Hospital Reimbursement in the Presence of Cherry Picking and Upcoding

Nicos Savva
London Business School
nsavva@london.edu

Laurens Debo, Robert A. Shumsky
Tuck School of Business - Dartmouth College
laurens.g.debo@tuck.dartmouth.edu, robert.a.shumsky@tuck.dartmouth.edu

Hospitals throughout the developed world are reimbursed on the basis of diagnosis-related groups (DRGs). Under this scheme, patients are divided into clinically meaningful groups, and hospitals receive a fixed fee per patient episode tied to the patient DRG. The fee is based on the average cost of providing care to patients who belong to the same DRG across all hospitals. This scheme, sometimes referred to as 'yardstick competition', provides incentives for cost reduction, as no hospital wants to operate at a higher cost than average, and can be implemented using accounting data alone. Nevertheless, if costs within a DRG are heterogeneous, this scheme may give rise to cherry-picking incentives, where providers 'drop' patients who are more expensive to treat than average. To address this problem, regulators have tried to reduce within-DRG cost heterogeneity by expanding the number of DRG classes. In this paper, we show that even if cost heterogeneity is eliminated, such expansion will fail to completely eliminate patient cherry picking. In equilibrium, the market will bifurcate into two groups, one of which will continue to cherry-pick patients and underinvest in cost reduction, while the other group treats all patients. Furthermore, we show that DRG expansion is particularly problematic if hospitals are also able to 'upcode' patients, i.e., intentionally assign patients to a more resource-intensive DRG than needed to increase income. Upcoding increases within-DRG cost heterogeneity and amplifies cherry-picking incentives. We examine potential solutions involving yardstick competition based on input statistics.

Key words: Yardstick competition, credence goods, hospital regulation, upcoding, cherry picking.
History: Revised April 7, 2023

1. Introduction

Hospitals throughout the developed world are reimbursed based on diagnosis related groups (DRGs), a patient classification system first developed by the Operations Research Department of Yale University (Fetter 1991). Under the DRG system, patients are divided into a small number of clinically meaningful groups, which can be thought of as *hospital products*, such that resources and services required to treat patients within a group are as homogeneous as possible.[1] Hospitals are

[1] The original goal of the classification system was to allow hospitals to better measure what they "produce" and serve a tool for "budgeting, cost control, and quality control." In 1990, Professor Fetter, the inventor of the DRG classification system, was honored with the Franz Edelman Award (Fetter 1991).

then reimbursed a fixed fee for every patient episode, which is set at to the average cost of treating patients of the same DRG across other similar hospitals, after applying local adjustments, e.g., for differences in labor costs. The Inpatient Prospective Payment System (IPPS) used by the Centers for Medicare & Medicaid Services in the United States is an example of this reimbursement model (CMS.gov 2021a).

This form of reimbursement was first introduced by CMS in 1983 and similar versions were soon adopted by the private sector and internationally (Mayes 2007, Busse et al. 2013). In contrast to the retrospective cost-based reimbursement system used in the US before 1983, this payment innovation was prospective in the sense that it separated hospital pay from the intensity of services provided, thus generating incentives for cost efficiency. Hospitals treating patients at a cost lower than the DRG fixed fee would be making a profit and inefficient hospitals would have a strong incentive to improve. Furthermore, by linking the payment of each hospital to the cost of treatment of similar patients at other hospitals, the DRG system induced a form of indirect competition between hospitals, sometimes referred to as 'yardstick competition' – for any DRG, no hospital would want to operate at a cost higher than the average of all other hospitals and, as a result, the average cost itself would, in equilibrium, be reduced to the efficient level (Shleifer 1985). Importantly, this scheme can be implemented through retrospective cost accounting data alone, placing a relatively small informational burden on the regulator. For these reasons, the DRG payment system has been described as "revolutionary" (Mayes 2007).

One concern with the initial implementation of the DRG system was that it was vulnerable to cherry picking (Newhouse 1996). More specifically, because DRG definitions are based on relatively coarse clinical diagnoses, there may be heterogeneity in the complexity of patient conditions within a DRG and, therefore, in the resources required to effectively treat them. To the extent that providers can observe patient complexity and predict patient profitability, they have an incentive to selectively treat those patients with lower-than-average cost (cherry picking) and avoid treating those with higher-than-average cost (lemon dropping).[2] Empirical evidence supports this hypothesis. For example, Newhouse (1989) finds disproportionately more high-costs patients to be treated in hospitals of last resort that do not have the option to turn patients away and KC and Terwiesch (2011) find evidence that focused cardiology hospitals selectively admit "easy-to-treat" patients. In a similar vain, Shactman (2005) reports that "specialty hospitals [...] concentrate on certain DRGs and treat relatively low-severity cases within them." Alexander (2020) exploits a natural experiment to show that physician financial incentives is one mechanism that gives rise to such cherry picking behaviour.

[2] Throughout the paper we refer to the activity of overly selecting low-cost patients as cherry picking and the reciprocal activity of dropping high-cost patients as lemon dropping.

To address the cherry-picking problem, the DRG system has been refined over the past 30 years, primarily by increasing the number of DRG classes to better reflect patient severity (Shactman 2005). A larger number of DRG classes, with patients of higher severity being allocated to a different DRG class than patients of lower severity, should reduce cost heterogeneity within a DRG and limit providers' ability to cherry pick (or lemon drop) profitable (unprofitable) patients. For instnance, in 1983 there were only two DRGs associated with concussion – 31 for concussion with complications and comorbidities (CCs) and 32 for concussion without CCs, with payment increasing by 33% for treating patients with CC compared to treating those without (Latta and Helbing 1991). In 2020, there were three DRGs associated with concussion – 090 for concussion without CCs, 089 for concussion with CCs, and 088 for concussion with multiple CCs, with payment increasing by 16% and 64% from 090 to 089 and 088, respectively.[3] Indeed, when the DRG system was first implemented in the US in 1983 there were 467 DRGs. The number in 2020 was 761 (a relative increase of 63%). A similar trend can also be observed in Europe, where between 2005 and 2011 the number of DRGs increased by 36% in Germany, 127% in the UK, and 239% in France (Busse et al. 2013).

A second issue with the DRG system is 'upcoding,' also referred to as 'DRG creep,' where providers modify patient diagnosis (e.g. by including additional CCs) or even provide unnecessary treatments in order to push the DRG assignment to one that commands a higher fee.[4] This problem arises due to the fact that care provision is essentially a credence good, where the regulator and/or the patient are not able to assess if the diagnoses and/or treatments provided accurately reflect the needs of the patient (Dulleck and Kerschbamer 2006). The work of Jürges and Köberlein (2015) provides an interesting instance of upcoding in neonatology. Extremely premature babies are expensive to treat and, naturally, are assigned to a higher-paying DRG compared to babies at or close to full term. Due to the approximate nature of estimating gestation term, the assignment to DRGs associated with prematurity in Germany is based on birth weight, with substantial discontinuities at specific weights (e.g. 1000g and 1500g). The authors compare the distribution of reported birth weights before the DRG system was implemented to birth weights after implementation. Under the new DRG system, the number of babies reported to have been born weighing just below the DRG thresholds for prematurity increased substantially, with a similar drop in the number of babies reported just above the threshold. The authors estimate that, in the period 2003–2010, upcoding generated additional reimbursement of 100M Euro ($ 133M using the average 2010 exchange rate).

[3] See Table 5 `https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/Downloads/FY2020-FR-Table-5.zip`

[4] For the purposes of this research we use the term 'upcoding' to refer to both the process of modifying patient diagnosis and the process of providing unnecessary treatment done for reimbursement purposes.

Beyond neonatology in Germany, Silverman and Skinner (2004) find evidence that, following the implementation of DRG payments in the US, patients with respiratory conditions were more likely to be upcoded to the more generous DRG for pneumonia and more so at for-profit hospitals. Dafny (2005) find that following a change in DRG prices hospitals responded by upcoding patients to DRGs with the largest price increase, and Psaty et al. (1999) estimate the cost of upcoding in heart failure in the US to be as high as $993M per year.

The focus of this paper is on the interaction between the two phenomena described above: Patient cherry picking and upcoding. In particular, we develop a unifying theoretical framework that extends the classic model of yardstick competition (Shleifer 1985) by explicitly modeling patient heterogeneity within DRGs and allowing for cherry picking and upcoding. The model confirms that, under the standard yardstick competition scheme applied in practice, providers will indeed have every incentive to lemon drop patients whose cost of treatment is higher than the DRG average. Furthermore, dropping patients will distort investment incentives, resulting in higher costs compared to socially optimal levels. More importantly, we show that, in the absence of upcoding, increasing the number of DRGs so that costs within DRG become homogeneous is effective in improving welfare but it does not completely eliminate cherry-picking incentives. In fact, we show that the unique equilibrium outcome of the yardstick competition game is asymmetric – the market bifurcates into two groups of providers: one group that chooses to specialize in treating patients with relatively minor needs (and drop patients with complex needs); and a second generalist group that treats everyone, irrespective of their care needs. Compared to the second group, the first group's treatment costs are relatively high, but nevertheless both groups make a positive profit – the first loses money on treating patients but this loss is more than offset by reduced investment costs, and the second group makes a profit on treating patients but this profit is partially eroded by higher investment costs. This result may provide a complementary explanation for the recent proliferation of specialist providers, which are often found to cherry-pick less-complex patients (Shactman 2005, KC and Terwiesch 2011) – in addition to the benefits of focus that have been identified by the literature (Freeman et al. 2020), our model suggests that the emergence of such cherry-picking providers, which co-exist with efficient all-purpose providers, may be the rational response of the industry to the expansion in the number of DRG classes within a condition.

Turning to the case where providers can also upcode patients, we show that expanding the number of DRGs is much less effective. Because the practice of upcoding 'mixes' some low-cost patients into the high-cost category, costs within a DRG become once again heterogeneous. As a result, providers have an even stronger incentive to drop high-cost patients and, in most cases, underinvest in cost reduction compared to socially optimal level levels. In fact, we show that in the

presence of upcoding, under some conditions, increasing the number of DRGs will actually reduce total welfare.

We conclude by investigating solutions to this problem. One potential solution is to move from prospective payments (DRG-system) to retrospective (or cost-of-service) reimbursement. We show that such a move eliminates the incentives inherent in prospective payment systems to cherry pick and upcode, but at the expense of removing incentives for cost reduction. Instead, we show that there exists an alternative yardstick competition scheme, which can be implemented with currently availably information based on 'input statics,' that can eliminate upcoding and can also reduce (and sometimes eliminate) the problem of cherry picking. More specifically, the regulator could add an additional payment that is proportional to the difference between the number of patients actually treated and the expected number of patients treated at each provider for every DRG. The expected number of patients could be calculated using appropriately adjusted benchmarks based on patient statistics drawn from all providers, as done in other regulatory schemes used in practice (e.g., the Hospital Readmissions Reduction Program (HRRP) introduced by the Centers of Medicare & Medicaid Services (CMS) in 2012, where each hospital's observed readmission rates in a number of monitored conditions are compared against a benchmark of expected readmission rates constructed using a national panel of hospital readmissions (Chen and Savva 2018)). This new payment will be positive (a bonus) if the provider treats more patients relative to other providers and negative (a penalty) if the provider treats relatively fewer patients. Therefore, it generates yardstick competition incentives where no provider wants to treat fewer patients than average for any DRG. As we show, it is possible to use information on the costs of other providers to set the penalties in a way that completely eliminates upcoding. It is more difficult to set the penalty sufficiently high in order to eliminate cherry picking, but we show that any positive penalty reduces the aggregate number of patients lemon dropped and, thus, improves welfare.

Finally, in a series of extensions, we show that our findings are robust to alternative model specifications. More specifically, we show that under quite mild conditions providers will not find it optimal to downcode patients (i.e., assign a lower complexity DRG tfrehan warranted based on the patient diagnosis), and examine the case where i) engaging in patient cherry picking or upcoding is itself costly; ii) transfer payments between the payer and the health care providers are not allowed; iii) there are two providers that are asymmetric in the number and the proportion of high complexity patients they treat.

2. Literature Review

The use of relative benchmarks (i.e., yardstick competition) in the context of hospital reimbursement was first analyzed by Shleifer (1985) – see also Holmstrom (1982), Nalebuff and Stiglitz (1983),

Laffont and Tirole (1993) for the use of relative benchmarks in other settings. This early work demonstrates that yardstrick competition is effective in inducing regulated firms (e.g., healthcare providers) to exert more effort (e.g., for cost cutting) in a setting where they are privately informed about the cost of effort but the outcomes (e.g., costs) are ex post verifiable. Sobel (1999) examines yardstick competition in a setting where in addition to effort in cost reduction the providers need to also invest in ex ante cost-reducing innovation. The paper shows that yardstick competition is effective in incentivizing effort but may fail to incentivize innovation. Lefouili (2015) extends this analysis to the case were there are no transfer payments. A common feature of these studies is that the quality of medical care is assumed to be exogenous and fixed. Tangerås (2009) provides an extension where, in addition to costs, quality is also endogenous and hospitals compete directly based on quality (but not costs). The paper finds that the use of yardstick competition reduces informational rents (compared to individual regulation) but it is fairly complex and thus difficult to implement. Continuing with the theme of exploring the implications of yardstick competition on dimensions other than costs, Savva et al. (2019) examine patient waiting times – a dimension of service quality that involves an externality (see also Naor (1969)). The authors show that the form of yardstick competition implemented in practice fails to incentivize wait-time reduction and propose modifications that better serve this purpose. More recently, the literature has examined yardstick competition as applied in HRRP. The program penalizes hospitals whose (risk-adjusted) readmission rates in a number of targeted conditions is higher than the national average. The research points out shortcomings with the current HRRP implementation that may lead to equilibrium outcomes where some hospitals are not incentivized to improve or an equilibrium might not exist (Zhang et al. 2016, Arifoglu et al. 2021).

In parallel to the work that examines the properties of prospective payments set through yardstick competition, another stream of literature has focused on comparing the performance of exogenously-set prospective payments against cost-based retrospective payments in settings where healthcare providers are local monopolists (Ellis and McGuire 1986, Dranove 1987, Ma 1994) or compete directly for patients on the basis of the quality offered (Pope 1989, Ellis 1998). For example, Dranove (1987) shows that if patient costs are heterogeneous, inefficient hospitals will "dump" high-cost patients, thus raising the cost of care of efficient hospitals. Ma (1994) shows that, if quality is endogenous and hospitals can dump patients, then prospective payments may provide incentives to reduce costs but at the same time, if there is cost heterogeneity, they will induce patient cherry picking which distorts investment incentives. Ellis (1998) finds that, in addition to patient dumping, prospective payments may induce hospitals to overinvest in quality for low cost patients (in order to attract more of them) or choose to underinvest in quality for high-cost patients, phenomena he dubs "creaming" and "skimping", respectively. A general conclusion from

this literature is that measures that reduce patient cost heterogeneity, such as increasing the number of DRGs, can mitigate the adverse incentives created by prospective payments (e.g., patient dumping) – a conclusion that regulators have implemented in earnest.

In contrast to the stream of literature introduced in the first paragraph, the stream of literature discussed in the second paragraph does not focus on problems of asymmetric information on costs – it either assumes costs are fixed (e.g., Dranove 1987, Ellis 1998) or that the regulator knows the cost technology and is able to set prospective payments exogenously (e.g., Ma 1994). Our work brings together elements from the previous two streams of literature by examining how prospective payments, which due to asymmetric information on costs are set endogenously via yardstick competition, affect patient cherry picking and investment incentives.

More importantly, the literature described above assumes that the patients' medical needs are verifiable at least ex post and, therefore, the provider administers and gets paid for the treatment needed – in other words, a patient without complications or comorbidities (i.e., has minor care needs) will not receive (or be billed as having received) care intended for patients with complications and comorbidities (a major service). In this paper we relax this assumption. In this sense this work borrows ideas from the literature of credence goods, first introduced by Darby and Karni (1973) as an additional category to complement search and experience goods. Unlike search goods, where the customer (or, in our case, the payer) can assess the value of a product before purchase (e.g., a new pair of jeans), or experience goods, where the customer finds out the value of the good after purchase and consumption (e.g., a bottle of wine), in the case of credence goods the customer is not able to tell even after purchase what the value of the good was. A canonical example of a credence good is healthcare, where a customer's problem is diagnosed and treated by an expert who is better informed than the customer (see e.g., Gottschalk et al. 2020). As a result, even after the service is provided the customer is not able to judge if it was appropriate (or indeed in some cases whether it was actually provided). The literature on credence goods is excellently summarized by Dulleck and Kerschbamer (2006). A general finding is that the expert has an incentive to overtreat (provide a higher intensity treatment than needed) and overcharge (charge for a treatment of higher intensity than the one provided). As Debo et al. (2008) show, the incentive to overtreat is more pronounced when the expert's workload level is relatively low. Our work adds to this literature by examining how yardstick competition, which naturally generates incentives for cost reduction, interacts with the credence character of healthcare provision. Importantly, our work shows that once the credence nature of healthcare is accounted for, the conclusion that reducing cost heterogeneity (by increasing the number of DRGs) will ameliorate cherry picking problems is no longer valid. The process of overtreating/overcharging patients reintroduces cost heterogeneity, which amplifies patient cherry picking incentives.

Finally, our work is also related to recent literature in Operations Management that examines the problem of designing payment schemes for healthcare. This includes So and Tang (2000), Lee and Zenios (2012), Jiang et al. (2012), Gupta and Mehrotra (2015), Adida et al. (2017), Zorc et al. (2017), Guo et al. (2019), Adida and Bravo (2019), Aswani et al. (2019), Jiang et al. (2020), Delana et al. (2021), Nassiri et al. (2021), amongst others. Typically, this literature does not examine the credence-good character of healthcare services or reimbursement via yardstick competition.

