The Business of Healthcare: Physician Integration in Bundled Payments

**Problem Definition:** Under the prevailing fee-for-service payments (FFS), hospitals receive a fixed fee while physicians receive payments for each treatment or procedure performed. Under FFS, incentives of hospitals and physicians are misaligned, leading to large inefficiencies. Bundled payments, an alternative to FFS unifying payments to the hospital and physician, is expected to encourage care coordination and reduce costs. However, as hospitals differ in their relationships with physicians in influencing care (level of physician alignment or physician integration), it remains unclear what spectrum of physician integration will facilitate bundling.

**Academic/Practical Relevance:** There is a lack of both academic and practical understanding of hospitals’ and physicians’ bundling incentives. Our study builds on and expands the recent OM literature on alternative payment models.

**Methodology:** We formulate game-theoretic models to study the impact of alignment level between the hospital and physicians, and cost and quality characteristics of a care context that facilitate bundling.

**Results:** We find that (1) hospitals with high physician alignment or with low physician alignment are less likely to gain from bundled payments, while the hospitals that lie in between these two cases will benefit the most; (2) although bundled payments decrease costs, quality may also decrease; (3) initiatives that promote quality awareness in hospitals may dampen the incentives for bundling in hospitals with independent physicians, whereas they are likely to enhance incentives for bundling in hospitals with salaried physicians.

**Managerial Implications:** Our findings have important managerial implications for both hospitals and payers: (1) in deciding whether to enroll or not in bundled payments, hospitals should take their level of physician alignments into account, (2) payers should be aware of and account for potential negative effects of bundling, including a possible quality reduction, and even a cost increase in certain cases.
“Hospitals cannot [...] execute any of the new alternative payment methodologies without physician participation.” (Goldsmith et al. 2016)

1. Introduction

The U.S. healthcare system suffers from poor coordination, large inefficiencies, and misaligned incentives. As a result, the healthcare system in the U.S. incurs high costs and often delivers low quality of care compared with many other developed countries (Davis et al. 2014). The U.S. healthcare spending is estimated to range from 3 to 4 trillion dollars (Patel 2016), of which about one third was being wasted (Berwick and Hackbarth 2012). From an efficiency standpoint, new payment models are viewed as important tools in reducing the overall healthcare costs and improving quality. However, while promising in value, many of the new payment models are unproven and the true impact on providers and overall healthcare system is yet to be seen. Therefore, understanding the financial incentives of providers and characterizing conditions that are essential for achieving the aimed objectives of these new payment models is critical (Schoen 2016).

The prevailing healthcare payment model in the U.S. has been fee-for-service (FFS). Under FFS, the hospital receives a fixed payment per visit specific to a standardized grouping of diseases, called the diagnosis-related groups (DRGs). In contrast, physicians are reimbursed separately for every service provided and procedure performed around that visit. As a result of this disparate payment setup, incentives of hospitals and physicians are misaligned. The hospital can be financially better off by eliminating services to reduce costs, while the physicians can be financially better off by increasing the number of services. The overuse of medical services is associated with higher costs, while there continues to be a heated debate on whether the resulting quality is better or worse (Fisher et al. 2003, Grady and Redberg 2010). Under such misaligned incentives, physicians and hospitals often fail to coordinate care and subsequently forego opportunities to improve quality and decrease costs (Mehrotra and Hussey 2015).

To better align the incentives, the Center for Medicare and Medicaid Services (CMS), the largest payer of healthcare services in the U.S., introduced bundled payments, which aims to combine hospital and physician reimbursements for an entire episode of care into one payment. For example, for an episode of knee surgery, the payer pays a fixed, bundled amount to the hospital (or “convener”) to cover all related services and procedures, including tests, treatments, as well as physician fees. Such a “bundle” typically includes fees of the surgeon, the anesthesiologist, the hospital, and the costs of rehabilitation, implants, and other medical devices. The hospital bears the financial risk due to uncertainty in the costs incurred during the episode and is also responsible for coordinating and reimbursing the physicians. In return, the hospital may keep any savings or share these with
Physicians using gainsharing, a mechanism that allows hospitals to induce physicians to make cost-conscious decisions in alignment with hospital’s incentives, which is not allowed in FFS-based models (Froimson et al. 2013).

