in our model can be seen as a more stringent threshold for what cases can be legitimately classified into the top tier. For instance, according to Sacarny (2018), the disease code "congestive heart failure, unspecified" that was treated as high severity before the policy change was recategorized into low severity after the reform, because patients with this (somewhat vague) code were not found to result in greater treatment costs than average as patients with codes that gave more information about the nature of the condition. The recent revision of the CC/MCC list shares a similar idea.

[15] Appendix I provides more detail on the effect on $\tilde{s}^u$ and $\tilde{s}^a$ as $\underline{s}^u$ varies.

Preposition 2 suggests that the predicted change in top bill coding might be in the opposite direction to that in Proposition 1. Indeed, if the response to the increase in ω^h is smaller than the deterring effect from greater $\underline{s}^u$, then the model implies that an increase in ω^h will lead to a *decline* in admissions into the top tier within a DRG. Prior studies have shown empirical evidence consistent with this prediction for specific DRGs and at particular points in time. For example, Sacarny (2018) found that hospitals only captured half of the revenue in a code reclassification on heart failure, suggesting that hospitals may face large barriers in meeting the documentation requirements. Gowrisankaran et al. (2019) found relatively fewer top bill codes among DRGs that experienced greater increments in the DRG weight of the high severity subclass, after the implementation of the MS-DRG system in 2007.

3.2.2 Compliance Utility Model

We extend the model in Section 3.2.1 by incorporating a preference for compliance into the utility function. Moreover, let $m^a = \tilde{s}^a$ and $m^u = \tilde{s}^u - \underline{s}^a$, where m^a and m^u denote the level of compliance in admission and tier coding assignment, respectively. In Figure 1 the white area to left of the light-shaded area captures compliance in admissions, m^a. Note that m^a includes a) units of compliance driven strictly by profit maximization, where, given the expected cost of an audit, complying is the choice that generates higher expected profit and b) additional units of compliance that stem from a desire to further increase utility from compliance even at the cost of forgone profits. The white area in between the two shaded regions captures compliance in bill coding, m^u. Again, compliance demand for m^u includes both profit-driven and compliance utility-driven units. Therefore, equilibrium compliance in admissions and coding is higher in this model than it is in the model without compliance utility. Assuming hospital capacity is sufficiently large, the maximization problem becomes

$$\max_{\{m^a,\, m^u\}} E[U(c,\, m^a,\, m^u)] \tag{4}$$

subject to the budget constrain

$$c = \int_{m^a}^{\underline{s}^a} \left(\pi^\ell - \phi^{\text{AUDIT}} \times \phi^a(\underline{s}^a - s) \times \xi^a\right) ds + \int_{m^u+\underline{s}^a}^{\underline{s}^u} \left(\pi^h - \pi^\ell - \phi^{\text{AUDIT}} \times \phi^u(\underline{s}^u - s) \times \xi^u\right) ds + \Pi^P \tag{5}$$

After some simple re-arrangements of the terms in the budget constraint we can rewrite the model as follows:

$$
\begin{aligned}
&\max_{\{m^a,\, m^u\}} E[U(c,\, m^a,\, m^u)] \\
&\text{s.t. } c + m^a\left(\pi^\ell - f(0, m^a)\right) + m^u \times \left(\pi^h - \pi^\ell - g(\underline{s}^a, m^u + \underline{s}^a)\right) = \tilde{\Pi}^I + \Pi^P, \\
&\quad \tilde{\Pi}^I = \underline{s}^a \times \left(\pi^\ell - f(0, \underline{s}^a)\right) + (\underline{s}^u - \underline{s}^a) \times \left(\pi^h - \pi^\ell - g(\underline{s}^a, \underline{s}^u)\right), \\
&\quad \Pi^P = (\underline{s}^u - \underline{s}^a)\pi^\ell + (1 - \underline{s}^u)\pi^h.
\end{aligned}
\tag{6}
$$

where $\tilde{\Pi}^I$ denotes the "full improper profits" associated with the maximum level of non-compliance that's in principle feasible: admitting everyone below $\underline{s}^a$ and upcoding everyone within $[\underline{s}^a, \underline{s}^u]$.[16]

The formulation in Equation (6) provides additional insight as it reflects "expenditures" in the two types of compliance on the left hand side of the budget constraint. Further, each of these compliance expenditures is factored into the corresponding quantities times the average price per unit. Note that the average "price," $\pi^\ell - f(0, m^a)$, for a unit of compliance in admission (i.e., a unit of m^a) equals the profit π^ℓ that the hospital could have received by admitting a patient that is not supposed to be admitted, minus $f(0, m^a)$, the average unit expected cost from potential detection in audits associated with the (hypothetical) inappropriate admission of patients in the range $[0, m^a]$. Similarly, the "price," $\pi^h - \pi^\ell - g(\underline{s}^a, m^u + \underline{s}^a)$, per unit of compliance with the coding assignment protocols (i.e., a unit of m^u) is the forgone incremental profits from upcoding a patient $(\pi^h - \pi^\ell)$ net of the average expected cost associated with audit detection from hypothetical non-compliance in the compliant coding range $[\underline{s}^a, m^u + \underline{s}^a]$, given by $g(\underline{s}^a, m^u + \underline{s}^a)$. The right-hand side of the budget constraint represents the sum of the proper profits Π^P and the improper profits $\tilde{\Pi}^I$ the hospital would make if it admitted everyone and coded everyone into the top tier net of the costs from the audits. Moreover, the average compliance "prices" associated with m^a and m^u—the ones on the left-hand side of the budget constraint—are lower than the ones associated with a maximum level of unnecessary admissions and upcoding—the ones on the right-hand side. This is because compliance is always on the units that are more likely to be detected and expected detection costs subtract from the compliance prices, making compliance on units with high expected detection costs effectively cheaper. Therefore the units of compliance that the hospital decision maker chooses to "buy" are always cheaper than the ones it chooses not to "buy."

[16] $\tilde{\Pi}^I$ does not necessarily represent the maximum expected profits because the expected profit from admitting patients with s close to zero could be negative as they have high detection probabilities. Similarly the payoff from upcoding patients with s close to $\underline{s}^a$ could be negative as they too have high detection probability in case of an audit.

Having formulated the optimization problem in the model augmented with compliance in the utility function, we now consider the effects of increases in the reimbursement for top tier coding ω^h which implies an increase in π^h. On the one hand, it becomes more "expensive" to follow coding guidelines and this leads to more inappropriate coding behavior (*substitution* effect). But on the other hand, reimbursement rate becomes higher, and thus, decision makers in hospitals can "afford" to engage in more legitimate coding and indulge their preference for compliance (*income* effect). Ultimately, the effect of a higher ω^h on the consumption of m^u or the level of upcoding (i.e., $\underline{s}^u - \tilde{s}^u$) depends on the interplay between the income and substitution effects: If the substitution effect is smaller than the income effect, hospitals will tend to be more compliant, (i.e., upcode less).