3. Model

We consider a setting where a Health Organization (HO) acts as a welfare-maximizing regulator and payer of healthcare provision and is responsible for $N \geq 2$ identical non-competing profit-maximizing service providers (e.g., hospitals) that provide service to a large population of patients as needed (e.g., when they fall ill). We expand on the objectives, decisions, and payments of each of the interacting parties below. We will present a summary of the main assumptions in Figure 1. Many modeling choices are made to ensure the model (and more importantly the results) are easily comparable to the classic model of yardstick competition by Shleifer (1985).

3.1. Patients

In the catchment area of each service provider there is a large volume of patients λ that may independently visit the service provider in case they experience a service need (e.g., a healthcare problem). The latter happens with a relatively small probability q per time period (e.g., a year), resulting in demand of λq. (In Appendix 3 we examine the case where the demand λq is asymmetric across providers.) The provider will restore the patient's health and will generate a positive welfare for the patient which we denote by U_0.

To focus the analysis on the actions of the providers rather than the patients, throughout this work we make two simplifying assumptions. First, that patients co-pays are zero, as they would be in some insurance plans or in national health care systems where care is free at the point of access. Second, to help with analytical tractability we assume that patients do not exercise choice and visit only one default provider. While this is clearly a simplification, empirical evidence suggests that in some settings this may not be far from reality. For example, Victoor et al. (2016) provide evidence based on semi-structured interviews that "most patients tend to visit the default [hospital] without being concerned about choice." We nevertheless acknowledge that it would be interesting for future work to extend our analysis to the case where patients exhibit choice.

3.2. Service providers

Provider costs: Once a patient develops a health problem they arrive at one of the $N \geq 2$ providers and they receive service that fully resolves their health problems. For this to happen,

some of the patients (proportion $1-h$) require a minor intervention which costs the provider on average $c_m = c$ and some (proportion h) require a major intervention which costs on average $c_M = c+\delta$, where $c, \delta > 0$. We will refer to these two groups of patients as low- and high-complexity, respectively. (In Appendix 3 we examine the case where providers are asymmetric in the proportion of high-complexity patients.) The costs assumptions reflect the fact that the service provided to low-complexity patients is a subset of the services provided to high-complexity patients.

The provider starts from historical average cost levels c_0, δ_0 and may decide to invest to reduce the average cost of treatment to some lower level c, δ. The cost of such investment is represented by $R_c(.)$ and $R_\delta(.)$, with $R_c(c_0) = 0$, $\frac{d}{dc}R_c() < 0$, $\frac{d^2}{d^2c}R_c() > 0$ and similarly for $R_\delta(.)$. This formulation assumes that reducing costs is relatively cheap when costs are high but it becomes increasingly more expensive to continue doing so. It also captures positive spillovers from low- to high-complexity patients, but not vice versa, in the sense that any investment to reduce the cost of the minor component of the treatment c will reduce the cost of both patients needing the minor and major intervention (c_m and c_M, respectively) but any intervention that reduces the cost of the major component of the treatment δ will not have an impact on the costs of the minor treatment. Furthermore, this specification ensures that the average cost of the major treatment is always greater than the average cost of the minor treatment.

Upcoding: The provider, but not the patient or the HO, is able to determine the type of service required by the patient – the service has a credence character. Although the provider may not undertreat patients (i.e., may not offer the minor service to a complex patient), it may upcode a proportion of patients $0 \leq \alpha \leq \bar{\alpha}$ (e.g., as described in the case of coding premature babies in Germany, presented in the introduction). The assumption that undertreatment is not possible is consistent with the fact that providers may be held financially and criminally responsible if patients are systematically discharged without being offered the appropriate level of treatment and it is not consistent with the Hippocratic Oath. In the credence goods literature it is sometimes referred to as the "liability rule" (Dulleck and Kerschbamer 2006). The cost associated with low-complexity patients who are upcoded is on average $c'_m = c_m + \beta(c_M - c_m) = c + \beta\delta$, where $0 < \beta < 1$. This formulation for upcoding nests two special cases: i) overbilling ($\beta \to 0$), where upcoded patients actually receive the minor treatment or a treatment very close to it in terms of costs ; ii) overtreatment ($0 < \beta < 1$), where at least some elements of the major treatment are offered to low-complexity patients. The fact that the average cost of overtreatment is lower than the average cost of the major treatment (i.e., $\beta < 1$) reflects the fact that upcoded patients are relatively healthier and are offered only a subset of the more expensive major treatment, or may be less time-consuming to treat and experience fewer complications. For the main analysis we assume that providers do

not engage in 'downcoding' (i.e., do not code patients that receive the major intervention as having received the minor intervention for reimbursement purposes) – as we show in Appendix A2.1 under reasonable conditions downcoding is not optimal.

Cherry picking and lemon dropping: The provider will treat all low-complexity patients but may select not to treat a fraction $0 \leq \gamma \leq \bar{\gamma}$ of the complex patients. If a high-complexity patient is denied service, they will have to seek service outside the system – for example in taxpayer supported safety-net hospitals or academic medical centers at a higher average cost $c_{out} > c_0 + \delta_0$ covered by the HO. The assumption that providers will not deny service to low-complexity patients is consistent with practice – providers would find it difficult to justify denying service to patients with relatively simple needs. We will refer to the fraction γ of patients dropped by the provider as the lemon-dropping rate. Such lemon dropping could take place through several mechanisms, such as offering physicians direct or indirect financial incentives to select low-cost patients (as empirically documented in Alexander (2020)) or ex ante through the hospital's investment decisions (e.g., specialist hospitals that choose not to invest in the staff and resources required to treat patients needing a major intervention (Shactman 2005)). The increased average cost c_{out} of treating such patients could be due to the fact that the 'hospital of last resort' operates with a less cost-efficient technology, or it could represent the direct reduction in patient welfare associated with the inconvenience of not receiving care at their hospital of first choice.[5]

We note that in the above formulation, the costs associated with treating patients (c, $c+\delta$, $c+\beta\delta$, c_{out}) are averages and the realized costs are subject to substantial (ex post) uncertainty (i.e., some patients requiring the major treatment will end up costing more than the average cost $c+\delta$ and others less). Nevertheless, providers are not able to estimate realized cost ex ante and, therefore, cannot lemon drop patients based on realized costs (i.e., they can only lemon drop patients based on complexity). In addition, despite the fact that the average cost of an upcoded patient may be lower than the average cost of providing the major treatment, the residual (ex post) cost uncertainty makes it impossible for the HO to determine with certainty that upcoding is taking place (i.e., the HO cannot rule out that a typical upcoded patient is a high-complexity patient that turned out to be cheaper than average). Furthermore, since the HO and providers are risk neutral, for the rest of the analysis we will ignore this ex post cost uncertainty and work directly with average costs.

[5] Instead of assuming that lemon dropped patients are treated outside the system it is possible to assume that (at least) some of them receive treatment at another provider within the system. The main results of the paper remain qualitatively unchanged.

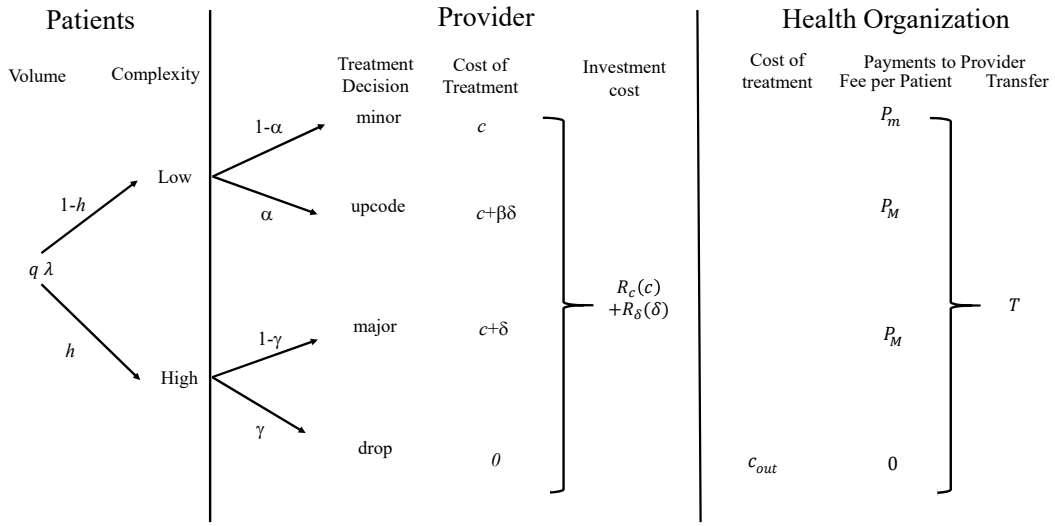

Figure 1 **Main modeling assumptions:** A large population λ may develop a medical need with probability q and seek care at one of N providers. A proportion h of patients is of high complexity and may be given the major treatment at an average cost $c+\delta$ or may be 'dropped' in which case they will receive treatment outside the system at an average cost c_{out} paid for by the HO. A proportion $1-h$ of patients is of low complexity and may be given the minor treatment at an average cost of c or may be 'upcoded' to the major treatment at an average cost of $c+\beta\delta$, where $0<\beta<1$. Providers need to decide on the cost level c and δ they want to operate at by investing $R(c)$ and $R(\delta)$, respectively, and the rate of patient upcoding α (up to a limit $\bar{\alpha}$) and dropping γ (up to a limit $\bar{\gamma}$). The HO decides on reimbursement rates for minor and major treatments (P_m and P_M, respectively) and transfer payment (T) to be paid to the providers.

Provider Payments and Objective: Providers receive payments p_M and p_m from the HO for each treatment provided in a period (typically a year). The payments may depend on the type of treatment, major or minor respectively, but not on the type of patient (low- or high-complexity) as the former is observable by the HO but not the latter. Providers may also receive a direct transfer payment $T \geq 0$ from the HO. For example, in the UK alongside the prospective payment system based on DRGs (referred to as the National Tariff) hospitals have also received transfer payments (referred to as block contracts). In Appendix A2.4 we extend the model to the case where transfer payments are not allowed. Providers are assumed to be profit maximizers and their profit is given by:

$$\begin{aligned}\pi(c,\delta,\alpha,\gamma) = T &+ \lambda q\left((h(1-\gamma)+(1-h)\alpha)\,p_M + (1-h)(1-\alpha)p_m\right) \\ &-\lambda q\left((h(1-\gamma)+(1-h)\alpha\beta)\delta + (1-h\gamma)c\right) - R_c(c) - R_\delta(\delta).\end{aligned} \tag{1}$$

The first line represents the provider's revenues, which include a transfer payment and fee for every treatment provided (including upcoding) and the second line represents the provider's costs, including the cost of providing treatments and the investment in cost reduction. We note that,

in the preceding discussion, we have effectively assumed that lemon dropping and upcoding are costless up to an exogenous threshold ($\bar{\gamma}$ and $\bar{\alpha}$, respectively) and infinitely expensive after this limit. In Appendix A2.3 we extend the model by presenting a continuously increasing cost (or, equivalently, risk of detection) in the proportion of upcoding or lemon dropping. We also note that except for costs c, δ and rates of upcoding and lemon-dropping α, γ, providers are identical – in particular, they are not differentiated on any dimension of clinical or service quality.

3.3. The Health Organization

The HO, such as the CMS in the US and the NHS in the UK, acts as the regulator of and payer for healthcare provision. We make the usual assumption in this literature that the HO's objective is to maximize total welfare subject to the constraint that providers break even ($\pi \geq 0$). Total welfare is the sum of the patient welfare and the provider profit within the regulated period (e.g., a year). As typical in such models, we assume that all within system payments (i.e., the transfer payments T, the per patient fees p_M and p_m) can be collected frictionlessly and therefore do not affect welfare directly. Of course, they affect welfare indirectly as they may influence provider actions.

The HO's objective function (for each provider) is given by:

$$U(c, \delta, \alpha, \gamma) = \lambda q U_0 - \left[(h(1-\gamma) + (1-h)\alpha\beta)\delta + (1 - h\gamma)c\right]\lambda q - R_c(c) - R_\delta(\delta) - \gamma h \lambda q c_{out}, \quad (2)$$

where U_0 is the patient welfare generated by having their health problem resolved irrespective of where the treatment is provider or of the complexity of the treatment. The next terms represent the provider cost (including both the variable cost and the investment cost) and the final term represents the additional cost (or the direct patient welfare loss) associated with providing treatment to lemon-dropped patients. We note that more general formulations, where patient welfare is a non-linear decreasing function of the lemon-dropping rate (γ) and/or the upcoding rate (α) would generate similar results. The total welfare over all N providers is simply the sum of (2) over all providers.

3.4. HO's (first-best) solution

The first-best solution, which maximizes the HO's objective subject to the constraint that the providers break even, is given by the following equations, which derive from the problem's first order conditions:

$$\gamma^* = \alpha^* = 0, -\frac{d}{dc}R_c(c^*) = \lambda q, -\frac{d}{d\delta}R_\delta(\delta^*) = \lambda q h,$$

and any combination of payments (T, p_M, p_m) that satisfies $T \geq R_c(c^*) + R_\delta(\delta^*) - \lambda q(h(p_M - c^* - \delta^*) + (1-h)(p_m - c^*))$. The second-order conditions $\frac{d^2}{d^2 c}R_c() > 0$, $\frac{d^2}{d^2\delta}R_\delta() > 0$ ensure that the first order conditions are necessary and sufficient to characterize the optimal solution. For the rest of

this work we assume that they are both satisfied and we focus on the more interesting case where the initial costs satisfy $c_0 > c^*$ and $\delta_0 > \delta^*$. We also assume that the HO will choose the minimum transfer payment that satisfies the providers' participation constraint ($\pi = 0$).

The solution makes intuitive sense. Since lemon-dropping patients generates additional costs ($c_{out} > c_0 + \delta_0$) the HO finds it optimal not to do it ($\gamma^* = 0$). Similarly, since upcoding patients increases costs ($\beta > 0$) without conferring any benefit to the patient, the HO finds it optimal not to upcode ($\alpha^* = 0$). The last two conditions determine the optimal costs of providing care – the HO finds it optimal to reduce the cost of care c such that the marginal benefit of reducing the cost by Δc for the λq patients requiring treatment is equal to the investment cost associated with such a reduction ($-\frac{d}{dc}R_c(c)$), and similarly for the proportion $h\lambda q$ that require major treatment. Note that, as consequence of the assumption that $\frac{d^2}{di^2}R_i(.) > 0$, investment in cost reduction exhibit economies of scale – the more patients a provider treats the first best solution shifts to lower costs. This is consistent with empirical evidence in this industry (Freeman et al. 2020).

Clearly, for the HO to implement this solution they need to know the cost functions $R_c(.), R_\delta(.)$, the disease incident rate (q) and the proportion of patients that require major treatment (h). Given the complex and ever-changing nature of healthcare costs and medical technology, this is a tall order. For the rest of the paper, we investigate payment schemes that do not place this unrealistic informational burden on the HO.

We note that, when the costs of treating patients are ex ante homogeneous (i.e., if $\delta_0 = 0$), or if it is impossible to lemon drop or upcode patients (i.e., if $\bar{\gamma} = \bar{\alpha} = 0$) then the model reduces to one equivalent to Shleifer (1985) albeit with the difference that we assume that patient demand is exogenous to prices. This is a simplification that is realistic in the context of national health systems (e.g., Medicare in the US, NHS in the UK).

4. Cost-based yardstick competition

Since implementing the first best solution directly places an unreasonable informational burden on the HO, in this section we will investigate the effectiveness of yardstick competition, a regulatory scheme that has been implemented in practice as it requires information more readily available to the HO (Shleifer 1985). Throughout this section we will not assume that the HO knows the provider's cost functions ($R_c(.), R_\delta(.)$) or anything relating to the disease (e.g., q or h). Instead, we will assume that the HO has access to patient-level ex post costing data, which are audited for accuracy, and is able to accurately estimate the realized average cost of treating patients within a DRG at least at the provider level. This is indeed the case in the UK (see NHS Digital (2021) where the average cost data per DRG are publicly available) and in the US as described in CMS.gov (2021a)). The regulator can use this cost information when setting provider payments.

To make this more exact, we need to amend the notation by adding the subscript i on each provider's decision variable (costs c_i and δ_i, dropping rate γ_i, and upcoding rate α_i). We need to distinguish two possible cases. In the first, the HO covers the costs of provider i by paying a single fee per patient episode (p_i) irrespective of the treatment provider (i.e., $p_i = p_{Mi} = p_{mi}$) – this is equivalent to assuming that there is only one coarse DRG associated with the condition. In this case, the price p_i is set at the observed average cost per episode at all other providers, which is given by[6]

$$\bar{c}_i := \frac{1}{N-1} \sum_{j \neq i} \left[c_j + \delta_j \frac{h(1-\gamma_j) + (1-h)\alpha_j \beta}{1 - h\gamma_j} \right]. \tag{3}$$

In the second case, the HO breaks the condition into two distinct DRGs – one for the minor and another for the major treatment and covers the providers' costs by paying tow distinct fees, one for each DRGs. In this case, the fee per patient episode are given by $p_{Mi} = \bar{c}_{Mi}$, $p_{mi} = \bar{c}_{mi}$ where $\bar{c}_{Mi}$ and $\bar{c}_{mi}$ are the average costs of providing the major and minor treatments at all other providers given by

$$\bar{c}_{Mi} := \frac{1}{N-1} \sum_{j \neq i} \left[c_j + \delta_j \frac{h(1-\gamma_j) + (1-h)\alpha_j \beta}{h(1-\gamma_j) + (1-h)\alpha_j} \right] \text{ and } \bar{c}_{mi} := \frac{1}{N-1} \sum_{j \neq i} c_j, \tag{4}$$

respectively. In both cases, the realized average investment cost incurred by all other providers is given by:

$$\bar{R}_i := \frac{1}{N-1} \sum_{j \neq i} [R_c(c_j) + R_\delta(\delta_j)] \tag{5}$$

and the providers are paid a transfer payment equal to $T_i = \bar{R}_i$.