Gainsharing has long been promoted by experts to support aligning hospitals and physicians (Wilensky et al. 2007, Grandusky and Kronenberg 2006). Through gainsharing, hospitals reward physicians for the savings realized when physicians use more standardized supplies, choose cheaper devices when appropriate, and, more generally, help reduce unnecessary utilization. In contrast, under the standard FFS model, the physicians often lack incentives to decrease the utilization; indeed, a higher utilization means larger financial gains for them. Despite its promise, both CMS and related legislation such as the Stark Laws and the Civil Monetary Penalties Law have prohibited or discouraged gainsharing in the past because of antitrust concerns. Recently, CMS authorized limited use of gainsharing (e.g., capping the payments and placing constraints on quality) in demonstration initiatives for bundled payments (Froimson et al. 2013).

Bundled payments offer some opportunities in improving healthcare services and delivery. The bundling of payments is expected to realign the incentives for hospital and physicians, misalignment of which is currently plaguing the healthcare industry. Engaging physicians in containing hospital costs could mitigate the overuse problem, and therefore help in reducing the excessive healthcare costs. In addition, by eliminating one of the two separate billing systems, one for hospital payment and one for physician payment, the lump sum payments to involved parties could also decrease the high administrative costs, currently accounting for a quarter of all hospital spending (Mehrotra and Hussey 2015).

However, despite their promise of improved efficiency, bundled payments are often resisted by the physicians, which typically deters hospitals from more widely enrolling in bundled payments (Tsai et al. 2015). Physicians resist because they believe bundled payments may encroach on their autonomy, constrain how they practice medicine, and possibly reduce their profits, which is why physicians have been previously excluded from the hospital DRG-based payments in FFS models (Mehrotra and Hussey 2015).

Hospitals differ in their relationships with physicians in influencing care (typically referred to as level of alignment or level of integration in care coordination), and subsequently in their influence on care intensity. Some hospitals constrain physicians by developing care protocols thus substantially limiting physician autonomy while other hospitals leave most treatment decisions completely to physicians (Burns and Muller 2008). For example, integrated hospital systems and larger hospitals typically have more standardized care protocols as well as more aligned physicians (Bloom et al. 2013). On the other hand, stand-alone nonprofit hospitals are de facto physician-controlled, and therefore the management in such hospitals may find it difficult to restrain physicians’ excessive treatment
choices (Pauly and Redisch 1973, Sloan 2000). The hospital’s capability to influence physician’s care intensity, which we henceforth refer to as physician alignment, determines the cost and quality of care. While bundled payments are expected to promote cost reduction, it remains unclear whether they are suitable for all hospitals or only for certain types of hospitals in the spectrum of alignment levels.

Recent Operations Management (OM) literature explored various aspects of payment models in the healthcare context. Ata et al. (2013) investigate how existing CMS policies for hospice reimbursement tie to providers’ patient selection decisions, and they propose an improvement over the current policy. Several studies investigate the role of performance-based payment models in various contexts. Taking a social planner’s perspective, Lee and Zenios (2012) find the socially optimal decisions in regards to implementing payment policies based on process compliance vs. outcomes for the End-Stage Renal Disease (ESRD) patients. In a principal-agent framework, Jiang et al. (2012) propose a penalty-based contract for coordinating providers’ capacity allocation decisions to ensure timely access to outpatient services while Zhang et al. (2016) study readmission penalties implemented under the Hospital Readmissions Reduction Program and show the unintended consequences of benchmarking when setting the penalty thresholds.