Note that variation in ω^h might also affect the consumption of m^a or the number of inappropriate admissions (through a *cross* price effect), which also contains the income and substitution effects. Similarly, if the income effects dominates, a higher ω^h will result in greater consumption of m^a. However, if the income effect does not dominate, the prediction of the cross price effect would be ambiguous. We summarize how ω^h affects $\tilde{s}^a$ and $\tilde{s}^u$ in the following proposition:

Proposition 3 With greater ω^h, if the income effect dominates the substitution effect, we expect fewer total admissions within the base DRG and fewer admissions to the high-severity level.

Proof. See Appendix I.[17]

3.2.3 Capacity Constraint Model

In the following, we consider a case where hospitals comply with all the admission and coding protocols, (i.e. they engage in no unnecessary admissions or inappropriate coding whatsoever). However, hospitals face a capacity constraint, $\bar{u}$, in accommodating *all* types of patients. Kleiner (2019) pointed out while the average occupancy rate in U.S. hospitals is between 60% and 70%, there are 25% – 66% of them reporting the need to divert patients due to hospital overcrowding. Demand for hospital care could be highly variable (Sharma et al., 2008), involving substantial unanticipated demand shocks (Evans and Kim, 2006). Important factors contributing to the limited hospital capacity include nurse shortage and emergency department crowding (Buerhaus et al., 2007). The presence of a capacity constraint results in restrictions on inpatient admissions. However, the capacity could be

[17]Note that we illustrate Proposition 3 using a constant elasticity of substitution (CES) utility function with two goods. It is a simplified version of our model, but we expect the implications hold in a more general setup.

enlarged over time as hospitals expand staff/bed size, introduce advanced treatment technology, and adopt better strategy for resource allocation (Litvak and Bisognano, 2011). As hospitals become more efficient, i.e., fewer resources are required, to treat patients with a given condition, the reimbursement rate for this condition set by CMS could also go down. DRG weights are recalculated periodically to reflect the average resource consumption for the given condition.[18] As a result, we consider a model where hospitals are fully compliant with all admission and coding protocols, but have to adjust admission and coding intensity in response to the expanded capacity along with a decreasing reimbursement rate.

Let $\mathbb{r}^\ell$ denote the proportion of admitted patients whose severity level is between $[\underline{s}^a, \underline{s}^u]$ among all the patients falling into this range. Let $\mathbb{r}^h$ denote the proportion of top-coded patients among those with severity level above $\underline{s}^u$. Note that in an unconstrained model we would have $\mathbb{r}^\ell = \mathbb{r}^h = 1$. Because of the capacity constraint, though, we have $\mathbb{r}^\ell < 1$ and $\mathbb{r}^h < 1$. Assume that patients arriving at hospitals follows a random process, and hospitals cannot "select" whom to admit and will admit and top-code all qualified patients upon their arrival, when the capacity constraint is not binding.[19] Hence, hospital decision-makers decide on when to stop admitting or top-coding patients, i.e., the size of $\mathbb{r}^h$ and $\mathbb{r}^\ell$, according to the constraint they face. Moreover, since patients' arrival is purely random and there is no selection/sorting in the decisions of admission and coding assignment, the admitted patients constitute a random sample among all patients that should be admitted, i.e., those with severity levels in $[\underline{s}^a, 1]$. Assuming that this sample is representative of the patient population in terms of the distribution of patient severity, we have $\mathbb{r}^\ell = \mathbb{r}^h = \bar{u}/(1-\underline{s}^a)$. In case the constraint is binding, we have $\mathbb{r}^\ell(\underline{s}^u - \underline{s}^a) + \mathbb{r}^h(1-\underline{s}^u) = \bar{u}$. We formalize the problem as follows:

$$\begin{aligned} &\max_{\mathbb{r}^\ell,\, \mathbb{r}^h} E[U(c)] \\ \text{s.t. } & c = \mathbb{r}^\ell(\underline{s}^u - \underline{s}^a)\pi^\ell + \mathbb{r}^h(1-\underline{s}^u)\pi^h; \\ & \mathbb{r}^\ell = \mathbb{r}^h = \max\left\{\frac{\bar{u}}{1-\underline{s}^a},\ 1\right\}. \end{aligned}$$

In case the constraint is binding, $\mathbb{r}^\ell = \mathbb{r}^h = \bar{u}/(1-\underline{s}^a)$. Thus, as $\bar{u}$ increases with a simultaneous decrease in ω^h, both $\mathbb{r}^\ell$ and $\mathbb{r}^h$ will go up. When $\bar{u}$ is large enough, $\mathbb{r}^\ell = \mathbb{r}^h = 1$ and will not change even when $\bar{u}$ or ω^h vary. Thus, we present the fourth testable implication as follows:

[18] See https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/MS-DRG-Classifications-and-Software.

[19] For instance, hospitals cannot prioritize the admission of more severe patients and will accommodate qualified patients based on the first-come-first-served principle before reaching the limit.

Proposition 4 When the capacity constraint is met, we expect more admissions within the base DRG and more patients coded into the top tier as efficiencies are realized and ω^h drops. When the constraint is non-binding, we expect no changes in total admissions within the base DRG and top bill coding, even if ω^h varies.

To summarize, the various special cases have different testable implications in the number of total admissions and top-coded patients. The following table shows the implications of these 4 specialized mechanisms:

Table 1: Testable Implications of $\uparrow \omega^h$ on Aggregate DRG-level Claims Count

Model	Admissions	Top-tier coding
Financial-driven mechanism	No effect	More
Financial-driven mechanism enhanced by coding requirements	No effect	Fewer if the effect from the change in $\underline{s}^u$ dominates or more otherwise
Compliance utility model	Fewer if income effect dominates	Fewer if income effect dominates
Capacity constraint model	Fewer if the capacity constraint is binding	Fewer if the capacity constraint is binding

4 Data

4.1 Aggregate Medicare Part A Claims

Our first dataset comes from the files for the IPPS final rules and correction notice published by CMS, which include wage indices, the list of DRGs, DRG weights, mean length of stay, the number of discharges for each DRG, and IPPS operating and capital statewide average cost-to-charge-ratios. We construct a base DRG by grouping together DRGs sharing a common primary diagnosis/procedure. We use the aggregate counts of discharges at the DRG level to construct the two dependent variables: total admissions within a base DRG and the number of admissions to the top severity subclass. The key variable of interest is the DRG weight, which approximately measures the revenue hospitals receive from treating a patient that is assigned to the particular DRG.

Our analysis focuses on base DRGs with at least two severity subsclasses. On average,

there are 748 DRGs, about 682 of which belong to base DRGs with multiple tiers. Table 2 shows the summary statistics for these DRGs. Although the mean DRG weight is relatively stable over time, the changes in weights between a pair of successive years—denoted by Δwt—vary a lot across DRGs and years, creating important variation for our identification.

Table 2: Summary statistics for aggregate claims data

Year	# DRGs	DRG wt	Δwt	Average discharges per DRG
2012	679	2.04	–	14,419
		(1.85)	–	(34,396)
2013	679	2.05	0.0067	14,457
		(1.87)	(0.12)	(34,765)
2014	680	2.06	0.0075	13,847
		(1.89)	(0.13)	(33,768)
2015	680	2.09	0.0101	13,481
		(1.92)	(0.10)	(34,289)
2016	685	2.1	0.0086	12,861
		(1.92)	(0.13)	(34,068)
2017	684	2.12	0.0145	13,039
		(1.95)	(0.12)	(36,084)
2018	682	2.12	-0.00060	12,932
		(1.87)	(0.19)	(36,867)
2019	694	2.12	0.0163	12,758
		(1.87)	(0.15)	(38,666)

Note: Table reports the mean value, with the standard deviation in parenthesis.