Note that, in order to implement the payments defined above, the HO must credibly commit to setting payments equal to the average costs realized by the other service providers. Furthermore, since providers are identical in our setting the average cost of other providers should be a good proxy of the cost of provider i. If, however, providers are heterogeneous in observable and exogenous dimensions (e.g., labor costs, teaching status, patient demographic characteristics) these costs could be adjusted accordingly (see Shleifer (1985) for a theoretical treatment and CMS.gov (2021a) for an description on how this is done in practice).

Given these terms, each provider decides on the cost-reduction investment (which determines the costs c_i and δ_i they will operate at) and whether they want to engage in any patient upcoding α_i or lemon dropping γ_i. Note that due to the relative benchmarking associated with yardstick

[6] To help the reader understand this formula, note that the cost of providing the minor service to patients who are not upcoded at a provider j is $\lambda q(1-h)(1-\alpha_j)c_j$, the cost of providing the major service to patients who are not lemon dropped is $\lambda q h(1-\gamma_j)(c_j + \delta_j)$ and the cost of upcoding α_j minor patients is $\lambda q(1-h)\alpha_j(c_j + \beta\delta_j)$. The total volume of patients treated, is $\lambda q(1 - h\gamma_j)$. The sum of the first three expressions divided by the fourth gives the summand.

competition, each provider's payoff depends on the actions of other providers through the fees p_i and transfer payment T_i. Therefore, even though providers do not engage in direct competition, the reimbursement mechanism described above forces them to engage in a simultaneous-move game. We will characterize the equilibrium outcome of this game for different cases.

4.1. In the absence of upcoding and cherry picking.

We begin the analysis by examining the simpler case where patient upcoding is not possible (e.g., because patient needs are observable and verifiable by the HO thus alleviating any credence goods concerns) and that providers are not able to select patients (e.g., because all hospitals have to treat every patient). This case is equivalent to setting $\bar{\gamma} = \bar{\alpha} = 0$.

Proposition 1 *Under yardstick competition, and irrespective of the number of DRGs used (one or two), in the absence of patient cherry picking and upcoding ($\bar{\gamma} = \bar{\alpha} = 0$) there exists a unique Nash equilibrium which is symmetric. All providers invest in cost reduction optimally (i.e., c^*, δ^*) and they all break even.*

The proposition above demonstrates the power of yardstick competition as first described by Shleifer (1985). Without knowing the providers' cost structure, just by observing the providers' realized (average) costs, the HO can incentivize first-best investment in cost reduction. The relative benchmark forces the providers to engage in indirect competition – no provider wants to operate at a cost that is higher than average – and, as a result, the average cost of treatment is set at the level the HO would have chosen. Furthermore, the HO can achieve this without surrendering rents to providers – the prices and transfer payments are set so as to break even. More importantly, for the purposes of this work, if cherry picking and upcoding are not possible, it does not matter if the patient categories are finely or coarsely defined – having one or two DRGs in this case makes no difference. As we shall see in the rest of this section, this is no longer the case if cherry picking and/or upcoding are a possibility.

4.2. Patient cherry picking and category expansion.

To isolate the impact of cherry picking on yardstick competition, we proceed by assuming that upcoding is not possible ($\bar{\alpha} = 0$) but providers may engage in patient cherry picking if it is profitable to do so (i.e., they may choose to drop a proportion of high-complexity patients $0 \leq \gamma_i \leq \bar{\gamma}$). If the HO implements yardstick competition without differentiating between major or minor treatments (i.e., pays the providers on the basis of a single DRG, where the payment is given by $p_{Mi} = p_{mi} = \bar{c}_i$ as defined in (3)), the providers engage in a simultaneous move game. Each provider's strategy is characterized by the tuple $(\gamma_i, c_i, \delta_i)$.

Before we characterize the equilibrium of this game, it is convenient to define the costs c^{e1}, δ^{e1} as the unique solutions to the equations

$$-\frac{d}{dc}R_c(c) = \lambda q(1-h\bar{\gamma}), \quad -\frac{d}{d\delta}R_\delta(\delta) = \lambda q h(1-\bar{\gamma}), \tag{6}$$

respectively. These are the cost levels that the HO would have chosen for a provider who engages fully in lemon dropping (i.e., given the provider's decision to drop $\bar{\gamma}$ high-complexity patients, these cost levels maximize welfare). For this reason, we will refer to them as the *cherry-picking-best* costs.

Note that the cherry-picking-best variable costs are higher than first-best variable costs (i.e., $c^{e1} > c^*$ and $\delta^{e1} > \delta^*$) – this is a consequence of the fact that a provider that engages in lemon dropping will treat fewer patients, and, because investment in cost reduction is characterized by economies of scale, the provider will find it optimal to invest less compared to first best. Although in theory it could be possible for the cost distortions associated with lemon dropping to be so extreme that they render lemon dropping unprofitable (i.e., lemon dropping becomes unprofitable if the average cost of a major treatment at cherry-picking-best costs is greater than the cost of a major treatment at first-best cost, $c^{e1} + \delta^{e1} h\frac{1-\bar{\gamma}}{1-h\bar{\gamma}} \geq c^* + \delta^*$), for ease of exposition we will make the assumption that this is not the case. (We explore the implications of such extreme cherry-picking-best costs in Appendix A2.2.)

Proposition 2 *In the absence of upcoding ($\bar{\alpha} = 0$), if the HO implements yardstick competition based on a single DRG and cherry-picking-best costs are not too extreme (i.e., $c^{e1} + \delta^{e1} h\frac{1-\bar{\gamma}}{1-h\bar{\gamma}} < c^* + \delta^*$), then there exists a unique Nash equilibrium which is symmetric. Providers drop as many patients as possible and invest in cherry-picking-best costs (i.e., all providers choose $(\bar{\gamma}, c^{e1}, \delta^{e1})$, and $c^{e1} > c^*$, $\delta^{e1} > \delta^*$). All providers break even.*

In sharp contrast to the case in which cherry picking was not possible (see Proposition 2), when providers have the ability to select patients based on their cost of treatment, the use of yardstick competition payments with a single coarse DRG is problematic in the sense that it encourages providers to drop high-complexity patients and underinvestment in cost reduction. The crux of the problem is cost heterogeneity within a DRG. Providers are reimbursed with a single fee for treating patients, irrespective of the patients' needs. This single fee is designed to cover the average cost of providing treatment and, as a result, a provider makes a profit when treating low-complexity patients and a loss when treating high-complexity patients. This naturally generates strong cherry-picking and lemon-dropping incentives. Furthermore, the best response of a provider who knows that all other providers will drop the maximum number of patients possible is to also drop as

many patients as possible – yardstick competition does nothing to ameliorate such an incentive. Nevertheless, yardstick competition still provides incentives for cost reduction – given the providers' cherry picking behavior the cost levels chosen by the providers (c^{e1}, δ^{e1}) are those that the HO would have chosen. Finally, in the symmetric equilibrium all providers receive a transfer payment (equal to the investment cost of other providers) and, thus, they break even.

We next investigate whether expanding the number of DRGs, as has been done in practice, can improve the equilibrium outcome. More specifically, the HO may use two distinct DRGs: one associated with the major and another with the minor treatment according to $p_{Mi} = \bar{c}_{Mi}$, $p_{mi} = \bar{c}_{mi}$ as defined in (4). In this case the DRG expansion is able to completely remove ex ante cost heterogeneity within DRGs.

Before we characterize the equilibrium outcome of this competition, we define the following two quantities:

$$v_1 := \lambda q(h(\delta^{e1} + c^{e1} - \delta^* - c^*) + (1-h)(c^{e1} - c^*)) + R_c(c^{e1}) + R_\delta(\delta^{e1}) - R_c(c^*) - R_\delta(\delta^*)$$
$$u_1 := \lambda q(h(1-\bar{\gamma})(\delta^* + c^* - \delta^{e1} - c^{e1}) + (1-h)(c^* - c^{e1})) + R_c(c^*) + R_\delta(\delta^*) - R_c(c^{e1}) - R_\delta(\delta^{e1})$$

The quantity v_1 is the profit of a provider who is paid according to yardstick competition based on two DRGs and chooses $(0, c^*, \delta^*)$ when all other providers choose $(\bar{\gamma}, c^{e1}, \delta^{e1})$ and vice versa for the quantity u_1. Note that $v_1 > 0$ and $u_1 > 0$ (see proof of Proposition 4).

Proposition 3 *In the absence of upcoding ($\bar{\alpha} = 0$), if the HO implements yardstick competition based on two DRGs, there exists a unique Nash equilibrium which is asymmetric: $N - \theta_1$ providers drop as many patients as possible and choose cherry-picking-best costs (i.e., these providers choose $(\bar{\gamma}, c^{e1}, \delta^{e1})$ and $c^{e1} > c^*$, $\delta^{e1} > \delta^*$) and θ_1 providers do not drop patients and invest efficiently (first-best) in cost reduction (i.e., these providers choose $(0, c^*, \delta^*)$). The number of efficient providers θ_1 is the unique integer in the interval $\left[\frac{(N-1)v_1}{v_1+u_1}, \frac{Nv_1+u_1}{v_1+u_1}\right]$. All providers receive a positive rent and total welfare is higher under two DRG compared to one DRG.*

The proposition shows that expanding the number of DRGs associated with a condition has a desirable effect. Compared to the equilibrium outcome when the HO used a single DRG, the symmetric equilibrium where all providers engage in lemon dropping and underinvest in cost reduction (see Proposition 3) disappears. Nevertheless, and quite surprisingly, eliminating cost heterogeneity is not completely effective in eliminating lemon dropping or restoring first-best investment incentives across the whole market – some providers will find it optimal to continue dropping high-complexity patients and underinvest in cost reduction.

First, we will explain why no symmetric equilibrium exists despite the fact that all providers are identical and are subject to a symmetric payment mechanism. The explanation is also a rough

sketch of the proof of Proposition 4. In a symmetric equilibrium, had one existed, all providers would have chosen the same costs (c, δ) and to drop the same proportion of high-complexity patients (γ). In contrast to the case where the HO deployed one coarse DRG definition, in this case the use of two DRGs eliminates cost heterogeneity within DRG and ensures that, in any symmetric equilibrium, providers will be paid a higher fee for treating high-complexity patients compared to low-complexity patients (i.e., $p_M > p_m$), and that these fees will cover the costs of providing treatment for both types of patients (i.e., $p_m = c$ and $p_M = c + \delta$). The transfer payment would ensure that all providers break even. Note that, in any symmetric equilibrium candidate where providers drop patients (i.e., $\gamma > 0$), there would be underinvestment in cost reduction compared to first best (i.e., $c > c^*$ and $\delta > \delta^*$) – this is a consequence of the fact that investment in cost reduction is characterized by economies of scale. A symmetric equilibrium where providers drop patients (i.e., $\gamma > 0$, $c > c^*$, and $\delta > \delta^*$) can be ruled out, as at least one provider will find it profitable to deviate. Note that deviating to any action will not change the fees the provider is paid (these only depend on the actions of other providers). Furthermore, since treating high-complexity patients is not loss making (i.e., $p_M = c + \delta$), if a deviating provider stopped dropping patients their profit would not change. But this deviating provider is now treating more patients so they would also find it optimal to invest more in cost reduction (to benefit from economies of scale) which would in turn increase their profit. Therefore, the only symmetric equilibrium candidate that survives such deviation is one where all providers take first-best actions (i.e., all providers choose $\gamma = 0$, $c = c^*$, and $\delta = \delta^*$). Nevertheless, this is not an equilibrium outcome either because the best response of a provider that knows that everyone else will not drop patients and invest in first-best cost reduction is to drop some high-complexity patients and underinvest in cost reduction compared to first best. Again, such a deviation will not change the fees the provider is paid (as these depend on the actions of other providers only). Therefore, this provider, who will now operate at a higher cost than everyone else, will experience a loss in treating any patient that they do not drop. Nevertheless, this loss is more than offset by the lower investment cost. Since such a deviation is profitable, first-best actions by all providers cannot be sustained as a symmetric equilibrium outcome and, more generally, no symmetric equilibrium outcome exists.

Instead, what emerges is an asymmetric equilibrium where the market for providing care bifurcates into two distinct groups. The first group of 'specialist' providers drops the maximum number of high-complexity patients and focuses in treating low-complexity patients. As a result of treating fewer patients than first best, these providers underinvest in cost reduction compared to first best. The second group of 'generalist' providers treat both low- and high-complexity patients (i.e., do not drop any patients) and invest optimally in cost reduction. The source of profit for these two groups are diametrically opposed. Since the generalists have a lower treatment cost than the specialists,

and since the fee per patient is equal to the average cost across the market, generalists are making a profit from treating patients and specialists are making a loss. This explains why specialists drop patients and generalists do not. In contrast, since specialists invest less in cost reduction compared to generalists, and the transfer payment is equal to the average investment cost across the market, specialists experience an investment profit and generalists an investment loss. Naturally, the more specialist providers (i.e., the smaller θ_1) the higher the fee-per-patient will be and the lower the transfer payment – this favours the generalists who treat patients profitably and harms the specialists who draw their profit from the transfer payments. As a result, the equilibrium number of providers is reached when the profit of providers in each of the two groups are roughly equal – or to be more precise, sufficiently close so that none of the providers find it optimal to deviate to the other group.

Despite the fact that expanding the number of DRGs does not restore first-best outcomes, total welfare is clearly higher if the HO uses two DRGs instead of one. This observation follows from the fact that, in contrast to the case of one DRG where all providers engaged in lemon dropping and underinvest in cost reduction, in this case only a fraction $1 - \frac{\theta_1}{N}$ does so.

Interestingly, Proposition 3 may provide an additional and complementary explanation for the relatively recent emergence of specialized providers. Such providers are often found to cherry-pick low-complexity patients (and actively avoid treating patients with comorbidities and other complications that are expensive to treat) compared to generalist providers who treat everyone (Shactman 2005, Greenwald et al. 2006, KC and Terwiesch 2011). While benefits of focus and scope may certainly play a role in the emergence of such providers (e.g., as described in the case of the Shouldice Hospital, a specialist provider of hernia operations, the focus on treating low-complexity patients has allowed the hospital to implement process standardization that generates cost efficiency and better patient outcomes (Heskett 2003), see also Freeman et al. (2020) for more large-scale empirical evidence) the result above suggests that it may be the rational response of the market to the increase in the number of DRGs associated with a given condition. Specialist providers in our model do not need to have an advantage over generalist providers in order to emerge as an equilibrium outcome. Furthermore, they generate a deadweight loss for the HO (both because they underinvest in cost reduction and because dropped patients have to be treated at an higher cost c_{out}).

In summary, the preceding analysis only partially confirms the rationale behind the progressive increase in the number of DRGs observed in multiple health systems over the previous years – removing cost heterogeneity within DRGs improves incentives but is not a panacea.

4.3. Cost-based yardstick competition with patient cherry picking and upcoding

We proceed by examining the case where, in addition to lemon dropping, providers may also engage in patient upcoding (i.e., $0 \leq \gamma_i \leq \bar{\gamma}$, $0 \leq \alpha_i \leq \bar{\alpha}$). Clearly, upcoding is meaningless in the case where there is only one coarse DRG definition as there is no higher-paying DRG code to upcode to and the results of Proposition 3 would apply. For this reason, we will focus on the case where the HO uses two distinct DRGs. The presence of two DRGs, one of which is reimbursed at a higher rate, should generate incentives for providers to upcode. What is less clear is how upcoding interacts with cherry-picking incentives.

Each provider's strategy is characterized by the 4-tuple $(\alpha_i, \gamma_i, c_i, \delta_i)$. Before we present results, we first define the costs δ^{e2}, δ^{e3} as the unique solution to the equations

$$-\frac{d}{d\delta}R_\delta(\delta) = \lambda q(h(1-\bar{\gamma}) + (1-h)\bar{\alpha}\beta), -\frac{d}{d\delta}R_\delta(\delta) = \lambda q(h + (1-h)\bar{\alpha}\beta),$$

respectively. The costs (c^{e1}, δ^{e2}) represent the costs levels that maximize welfare for a provider who engages fully in both upcoding and lemon dropping. For this reason we will refer to these costs as 'cherry-picking-upcoding-best' (CPU-best) costs. In contrast, the "upcoding-best" costs (c^*, δ^{e3}) correspond to the cost levels that maximize welfare for a provider who engages only in upcoding. It is important to note that, unlike lemon dropping, upcoding does not affect the cost of the minor treatment. Therefore, the CPU-best cost c^{e1} is the same as the one defined earlier for the case where no providers engaged in upcoding. Similarly, the upcoding-best cost c^* is the same as the first best case where no providers engage in either upcoding or lemon dropping. Finally, it is worth mentioning that the welfare-maximizing cost of the major component of treatment in the presence of upcoding is lower than the corresponding value in the absence of upcoding, i.e., $\delta^{e3} < \delta^*$ and $\delta^{e2} < \delta^{e1}$.