Because the interest in bundled payments has reemerged only recently with a perceived commitment from the CMS for implementing them, there has been an increasing interest in the subject by the OM researchers. Gupta and Mehrotra (2014) take the payer’s perspective and examine how bundled payment contracts that providers propose should be selected by CMS, the major bundled payments contractor. On the other hand, Adida et al. (2016) consider how contending healthcare payment models (including bundled payments) appear from the perspective of a single integrated risk-averse provider and how the models impact patient selection, intensity of care, and the system payoff. A working paper by Andritsos and Tang (2015) compares how readmissions occur under fee-for-service, pay for performance, and bundled payments. Finally, Arikan et al. (2017) consider the strategic interaction of hospitals when determining care quality under bundled payments. While we also consider bundled payments, our focus is very different than the above mentioned studies. In particular, we study the interaction between the hospital and physician and focus on the implications of this interaction on bundling decisions and corresponding outcomes. Considering the role that physicians play in determining the financial bottom line of hospitals and the idiosyncrasies of hospital-physician relations that we discussed earlier, our work makes the first attempt in healthcare operations literature to study the interdependency between hospitals and physicians, the resulting trade-offs, and the obstacles for setting up a bundled payment model for the payers.

Our findings suggest the followings. First, in regards to who will benefit and who will lose in a bundled payment environment, we confirm the experts’ expectation that hospitals with very loosely
aligned physicians would not benefit from bundling and be better off under FFS. However, somewhat unexpectedly, we also find that the hospitals with highly-aligned physicians in general are less likely to benefit as well; and that those hospitals which lie in between these two cases in the spectrum of alignment levels will benefit the most (Theorem 1 and Proposition 2). Second, in regards to quality implications of bundling, we characterize hospital care contexts where the quality will deteriorate under bundled payments (Proposition 3). We demonstrate how the payer can employ quality constraints to prevent the quality from decreasing excessively and, we find that FFS can sometimes be the better choice when the payer worries about quality (Theorem 2). Further, we show that quality initiatives that motivate hospitals to safeguard quality, such as the ongoing Hospital Readmissions Reduction Program, might actually demotivate hospitals from adopting bundled payments (Corollary 2). Finally, we extend our model to capture a setting where physicians are hospital employees (salaried physicians), as opposed to being paid for care services independently from hospitals—a specific case of highly aligned physicians. We find that initiatives that further hospitals’ accountability for care quality may dampen the incentives for bundling in hospitals with independent physicians, whereas they are likely to enhance incentives for bundling in hospitals with salaried physicians. (Proposition 5 vs. Corollary 2).

The rest of this paper is organized as follows. In Section 2, we introduce some key concepts that will be present throughout the paper and also review the relevant literature. In Section 3, we introduce and analyze our base model. We extend the base model in Section 4 to consider a quality-aware model and a model with salaried physicians. In Section 5, using real data, we present a data-driven approach to construct clinical pathways, a key concept in our analysis. Finally, we summarize and conclude in Section 6.

2. Background

In this section, we provide background information, summarize some of the key concepts related to bundled payments, and review related literature.

Current Bundling Initiatives: The recent CMS initiatives include the Bundled Payments for Care Improvement (BPCI) and Comprehensive Care for Joint Replacement (CCJR), where the former is a voluntary model and the latter is a mandatory one (Mechanic 2015). Under the voluntary BPCI program, typically a hospital acts as the convener and proposes a contract to bundle all the services in an episode of care, where the total proposed payment for an episode should include a mandatory discount as compared with the total expected FFS payments from the most recent year. BPCI is currently the predominant bundled payments model, and it has been in effect since 2013, allowing hospitals to bundle up to 48 different DRGs (e.g., major joint replacement of the lower extremity, acute myocardial infarction, congestive heart failure, or simple pneumonia and respiratory
infections). In contrast, CCJR is in experimental stage and introduces bundled payments only for two DRGs but across a broad cross-section of hospitals and is effective as of April 2016. As a part of the mandatory CCJR model, all hospitals in selected metropolitan areas that provide services for DRG 469 and 470 (joint replacement of lower extremity with and without complications, respectively) are required to contract with CMS for bundling such payments.

**Retrospective vs. Prospective Bundling:** To date, bundled contracts have been mostly retrospective, that is, all providers continue to receive individual FFS payments based on their standard reimbursement rates. Under this model, at the end of the year, CMS (payer) compares the total reimbursed amount with the pre-established, discounted bundle price (benchmark). If the providers succeeded in care redesign, and the amount spent is less than the benchmark, CMS pays the difference to the providers. Otherwise, the providers have to repay CMS. Since the retrospective payments allow providers to experiment with bundling without substantially changing their existing operations, retrospectively paying the FFS rate under bundled payments has been the default, and by far the most common among the current bundled payment models (CMS 2014, Dummit et al. 2015). As such, we consider and focus on retrospective models in this study. On the other hand, under a prospective bundling model, the hospital receives a single, lump sum payment from CMS and then distributes the payment among all the providers involved in the episode of care. Clearly, implementing a prospective bundling model is more costly to set up; therefore compared with the retrospective models, prospective models have been less common so far.