4.2 CERT Microdata

The second dataset we rely on is the Medicare FFS CERT improper payment data in 2012-2019, a *random* sample of Medicare FFS claims that are reviewed by CERT auditors. In particular, we focus on Part A IPPS claims. For each claim, the data contains information on provider type, type of bill, assigned DRG and procedure codes, and whether a determination of improper payment was made due to one of the five error types specified by CMS. We focus on upcoding-related errors, which occur if the claim is determined to have incurred in one of the following four errors: (1) incorrect coding, (2) insufficient documentation, (3) no documentation, and (4) others. The definition relies on these error types, as the occurrence of any of them—such as inadequate justification for the reported code due to no or insufficient

documentation, assigning the incorrect codes, or other potential miscellaneous causes—could all be associated with upcoding practices.

Table 3 reports the summary statistics based on the CERT data. Despite a relatively small number of claims in the data compared with the entire Part A IPPS population, the sampling strategy in CERT follows a stratification plan that enables a good representation of the overall population. Claims with upcoding-related errors account for 7% – 9% of the entire sample.

Table 3: Summary statistics for CERT microdata

Year	# claims	Upcoding (%)
2012	10,744	8.7
2013	12,519	9.3
2014	10,936	8.8
2015	12,577	5.4
2016	12,362	7.5
2017	11,166	7.5
2018	11,926	7.1
2019	11,768	8.7

Note: Table reports the mean value in terms of frequency. Column 3 reports the proportion of audited claims that were flagged with upcoding-related, improper payment errors.

5 Empirical Analysis of Aggregate Medicare Part A Claims

In this section, we examine the testable implications summarized in Table 1 using the aggregate Medicare Part A claims data, to identify which mechanism seems most important during our study period. We first describe our fixed effects specification and then discuss the empirical results.

5.1 Empirical strategy

We exploit within base DRG variation in reimbursement over time by estimating the following model:

$$Y_{dt} = \alpha + \beta^h \omega_{dt}^h + \beta^\ell \omega_{dt}^\ell + \lambda_d + \lambda_t + \mu_d \cdot t + \varepsilon_{dt} \tag{7}$$

where Y_{dt} denotes the outcome of interest, which could be total admissions within a base DRG, d, or the number of admissions to the top severity subclass in this base DRG.[20] In the model notation from Section 3.1, total admissions at the base DRG level are captured by $1 - \tilde{s}^a$ and admissions into the top tier DRG correspond to $1 - \tilde{s}^u$. In the models for both of these dependent variables, we use as key regressor the DRG weight at the top severity level, ω_{dt}^h. In the analysis for total admissions, we use the DRG weight at the bottom level as empirical measure of ω_{dt}^ℓ. In the analysis of admissions into the top tier we use the weight at the next lower level as an empirical measure of ω_{dt}^ℓ. In particular, we use the bottom level among base DRGs with two tiers and the middle level among base DRGs with three tiers. The rationale for these choices is that the bottom-tier DRG weight is more relevant to the admission decision, i.e., $\tilde{s}^a$, whereas the DRG weight of the next lower severity level reflects the incremental gain, and matters more in deciding whether to code a patient into the true severity level or the next (higher) severity level. λ_d and λ_t, respectively, denote base DRG and year fixed effects. $\mu_d \cdot t$ captures base DRG-specific linear trends.

Our regression analysis first explores how the number of admissions to the top bill code varies in response to the change in revenues, captured by DRG weights. Our specification differs from that in related studies (Dafny, 2005; Li, 2014; Gowrisankaran et al., 2019; Ganju et al., 2021) in the following aspects. First, we include the DRG weights of both the top and lower levels in the regression instead of using the spread—the difference in DRG weights between the top and low levels. By doing so, we allow for a more flexible functional form to capture the effect of revenue changes on coding intensity. Second, we use the (log of the) *number* of admissions to the entire base DRG as dependent variable.

The key variables of interest are the DRG weights, the coefficients for which, β^ℓ and β^h, measure the marginal effects of the weights. Our identification relies on variation within a base DRG across time. If upcoding is mainly driven by financial incentives as suggested in the literature, we expect β^h to be positive and/or β^ℓ to be negative, which is equivalent to a positive effect of spread. In contrast, a negative β^h and/or a positive β^ℓ might suggest other mechanisms proposed in Section 3.2. While we mainly focus on coding intensity in response to the changes in financial payoff, we also examine another outcome—total admissions within a base DRG, which helps us distinguish across mechanisms. Our identifying assumption is that the *changes* in DRG weights are exogenous to admission/coding decisions after controlling for the various fixed effects and "base DRG"-specific trends. Below in Section 7 we supplement our current analysis in a way that supports the validity of this assumption. We weight our regressions by the mean number of discharges over time within a base DRG and cluster standard errors at the same level.

[20] For base DRGs with three severity subclasses, we do not consider the middle tier.

5.2 Results

We present the estimated coefficients for the key variables of interest in Table 4. The first coefficient column reports the estimates for β^h and β^l based on the entire sample. It suggests that the *greater* the DRG weight for the top tier is, the *fewer* admissions there will be assigned to that tier. Specifically, we expect an average reduction of 5.8% in the admissions to the top tier, as the corresponding DRG weight increases by 0.2, approximately 10% of the average DRG weights across all DRGs in multiple-tier base DRGs. The coefficient for the DRG weight at the next low level DRG is positive but not significant. This is different from the prediction in the financial-driven mechanism in Table 1, suggesting that while the financial incentive to both, unnecessarily admit and upcode patients, is still present and prevents a fully compliant behavior, other mechanisms could explain why we see fewer admissions into a DRG when its reimbursement rate increases.

Table 4: Effect of DRG weights on Log(Admissions) to the top tier within a base DRG

	All	High weights	Low weights	Medium weights	Common DRGs	Less common DRGs
Top-tier DRG weight	-0.2906**	-0.1552**	-2.0508*	-0.1912***	-0.6948*	-0.1117*
	(0.1228)	(0.0602)	(1.1446)	(0.0720)	(0.3806)	(0.0579)
Next lower-tier DRG weight	0.0039	-0.0766*	6.0525**	-0.1493	0.2047	-0.0218
	(0.1061)	(0.0449)	(2.7375)	(0.2505)	(0.4485)	(0.0968)
N	2110	209	190	1711	526	1584
R^2	0.998	0.998	0.993	0.999	0.997	0.979
MeanDepVar	8.155	8.088	8.618	8.112	10.118	7.504

Note: Analysis based on Medicare Part A Population data. Unit of observation is DRG/year. Other regressors include DRG fixed effects, year fixed effects, and DRG-specific time trends. Each observation is a DRG-year combination. Regression weighted by total number of discharges per base DRG. Standard errors in parentheses, clustered at base DRGs.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