With these definitions in place, we are in a position to examine the basic economics of upcoding under yardstick competition. On the one hand, if a provider chooses to upcode a patient, they will receive a fee equal to the average cost of treating major patients at other providers instead of a fee equal to the average cost of treating minor patients at other providers. Therefore, upcoding will generate an additional revenue which depends on how cost efficient other providers are – the more other providers invest in cost reduction for the major treatment, the lower this additional revenue will be. On the other hand, by upcoding a patient, a provider will incur the additional cost associated with overtreatment $(\beta\delta_i)$. This additional cost depends on how cost efficient the upcoding provider is – the less they invest in cost reduction the higher this additional cost will be. It is conceivable that a provider may be so inefficient compared to other providers, so that the net effect of upcoding would be negative. Such a provider would not engage in upcoding. Since

we want to focus on upcoding, we would like to rule out such scenarios. The following inequality ensures that this is the case

$$\beta < \frac{h\delta^{e3}}{h\delta^{e1} + (\delta^{e1} - \delta^{e3})(1-h)\bar{\alpha}}. \tag{7}$$

Clearly, this is always satisfied if upcoding is primarily due to overbilling as opposed to overtreatment (i.e., β is sufficiently close to 0). For the rest of this section we assume this inequality holds.

We also define the following two quantities:

$$\begin{aligned} v_2 :=\ & \lambda q \left((h + (1-h)\bar{\alpha}) \frac{(h(1-\bar{\gamma}) + (1-h)\bar{\alpha}\beta)}{h(1-\bar{\gamma}) + (1-h)\bar{\alpha}} \delta^{e2} - (h + (1-h)\bar{\alpha}\beta)\delta^{e3} + c^{e1} - c^* \right) \\ & + R_c(c^{e1}) + R_\delta(\delta^{e2}) - R_c(c^*) - R_\delta(\delta^{e3}), \\ u_2 :=\ & \lambda q \left(\delta^{e3}(h + (1-h)\bar{\alpha}\beta) \frac{h(1-\bar{\gamma}) + (1-h)\bar{\alpha}}{h + (1-h)\bar{\alpha}} - (h(1-\bar{\gamma}) + (1-h)\bar{\alpha}\beta)\delta^{e2} + (1-\bar{\gamma}h)(c^* - c^{e1}) \right) \\ & + R_c(c^*) + R_\delta(\delta^{e3}) - R_c(c^{e1}) - R_\delta(\delta^{e2}). \end{aligned}$$

The quantity v_2 is the profit of a provider who is paid according to yardstick competition with two DRGs and chooses $(\bar{\alpha}, 0, c^*, \delta^{e3})$ when all other providers choose $(\bar{\alpha}, \bar{\gamma}, c^{e1}, \delta^{e2})$ and vice versa for the quantity u_2. Note that in this case $u_2 > 0$, but the sign of v_2 will depend on model parameters (see Proof of Proposition 4).

Finally, we will call the CPU-best costs as 'comparable' to upcoding-best costs if either of the following two conditions hold:

- The upcoding-best cost is not less than the CPU-best costs (i.e., $c^* + \delta^{e3} > c^{e1} + \delta^{e2}\frac{h(1-\bar{\gamma})+(1-h)\bar{\alpha}\beta}{h(1-\bar{\gamma})+(1-h)\bar{\alpha}}$).
- The profit of a provider who is paid according to yardstick competition based on two DRGs and chooses $(\bar{\alpha}, 0, c^*, \delta^{e3})$ when all other providers choose $(\bar{\alpha}, \bar{\gamma}, c^{e1}, \delta^{e2})$ is negative (i.e., $v_2 < 0$).

Proposition 4 *If both cherry picking and upcoding are possible, if the HO implements yardstick competition based on two DRGs, then there exists a unique Nash equilibrium:*

A. If CPU-best costs are comparable to upcoding-best costs, then the equilibrium is symmetric. Providers upcode and lemon drop as many patients as possible and invest in upcoding-cherry-picking-best costs (i.e., all providers choose $(\bar{\alpha}, \bar{\gamma}, c^{e1}, \delta^{e2})$). Furthermore, there is underinvestment in cost reduction compared to first best for the minor treatment ($c^{e1} > c^*$) and if $\beta < \frac{h}{1-h}\bar{\gamma}\bar{\alpha}$ then there is also underinvestment for the major treatment ($\delta^{e2} > \delta^*$), otherwise there is overinvestment. All providers break even.

B. Otherwise, the equilibrium is asymmetric: $N - \theta_2$ providers upcode and drop as many patients as possible and choose cherry-picking-upcoding-best costs (i.e., these providers choose $(\bar{\alpha}, \bar{\gamma}, c^{e1}, \delta^{e2})$) and θ_2 providers upcode as many patients as possible but do not engage in lemon dropping and invest in upcoding best costs (i.e., these providers choose $(\bar{\alpha}, 0, c^*, \delta^{e3})$). The number θ_2 is the only integer in the interval $\left[\frac{(N-1)v_2}{u_2+v_2}, \frac{Nv_2+u_2}{u_2+v_2}\right]$. All providers receive positive rents.

The proposition above shows that upcoding is deeply problematic. Not surprisingly, if upcoding is possible, providers will fully engage in it – yardstick competition does nothing to curtail the providers' incentives to upcode patients. What is more interesting is the finding that, in the presence of upcoding, increasing the number of DRGs does little to restrain providers from lemon dropping. We discuss why below.

Depending on model parameters, the structure of the equilibrium outcome in the presence of upcoding (Proposition 4) is sometimes similar to that of the symmetric equilibrium of Proposition 2 (Case A), and sometimes similar to the asymmetric equilibrium of Proposition 3 (Case B).

In the first case (Case A), which emerges whenever the CPU-best costs are not too extreme, all providers engage in upcoding low-complexity patients and lemon-droping high-complexity patients, despite the fact that the HO uses two DRGs. This result occurs because upcoding essentially *mixes* low-complexity patients into the high-cost major treatment category. As a result of this mixing, the average cost of patients treated under the major DRG is reduced to a level below the cost of treating high-complexity patients. And since under yardstick competition the payment is, in equilibrium, equal to the average cost (see Equation (4)), the payment will not be enough to cover the cost of treating high-complexity patients (i.e., since $\beta < 1$, for any $\gamma \geq 0$ the payment for providing the major treatment $c + \delta \frac{h(1-\gamma)+(1-h)\bar{\alpha}\beta}{h(1-\gamma)+(1-h)\bar{\alpha}}$ is lower than the cost of treating high-complexity patients $c + \delta$). Therefore, dropping high-complexity patients is now profitable. In other words, upcoding reintroduces cost heterogeneity within DRGs and, therefore, reinstates the incentive for patient lemon dropping. Furthermore, the practice of lemon dropping and upcoding also distort providers' investment incentives. For the minor component of the treatment the distortion is always towards underinvestment – since in equilibrium $h\bar{\gamma}$ patients are dropped there is a corresponding reduction in the minor component of the treatment administered and, since investment is characterized by economies of scale, there will be less investment in cost reduction. For the major component of the treatment, lemon dropping and upcoding interact in a more complex manner. On the one hand, $h\bar{\gamma}$ fewer major patients will be treated leading to a reduction in activity. On the other hand, $(1-h)\bar{\alpha}$ minor patients will receive a fraction β of the major treatment, leading to an increase in activity. Naturally, if the aggregate impact is a decrease in major activity (i.e., $h\bar{\gamma} > (1-h)\bar{\alpha}\beta$) then there will be underinvestment compared to first best and vice versa. Note that, all things being equal, when upcoding is largely due to overbilling (i.e., β is sufficiently low) the equilibrium outcome will be characterised by underinvestment.

In the second case (case B) of Proposition 4, where provider costs are inflated by lemon dropping to a high degree (i.e., CPU-best costs are no longer comparable to upcoding-best costs), the symmetric outcome described above is no longer sustained in equilibrium. The best response of a provider that knows that all other providers will drop the maximum number of patients and

underinvest in cost reduction will be to invest heavily in cost reduction so that they no longer make a loss when treating patients requiring the major treatment. Instead the market bifurcates into two groups of providers, as was the case when upcoding was not possible (see Proposition 3). All providers will continue to upcode low-complexity patients as much as possible, and some $(N-\theta_2)$ will also continue to lemon-drop high-complexity patients while the rest (θ_2) will find it optimal to treat everyone and invest even more than first-best in cost reduction. Compared to the first group of providers, the second group will make a profit out of treating patients, but this profit will be largely eroded by the the inflated investment in cost reduction. As in the case of Proposition 3, the number of providers in each group is set so that their profits are roughly equal.

In contrast to the case where upcoding is not possible, the results of Proposition 4 cast doubt on whether it is useful to expand the number of DRGs in the presence of upcoding. It seems that, in addition to the welfare loss associated with upcoding itself, upcoding reinstates providers' lemon-dropping incentives that the DRG expansion was meant to eliminate. We examine this more formally with the proposition below.

Proposition 5 *If both cherry picking and upcoding are possible and if cherry-picking-best costs are not too extreme (i.e.,* $c^{e1}+\delta^{e1}h\frac{1-\bar{\gamma}}{1-h\bar{\gamma}} < c^*+\delta^*$*) and CPU-best-costs are comparable to upcoding-best costs, then total welfare is (weakly) higher if the HO implements yardstick competition based on one DRG instead of two DRGs.*

Proposition 5 confirms that, in the presence of upcoding, the HO is better off with coarser DRG definitions, at least for the case where the cost distortions associated with cherry picking and upcoding are not too extreme (i.e., when the equilibria are symmetric). Although coarser DRG definitions do nothing to curtail cherry picking, at least they do not incentivize upcoding.

5. Numerical Example

In this section we illustrate the inefficiencies associated with lemon dropping and upcoding using a numerical example. We use publicly available data[7] and we focus on the DRGs associated with concussion, which were briefly mentioned in the introduction. In 2020 there were three DRGs associated with concussion: 088, 089, and 090 with weights 1.3891, 0.9863, and 0.8483, and with 753, 1887, and 605 discharges, respectively. The last two DRGs (089 and 090) refer to patients with at most one CC, and since they have relatively similar weights we will treat them as the minor treatment, with a volume weighted DRG weight of 0.9469 and a total number of discharges of 2640. DRG 088 refers to patients with multiple CCs and, therefore, we will treat it as the major treatment,

[7] We use Tables 5 and 7 reported here `https://www.medpac.gov/wp-content/uploads/2021/11/medpac_payment_basics_21_hospital_final_sec.pdf`

with 753 discharges. Assuming no cherry picking or upcoding, this suggests that the proportion of high complexity patients is $h = 18.64\%$. There are 62.6 million Medicare enrolles, which suggests that the rate of concussion per year is $q = 0.0052\%$. There are $N = 5,141$ US hospitals which serve on average $\lambda = 62,000$ patients each. Turning to costs, the Medicare base rate is \$6,555, which implies the average cost for the minor treatment is \$6,197 and the additional cost for the major treatment is \$2,894. We take these to be the first best costs under yardstick competition and assume that in the absence of yardstick competition the costs would have been m times higher (i.e., $c_0 = mc^*$ and $\delta_0 = m\delta^*$). The cost-multiplier $m > 1$ is a measure of the efficiency gains that can be achieved via yardstick competition (e.g., when $m = 1$ there is no need to implement yardstick competition). We parameterize the cost functions with $R_c(c) = \zeta_c(c - c_0)^2$ and $R_\delta(\delta) = \zeta_\delta(\delta - \delta_0)^2$ as in Savva et al. (2019). This specification implies that $\zeta_c = \frac{\lambda q}{2(c_0 - c^*)}$ and $\zeta_\delta = \frac{\lambda q h}{2(\delta_0 - \delta^*)}$. In the base case, we set $\bar{\gamma} = 0.2$, $\bar{\alpha} = 0.2$, $\beta = 0.2$, and $m = 1.5$, but we vary these parameters in the following intervals $\bar{\gamma} \in [0.01, 0.4]$, $\alpha \in [0.01, 0.25]$, $\beta \in [0.01, 0.22]$, $m \in [1.2, 4]$. The range of parameters was chosen to be as wide as possible around the base-case without violating the condition (7). For all parameters, cherry-picking-best costs are not extreme (see Proposition 2).

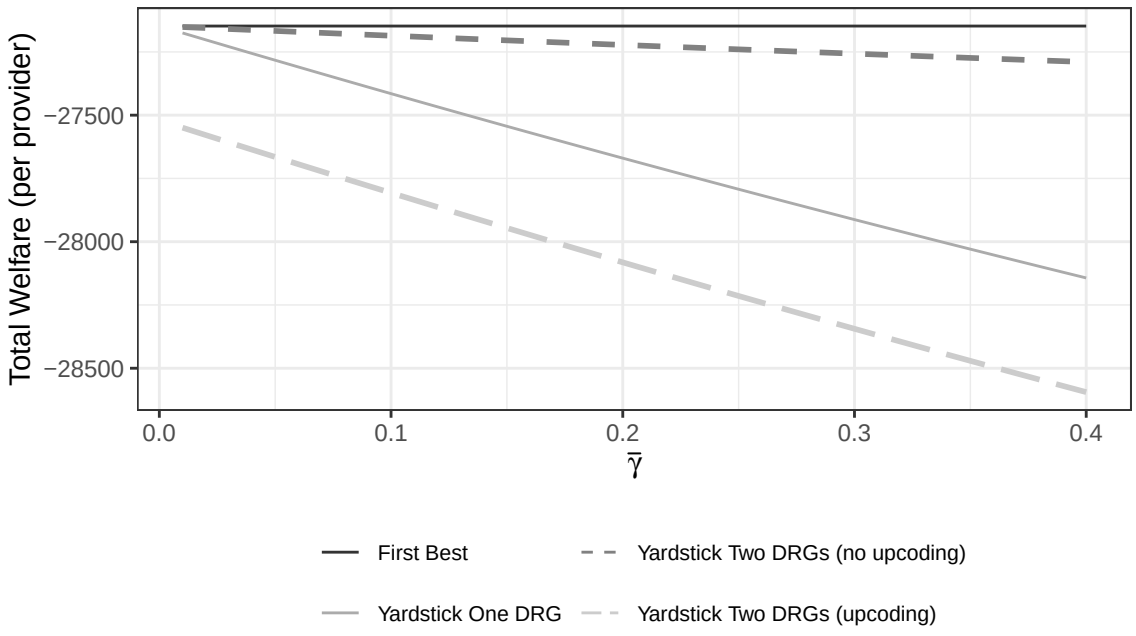

Figure 2 **Numerical example:** total welfare as a function of the patient maximum lemon-dropping rate $\bar{\gamma}$ under different scenarios (see figure legend). All other parameters are set to base value.

We normalize the utility from receiving treatment to zero ($U_0 = 0$), therefore the total welfare (per provider) depicted in Figures 2 and 3 is equivalent to costs. From Figure 2, it is clear that

if lemon dropping is possible (i.e., if $\bar{\gamma} > 0$) using one DRG generates a welfare loss which is increasing in $\bar{\gamma}$. This loss is reduced but not eliminated if the HO uses two DRGs, but only in the absence of upcoding (i.e., if $\bar{\alpha} = 0$). In this case, most providers are generalists (θ_1 =86%) (i.e., treat everyone and invest optimally in cost reduction) while the rest are specialists (i.e., lemon drop high-complexity patients and undervest in cost reduction). In the presence of upcoding, using two DRGs is counterproductive – the resulting equilibrium is symmetric and all providers upcode and, as a result, lemon drop as much as possible (i.e., $\theta_2 = 0$).

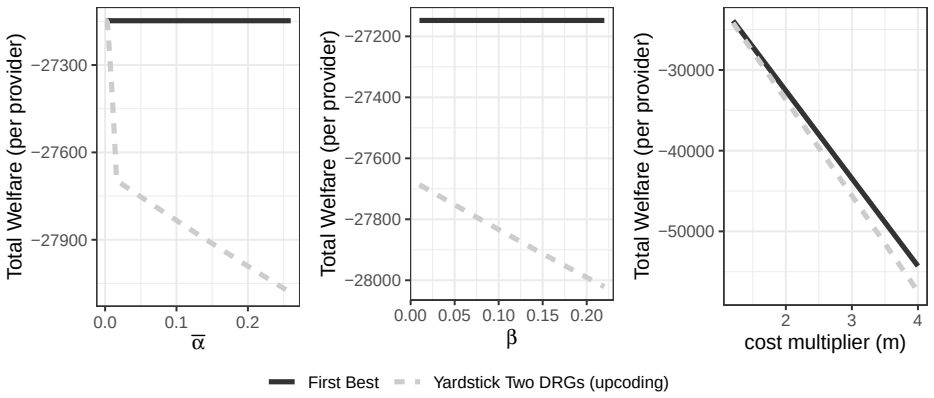

Figure 3 **Numerical example:** total welfare as a function of the patient maximum upcoding rate ($\bar{\alpha}$), cost of upcoding (β), and cost multiplier (m). All other parameters are set to base value.

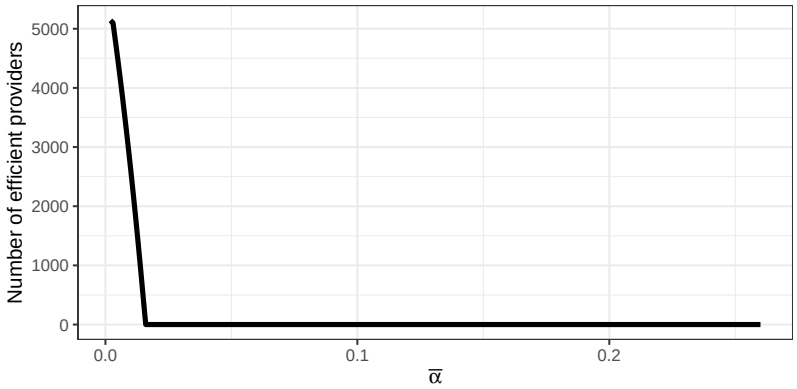

Figure 4 **Numerical example:** Number of efficient providers (θ_2) as a function of the patient maximum upcoding rate $\bar{\alpha}$. All other parameters are set to base value.