### 2.1. Other Relevant Literature

In addition to the OM literature that we covered in the introduction, we describe other relevant literature in this section: 1) health economics literature on care coordination among providers, and 2) medical and health policy literature on the expectations and organizational perspectives about bundled payments.

#### 2.1.1. Health Economics Literature

In the health economics literature, several papers study care coordination among providers. While these studies are relevant to ours from a modeling perspective, our work is distinct as we explore when bundling occurs and what is needed for care coordination, and discover how the hospital and the physicians are likely to react to the most recent bundled payment initiatives. Harris (1977) provides a modeling framework to study incentive relationships and interactions between hospital management and physicians. Ma (1994) wrote a seminal paper that theoretically models care delivery coordination by hospitals and physicians, which also partially motivates our model setup. Crainich et al. (2008) extend this work by incorporating features from international health systems. Custer et al. (1990) propose a model of how physicians react to the Prospective Payment System (implemented in the 80s), and how this affects the hospital under several
different modes of hospital-physician coordination. Boadway et al. (2004) model the two-way contracts among doctors, hospitals, and a social planner where the purpose of the contracts is to address the inefficiencies due to information asymmetry around patient severity. Huang and McCarthy (2015) explore how coordination between the hospital and physicians changes as the insurance market shifts.

While relevant from a modeling perspective, all of these studies are concerned with prior payment mechanisms and are not directly applicable to current bundled payment models. In this regard, two studies published in the 1990s and 2000s in the wake of earlier bundled payment discussions are relevant to our work. The first one by Jelovac and Macho-Stadler (2002) focuses on the payer’s problem in terms of the contract design with the hospital and the physicians. The second one, a short note by Dor and Watson (1995), examines the coordinating split of a bundled fee for an efficient outcome. The authors conclude that there is a clear need for investigating i) the role of varying hospital-physician alignment in bundling decisions, ii) the efficiency implications of quality contracts when offered alongside bundling, which, in some sense, motivate our work.

### 2.1.2. Medical and Health Policy Literature

Bundled payments as an alternative to FFS have for long interested health-policy researchers. While most of this literature is qualitative, we summarize the relevant findings concerning bundled payment models and physician-hospital relationships.

**Bundled Payments and Expectations** Bundled payments are expected to better control care intensity (utilization), encourage high quality, promote provider coordination and integration but can be readily implemented (Mechanic and Altman 2009). These expectations are largely speculative. For instance, Hussey et al. (2012) conduct a thorough literature review and find that almost all existing work were observational or descriptive. The study reports a consistent drop in intensity with ambiguous quality impact among the reviewed articles, however, the body of evidence was rated as low due to concerns about bias, confounding, a lack of design and contextual factors.

Another stream of research addresses questions around how to form bundles. For instance, Dobson et al. (2012) analyze historical Medicare claims data and provide widely-cited guidelines on how to define, price, and manage a bundle. Sood et al. (2011) review the existing evidence and focus on which conditions to choose for bundling and how to choose the bundle episode length. Ridgely et al. (2014) describe an unsuccessful bundled payments initiative, Prometheus in California, to inform future bundled payments initiatives. Finally, many studies from physician communities discuss how bundled payments will affect the physicians and how physicians should react (e.g., Bozic et al. 2014, Shih et al. 2015, Mukherji and Fockler 2014, Brill et al. 2014).
Bundled Payments and Organizational Perspectives Health policy studies also explore how new payment models relate to physician-hospital relationships. The new payment models are expected to further encourage ongoing provider integration (Mechanic and Altman 2009). Gaynor and Town (2012) review the effects of such consolidation under the employment model, while Casalino et al. (2008) and Berenson et al. (2007) highlight the rise of physician-owned facilities, particularly ambulatory surgery centers and physician-owned hospitals that directly compete with traditional hospitals. Friedberg et al. (2015) survey physician practices about how they perceive and worry about new payment models. They find that the new payment models may induce some physicians to more closely collaborate with hospitals, force them to face higher expectations, but possibly also provide them with opportunities to improve care quality.