The rest of the columns in Table 4 report results that explore heterogeneity across two dimensions. We first decompose the sample into DRGs with high weights (the DRGs with the top 10% highest weights), DRGs with low weights (the bottom 10%), and the remaining DRGs (those with weights within the 10% to 90% range). The general finding holds: fewer admissions will be assigned to the top tier when the corresponding weight becomes higher. Interestingly, among the bottom 10% DRGs, the prediction is the same when the weight of the lower level DRG rises, suggesting that a greater spread—that could arise from the increase in DRG weight from the top level and/or the reduction from the lower level—results in fewer top bill codes. The last two columns show the results separately for common DRGs (the top 30% in terms of the average number of discharges across years) and the less common

ones (the remaining DRGs), and the main results hold.[21]

To sum up, the negative effect of top-tier DRG weights on admissions into the top tier suggests that mechanisms other than financial incentives alone might underlie the coding behavior during our sample period. As our model proposed, this could be related to a tied increase in coding requirements, the dominating income effect in the demand for compliance among hospital decision-makers that care about some "internal conscience," or the relaxing capacity constraints due to improved operation efficiency. In the following, we explore these other mechanisms by examining the effect of financial incentives on a different outcome.

Table 5: Effect of DRG weights on Log(Total Admissions) in a base DRG

	All	High weights	Low weights	Medium weights	Common DRGs	Less common DRGs
Top-tier DRG weight	-0.117*	-0.141	0.073	-0.080	-0.267**	-0.034
	(0.069)	(0.100)	(0.553)	(0.070)	(0.106)	(0.072)
Bottom-tier DRG weight	-0.170	-0.132	2.656	-0.324	-0.200	-0.174
	(0.127)	(0.133)	(1.633)	(0.292)	(0.138)	(0.163)
N	2110	209	190	1711	526	1584
R^2	0.998	0.996	0.996	0.999	0.998	0.989
MeanDepVar	9.292	9.031	9.607	9.288	11.157	8.672

Note: Analysis based on Medicare Part A Population data. Other regressors include base DRG fixed effects, year fixed effects, and linear time trends by base DRGs. Each observation is a combination of a base DRG and year. Regression weighted by total number of discharges per base DRG. Standard errors in parentheses, clustered at base DRGs.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 5 reports the estimated effect of DRG weights on total admissions in a similar format to Table 4. The analysis based on the entire sample shows a negative effect of the top-tier DRG weight on the total admissions, yet with less significance and magnitude. The coefficients for the weight of the top bill code turn insignificant when we break down the sample by DRG weight magnitude groups, but it is still large and significantly negative in the subsample of common DRGs, suggesting that most of the effect might be driven by this group. The finding here suggests that an increase in financial incentives that is tied with increased stringency in coding requirements cannot be the only explanation behind hospitals' admission and coding practices. This leaves us with the model of financial incentives augmented with compliance utility and the capacity constraint model. Notice that the capacity constraint

[21]Note that the number of observations among common DRGs is smaller because the unit of observation is a unique DRG and year cell. The analysis is based on a group of DRGs ranked top 30% in terms of average discharges.

model is one specific formalization of a more general class of possible models consistent with our findings where technological changes drive both increases in admissions and reductions in ω^h. To sort out across models where financial incentives do and do not matter one possibility at this stage to look for an instrumental variable that would generate exogenous variation in reimbursements within a DRG over time. An alternative to which we now turn is to explore the different implications these models have in a random sample of audits.

6 Testable Implications in a Random Sample of Audits

To separate the financial incentives model augmented with compliance utility from the capacity constraint model and more generally models of technological change where financial incentives do not play a role, we rely on their different implications in a random sample of cases that were audited by the CERT program. Specifically, we focus on the different predictions these two models have on top code billing. In short, hospital decision makers who are concerned about the disutility of noncompliance should vary the extent of inappropriate coding in response to the change in DRG weights and this should, in turn, be reflected in the sample of audits. Models where financial incentives do not matter, such as the capacity constraint model, on the other hand, will not generate the same implication.

All individuals whose severity level s is such that $s > \tilde{s}^a$ are admitted to hospitals and constitute the universe of admissions. The CERT program takes a random sample from this universe and examines whether each case is subject to improper billing after auditors evaluate all the relevant information. The extent to which hospitals upcode cases can be represented by $|\tilde{s}^u - \underline{s}^u|$. Since $\underline{s}^u$ is simply another parameter and we only consider the change in financial incentives here, i.e., $cov(\underline{s}^u, \omega^h) = 0$, the compliance utility model with a dominant income effect implies that we should expect to see fewer claims flagged with upcoding-related errors in the CERT data when reimbursements increase within a DRG. In the mechanism driven by capacity constraint, hospitals top-code all qualified patients, subject to the capacity limit. One should not expect any changes in the incidence of upcoding-related errors as the top-tier DRG weight varies. Note that the CERT audit data may not detect all upcoding but simply reflects the level of upcoding that was detected. This is captured in our model by the fact that $\phi^u < 1$ in the range of s where the hospital may upcode, $(\underline{s}^a, \underline{s}^u)$. However, given that it is a random sample of the population claims, it could still inform the *directional* changes in coding intensity as the financial payoff varies. The following table summarizes the testable implications in CERT for the both models:

Table 6: Summary of Testable Implications of $\uparrow \omega^h$ in Audits Data

Model	Upcoding-related errors
Compliance utility model	Fewer if income effect dominates
Capacity constraint model	No effect

7 Empirical Analysis of Audit Microdata

To test the additional implications available in a sample of audits, we estimate a similar specification as in Equation (7) but now using the CERT microdata and using as dependent variable a binary indicator which equals one when an audited claim is flagged with upcoding-related errors. In essence the model is a linear probability model for upcoding. The model can be used to estimate how the probability that a randomly selected claim gets flagged with an "upcoding" determination depends on the DRG weight. Since we retain the DRG fixed effects specification, identification still comes from "within DRG" variation in reimbursements over time. Thus, we identify the effect of financial incentives based on the changes in the top-tier DRG weights and the changing incidence of audited claims with upcoding flags within a base DRG. The unit of observation is a claim-year observation, and to account for stratification, we weight each observation with the relative frequency with which claims from a given DRG in the CERT sample compare with the aggregate claims data.

Table 7: Effect of DRG weights on upcoding-related errors

	All	High weights	Low weights	Medium weights	Common DRGs	Less common DRGs
Top-tier DRG weight	-0.1235**	-0.1736*	-0.3610**	-0.0433	-0.2711**	-0.0273
	(0.0558)	(0.0949)	(0.1322)	(0.0875)	(0.1094)	(0.0822)
Next lower-tier DRG weight	0.0142	0.0953	0.6314	-0.1591	0.2333	-0.0289
	(0.0571)	(0.0841)	(0.4641)	(0.1418)	(0.2052)	(0.0489)
N	25276	2779	3263	19234	18539	6737
R^2	0.052	0.152	0.022	0.057	0.030	0.181
MeanDepVar	0.083	0.091	0.082	0.083	0.081	0.091

Note: Analysis based on CERT microdata. Other regressors include DRG fixed effects, year fixed effects, and DRG-specific time trends. Each observation is a claim-year combination. Regression weighted by the relative frequency per DRG per year. Standard errors in parentheses, clustered at base DRGs.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 7 summarizes the results in a similar format to Table 4. The first column suggests that relatively fewer upcoding-related errors are detected among DRGs with an increasing weight. It is equivalent to a reduction of 2.47 percentage point in the incidence of these

errors as the corresponding weight increases by 0.2. When we break down the DRGs by DRG weights, both of those standing above the 90 percentile and below the bottom 10 percentile experience fewer upcoding-related errors as the top-tier DRG weight gets higher. Also, such a negative effect seems to mainly arise from the most frequent DRGs among all admissions.