Figure 3 confirms that the welfare lost due to upcoding and lemon dropping increases as i) the maximum proportion of patient upcoding ($\bar{\alpha}$) increases; ii) the cost of upcoding patients (β) increases; iii) the cost multiplier (m) increases. Interestingly, the slope of the total welfare with respect to $\bar{\alpha}$ is larger when $\bar{\alpha}$ is small and decreases at larger values of $\bar{\alpha}$. This happens because

when $\bar{\alpha}$ is small, the equilibrium is asymmetric – most providers are efficient generalists that do not lemon drop patients and only a fraction of providers are innefficient specialists. As $\bar{\alpha}$ increases, welfare decreases for two reasons. First, the inefficient providers upcode more patients which reduces welfare. Second, the equilibrium number of inefficient providers increases which also reduces welfare. This is depicted in Figure 4, where we show the number of efficient generalist providers (θ_2) as a function of $\bar{\alpha}$. When $\bar{\alpha}$ is sufficiently high (about 0.017 in this example), the equilibrium becomes symmetric and all providers are inefficient. As a result, as $\bar{\alpha}$ increases further welfare reduces due to the first mechanism (i.e., inefficient providers become more inefficient) but not the second (as all providers are inefficient).

6. Potential Solution: Yardstick based on input statistics

Clearly, expanding the number of DRGs so that costs become more homogeneous within a DRG provides only a partial solution to the problem of cherry picking and only in the absence of upcoding. Upcoding reintroduces heterogeneity by mixing low-complexity patients into the DRG for the major treatment. This heterogeneity reduces the average cost and, in a yardstick competition world, also reduces the payment associated with the major treatment to a level that is below the cost of treating high-complexity patients. One obvious solution to this problem is to rigorously audit reimbursement to correct for upcoding ex post and impose penalties to deter it ex ante. This should also reduce (but not eliminate) cherry picking. A second solution is to move away from reimbursement based on average cost benchmarks (yardstick competition). For example, as we demonstrated in the Appendix A2.0, returning to a cost-of-service reimbursement model (similarly to how CMS used to reimburse hospitals before 1983), where providers are paid according to their reported costs, completely removes incentives for patient cherry picking and upcoding. However, cost-of-service reimbursement also removes incentives for cost reduction.

A more promising solution that does not increase the HO's informational burden is to implement yardstick competition based on input statistics – that is, monitor the number of patients treated for each DRG and link the providers' reimbursement to these numbers. If providers are indeed identical (or differ from each other in a number of observable features) then they should, on average, treat a similar composition of majors and minors (perhaps after controlling for observable differences in catchment size and patient composition). Any difference could be a sign of patient cherry picking or upcoding and penalties could be placed to discourage such behavior. We note that yardstick competition based on input statistics is similar in spirit to the HRRP implemented by CMS in 2012. The observed readmission rate of each hospital in a number of monitored conditions is compared to the expected readmission rate, which is estimated using data from all eligible US hospitals. Hospitals whose readmission rate is higher than the expected rate are then penalized (Chen and Savva 2018).

Using this idea, we can show that there exists a relatively simple mechanism that eliminates the problem of upcoding and could potentially also resolve the problem of cherry picking. In the case of two DRGs, this can be achieved by augmenting the cost-based yardstick competition payments ($p_{mi} = \bar{c}_{mi}, p_{Mi} = \bar{c}_{Mi}$ as defined in Equation (4)) with the following payment S_i for provider i:

$$S_i = \kappa_i(M_i - \bar{M}_i) + \phi_i(m_i - \bar{m}_i), \tag{8}$$

where M_i and m_i, are the number of major and minor patients treated by provider i, respectively, and $\bar{M}_i$, $\bar{m}_i$ is the average number of major and minor patients treated by all other providers, respectively. In other words, the HO awards a bonus (penalty) to any provider that treats more (fewer) patients than expected in each DRG based on a benchmark estimated using input statistics from other providers. The parameters κ_i and ϕ_i need to be chosen such that $\kappa_i \geq 0$ and $\phi_i - \kappa_i = \bar{\delta}_i - \epsilon$, where $0 < \epsilon < \beta\delta^{e3}$ and $\bar{\delta}_i := \bar{c}_{Mi} - \bar{c}_{mi}$.

Before we characterize the equilibrium outcome of the proposed scheme, we define

$$\bar{\kappa} := \min\left\{ c^{e1} + \delta^{e1} - c^* - \delta^*, \frac{u_1}{\lambda q h \bar{\gamma}} \right\}.$$

Proposition 6 *Under the two-DRG payment scheme with input statistics described above, with $\kappa_i = \kappa$ for all i, there exists a unique Nash equilibrium:*

- *If $\kappa > \bar{\kappa}$, the equilibrium is symmetric in which all providers choose first-best actions (i.e., all providers choose $(0, 0, c^*, \delta^*)$).*
- *If $0 \leq \kappa \leq \bar{\kappa}$, the equilibrium is asymmetric. No provider engages in upcoding and $N - \theta_3$ providers drop as many patients as possible and choose cherry-picking-best costs (i.e., these providers choose $(0, \bar{\gamma}, c^{e1}, \delta^{e1})$ and $c^{e1} > c^*$, $\delta^{e1} > \delta^*$) and θ_3 providers do not engage in cherry picking and choose first-best costs (i.e., these providers choose $(0, 0, c^*, \delta^*)$). The number of efficient providers is the only integer in the interval $[\frac{(N-1)(v_1+\lambda q h \bar{\gamma}\kappa)}{v_1+u_1}, \frac{N v_1+u_1+(N-1)\lambda q h \bar{\gamma}\kappa}{v_1+u_1}]$, satisfies $\theta_3 \geq \theta_1$ and is non-decreasing in κ.*

Intuitively, the additional yardstick competition payment creates indirect competition between providers in one additional dimension – the number of patients treated. More specifically, no provider wants to treat fewer patients than average in each DRG, as doing so will trigger a penalty; conversely, every provider wants to treat more patients than average in each DRG, as doing so will trigger a bonus payment. By choosing the parameters (κ_i, ϕ_i) wisely, the HO can eliminate upcoding and reduce if not completely eliminate cherry picking.

To understand how this works, and how high these parameters need to be set, let's consider an equilibrium outcome where no provider engages in upcoding. If a provider i decided to deviate from this outcome by upcoding one low-complexity patient, then, on the negative side they would

incur the additional cost associated with upcoding ($\beta\delta_i$) and, since they would be providing one fewer minor treatment than anyone else, they would incur a penalty ϕ_i. On the positive side, they would receive the additional reimbursement associated with the major treatment ($\bar{\delta}_i = \bar{c}_{Mi} - \bar{c}_{mi}$) and, since they are providing one additional major treatment, they would also receive a bonus κ_i. So, provided that $\phi_i > \kappa_i + \bar{\delta}_i - \beta\delta_i$, then there is no incentive to upcode. The HO can always ensure that this is the case by setting $\phi_i > \kappa_i + \bar{\delta}_i - \beta\delta^{e3}$ (since in equilibrium $\delta_i \geq \delta^{e3}$). Therefore the scheme proposed above eliminates incentives to upcode.

As the discussion above makes clear, by choosing the parameter ϕ_i appropriately, the HO can eliminate upcoding. Therefore, by setting $\kappa_i = 0$ for all i the regulator can achieve the results of Proposition 4 (where upcoding was not possible), which are already better from a welfare perspective than the results of Proposition 5 (where upcoding was possible). However, the regulator can do even better by increasing κ_i. To see why consider that, for every patient dropped by provider i, on the one hand, they would have to pay a penalty κ_i and would forgo a fee of p_{Mi}. On the other hand, they would not incur the costs $c_i + \delta_i$. Therefore, increasing the parameter κ_i reduces the value of dropping high-complexity patients. As a result, if $\kappa_i = \kappa$ for all providers, then by increasing κ from 0 to $\bar{\kappa}$ for all providers, the asymmetric equilibrium will involve more providers who choose not to lemon drop patients and invest in first-best cost reduction (i.e., θ_3 is non-decreasing in κ). When κ exceeds the threshold $\bar{\kappa}$, all providers choose not to lemon drop and to implement first-best investment in cost reduction and the equilibrium becomes symmetric.

It is important to note that the HO can eliminate upcoding (by setting the parameter ϕ_i appropriately) without using any information not already available – relative benchmarks ensure that no provider wants to treat fewer minor patients than everyone else. Furthermore, in equilibrium, the solution to the upcoding problem is free. No penalties or rewards would have to be paid – the threat of such penalties (promise of rewards) is enough to discourage upcoding. In contrast, to ensure that the penalty κ_i is set sufficiently high to eliminate lemon dropping, the HO would need to be able to calculate $\bar{\kappa}$, which in turn depends on the providers' cost functions $R_c(.)$ and $R_\delta(.)$ that the HO does not have access to. Therefore, one should treat the parameter κ_i as a lever which could be set by the HO at a tentative level – any $\kappa > 0$ improves the equilibrium outcome in the sense that more providers choose not to lemon drop compared to setting $\kappa = 0$. Furthermore, in the asymmetric equilibrium (i.e., if $\kappa \leq \bar{\kappa}$) there will be penalties applied to inefficient providers and bonuses paid to efficient providers, while in the symmetric equilibrium (i.e., if $\kappa > \bar{\kappa}$) no such penalties (or bonuses) will actually need to be paid.

A potential concern with this scheme is that it may incentivize providers to downcode patients (i.e., provide the major treatment to high-complexity patients and then code them as having received the minor treatment for reimbursement purposes). This would be the case if the difference

$\phi_i - \kappa_i > \bar{\delta}_i$, as in this case the bonus associated with treating one additional minor patient is greater than the penalty associated with treating one fewer major patient. Nevertheless, this can be avoided by ensuring that $\phi_i - \kappa_i = \bar{\delta}_i - \epsilon$ where $\epsilon > 0$. If in addition $\epsilon < \beta\delta^{e3}$, then upcoding continues not to be optimal. We prove this more rigorously in Appendix A2.1.

A drawback of yardstick competition based on input statistics is that it adds complexity to an already complex system. Particularly, if hospitals are serving catchment areas of different size and with substantial case mix heterogeneity, taking the average number of patients treated in other hospitals may not be a sensible benchmark. Nevertheless, if these differences are based on observable and exogenous characteristics – for example due to differences in the catchment areas in terms of demographics, size, or prevalence of complications and comorbidities – then appropriate adjustments could be accommodated. More specifically, the benchmarks $\bar{m}_i$ and $\bar{M}_i$ could be estimated using information about the number of patients of all providers as well as other observable covariates that may influence the number of patients and the mixture between minor and major cases that seek care in provider i. To the extent that these covariates successfully account for heterogeneity and cannot be directly influenced by the provider (i.e., they are exogenous), then the reimbursement scheme proposed in the section would provide the right incentives (see Shleifer (1985) and Savva et al. (2019) for more formal arguments). We note that such adjustments are often used in practice in different yardstick competition schemes to account for provider heterogeneity – for example hospital benchmark costs used in IPPS are adjusted for local market conditions (CMS.gov 2021a), the benchmark readmission rate used in HRRP is adjusted for patient risk factors (CMS.gov 2021b).

7. Conclusions

The creation of DRGs is arguably one the most impactful innovations to have come out of an Operations Research group since World War II – the use of DRGs, coupled with payments based on relative benchmarks, has been credited with saving Medicare billions in the US (Fetter 1991) and has been copied extensively in Europe and elsewhere (Busse et al. 2013). Over the past 30 years, as more data and better coding practices became available, the system has been refined to increase the number of patient categories. At least in part, the motivation behind these successive refinements has been a desire to reduce patient cherry picking incentives; with a larger number of DRGs the cost heterogeneity within a DRG is reduced, leaving no 'cherries' for providers to pick or 'lemons' to drop. This work casts doubt on the effectiveness of these successive expansions in reducing cherry picking incentives. In particular, we show that increasing the number of DRGs leads to the market bifurcating into two groups of providers – one that treats patients efficiently (as proponents of increase in the number of DRGs intended) and another that drops expensive

patients and underinvests in cost reduction. In addition, if providers engage in upcoding, where low-complexity patients are coded as belonging to a more expensive (and therefore higher reward) category, they effectively reintroduce cost heterogeneity within a DRG, thus exacerbating patient cherry-picking incentives. Finally, this work has proposed a potential solution based on input statistics. While the solution requires little additional information to that already collected, it will certainly increase the complexity of the reimbursement system and, if implemented, should be carefully monitored to ensure it does not induce any unintended adverse effects.

In an online Appendix we extend the main model to show that under realistic conditions down-coding is not optimal and also to allow for i) continuous increasing costs of upcoding and cherry picking; ii) the absence of a transfer payment; iii) asymmetric providers. The main conclusion continues to hold; expanding the number of DRGs to tackle the problem of cherry picking improves welfare but is less effective in the presence of patient upcoding. Moreover, we show that yardstick competition based on input statistics provides a solution in these cases as well.

References

Adida, E., F. Bravo. 2019. Contracts for healthcare referral services: Coordination via outcome-based penalty contracts. *Management Science* **65**(3) 1322–1341.

Adida, E., H. Mamani, S. Nassiri. 2017. Bundled payment vs. fee-for-service: Impact of payment scheme on performance. *Management Science* **63**(5) 1606–1624.

Alexander, D. 2020. How do doctors respond to incentives? Unintended consequences of paying doctors to reduce costs. *Journal of Political Economy* doi:10.1086/710334.

Arifoglu, K., H. Ren, T. Tezcan. 2021. Hospital Readmissions Reduction Rrogram does not provide the right incentives: Issues and remedies. *Management Science* .

Aswani, A., Z.-J. M. Shen, A. Siddiq. 2019. Data-driven incentive design in the medicare shared savings program. *Operations Research* **67**(4) 1002–1026.

Busse, R., A. Geissler, A. Aaviksoo, et al. 2013. Diagnosis related groups in Europe: moving towards transparency, efficiency, and quality in hospitals? *BMJ* **346**. doi:10.1136/bmj.f3197.

Chen, C., N. Savva. 2018. Unintended consequences of hospital regulation: The case of the hospital readmissions reduction program. *Available at SSRN 3236983* .

CMS.gov. 2021a. Acute inpatient pps. https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS.

CMS.gov. 2021b. Hrrp readmission measures. https://qualitynet.cms.gov/inpatient/measures/readmission/methodology.

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

Darby, M.R., E. Karni. 1973. Free competition and the optimal amount of fraud. *The Journal of Law and Economics* **16**(1) 67–88.

Debo, L.G., L.B. Toktay, L.N. Van Wassenhove. 2008. Queuing for expert services. *Management Science* **54**(8) 1497–1512.

Delana, K., N. Savva, T. Tezcan. 2021. Proactive customer service: Operational benefits and economic frictions. *Manufacturing & Service Operations Management* **23**(1) 70–87.

Dranove, D. 1987. Rate-setting by diagnosis related groups and hospital specialization. *The RAND Journal of Economics* **18**(3) 417–427.

Dulleck, U., R. Kerschbamer. 2006. On doctors, mechanics, and computer specialists: The economics of credence goods. *Journal of Economic literature* **44**(1) 5–42.

Ellis, R. P. 1998. Creaming, skimping and dumping: provider competition on the intensive and extensive margins. *Journal of Health Economics* **17**(5) 537–555.

Ellis, R. P., T. G. McGuire. 1986. Provider behavior under prospective reimbursement: Cost sharing and supply. *Journal of Health Economics* **5**(2) 129–151.

Fetter, R. B. 1991. Diagnosis related groups: Understanding hospital performance. *Interfaces* **21**(1) 6–26.

Freeman, M., N. Savva, S. Scholtes. 2020. Economies of scale and scope in hospitals: An empirical study of volume spillovers. *Management Science* **67**(2) 3147–3167.

Gottschalk, F., W. Mimra, C. Waibel. 2020. Health services as credence goods: A field experiment. *The Economic Journal* **130**(629) 1346–1383.

Greenwald, L., J. Cromwell, W. Adamache, S. Bernard, E. Drozd, E. Root, K. Devers. 2006. Specialty versus community hospitals: referrals, quality, and community benefits. *Health Affairs* **25**(1) 106–118.

Guo, P., C.S. Tang, Y. Wang, M. Zhao. 2019. The impact of reimbursement policy on social welfare, revisit rate, and waiting time in a public healthcare system: Fee-for-service versus bundled payment. *Manufacturing & Service Operations Management* **21**(1) 154–170.

Gupta, D., M. Mehrotra. 2015. Bundled payments for healthcare services: Proposer selection and information sharing. *Operations Research* **63**(4) 772–788.

Heskett, J. L. 2003. Shouldice Hospital Limited. HBS Case 9-683-068. Harvard Business School, Boston.

Holmstrom, B. 1982. Moral hazard in teams. *The Bell Journal of Economics* **13**(2) 324–340.

Jiang, H., Z. Pang, S. Savin. 2012. Performance-based contracts for outpatient medical services. *Manufacturing & Service Operations Management* **14**(4) 654–669.

Jiang, H., Z. Pang, S. Savin. 2020. Performance incentives and competition in health care markets. *Production and Operations Management* **29**(5) 1145–1164.

Jürges, H., J. Köberlein. 2015. What explains drg upcoding in neonatology? the roles of financial incentives and infant health. *Journal of Health Economics* **43** 13–26. doi:https://doi.org/10.1016/j.jhealeco.2015.06.001. URL `https://www.sciencedirect.com/science/article/pii/S0167629615000557`.

KC, D., C. Terwiesch. 2011. The effects of focus on performance: Evidence from california hospitals. *Management Science* **57**(11) 1897–1912.