To summarize, the significantly negative effect of financial incentives on the occurrence of upcoding-related errors from the audit data lends support to the compliance-utility model. It is consistent with the mechanism where hospital decision-makers pay attention to and care about financial incentives but show a preference for compliance in admission and billing practices. This occurs when the income effect out-weights the substitution effect in their decision-making. This desire for compliance only limits, but does not eliminate the role financial incentives play in inducing upcoding.

8 Conclusion

We study to what extent inappropriate coding practices respond to increases in reimbursement rates in recent years in the context of Medicare Part A. To guide our empirical investigation and explore the possible underlying mechanisms, we consider a general model characterizing hospitals' decisions regarding the admission of patients and their billing coding, incorporating both the financial incentives and a preference for compliant behavior into hospital decision makers' utility function. Moreover, we develop several specialized models that feature particular mechanisms from the general model which have been emphasized in the literature. We test these specialized models by analyzing claim counts from Medicare at the DRG level using changes in reimbursement within DRGs over time as the key source of identifying variation. While the aggregate claims data allows us to distinguish across models, it is in principle consistent with models where the financial incentives do not matter. To further test these two types of models we rely on a novel empirical strategy based on microdata from a random sample of audited claims. We find a *negative* effect of the top-tier DRG weight on the number of admissions into the top tier of those DRGs and on the overall number of admissions into the base DRG. We show that a model where financial incentives drive improper admission and coding practices supplemented with a compliance utility feature where demand for compliance is dominated by a strong income effect might be consistent with our evidence from both aggregate claims and the audit sample. Our model brings a well known mechanism from the physician behavior models into an important literature that seeks to explain how hospitals engage in coding practices in response to changing financial incentives. This new mechanism seems important to understand hospital behavior

in recent years. Our finding also imply that the decisions on admission and coding by hospitals could be quite sophisticated and as a result, policymakers might consider imposing more comprehensive reward/penalty mechanisms to encourage proper billing.

References

Buerhaus, P. I., K. Donelan, B. T. Ulrich, L. Norman, C. DesRoches, and R. Dittus (2007). Impact of the nurse shortage on hospital patient care: Comparative perspectives. *Health affairs 26*(3), 853–862. 3.2.3

CERT Program (2011-2019). Medicare fee-for-service improper payments report. *Centers for Medicare and Medicaid Services*. 7

Cook, A. and S. Averett (2020). Do hospitals respond to changing incentive structures? Evidence from medicare's 2007 drg restructuring. *Journal of Health Economics 73*, 102319. 1

Dafny, L. (2005). How do hospitals respond to price changes? *American Economic Review 95*(5), 1525–1547. 1, 3.2.1, 3.2.1, 5.1

Dranove, D. (1988). Demand inducement and the physician/patient relationship. *Economic inquiry 26*(2), 281–298. 1

Evans, W. N. and B. Kim (2006). Patient outcomes when hospitals experience a surge in admissions. *Journal of Health Economics 25*(2), 365–388. 3.2.3

Fang, H. and Q. Gong (2017). Detecting potential overbilling in medicare reimbursement via hours worked. *American Economic Review 107*(2), 562–91. 1

Finkelstein, A., M. Gentzkow, P. Hull, and H. Williams (2017). Adjusting risk adjustment–accounting for variation in diagnostic intensity. *The New England journal of medicine 376*(7), 608. 1

Ganju, K. K., H. Atasoy, and P. A. Pavlou (2021). Do electronic health record systems increase medicare reimbursements? The moderating effect of the recovery audit program. *Management Science*. 5.1

Geruso, M. and T. Layton (2020). Upcoding: Evidence from medicare on squishy risk adjustment. *Journal of Political Economy 128*(3), 984–1026. 1

Gowrisankaran, G., K. Joiner, and J. Lin (2019). How do hospitals respond to Medicare payment reforms? *NBER Working Paper No. 26455*. 3.2.1, 5.1

Gruber, J. and M. Owings (1996). Physician financial incentives and cesarean section delivery. *RAND Journal of Economics 27*(1), 99–123. 1, 3.1

Jürges, H. and J. Köberlein (2015). What explains drg upcoding in neonatology? The roles of financial incentives and infant health. *Journal of health economics 43*, 13–26. 3.2.1

Kleiner, S. A. (2019). Hospital treatment and patient outcomes: Evidence from capacity constraints. *Journal of Public Economics 175*, 94–118. 3.2.3

Li, B. (2014). Cracking the codes: Do electronic medical records facilitate hospital revenue enhancement. *Working paper*. 5.1

Litvak, E. and M. Bisognano (2011). More patients, less payment: Increasing hospital efficiency in the aftermath of health reform. *Health Affairs 30*(1), 76–80. 3.2.3

McGuire, T. G. and M. V. Pauly (1991). Physician response to fee changes with multiple payers. *Journal of health economics 10*(4), 385–410. 1, 3.1

Office of the Federal Register and National Archives and Records Service (2007). Medicare program; changes to the hospital inpatient prospective payment systems and fiscal year 2008 rates. *Federal Register (Wednesday, August 22, 2007) 72*(162), 47129–48175. From the Federal Register Online via the Government Publishing Office [www.gpo.gov], FR Doc No: 07-3820. 4

Sacarny, A. (2018). Adoption and learning across hospitals: The case of a revenue-generating practice. *Journal of Health Economics 60*, 142–164. 3.2.1, 14

Sharma, R., M. Stano, and R. Gehring (2008). Short-term fluctuations in hospital demand: Implications for admission, discharge, and discriminatory behavior. *The Rand journal of economics 39*(2), 586–606. 3.2.3

Shi, M. (2021). The costs and benefits of monitoring providers: Evidence from medicare audits. *Working paper*. 1

Silverman, E. and J. Skinner (2004). Medicare upcoding and hospital ownership. *Journal of health economics 23*(2), 369–389. 1, 3.2.1

Appendix I

Proof of Proposition 1

In this specialized model, the hospital decision-maker aims to

$$\begin{aligned}&\max_{\tilde{s}^a,\,\tilde{s}^u} E[U(c)]\\ \text{s.t. } c = &(\underline{s}^a - \tilde{s}^a)\left(\pi^\ell - f(\tilde{s}^a, \underline{s}^a)\right) + (\underline{s}^u - \tilde{s}^u)\left(\pi^h - \pi^\ell - g(\tilde{s}^u, \underline{s}^u)\right)\\ &+ (\underline{s}^u - \underline{s}^a)\pi^\ell + (1 - \underline{s}^u)\pi^h\end{aligned}$$

Consider an increase in ω^h, which leads to an increase in π^h. We are interested in how the admission and coding decision varies as π^h increases. In the following, we describe and prove the first proposition.

Proposition 1 As ω^h and thus π^h gets higher, we expect no changes in total admissions within the base DRG, i.e., $\partial\tilde{s}^a/\partial\pi^h = 0$, and more admissions to the high-severity level, i.e., $\partial\tilde{s}^u/\partial\pi^h < 0$.

Proof. The first order conditions are

$$\begin{aligned}\frac{\partial U}{\partial \tilde{s}^a} &= U_c\left[(\underline{s}^a - \tilde{s}^a)(-f_1) + (\pi^\ell - f(\tilde{s}^a, \underline{s}^a))(-1)\right] = 0;\\ \frac{\partial U}{\partial \tilde{s}^u} &= U_c\left[(\underline{s}^u - \tilde{s}^u)(-g_1) + (\pi^h - \pi^\ell - g(\tilde{s}^u, \underline{s}^u))(-1)\right] = 0\end{aligned}$$

where $f_1 = \partial f(\tilde{s}^a, \underline{s}^a)/\partial\tilde{s}^a$ and $g_1 = \partial g(\tilde{s}^u, \underline{s}^u)/\partial\tilde{s}^u$. After simplification,

$$\begin{aligned}f - \pi^\ell - f_1(\underline{s}^a - \tilde{s}^a) = 0;\\ g - \pi^h + \pi^\ell - g_1(\underline{s}^u - \tilde{s}^u) = 0.\end{aligned} \tag{8}$$

By assuming that $\Delta B\omega^h > \Delta c^h$, we know that the net change leads to higher π^h. Take derivatives w.r.t. π^h on both sides:

$$\begin{aligned}f_1\frac{\partial\tilde{s}^a}{\partial\pi^h} - f_{11}(\underline{s}^a - \tilde{s}^a)\frac{\partial\tilde{s}^a}{\partial\pi^h} + f_1\frac{\partial\tilde{s}^a}{\partial\pi^h} = 0;\\ g_1\frac{\partial\tilde{s}^u}{\partial\pi^h} - 1 - g_{11}(\underline{s}^u - \tilde{s}^u)\frac{\partial\tilde{s}^u}{\partial\pi^h} + g_1\frac{\partial\tilde{s}^u}{\partial\pi^h} = 0.\end{aligned}$$

$\Longrightarrow$

$$\begin{aligned}\frac{\partial\tilde{s}^a}{\partial\pi^h}\left[2f_1 - f_{11}(\underline{s}^a - \tilde{s}^a)\right] = 0 \;\rightarrow\; \frac{\partial\tilde{s}^a}{\partial\pi^h} = 0\\ \frac{\partial\tilde{s}^u}{\partial\pi^h}\left[2g_1 - g_{11}(\underline{s}^u - \tilde{s}^u)\right] = 1 \;\rightarrow\; \frac{\partial\tilde{s}^u}{\partial\pi^h} = \frac{1}{2g_1 - g_{11}(\underline{s}^u - \tilde{s}^u)} < 0\end{aligned}$$

As a result, the testable implications for a higher ω^h are no change in admissions and more top-tier coding in claims.

Proof of Proposition 2

Based on the same setup as the previous proposition, we consider the change in $\underline{s}^u$. Take derivatives w.r.t. $\underline{s}^u$ on both sides of Equations (8):

$$f_1\frac{\partial\tilde{s}^a}{\partial\underline{s}^u} - f_{11}(\underline{s}^a-\tilde{s}^a)\frac{\partial\tilde{s}^a}{\partial\underline{s}^u} + f_1\frac{\partial\tilde{s}^a}{\partial\underline{s}^u} = 0;$$
$$g_1\frac{\partial\tilde{s}^u}{\partial\underline{s}^u} + g_2 - g_{11}(\underline{s}^u-\tilde{s}^u)\frac{\partial\tilde{s}^u}{\partial\underline{s}^u} - g_{12}(\underline{s}^u-\tilde{s}^u) - g_1 + g_1\frac{\partial\tilde{s}^u}{\partial\underline{s}^u} = 0.$$

$\Longrightarrow$

$$\frac{\partial\tilde{s}^a}{\partial\underline{s}^u}\left[2f_1 - f_{11}(\underline{s}^a-\tilde{s}^a)\right] = 0 \;\rightarrow\; \frac{\partial\tilde{s}^a}{\partial\underline{s}^u} = 0;$$
$$\frac{\partial\tilde{s}^u}{\partial\underline{s}^u}\left[2g_1 - g_{11}(\underline{s}^u-\tilde{s}^u)\right] = g_1 - g_2 + g_{12}(\underline{s}^u-\tilde{s}^u) \;\rightarrow\; \frac{\partial\tilde{s}^u}{\partial\underline{s}^u} = \frac{g_1-g_2+g_{12}(\underline{s}^u-\tilde{s}^u)}{2g_1-g_{11}(\underline{s}^u-\tilde{s}^u)} > 0 \text{ if } g_{12}\leq 0.$$

For instance, $g_{12}\leq 0$ if $\phi^u(\cdot)$ is a constant or a polynomial function of $(\underline{s}^u-\tilde{s}^u)$, with at least one degree.

Since $cov(\omega^h,\underline{s}^u)>0$, a higher ω^h leads to more top-tier coding/upcoding but a simultaneous increase in $\underline{s}^u_j$ produces an opposite effect. The ultimate effect depends on the force that dominates.

Illustration of Proposition 3

Below we use a CES utility function for illustration with the following simplifications: (i) Only two goods are included; and (ii) the price of compliance does *not* vary by the level of compliance. Next, we go through the same exercise after relaxing (ii) by using a linear function for the probability of getting caught, i.e., $\phi^a(\underline{s}^a-s)$ is linear in $(\underline{s}^a-s)$ and $\phi^u(\underline{s}^u-s)$ is linear in $(\underline{s}^u-s)$.

Consider the following problem:

$$\max_{\{\mathbb{m}_a,\,\mathbb{m}_u\}} U(\mathbb{m}_a,\mathbb{m}_u) = (\mathbb{m}_a^\rho+\mathbb{m}_u^\rho)^{1/\rho}$$
$$\text{s.t. } p_a\mathbb{m}_a + p_u\mathbb{m}_u = p_a+p_u-\eta,$$
$$\text{where } \mathbb{m}_a\in(0,1),\ \mathbb{m}_u\in(0,1),\ \text{and } \eta>0.$$

In this model, hospitals decide on $\mathbb{m}_a$ and $\mathbb{m}_u$, with p_a and p_u denoting the price for each of

them, respectively. Although the model does not include consumption in the utility function, it is implicitly captured in η. Unlike in the original model where hospitals balance the trade-off between consumption, compliance in admission decision, and compliance in top-coding, we show the decision making on the two compliance levels by holding consumption constant in this model.