Laffont, J. J., J. Tirole. 1993. *A Theory of Incentives in Procurement and Regulation*. MIT Press, Cambridge, MA.

Latta, V.B., C. Helbing. 1991. Medicare short-stay hospital services by diagnosis-related groups. *Health Care Financing Review* **12**(4) 105.

Lee, D. K. K., S. A. Zenios. 2012. An evidence-based incentive system for Medicare's End-Stage Renal Disease program. *Management Science* **58**(6) 1092–1105.

Lefouili, Y. 2015. Does competition spur innovation? The case of yardstick competition. *Economics Letters* **137** 135–139.

Ma, C. A. 1994. Health care payment systems: Cost and quality incentives. *Journal of Economics & Management Strategy* **3**(1) 93–112.

Mayes, R. 2007. The origins, development, and passage of Medicare's revolutionary prospective payment system. *Journal of the History of Medicine and Allied Sciences* **62**(1) 21–55.

Nalebuff, B. J., J. E. Stiglitz. 1983. Information, competition, and markets. *The American Economic Review* **73**(2) 278–283.

Naor, P. 1969. The regulation of queue size by levying tolls. *Econometrica* **37**(1) 15–24.

Nassiri, S., E. Adida, H. Mamani. 2021. Reference pricing for healthcare services. *Manufacturing & Service Operations Management* .

Newhouse, J.P. 1989. Do unprofitable patients face access problems? *Health Care Financing Review* **11**(2) 33.

Newhouse, J.P. 1996. Reimbursing health plans and health providers: Efficiency in production versus selection. *Journal of Economic Literature* **34**(3) 1236–1263.

NHS Digital. 2021. Patient level information and costing system (plics) data collections. https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/patient-level-information-and-costing-system-plics-data-collections.

Pope, G. C. 1989. Hospital nonprice competition and Medicare reimbursement policy. *Journal of Health Economics* **8**(2) 147–172.

Psaty, B.M., R. Boineau, L.H. Kuller, R.V. Luepker. 1999. The potential costs of upcoding for heart failure in the United States. *American Journal of Cardiology* **84**(1) 108–109.

Savva, N., T. Tezcan, Ö. Yıldız. 2019. Can yardstick competition reduce waiting times? *Management Science* **65**(7) 3196–3215.

Shactman, D. 2005. Specialty hospitals, ambulatory surgery centers, and general hospitals: charting a wise public policy course. *Health Affairs* **24**(3) 868–873.

Shleifer, A. 1985. A theory of yardstick competition. *The RAND Journal of Economics* **16**(3) 319–327.

Silverman, E., J. Skinner. 2004. Medicare upcoding and hospital ownership. *Journal of Health Economics* **23**(2) 369–389.

So, K. C., C. S. Tang. 2000. Modeling the impact of an outcome-oriented reimbursement policy on clinic, patients, and pharmaceutical firms. *Management Science* **46**(7) 875–892.

Sobel, J. 1999. A reexamination of yardstick competition. *Journal of Economics & Management Strategy* **8**(1) 33–60.

Tangerås, T. P. 2009. Yardstick competition and quality. *Journal of Economics & Management Strategy* **18**(2) 589–613.

Victoor, A., D. Delnoij, R. Friele, J. Rademakers. 2016. Why patients may not exercise their choice when referred for hospital care. an exploratory study based on interviews with patients. *Health Expectations* **19**(3) 667–678.

Zhang, D.J., I. Gurvich, J.A. Van Mieghem, E. Park, R. S. Young, M. V. Williams. 2016. Hospital readmissions reduction program: An economic and operational analysis. *Management Science* **62**(11) 3351–3371.

Zorc, S., S.E. Chick, S. Hasija. 2017. Outcomes-based reimbursement policies for chronic care pathways Working Paper No. 2017/35/DSC/TOM, INSEAD, Fontainebleau, France.

Appendix 1: Proofs of Propositions

Proof of Proposition 1: If there is no upcoding or cherry picking the profit of provider i is given by

$$\pi_i(c_i, \delta_i, 0, 0) = \lambda q \left[h(p_{Mi} - \delta_i - c_i) + (1-h)(p_{mi} - c_i)\right] - R_c(c_i) - R_\delta(\delta_i) + T_i$$

(see Equation (1)). Since the provider's choice of c_i and δ_i does not affect the reimbursement received, and this is true irrespective of the number of DRGs used by the HO, the profit-maximizing choice of the provider is given by

$$-\frac{d}{dc}R_c(c_i) = \lambda q, \; -\frac{d}{d\delta}R_\delta(\delta_i) = \lambda q h.$$

Any values of c_i and δ_i that satisfy these $2 \times N$ conditions are equilibria. Naturally, if all providers choose $c_i = c^*$ and $\delta_i = \delta^*$ the conditions above are identical for all providers, and in fact reduce to the first order conditions of the welfare-maximization problem. In addition, the transfer payment ensures that all providers break even. Therefore, the first-best investment decisions constitute a symmetric Nash equilibrium that achieves first-best investment in cost reduction. Furthermore, since $R_c'' > 0$ and $R_\delta'' > 0$, the symmetric equilibrium is unique and no asymmetric equilibrium exists.□

Proof of Proposition 2: In the absence of upcoding ($\bar{\alpha} = 0$), under the yardstick competition scheme with a single DRG, the profit of provider i is given by

$$\pi_i(c_i, \delta_i, 0, \gamma_i) = \lambda q \left[(1 - h\gamma_i)\bar{c}_i - \delta_i h(1-\gamma_i) - c(1-h\gamma_i)\right] - R_c(c_i) - R_\delta(\delta_i) + \bar{R}_i,$$

where $\bar{c}_i = \frac{1}{N-1}\sum_{j \neq i}\left[c_j + \delta_j \frac{h(1-\gamma_j)}{1-h\gamma_j}\right]$ and $\bar{R}_i := \frac{1}{N-1}\sum_{j\neq i}\left[R_c(c_j) + R_\delta(\delta_j)\right]$ as defined in §4. The derivatives of the profit function of provider i are given by:

$$\frac{\partial}{\partial \gamma_i}\pi_i = \lambda q h(c_i + \delta_i - \bar{c}_i), \tag{A.1}$$

$$\frac{\partial}{\partial c_i}\pi_i = -\frac{d}{dc}R_c(c_i) - \lambda q(1 - h\gamma_i), \tag{A.2}$$

$$\frac{\partial}{\partial \delta_i}\pi_i = -\frac{d}{d\delta}R_\delta(\delta_i) - \lambda q h(1-\gamma_i). \tag{A.3}$$

In any equilibrium outcome the last two conditions will be equal to zero for all providers. Otherwise the provider for whom one of these conditions is not zero could increase their profit by changing c_i or δ_i. Furthermore, the conditions above imply that any two providers with $\gamma_i = \gamma_j$ will have the same costs $c_i = c_j$ and $\delta_i = \delta_j$. Furthermore, since $R_c'' > 0$ and $R_\delta'' > 0$, if a provider has $\gamma_i > \gamma_j$ then $c_i > c_j$, $\delta_i > \delta_j$ and the converse is also true – if a provider has costs such that $\delta_i > \delta_j$ (or $c_i > c_j$) then $\gamma_i > \gamma_j$. Furthermore, since $0 \leq \gamma_i \leq \bar{\gamma}$, from (A.2) and (A.3), provider costs must satisfy $c^* \leq c_i \leq c^{e1}, \delta^* \leq \delta_i \leq \delta^{e1}$.

Consider a symmetric equilibrium where all providers choose (γ, c, δ). In such a symmetric equilibrium then $\frac{\partial}{\partial \gamma_i}\pi_i = \lambda q h \delta \frac{1-h}{1-h\gamma} > 0$. Therefore, $\gamma = \bar{\gamma}$ which also implies that $c = c^{e1}, \delta = \delta^{e1}$ would constitute a candidate for a symmetric equilibrium. Furthermore, since $R_c'' > 0$ and $R_\delta'' > 0$ the symmetric equilibrium candidate is unique. We note that in this symmetric equilibrium candidate, the transfer payment ensures that all providers break even (i.e., make a profit of zero). For this to be an equilibrium outcome no provider must find it profitable to unilaterally deviate to a different strategy. Consider the payoff of one provider (labeled j) that chooses to deviate to a different strategy $(\gamma_j, c_j, \delta_j)$ when all other providers choose $(\bar{\gamma}, c^{e1}, \delta^{e1})$. The derivative of the profit function of provider j with respect to γ_j is given by $\frac{\partial}{\partial \gamma_i}\pi_i = \lambda q h (c_j + \delta_j - c^{e1} - \delta^{e1}\frac{h(1-\bar{\gamma})}{1-h\bar{\gamma}})$. Since $c_j + \delta_j \geq c^* + \delta^*$ then if $c^* + \delta^* > c^{e1} + \delta^{e1}\frac{h(1-\bar{\gamma})}{1-h\bar{\gamma}}$ (i.e., if cherry-picking-best costs are not too extreme) which we have assumed to be the case, then choosing $\gamma_j < \bar{\gamma}$ cannot be a profitable deviation. Therefore, $(\bar{\gamma}, c^{e1}, \delta^{e1})$ is the unique symmetric equilibrium.

We will next investigate the existence of asymmetric equilibria. If an asymmetric equilibrium exists, then at least one provider (labeled j) would have the highest $(\gamma_j, c_j, \delta_j)$ (i.e., $\gamma_j \geq \gamma_i$ for all i and the inequality is strict for at least one i, and similarly for c_j, δ_j). Therefore, $c_j + \delta_j > \bar{c}_j$ (recall that $\bar{c}_j$ is the average cost of all other providers and at least some of these providers will have lower costs). From (A.1), this implies that $\gamma_j = \bar{\gamma}$, which also implies that the costs $c_j = c^{e1}, \delta_j = \delta^{e1}$. Therefore, in any asymmetric equilibrium, some providers (at least one) will choose $(\bar{\gamma}, c^{e1}, \delta^{e1})$. The rest of the providers will have $\gamma_k < \bar{\gamma}, c_k < c^{e1}, \delta_k < \delta^{e1}$. For this to be an equilibrium outcome, from (A.1) it must be the case that $c_k + \delta_k - \bar{c}_k \leq 0$. If it was not the case then $\frac{\partial}{\partial \gamma_k}\pi_k > 0$, implying that the provider's profit could increase by increasing γ_k which is a contradiction. Consider a provider with $c_k + \delta_k - \bar{c}_k = 0$. The profit of this provider can be written as $[\lambda q\,[-\delta_k h - c_k] - R_c(c_k) - R_\delta(\delta_k)] - \lambda q \gamma_k h(\bar{c}_k - \delta_k - c_k) + C$, where C is an exogenous constant. Note that the first term is independent of γ_i and is maximized at c^* and δ^*. Consider a deviation from $(\gamma_k, c_k, \delta_k)$ to $(0, c^*, \delta^*)$. This deviation does not affect the second term (it is zero under both strategies) and increases the first term (the first term is maximized at c^*, δ^*). Therefore this deviation is profitable. This suggests than no provider with costs $c_k + \delta_k - \bar{c}_k = 0$ can exist, which implies that any provider with cost other that c^{e1}, δ^{e1} must satisfy $c_k + \delta_k - \bar{c}_k < 0$, which implies $\frac{\partial}{\partial \gamma_k}\pi_k < 0$. Therefore, this provider must choose $(0, c^*, \delta^*)$ (any other choice of $\gamma_k > 0$ cannot be an equilibrium outcome as provider k can increase their profit by reducing γ). Therefore the condition $c_k + \delta_k - \bar{c}_k < 0$ becomes $c^* + \delta^* < \bar{c}_k$, and note that $\bar{c}_k > c^{e1} + \delta^{e1}\frac{h(1-\bar{\gamma})}{1-h\bar{\gamma}}$. In words, in any asymmetric equilibrium, providers will divide in two groups, θ_0 providers will not drop any patients and choose to operate at a cost as low as first best $(0, c^*, \delta^*)$, and $N - \theta_0$ providers will drop the maximum number of patients and operate at a higher cost compared to first best $(\bar{\gamma}, c^{e1}, \delta^{e1})$.

Consider one of the θ_0 low-cost providers. The fee per patient treated by this provider will be given by $\bar{c}_k = \frac{\theta_0-1}{N-1}(c^* + h\delta^*) + \frac{N-\theta_0}{N-1}(c^{e1} + h\frac{1-\bar{\gamma}}{1-\bar{\gamma}h}\delta^{e1})$. The condition $c_k + \delta_k - \bar{c}_k < 0$ implies that $c^* + \delta^* < \nu(c^* + h\delta^*) + (1-\nu)(c^{e1} + h\frac{1-\bar{\gamma}}{1-\bar{\gamma}h}\delta^{e1})$, for $\nu = \frac{\theta_0-1}{N-1}$. This is a contradiction as $c^* + \delta^* > (c^* + h\delta^*)$ and since we have assume that cherry-picking-best costs are not extreme, $c^* + \delta^* > c^{e1} + h\frac{1-\bar{\gamma}}{1-\bar{\gamma}h}\delta^{e1}$. Therefore, an asymmetric equilibrium cannot exist. □

Proof of Proposition 3: In the absence of upcoding ($\bar{\alpha} = 0$), under the yardstick competition scheme with two DRGs, the profit of provider i is given by

$$\pi_i(c_i, \delta_i, 0, \gamma_i) = \lambda q\left[h(1-\gamma_i)(\bar{c}_{Mi} - \delta_i - c_i) + (1-h)(\bar{c}_{mi} - c_i)\right] - R_c(c_i) - R_\delta(\delta_i) + \bar{R}_i,$$

where $\bar{c}_{Mi} := \frac{1}{N-1}\sum_{j\neq i}[c_j + \delta_j]$, $\bar{c}_{mi} := \frac{1}{N-1}\sum_{j\neq i} c_j$, and $\bar{R}_i := \frac{1}{N-1}\sum_{j\neq i}[R_c(c_j) + R_\delta(\delta_j)]$ as defined in §4. The derivatives of the profit function of provider i are given by:

$$\frac{\partial}{\partial\gamma_i}\pi_i = \lambda q h(c_i + \delta_i - \bar{c}_{Mi}), \tag{A.4}$$

$$\frac{\partial}{\partial c_i}\pi_i = -\frac{d}{dc}R_c(c_i) - \lambda q(1 - h\gamma_i), \tag{A.5}$$

$$\frac{\partial}{\partial\delta_i}\pi_i = -\frac{d}{d\delta}R_\delta(\delta_i) - \lambda q h(1-\gamma_i). \tag{A.6}$$

In any equilibrium outcome, the last two conditions will be equal to zero for all providers. Otherwise the provider for whom one of these conditions is not zero could increase their profit by changing c_i or δ_i. Furthermore, the conditions above imply that any two providers with $\gamma_i = \gamma_j$ will have the same costs $c_i = c_j$ and $\delta_i = \delta_j$. Since $R_c'' > 0$ and $R_\delta'' > 0$, if a provider has $\gamma_i > \gamma_j$ then $c_i > c_j$, $\delta_i > \delta_j$ and the converse is also true – if a provider has costs such that $\delta_i > \delta_j$ (or $c_i > c_j$) then $\gamma_i > \gamma_j$. Furthermore, since $0 \le \gamma_i \le \bar{\gamma}$, from (A.5) and (A.6), provider costs satisfy $c^* \le c_i \le c^{e1}$, $\delta^* \le \delta_i \le \delta^{e1}$.

Clearly, in any symmetric equilibrium $\frac{\partial}{\partial\gamma_i}\pi_i = 0$. Therefore, any $\gamma_i = \gamma$ where $0 \le \gamma \le \bar{\gamma}$ along with $c_i = c$ and $\delta_i = \delta$ such that $-\frac{d}{dc}R_c(c) = \lambda q(1-h\gamma)$, $-\frac{d}{d\delta}R_\delta(\delta) = \lambda q h(1-\gamma)$ would be a candidate for a symmetric equilibrium outcome. For any such γ, the values of c and δ are unique and are increasing in γ (since $R_c'' > 0$ and $R_\delta'' > 0$). Now consider any such symmetric equilibrium candidate where $\gamma > 0$ and consider the profit of provider i which will be given by $\pi_i = [-\lambda q(h\delta_i + c_i) - R_c(c_i) - R_\delta(\delta_i)] - \lambda q h\gamma_i(\bar{c}_{Mi} - \delta_i - c_i) + C$, where C is an exogenous constant. Note that the first term is independent of γ_i and is maximized at c^* and δ^*. Consider a deviation from the symmetric equilibrium candidate (γ, c, δ) to $(0, c^*, \delta^*)$. This deviation will be profitable for provider i as it would increase the first term and leave the second term unaffected (it is zero under both strategies). Therefore, $\gamma > 0$ cannot be a symmetric equilibrium outcome. Therefore, the only symmetric equilibrium candidate that survives is $(0, c^*, \delta^*)$. For this to be an equilibrium outcome, no provider must find it profitable to unilaterally deviate to a different strategy. Consider the payoff of one provider (labeled j) that chooses to

deviate to a different strategy $(\gamma_j, c_j, \delta_j)$ when all other providers choose $(0, c^*, \delta^*)$. The derivative of the profit function of provider j with respect to γ_j is given by $\frac{\partial}{\partial \gamma_j}\pi_j = \lambda q h(c_j + \delta_j - c^* - \delta^*) > 0$. Therefore, this provider can improve their profit by deviating to a strategy where they drop some patients (i.e., $\gamma_j > 0$) and invest less in cost reduction, (i.e., $c_j > c^*, \delta_j > \delta^*$). Therefore no symmetric equilibrium outcome can exist.