Write down the Lagrangian: $\mathcal{L} = U(\mathbb{m}_a, \mathbb{m}_u) - \lambda(p_a\mathbb{m}_a + p_u\mathbb{m}_u - p_a - p_u + \eta)$. The first order conditions are

$$\begin{aligned} &U_a - \lambda^* p_a = 0 \\ &U_u - \lambda^* p_u = 0 \\ &-p_a\mathbb{m}_a^* - p_u\mathbb{m}_u^* + p_a + p_u - \eta = 0. \end{aligned}$$

Take the derivative w.r.t p_u on the first-order conditions:

$$\begin{bmatrix} U_{aa} & U_{au} & -p_a \\ U_{ua} & U_{uu} & -p_u \\ -p_a & -p_u & 0 \end{bmatrix} \times \begin{bmatrix} \frac{\partial \mathbb{m}_a^*}{\partial p_u} \\ \frac{\partial \mathbb{m}_u^*}{\partial p_u} \\ \frac{\partial \lambda^*}{\partial p_u} \end{bmatrix} = \begin{bmatrix} 0 \\ \lambda^* \\ \mathbb{m}_u^* - 1 \end{bmatrix}$$

Solving the system of equations above, the sign of $\partial \mathbb{m}_u^*/\partial p_u$ depends on the sign of

$$-\lambda^* p_a^2 - (1 - \mathbb{m}_u^*)\frac{p_u}{U_u}(U_u U_{aa} - U_a U_{au}),$$

and the sign of $\partial \mathbb{m}_a^*/\partial p_u$ depends on

$$\lambda^* p_a p_u + (1 - \mathbb{m}_u^*)\frac{p_u}{U_u}(U_u U_{au} - U_a U_{uu}).$$

If the income effect of the change in p_u dominates the substitution effect, then $\partial \mathbb{m}_u^*/\partial p_u > 0$, and thus, $U_u U_{aa} - U_a U_{au} < 0$, i.e., U_a/U_u decreases in $\mathbb{m}_a$. Since $U_a/U_u = (\mathbb{m}_u/\mathbb{m}_a)^{1-\rho}$, U_a/U_u decreases in $\mathbb{m}_a$ as long as $\rho < 1$. When $\rho < 1$, U_a/U_u increases in $\mathbb{m}_u$, i.e., $U_u U_{au} - U_a U_{uu} > 0 \implies \partial \mathbb{m}_a^*/\partial p_u > 0$.

Now we relax the assumption in (ii), and assume that $\phi^a(\underline{s}^a - s) = \gamma^a(\underline{s}^a - s) + \gamma^{a0}$ and $\phi^u(\underline{s}^u - s) = \gamma^u(\underline{s}^u - s) + \gamma^{u0}$, where $\gamma^a > 0$ and $\gamma^u > 0$. Rewrite the average expected

cost of receiving a negative audit determination (i.e., Equation (3)):

$$
\begin{aligned}
f(0,m^a) &= \frac{\phi^{\text{AUDIT}}\xi^a}{m^a}\int_0^{m^a}\left[\gamma^a(\underline{s}^a-s)+\gamma^{a0}\right]ds = \frac{\phi^{\text{AUDIT}}\xi^a}{m^a}\times\left[\gamma^a\underline{s}^a m^a-\frac{\gamma^a(m^a)^2}{2}+\gamma^{a0}m^a\right]\\
&= \phi^{\text{AUDIT}}\xi^a\gamma^{a0}+\phi^{\text{AUDIT}}\xi^a\gamma^a\underline{s}^a-\frac{\phi^{\text{AUDIT}}\xi^a\gamma^a m^a}{2};\\
f(0,\underline{s}^a) &= \frac{\phi^{\text{AUDIT}}\xi^a}{\underline{s}^a}\int_0^{\underline{s}^a}\left[\gamma^a(\underline{s}^a-s)+\gamma^{a0}\right]ds = \frac{\phi^{\text{AUDIT}}\xi^a}{\underline{s}^a}\times\left[\gamma^a(\underline{s}^a)^2-\frac{\gamma^a(\underline{s}^a)^2}{2}+\gamma^{a0}\underline{s}^a\right]\\
&= \phi^{\text{AUDIT}}\xi^a\gamma^{a0}+\phi^{\text{AUDIT}}\xi^a\gamma^a\underline{s}^a-\frac{\phi^{\text{AUDIT}}\xi^a\gamma^a\underline{s}^a}{2};\\
g(\underline{s}^a,m^u+\underline{s}^a) &= \frac{\phi^{\text{AUDIT}}\xi^u}{m^u}\int_{\underline{s}^a}^{m^u+\underline{s}^a}\left[\gamma^u(\underline{s}^u-s)+\gamma^{u0}\right]ds\\
&= \frac{\phi^{\text{AUDIT}}\xi^u}{m^u}\times\left[\gamma^u\underline{s}^u m^u-\frac{\gamma^u((m^u)^2+2m^u\underline{s}^a)}{2}+\gamma^{u0}m^u\right]\\
&= \phi^{\text{AUDIT}}\xi^u\gamma^{u0}+\phi^{\text{AUDIT}}\xi^u\gamma^u\underline{s}^u-\phi^{\text{AUDIT}}\xi^u\gamma^u\underline{s}^a-\frac{\phi^{\text{AUDIT}}\xi^u\gamma^u m^u}{2};\\
g(\underline{s}^a,\underline{s}^u) &= \frac{\phi^{\text{AUDIT}}\xi^u}{\underline{s}^u-\underline{s}^a}\int_{\underline{s}^a}^{\underline{s}^u}\left[\gamma^u(\underline{s}^u-s)+\gamma^{u0}\right]ds\\
&= \frac{\phi^{\text{AUDIT}}\xi^u}{\underline{s}^u-\underline{s}^a}\times\left[\gamma^u\underline{s}^u(\underline{s}^u-\underline{s}^a)-\frac{\gamma^u((\underline{s}^u)^2-(\underline{s}^a)^2)}{2}+\gamma^{u0}(\underline{s}^u-\underline{s}^a)\right]\\
&= \phi^{\text{AUDIT}}\xi^u\gamma^{u0}+\phi^{\text{AUDIT}}\xi^u\gamma^u\underline{s}^u-\phi^{\text{AUDIT}}\xi^u\gamma^u\underline{s}^a-\frac{\phi^{\text{AUDIT}}\xi^u\gamma^u(\underline{s}^u-\underline{s}^a)}{2};
\end{aligned}
$$

Notice that $f(0,m^a)>f(0,\underline{s}^a)$ and $g(\underline{s}^a,m^u+\underline{s}^a)>g(\underline{s}^a,\underline{s}^u)$. Using the equations above to replace all the $f(\cdot)$ and $g(\cdot)$, the budget constraint in Equation (**??**) becomes

$$
\begin{aligned}
&c+m^a\left(\pi^\ell-\phi^{\text{AUDIT}}\xi^a\gamma^{a0}-\phi^{\text{AUDIT}}\xi^a\gamma^a\underline{s}^a+\frac{\phi^{\text{AUDIT}}\xi^a\gamma^a m^a}{2}\right)\\
&+m^u\left(\pi^h-\pi^\ell-\phi^{\text{AUDIT}}\xi^u\gamma^{u0}-\phi^{\text{AUDIT}}\xi^u\gamma^u\underline{s}^u+\phi^{\text{AUDIT}}\xi^u\gamma^u\underline{s}^a+\frac{\phi^{\text{AUDIT}}\xi^u\gamma^u m^u}{2}\right)\\
&=\underline{s}^a\left(\pi^\ell-\phi^{\text{AUDIT}}\xi^a\gamma^{a0}-\phi^{\text{AUDIT}}\xi^a\gamma^a\underline{s}^a+\frac{\phi^{\text{AUDIT}}\xi^a\gamma^a\underline{s}^a}{2}\right)\\
&+(\underline{s}^u-\underline{s}^a)\left(\pi^h-\pi^\ell-\phi^{\text{AUDIT}}\xi^u\gamma^{u0}-\phi^{\text{AUDIT}}\xi^u\gamma^u\underline{s}^u+\phi^{\text{AUDIT}}\xi^u\gamma^u\underline{s}^a+\frac{\phi^{\text{AUDIT}}\xi^u\gamma^u(\underline{s}^u-\underline{s}^a)}{2}\right)+\Pi^p.
\end{aligned}
$$