We will now turn to asymmetric equilibria. In any asymmetric equilibrium, at least one provider (labeled j) would have the highest $(\gamma_j, c_j, \delta_j)$ (i.e., $\gamma_j \geq \gamma_i$ for all i and the inequality is strict for at least one i, and similarly for c_j, δ_j). Therefore, $c_j + \delta_j > \bar{c}_{Mj}$ (recall that $\bar{c}_{Mj}$ is the average cost of all other providers and at least some of these providers will have lower costs). From (A.4), this implies that $\gamma_j = \bar{\gamma}$, which also implies that the costs $c_j = c^{e1}, \delta_j = \delta^{e1}$. Conversely, at least one provider (labeled k) will have the lowest c_k, δ_k, γ_k (i.e., $c_k \leq c_i$ for all i and the inequality is strict for at least one i, and similarly for δ_k, γ_k). Therefore, $c_k + \delta_k < \bar{c}_{Mk}$ (recall that $\bar{c}_{Mk}$ is the average cost of all other providers and at least some of these providers will have higher costs). This implies that this provider will choose $\gamma_k = 0$ and $c_k = c^*, \delta_k = \delta^*$. Furthermore, consider a provider with costs other than c^* or c^{e1}, which we label as provider s. This provider must have costs c_s and δ_s such that $c_s + \delta_s = \bar{c}_{Ms}$ and a corresponding γ_s. Consider the profit of this provider, which can be written as $\pi_s = [-\lambda q(h\delta_s + c_s) - R_c(c_s) - R_\delta(\delta_s)] - \lambda q h \gamma_s(\bar{c}_{Ms} - \delta_s - c_s) + C$, where C is an exogenous constant. Note that the first term is independent of γ_s and is maximized at $c_s = c^*$ and $\delta_s = \delta^*$. Consider a deviation from $(\gamma_s, c_s, \delta_s)$ to $(0, c^*, \delta^*)$. This deviation does not affect the second term (it is zero under both strategies) and increases the first term. Therefore this deviation is profitable. This suggests than no provider with costs γ_s, c_s, δ_s can exist. In words, in any asymmetric equilibrium, providers will divide in two groups: θ_1 providers will not drop any patients and choose to operate at a cost as low as first best $(0, c^*, \delta^*)$ and $N - \theta_1$ providers will drop the maximum number of patients and operate at a higher cost compared to first best $(\bar{\gamma}, c^{e1}, \delta^{e1})$.

For such an asymmetric equilibrium to exist, the profit of the θ_1 low-cost providers and the profit of the $N - \theta_1$ high-cost providers need to be non-negative. Consider one of the θ_1 low-cost providers. The fee for providing the major treatment is given by $\bar{c}_{Mk} = \frac{N-\theta_1}{N-1}(\delta^{e1} + c^{e1}) + \frac{\theta_1 - 1}{N-1}(\delta^* + c^*)$, the minor treatment is given by $\bar{c}_{mk} = \frac{N-\theta_1}{N-1}c^{e1} + \frac{\theta_1 - 1}{N-1}(c^*)$, and the transfer payment they will receive is given by $\bar{T}_k = \frac{N-\theta_1}{N-1}(R_\delta(\delta^{e1}) + R_c(c^{e1})) + \frac{\theta_1 - 1}{N-1}(R_\delta(\delta^*) + R_c(c^*))$. After some algebra, the profit of the efficient provider can be written as $\frac{N-\theta_1}{N-1}v_1$, where

$$v_1 := \lambda q(h(\delta^{e1} + c^{e1} - \delta^* - c^*) + (1-h)(c^{e1} - c^*)) + R_c(c^{e1}) + R_\delta(\delta^{e1}) - R_c(c^*) - R_\delta(\delta^*).$$

Note that the expression $-\lambda q(h(\delta + c) + (1-h)c) - R_c(c) - R_\delta(\delta))$ is maximized at $c = c^*$ and $\delta = \delta^*$, therefore $v_1 > 0$. Similarly, the profit of one of the $N - \theta_1$ high cost providers can be written as $\frac{\theta_1}{N-1}u_1$, where

$$u_1 := \lambda q(h(1-\bar{\gamma})(\delta^* + c^* - \delta^{e1} - c^{e1}) + (1-h)(c^* - c^{e1})) + R_c(c^*) + R_\delta(\delta^*) - R_c(c^{e1}) - R_\delta(\delta^{e1}).$$

Note that the expression $-\lambda q(h(1-\bar{\gamma})(\delta+c)+(1-h)c)-R_c(c)-R_\delta(\delta))$ is maximized at $c=c^{e1}$ and $\delta=\delta^{e1}$, therefore $u_1>0$.

We will next determine the value of θ_1. For this to be an equilibrium outcome, it must be the case that the profit one of the low-cost providers makes by being low cost is greater than the profit they would make if they deviated to being a high-cost provider. After some algebra, this condition can be written as

$$\frac{N-\theta_1}{N-1}v_1 \geq \frac{\theta_1-1}{N-1}u_1.$$

Conversely, the profit of one of the high-cost providers must be greater than the payoff they would make if they deviated to being an low-cost provider. After some algebra, this condition reduces to

$$\frac{\theta_1}{N-1}u_1 \geq \frac{N-\theta_1-1}{N-1}v_1.$$

Together the last two inequalities imply that the number of efficient providers must satisfy

$$\frac{(N-1)v_1}{v_1+u_1} \leq \theta_1 \leq \frac{Nv_1+u_1}{v_1+u_1}.$$

Note that this interval contains exactly 1 integer as the difference $\frac{Nv_1+u_1}{v_1+u_1}-\frac{(N-1)v_1}{v_1+u_1}=1$. Furthermore, since $\frac{Nv_1+u_1}{v_1+u_1}<N$, this integer is always less than N. □

Proof of Proposition 4: Under the yardstick competition scheme with two DRGs, the profit of provider i is given by

$$\begin{aligned}\pi_i(c_i,\delta_i,\alpha_i,\gamma_i) = \lambda q\{&[h(1-\gamma_i)+(1-h)\alpha_i]\bar{c}_{Mi}+(1-h)(1-\alpha_i)\bar{c}_{mi}\\ &-[(h(1-\gamma_i)+(1-h)\alpha_i\beta)\delta_i+(1-h\gamma_i)c_i]\}-R_c(c_i)-R_\delta(\delta_i)+\bar{R}_i,\end{aligned}$$

where $\bar{c}_{Mi} := \frac{1}{N-1}\sum_{j\neq i}\left[c_j+\delta_j\frac{h(1-\gamma_j)+(1-h)\alpha_j\beta}{h(1-\gamma_j)+(1-h)\alpha_j}\right]$, $\bar{c}_{mi} := \frac{1}{N-1}\sum_{j\neq i}c_j$, and $\bar{R}_i := \frac{1}{N-1}\sum_{j\neq i}[R_c(c_j)+R_\delta(\delta_j)]$ as defined in §4. The derivatives of the profit function of provider i are given by:

$$\frac{\partial}{\partial\alpha_i}\pi_i = \lambda q(1-h)(\bar{c}_{Mi}-\bar{c}_{mi}-\beta\delta_i), \tag{A.7}$$

$$\frac{\partial}{\partial\gamma_i}\pi_i = \lambda qh(c_i+\delta_i-\bar{c}_{Mi}), \tag{A.8}$$

$$\frac{\partial}{\partial c_i}\pi_i = -\frac{d}{dc}R_c(c_i)-\lambda q(1-h\gamma_i), \tag{A.9}$$

$$\frac{\partial}{\partial\delta_i}\pi_i = -\frac{d}{d\delta}R_\delta(\delta_i)-\lambda q[h(1-\gamma_i)+(1-h)\alpha_i\beta]. \tag{A.10}$$

In any equilibrium outcome, the last two conditions will be equal to zero for all providers. Otherwise the provider for whom one of these conditions is not zero could increase their profit by changing c_i or δ_i. Furthermore, since $0\leq\gamma_i\leq\bar{\gamma}$, $0\leq\alpha_i\leq\bar{\alpha}$, and $R_c''>0$ and $R_\delta''>0$, from (A.9) and (A.10)

provider costs satisfy $c^* \le c_i \le c^{e1}, \delta^{e3} \le \delta_i \le \delta^{e1}$, where δ^{e3} is the unique solution to $-\frac{d}{d\delta}R_\delta(\delta_i) = \lambda q[h + (1-h)\bar{\alpha}\beta]$.

Turning to (A.7), this condition will be positive for provider i if $\frac{1}{N-1}\sum_{j\neq i}\left[\delta_j \frac{h(1-\gamma_j)+(1-h)\alpha_j\beta}{h(1-\gamma_j)+(1-h)\alpha_j}\right] > \beta\delta_i$. Note that the LHS is minimized when all providers other than i choose $\gamma_j = 0, \alpha_j = \bar{\alpha}, \delta_j = \delta^{e3}$. The RHS is maximized if provider i chooses $\delta_i = \delta^{e1}$. Therefore, a sufficient condition for (A.7) to be positive for all providers i is

$$\beta < \frac{h\delta^{e3}}{h\delta^{e1} + (\delta^{e1} - \delta^{e3})(1-h)\bar{\alpha}}, \tag{A.11}$$

which we have assumed holds. Therefore, in any equilibrium outcome, all providers will choose $\alpha_i = \bar{\alpha}$. Furthermore, (A.9) and (A.10) imply that any two providers with $\gamma_i = \gamma_j$ will have the same costs $c_i = c_j$ and $\delta_i = \delta_j$. Since $R_c'' > 0$ and $R_\delta'' > 0$, if a provider has $\gamma_i > \gamma_j$ then $c_i > c_j$, $\delta_i > \delta_j$ and the converse is also true – if a provider has costs such that $\delta_i > \delta_j$ (or $c_i > c_j$) then $\gamma_i > \gamma_j$.

Consider a symmetric equilibrium such that $\alpha_i = \alpha$, $\gamma_i = \gamma$, $c_i = c$ and $\delta_i = \delta$ for all i. In any symmetric equilibrium $\bar{c}_{Mi} - \bar{c}_{mi} - \beta\delta = \delta\frac{h(1-\gamma)(1-\beta)}{h(1-\gamma)+(1-h)\alpha} > 0$, which implies $\frac{\partial}{\partial\alpha_i}\pi_i > 0$. Therefore, in any symmetric equilibrium $\alpha_i = \bar{\alpha}$ for all i. This, implies that $c + \delta - \bar{c}_{Mi} = \frac{(1-h)(1-\beta)\delta\bar{\alpha}}{h(1-\gamma)+(1-h)\bar{\alpha}} > 0$, which also implies that $\frac{\partial}{\partial\gamma_i}\pi_i > 0$. Therefore in any symmetric equilibrium $\gamma = \bar{\gamma}$ for all i. The values of $c = c^{e1}$ and $\delta = \delta^{e2}$ are the solution to $-\frac{d}{dc}R_c(c) = \lambda q(1 - h\bar{\gamma})$, $-\frac{d}{d\delta}R_\delta(\delta) = \lambda q[h(1-\bar{\gamma}) + (1-h)\bar{\alpha}\beta])$, and they are unique (since $R_c'' > 0$ and $R_\delta'' > 0$). In addition, the transfer payment ensures that all providers break even. Furthermore, since $(1 - h\bar{\gamma}) < 1$, this implies that $c^{e1} > c^*$. If $h(1-\bar{\gamma}) + (1-h)\bar{\alpha}\beta < h$ then $\delta^{e2} > \delta^*$, otherwise the opposite holds. For this to be an equilibrium outcome, no provider must find it profitable to unilaterally deviate to a different strategy. Consider the payoff of one provider (labeled j) that chooses to deviate to a different strategy $(\alpha_j, \gamma_j, c_j, \delta_j)$ when all other providers choose $(\bar{\alpha}, \bar{\gamma}, c^{e1}, \delta^{e2})$. Due to condition (7), it is not profitable to choose any $\alpha_j < \bar{\alpha}$. The derivative of the profit function of provider j with respect to γ_j is given by $\frac{\partial}{\partial\gamma_j}\pi_j = \lambda q h(c_j + \delta_j - c^{e1} - \delta^{e2}\frac{h(1-\bar{\gamma})+(1-h)\bar{\alpha}\beta}{h(1-\bar{\gamma})+(1-h)\bar{\alpha}})$. Since $c_j + \delta_j \ge c^* + \delta^{e3}$ then if $c^* + \delta^{e3} \ge c^{e1} + \delta^{e2}\frac{h(1-\bar{\gamma})+(1-h)\bar{\alpha}\beta}{h(1-\bar{\gamma})+(1-h)\bar{\alpha}}$ then choosing $\gamma_j < \bar{\gamma}$ cannot be a profitable deviation. Therefore, $(\bar{\alpha}, \bar{\gamma}, c^{e1}, \delta^{e2})$ will constitute a symmetric equilibrium. If however, $c^* + \delta^{e3} \le c^{e1} + \delta^{e2}\frac{h(1-\bar{\gamma})+(1-h)\bar{\alpha}\beta}{h(1-\bar{\gamma})+(1-h)\bar{\alpha}}$ then it may be profitable to deviate to $(\bar{\alpha}, 0, c^*, \delta^{e3})$. For this to be the case, it must be the case that the profit of the provider who deviates to $(\bar{\alpha}, 0, c^*, \delta^{e3})$ when all other providers choose $(\bar{\alpha}, \bar{\gamma}, c^{e1}, \delta^{e2})$ is non-negative (as the profit associated with not deviating is zero). This condition can be written as $v_2 \ge 0$, where

$$\begin{aligned} v_2 := \lambda q \left((h + (1-h)\bar{\alpha})\frac{(h(1-\bar{\gamma}) + (1-h)\bar{\alpha}\beta)}{h(1-\bar{\gamma}) + (1-h)\bar{\alpha}}\delta^{e2} - (h + (1-h)\bar{\alpha}\beta)\delta^{e3} + c^{e1} - c^* \right) \\ + \; R_c(c^{e1}) + R_\delta(\delta^{e2}) - R_c(c^*) - R_\delta(\delta^{e3}). \end{aligned}$$

We will then consider asymmetric equilibria. Due to condition (7), in any asymmetric equilibrium, all providers will choose $\alpha_j = \bar{\alpha}$. At least one provider (labeled j) would have the highest $(\gamma_j, c_j, \delta_j)$

(i.e., $\gamma_j \geq \gamma_i$ for all i and the inequality is strict for at least one i, and similarly for c_j, δ_j). Therefore, $c_j + \delta_j > \bar{c}_{Mj}$ (recall that $\bar{c}_{Mj}$ is the average cost of all other providers and at least some of these providers will have lower costs). From (A.8), this implies that $\gamma_j = \bar{\gamma}$, which also implies that the costs $c_j = c^{e1}, \delta_j = \delta^{e2}$. Conversely, at least one provider (labeled k) will have the lowest $(\gamma_k, c_k, \delta_k)$ (i.e., $\gamma_k \leq \gamma_i$ for all i and the inequality is strict for at least one i, and similarly for c_k, δ_k). Therefore, $c_k + \delta_k < \bar{c}_{Mk}$ (recall that $\bar{c}_{Mk}$ is the average cost of all other providers and at least some of these providers will have higher costs). From (A.8), this implies that $\gamma_k = 0$ and $c_k = c^*, \delta_k = \delta^{e3}$. For this to be possible, it must be the case that $c^* + \delta^{e3} < c^{e1} + \delta^{e2}\frac{h(1-\bar{\gamma})+(1-h)\bar{\alpha}\beta}{h(1-\bar{\gamma})+(1-h)\bar{\alpha}}$.

Furthermore, consider a provider with γ_s other than 0 or $\bar{\gamma}$, which we label as provider s. Due to condition (7), this provider will still have $\alpha_s = \bar{\alpha}$, and from (A.8) must have costs c_s and δ_s such that $c_s + \delta_s = \bar{c}_{Ms}$. Consider the profit of this provider, which can be written as $\pi_s = [-\lambda q(h\delta_s + c_s) - R_c(c_s) - R_\delta(\delta_s)] - \lambda q h \gamma_s(\bar{c}_{Ms} - \delta_s - c_s) + C$, where C is a constant that does not depend on $(\gamma_s, c_s, \delta_s)$. Note that the first term is independent of γ_s and is maximized at $c_s = c^*$ and $\delta_s = \delta^*$. Consider a deviation from $(\gamma_s, c_s, \delta_s)$ to $(0, c^*, \delta^*)$. This deviation does not affect the second term (it is zero under both strategies) and increases the first term. Therefore this deviation is profitable. This suggests than no provider with costs γ_s, c_s, δ_s can exist. In words, in any asymmetric equilibrium, providers will divide in two groups: θ_2 providers will upcode the maximum number of patients, will not drop any patients and choose to operate at relatively low costs $(\bar{\alpha}, 0, c^*, \delta^{e3})$ and $N - \theta_2$ providers that will upcode the maximum number of patients, will drop the maximum number of patients and operate at a higher cost $(\bar{\alpha}, \bar{\gamma}, c^{e1}, \delta^{e2})$.