The price of compliance level in admission decision (tier coding assignment), i.e., m^a (m^u), is a linear function of the consumption of m^a (m^u). Note that the marginal effect of consuming m^a (m^u) on the price of m^a (m^u) is strictly positive, as $\phi^{\text{AUDIT}}\xi^a\gamma^a>0$ ($\phi^{\text{AUDIT}}\xi^u\gamma^u>0$).

Let the average compliance prices have the following forms:

$$\pi^{\ell} - \phi^{\text{AUDIT}}\xi^{a}\gamma^{a0} - \phi^{\text{AUDIT}}\xi^{a}\gamma^{a}\underline{s}^{a} + \frac{\phi^{\text{AUDIT}}\xi^{a}\gamma^{a}m^{a}}{2} = \alpha_0 + \alpha_1 m^{a};$$
$$\pi^{h} - \pi^{\ell} - \phi^{\text{AUDIT}}\xi^{u}\gamma^{u0} - \phi^{\text{AUDIT}}\xi^{u}\gamma^{u}\underline{s}^{u} + \phi^{\text{AUDIT}}\xi^{u}\gamma^{u}\underline{s}^{a} + \frac{\phi^{\text{AUDIT}}\xi^{u}\gamma^{u}m^{u}}{2} = \beta_0 + \beta_1 m^{u}.$$

An increase in ω^h leads to higher β_0.

Now let us illustrate how ω^h affects compliance levels, using the CES utility function:

$$\begin{aligned}
&\max_{\{\mathbb{m}_a,\, \mathbb{m}_u\}} U(\mathbb{m}_a, \mathbb{m}_u) = (\mathbb{m}_a^{\rho} + \mathbb{m}_u^{\rho})^{1/\rho} \\
&\text{s.t. } (\alpha_0 + \alpha_1\mathbb{m}_a)\mathbb{m}_a + (\beta_0 + \beta_1\mathbb{m}_u)\mathbb{m}_u = (\alpha_0 + \alpha_1) + (\beta_0 + \beta_1) - \eta, \\
&\text{where } \mathbb{m}_a \in (0,1),\ \mathbb{m}_u \in (0,1),\ \alpha_1 > 0,\ \beta_1 > 0, \text{ and } \eta > 0.
\end{aligned}$$

The Lagrangian becomes

$$\mathcal{L} = U(\mathbb{m}_a, \mathbb{m}_u) - \lambda\left[(\alpha_0 + \alpha_1\mathbb{m}_a)\mathbb{m}_a + (\beta_0 + \beta_1\mathbb{m}_u)\mathbb{m}_u - (\alpha_0 + \alpha_1) - (\beta_0 + \beta_1) + \eta\right].$$

The first order conditions are

$$\begin{aligned}
&U_a - \lambda^*(\alpha_0 + 2\alpha_1\mathbb{m}_a^*) = 0 \\
&U_u - \lambda^*(\beta_0 + 2\beta_1\mathbb{m}_u^*) = 0 \\
&-(\alpha_0 + \alpha_1\mathbb{m}_a^*)\mathbb{m}_a^* - (\beta_0 + \beta_1\mathbb{m}_u^*)\mathbb{m}_u^* + (\alpha_0 + \alpha_1) + (\beta_0 + \beta_1) - \eta = 0.
\end{aligned}$$

Take the derivative w.r.t β_0 on the first-order conditions:

$$\begin{bmatrix} U_{aa} - 2\lambda^*\alpha_1 & U_{au} & -(\alpha_0 + 2\alpha_1\mathbb{m}_a^*) \\ U_{ua} & U_{uu} - 2\lambda^*\beta_1 & -(\beta_0 + 2\beta_1\mathbb{m}_u^*) \\ -(\alpha_0 + 2\alpha_1\mathbb{m}_a^*) & -(\beta_0 + 2\beta_1\mathbb{m}_a^*) & 0 \end{bmatrix} \times \begin{bmatrix} \frac{\partial \mathbb{m}_a^*}{\partial \beta_0} \\ \frac{\partial \mathbb{m}_u^*}{\partial \beta_0} \\ \frac{\partial \lambda^*}{\partial \beta_0} \end{bmatrix} = \begin{bmatrix} 0 \\ \lambda^* \\ \mathbb{m}_u^* - 1 \end{bmatrix}$$

Let

$$\begin{aligned}
\tilde{U}_{aa} &= U_{aa} - 2\lambda^*\alpha_1; \\
\tilde{U}_{uu} &= U_{uu} - 2\lambda^*\beta_1; \\
\tilde{p}_a &= \alpha_0 + 2\alpha_1\mathbb{m}_a^*; \\
\tilde{p}_u &= \beta_0 + 2\beta_1\mathbb{m}_u^*.
\end{aligned}$$

Notice that the effective price of $\mathbb{m}_a$ ($\mathbb{m}_u$) becomes $\tilde{p}_a$ ($\tilde{p}_u$), which is strictly greater than the

original p_a (p_u). Thus, the system of equations has the following form:

$$\begin{bmatrix} \tilde{U}_{aa} & U_{au} & -\tilde{p}_a \\ U_{ua} & \tilde{U}_{uu} & -\tilde{p}_u \\ -\tilde{p}_a & -\tilde{p}_u & 0 \end{bmatrix} \times \begin{bmatrix} \frac{\partial \mathbb{m}_a^*}{\partial \beta_0} \\ \frac{\partial \mathbb{m}_u^*}{\partial \beta_0} \\ \frac{\partial \lambda^*}{\partial \beta_0} \end{bmatrix} = \begin{bmatrix} 0 \\ \lambda^* \\ \mathbb{m}_u^* - 1 \end{bmatrix}$$

The sign of $\partial \mathbb{m}_u^* / \partial \beta_0$ depends on the sign of

$$-\lambda^* \tilde{p}_a^2 - (1 - \mathbb{m}_u^*) \frac{\tilde{p}_u}{U_u} (U_u U_{aa} - U_a U_{au} - 2\lambda^* \alpha_1 U_u),$$

and the sign of $\partial \mathbb{m}_a^* / \partial \beta_0$ depends on

$$\lambda^* \tilde{p}_a \tilde{p}_u + (1 - \mathbb{m}_u^*) \frac{\tilde{p}_u}{U_u} (U_u U_{au} - U_a U_{uu} + 2\lambda^* \beta_1 U_a).$$

As a result, we have the same implications as those in the previous exercise.