For such an asymmetric equilibrium to exist, the profit of the θ_2 low-cost providers and the profit of the $N - \theta_2$ high-cost providers need to be non-negative. Consider one of the θ_2 low-cost providers. The fee they are paid for providing the major treatment is given by $\bar{c}_{Mk} = \frac{N-\theta_2}{N-1}(\delta^{e2}\frac{h(1-\bar{\gamma})+(1-h)\bar{\alpha}\beta}{h(1-\bar{\gamma})+(1-h)\bar{\alpha}} + c^{e1}) + \frac{\theta_2-1}{N-1}(\delta^{e3}\frac{h+(1-h)\bar{\alpha}\beta}{h+(1-h)\bar{\alpha}} + c^*)$, the fee for the minor treatment is given by $\bar{c}_{mk} = \frac{N-\theta_2}{N-1}c^{e1} + \frac{\theta_2-1}{N-1}c^*$, and the transfer payment they will receive is given by $\bar{T}_k = \frac{N-\theta_2}{N-1}(R_\delta(\delta^{e2}) + R_c(c^{e1})) + \frac{\theta_2-1}{N-1}(R_\delta(\delta^{e3}) + R_c(c^*))$. The profit of this provider will be given by $\frac{N-\theta_2}{N-1}v_2$. Similar algebra shows that the profit of one of the high-cost providers will be given by $\frac{\theta_2}{N-\theta_2}u_2$, where

$$\begin{aligned} u_2 :=\ & \lambda q(\delta^{e3}(h + (1-h)\bar{\alpha}\beta)\frac{h(1-\bar{\gamma}) + (1-h)\bar{\alpha}}{h + (1-h)\bar{\alpha}} - (h(1-\bar{\gamma}) + (1-h)\bar{\alpha}\beta)\delta^{e2} + (1-\bar{\gamma}h)(c^* - c^{e1})) \\ &+\ R_c(c^*) + R_\delta(\delta^{e3}) - R_c(c^{e1}) - R_\delta(\delta^{e2}). \end{aligned}$$

Therefore, for the asymmetric equilibrium to exist, it must be the case that $v_2 \geq 0$ and $u_2 \geq 0$. Note that $u_2 > 0$. To see this, note that u_2 can be written as

$$\begin{aligned} u_2 =\ & \left[-\lambda q\left((1-\bar{\gamma}h)c^{e1} + ((1-\bar{\gamma})h + (1-h)\bar{\alpha}\beta)\delta^{e2}\right) - R_c(c^{e1}) - R_\delta(\delta^{e2})\right] \\ &+\ \left[\lambda q\left((1-\bar{\gamma}h)c^* + ((1-\bar{\gamma})h + (1-h)\bar{\alpha}\beta)\delta^{e3}\right) - R_c(c^*) - R_\delta(\delta^{e3})\right] \end{aligned}$$

$$+\ h(1-h)\frac{\bar{\alpha}\bar{\gamma}(1-\beta)}{h+(1-h)\bar{\alpha}}\delta^*.$$

Note that the expression $[-\lambda q\,[(1-\bar{\gamma}h)c+((1-\bar{\gamma})h+(1-h)\bar{\alpha}\beta)\delta]-R_c(c)-R_\delta(\delta)]$ is maximized at c^{e1},δ^{e2}, therefore the sum of the terms in the first two brackets is positive. The third term is also positive, which implies that $u_2>0$. The sign of v_2 will depend on model parameters.

We next determine the value of θ_2. If one of the θ_2 low-cost providers was to deviate and become a high-cost provider then their profit would be given by $\frac{\theta_2-1}{N-1}u_2$ and if one of the $N-\theta_2$ high-cost providers was to deviate and become a low-cost provider it would be $\frac{N-\theta_2-1}{N-1}v_2$. In the symmetric equilibrium it must be the case that these deviations are not profitable. Therefore, $\frac{N-\theta_2}{N-1}v_2>\frac{\theta_2-1}{N-1}u_2$ and $\frac{\theta_2}{N-1}u_2\geq\frac{N-\theta_2-1}{N-1}v_2$. These conditions imply that θ_2 satisfies $\frac{(N-1)v_2}{u_2+v_2}\leq\theta_2\leq\frac{Nv_2+u_2}{u_2+v_2}$. If $v_2\geq 0$, this interval contains exactly 1 integer as the difference between the RHS and the LHS of the inequalities is $\frac{Nv_2+u_2}{u_2+v_2}-\frac{(N-1)v_2}{u_2+v_2}=1$ and this integer is always less than N. □

Proof of Proposition 5: Under one DRG, if upcoding is possible (i.e., $\bar{\alpha}>0$) the derivative of the profit of any provider with respect to α_i is given by $\frac{\partial}{\partial\alpha_i}\pi_i=-\lambda q(1-h)\beta\delta_i<0$. Therefore, in equilibrium no provider would choose to upcode and the equilibrium outcome is identical to that presented in Proposition 3 – namely, given the condition $c^{e1}+\delta^{e1}h\frac{1-\bar{\gamma}}{1-h\bar{\gamma}}<c^*+\delta^*$ the symmetric equilibrium is characterized by all providers choosing $(0,\bar{\gamma},c^{e1},\delta^{e1})$. Note that this is equivalent to the solution that maximizes total welfare (i.e., maximizes the objective of the HO as defined in (2)) under the constraint $\gamma=\bar{\gamma}$. Under two DRGs, the equilibrium outcome is given by Proposition 5. Namely, given that CPU-best costs are comparable to upcoding best costs, the equilibrium is symmetric and characterized by all providers choosing $(\bar{\alpha},\bar{\gamma},c^{e1},\delta^{e2})$. Note that this is the solution that maximizes total welfare (i.e., maximizes the objective of the HO as defined in (2)) under the constraints $\gamma=\bar{\gamma}$ and $\alpha=\bar{\alpha}$. Note that the feasible region of this welfare-maximization problem is a subset of the feasible region of the pervious welfare-maximization problem. Therefore welfare under two-DRG symmetric equilibrium cannot be greater than the welfare under the one-DRG equilibrium.□

Proof of Proposition 6: Under this yardstick competition scheme, the profit of provider i is given by

$$\begin{aligned}\pi_i(c_i,\delta_i,\alpha_i,\gamma_i)=\ &\lambda q\,([h(1-\gamma_i)+(1-h)\alpha_i]\bar{c}_{Mi}+(1-h)(1-\alpha_i)\bar{c}_{mi}\\ &-\ [(h(1-\gamma_i)+(1-h)\alpha_i\beta)\delta_i+(1-h\gamma_i)c_i]\ )\\ &-\ R_c(c)-R_\delta(\delta)+\bar{R}_i+\kappa(M_i-\bar{M}_i)+\phi_i(m_i-\bar{m}_i),\end{aligned}$$

where $M_i=\lambda q(h(1-\gamma_i)+(1-h)\alpha_i)$, $m_i=\lambda q(1-h)(1-\alpha_i)$, $\kappa\geq 0$, $\phi_i-\kappa>\bar{\delta}_i-\beta\delta^{e3}$, $\bar{\delta}_i:=\bar{c}_{Mi}-\bar{c}_{mi}$, $\bar{c}_{Mi}=\frac{1}{N-1}\sum_{j\neq i}\left[c_j+\delta_j\frac{h(1-\gamma_j)+(1-h)\alpha_j\beta}{h(1-\gamma_j)+(1-h)\alpha_j}\right]$, $\bar{c}_{mi}=\frac{1}{N-1}\sum_{j\neq i}c_j$, and $\bar{R}_i=\frac{1}{N-1}\sum_{j\neq i}[R_c(c_j)+R_\delta(\delta_j)]$.

The derivatives of the profit function of provider i are given by:

$$\frac{\partial}{\partial \alpha_i}\pi_i = \lambda q(1-h)(-\beta\delta_i + \bar{\delta}_i + \kappa - \phi_i), \tag{A.12}$$

$$\frac{\partial}{\partial \gamma_i}\pi_i = \lambda q h(c_i + \delta_i - \bar{c}_{Mi} - \kappa), \tag{A.13}$$

$$\frac{\partial}{\partial c_i}\pi_i = -\frac{d}{dc}R_c(c_i) - \lambda q(1 - h\gamma_i), \tag{A.14}$$

$$\frac{\partial}{\partial \delta_i}\pi_i = -\frac{d}{d\delta}R_\delta(\delta_i) - \lambda q[h(1-\gamma_i) + (1-h)\alpha_i\beta]. \tag{A.15}$$

In any equilibrium outcome, the expressions (A.14) and (A.15) have to be equal to zero for all providers. Otherwise the provider for whom one of these is not zero could increase their profit by changing c_i or δ_i. Furthermore, since $0 \le \gamma_i \le \bar{\gamma}$ and $0 \le \alpha_i \le \bar{\alpha}$, from (A.14) and (A.15) provider costs satisfy $c^* \le c_i \le c^{e1}, \delta^{e3} \le \delta_i \le \delta^{e1}$. Turning to (A.12), since $\phi_i - \kappa > \bar{\delta}_i - \beta\delta^{e3}$, in any equilibrium $\frac{\partial}{\partial \alpha_i}\pi_i < 0$ which implies that $\alpha_i = 0$ for all i. This implies that any two providers with $\gamma_i = \gamma_j$ will have the same costs $c_i = c_j$ and $\delta_i = \delta_j$ and since $R_c'' > 0$ and $R_\delta'' > 0$, if a provider has $\gamma_i > \gamma_j$ then $c_i > c_j$, $\delta_i > \delta_j$ and the converse is also true – if a provider has costs such that $\delta_i > \delta_j$ (or $c_i > c_j$) then $\gamma_i > \gamma_j$. Furthermore, since $\alpha_i = 0$ for all i we can use (A.15) to narrow down the range of possible costs δ_i to $\delta^* \le \delta_i \le \delta^{e1}$.

In any symmetric equilibrium $\frac{\partial}{\partial \gamma_i}\pi_i = -\kappa < 0$, therefore $\gamma_i = 0$, and $\delta_i = \delta^*$ and $c_i = c^*$ for all i. Therefore, the strategy $(0, 0, c^*, \delta^*)$ is the only candidate for a symmetric equilibrium outcome. For this to be an equilibrium outcome no provider must find it profitable to unilaterally deviate to a different strategy. Consider the payoff of one provider (labeled j) that chooses to deviate to a different strategy $(0, \gamma_j, c_j, \delta_j)$ when all other providers choose $(0, 0, c^*, \delta^*)$. The derivative of the profit function of provider j with respect to γ_j is given by $\frac{\partial}{\partial \gamma_j}\pi_j = \lambda q h(c_j + \delta_j - c^* - \delta^* - \kappa)$. If $\kappa > c^{e1} + \delta^{e1} - c^* - \delta^*$ then no profitable deviation can exist, therefore the strategy $(0, 0, c^*, \delta^*)$ is the unique symmetric equilibrium outcome. Otherwise, provider j may find it profitable to deviate to $(0, \bar{\gamma}, c^{e1}, \delta^{e1})$. In this case, the profit of provider j needs to be non-negative

$$\begin{aligned} u_4 &:= \lambda q(h(1-\bar{\gamma})(\delta^* + c^* - \delta^{e1} - c^{e1}) + (1-h)(c^* - c^{e1})) + R_c(c^*) + R_\delta(\delta^*) - R_c(c^{e1}) - R_\delta(\delta^{e1}) - \lambda q h\bar{\gamma}\kappa \\ &= u_1 - \lambda q h\bar{\gamma}\kappa. \end{aligned}$$

Note that $u_1 > 0$ (see Proof of Proposition 4). Therefore, if $\kappa < \min\{c^{e1} + \delta^{e1} - c^* - \delta^*, \frac{u_1}{\lambda q h\bar{\gamma}}\}$ then a profitable deviation will exist and no symmetric equilibrium outcome can exist.

We now turn to asymmetric equilibria. As shown above, in any asymmetric equilibrium $\alpha_i = 0$ for all i. Therefore, if $\kappa = 0$ for all i, then the problem of finding asymmetric equilibria reduces to that of finding equilibria in the case where there is cherry picking but not upcoding (see Proposition 4). We will consider the case where $\kappa > 0$. In this case, the provider with the highest γ_j will also be the

provider with the highest costs c_j and δ_j (see similar argument given in the Proof of Proposition 4). Consider one such provider. From (A.13), $\frac{\partial}{\partial \gamma_j}\pi_j = \lambda q h(c_j + \delta_j - \bar{c}_{Mj} - \kappa_j)$. Note that $c_j + \delta_j > \bar{c}_{Mj}$, nevertheless, if $\kappa \geq c^{e1} + \delta^{e1} - c^* - \delta^*$ then $\frac{\partial}{\partial \gamma_j}\pi_j < 0$, suggesting that lowering γ_j would increase the provider's profit. Therefore, for sufficiently high κ there cannot exist a provider with higher rate γ_i or higher costs than other providers, suggesting that an asymmetric equilibrium does not exist. If $\kappa < c^{e1} + \delta^{e1} - c^* - \delta^*$, then $\gamma_j = \bar{\gamma}$, which also implies that the costs $c_j = c^{e1}, \delta_j = \delta^{e1}$. Conversely, at least one provider (labeled k) will have the lowest c_k, δ_k, γ_k (i.e., $c_k \leq c_i$ for all i and the inequality is strict for at least one i, and similarly for δ_k, γ_k). Therefore, $c_k + \delta_k < \bar{c}_{Mk}$ (recall that $\bar{c}_{Mk}$ is the average cost of all other providers and at least some of these providers will have higher costs). This implies that this provider will choose $\gamma_k = 0$ and $c_k = c^*, \delta_k = \delta^*$. Furthermore, consider a provider with costs other than c^* or c^{e1}, which we label as provider s. This provider must have costs c_s and δ_s such that $c_s + \delta_s = \bar{c}_{Ms} + \kappa$ and a corresponding γ_s. Consider the profit of this provider, which can be written as $\pi_s = [-\lambda q(h\delta_s + c_s) - R_c(c_s) - R_\delta(\delta_s)] - \lambda q h \gamma_s(\bar{c}_{Ms} - \delta_s - c_s + \kappa) + C$, where C is an exogenous constant. Note that the first term is independent of γ_s and is maximized at $c_s = c^*$ and $\delta_s = \delta^*$. Consider a deviation from $(0, \gamma_s, c_s, \delta_s)$ to $(0, 0, c^*, \delta^*)$. This deviation does not affect the second term (it is zero under both strategies) and increases the first term. Therefore this deviation is profitable. This suggests than no provider with costs γ_s, c_s, δ_s can exist. In words, in any asymmetric equilibrium providers will divide in two groups: θ_3 providers will not drop any patients and choose to operate at a cost as low as first best $(0, 0, c^*, \delta^*)$ and $N - \theta_3$ providers will drop the maximum number of patients and operate at a higher cost compared to first best $(0, \bar{\gamma}, c^{e1}, \delta^{e1})$.

For such an asymmetric equilibrium to exist, the profit of the θ_3 low-cost providers and the profit of the $N - \theta_3$ high-cost providers need to be non-negative. Consider one of the θ_3 low-cost providers. The fee for providing the major treatment is given by $\bar{c}_{Mk} = \frac{N-\theta_1}{N-1}(\delta^{e1} + c^{e1}) + \frac{\theta_1 - 1}{N-1}(\delta^* + c^*)$, the minor treatment is given by $\bar{c}_{mk} = \frac{N-\theta_1}{N-1}c^{e1} + \frac{\theta_1 - 1}{N-1}(c^*)$, the number of major treatments provided by others $\bar{M}_i = \lambda q\left(\frac{N-\theta_1}{N-1}h(1-\bar{\gamma}) + \frac{\theta_1 - 1}{N-1}h\right)$, the number of minor treatments provided by others $\bar{m}_i = \lambda q\left(\frac{N-\theta_1}{N-1}(1 - h\bar{\gamma}) + \frac{\theta_1 - 1}{N-1}\right)$, and the transfer payment they will receive is given by $\bar{T}_k = \frac{N-\theta_1}{N-1}(R_\delta(\delta^{e1}) + R_c(c^{e1})) + \frac{\theta_1 - 1}{N-1}(R_\delta(\delta^*) + R_c(c^*))$. After some algebra, the profit of the low-cost provider can be written as $\frac{N-\theta_1}{N-1}v_4$, where

$$\begin{aligned} v_4 &:= \lambda q(h(\delta^{e1} + c^{e1} - \delta^* - c^*) + (1-h)(c^{e1} - c^*) + h\bar{\gamma}\kappa) + R_c(c^{e1}) + R_\delta(\delta^{e1}) - R_c(c^*) - R_\delta(\delta^*) \\ &= v_1 + \lambda q h \bar{\gamma} \kappa. \end{aligned}$$

Note that $v_1 > 0$ (see Proof of Proposition 4), therefore, $v_4 > 0$. Similarly, the profit of one of the $N - \theta_3$ high-cost providers can be written as $\frac{\theta_3}{N-1}u_4$, where $u_4 := u_1 - \lambda q h \bar{\gamma} \kappa$. Note that $u_1 > 0$ (see

Proof of Proposition 4). Therefore, for the asymmetric equilibrium to exist, it must be the case that $\kappa < \frac{u_1}{\lambda q h \bar{\gamma}}$.

We will next determine the value of θ_3. For this to be an equilibrium outcome the profit one of the low-cost providers makes by being low cost must be greater than the profit they would make if they deviated to being a high-cost provider. After some algebra this condition can be written as

$$\frac{N-\theta_3}{N-1} v_4 \geq \frac{\theta_3 - 1}{N-1} u_4.$$

Conversely, the profit of one of the high-cost providers must be greater than the payoff they would make if they deviated to being a low-cost provider. After some algebra this condition reduces to

$$\frac{\theta_3}{N-1} u_4 \geq \frac{N-\theta_3-1}{N-1} v_4.$$

Together the last two inequalities imply that the number of low-cost providers must satisfy

$$\frac{(N-1)(v_1+\lambda q h \bar{\gamma}\kappa)}{v_1+u_1} \leq \theta_3 \leq \frac{N v_1 + u_1 + (N-1)\lambda q h \bar{\gamma}\kappa}{v_1+u_1}.$$

Note that this interval contains exactly 1 integer as the difference between the RHS and the LHS of the inequalities is 1. Furthermore, θ_3 is non-decreasing in κ. Since $\theta_3 = \theta_1$ when $\kappa = 0$, it follows that $\theta_3 \geq \theta_1$. $\square$