3. Modeling the Alignment of Hospital and Physicians

As discussed earlier, under the prevailing FFS model, physician and hospital incentives do not align well, which leads to inefficiencies in the care delivery. The potential success of bundling, on the other hand, is argued to depend on the cooperation between the hospital and physicians in cost reduction and quality improvement efforts (Mechanic and Altman 2009). Our base model studies the incentive alignment problem between the hospital and physicians, and analyzes, in the spectrum of level of alignment/integration, when bundling becomes attractive for both parties. The model captures main dynamics of the coordination problem by considering index admissions\(^1\) in hospitals with non-salaried physicians, where the hospital is profit-driven and quality of care is predominantly determined by physician efforts. Later, we extend the base model to capture cases where a) the hospital is also cognizant of quality in addition to being profit-driven and the readmissions are also accounted for (§4.1), and b) physicians are salaried employees of the hospital (§4.2).

Without loss of generality, we build our model around payments made for a single medical condition such as knee-replacement, characterized by a DRG categorization. For example, a patient with “major joint replacement or reattachment of lower extremity without major complications” will be assigned DRG 470 and another patient with “simple pneumonia and pleurisy without complications” will be assigned DRG 195 for billing purposes. Under the FFS model, once the care is complete, the hospital will receive predetermined amounts for the respective DRGs regardless of the costs for treating the patient; whereas physician is paid separately for each service he provides. Bundled payments, on the other hand, brings the hospitals and physician payments together, by paying a single lump-sum amount for a given DRG (e.g., 470-knee replacement), which is then shared between the hospital and physician. Table 1 summarizes the variables that appear in our models. Next, we describe the key model components.

\(^{1}\) Index admission corresponds to the initial admission in a care episode.
A key feature in our analysis is the concept of a clinical pathway, which represents the set of medical procedures a patient in a given DRG category follows, including diagnostic tests, medications, and consultations, conducted during care delivery (De Bleser et al. 2006). In practice, when providers treat a condition, they often vary in terms of the clinical pathways chosen, resulting in different ranges of costs and health outcomes. Under FFS, because each service is billed for separately, hospitals and physicians tend to not worry much about identifying and coordinating on a common clinical pathway. In contrast, because cost accounting is a bigger concern under bundled payments, hospitals and physicians need to better understand their common clinical pathways and identify the most cost-effective ones.

For any given condition, without loss of generality, we consider two pathways, one being more intensive (and costly), and the other one being less intensive (and cheaper). For the ease of interpretation, we assume that the costlier pathway is a superset of the cheaper pathway, with more procedures performed. The corresponding costs to the hospital for these pathways are respectively captured by $c_1$, and $c_2$, where $c_1 > c_2$, and the corresponding reimbursements to the physician are respectively captured by $r_{1,p}$, and $r_{2,p}$, where $r_{1,p} > r_{2,p}$. We denote the payments to the hospital under FFS using $r_h$. Lastly, we remark that physician costs are not directly related to care intensity (Weiss 2003), and hence are not included in the analysis.

For each DRG, there is an optimal level of care intensity, $I_0 \in [0, 1]$, which captures the fraction of patients that should be assigned to the more expensive pathway when financial considerations are put aside, and the only objective is to maximize outcomes from patients’ perspective (e.g., $I_0 = 0.65$ for a specific DRG implies when the objective is to maximize care quality, 65% of patients should follow the more intense and costlier pathway). Typically, optimal patient outcomes are achieved at an
interior point of the support \([0,1]\) for care intensity, because both overtreatment and undertreatment lead to suboptimal patient outcomes (Fisher et al. 2003, Doyle et al. 2015). The practiced level of care intensity, \(I\), however often deviates from the optimal level of care intensity. In particular, given that the hospital is paid a fixed DRG-based rate under the FFS, it may try to reduce the cost by influencing physician practices through various means such as care protocols to reduce the intensity.\(^2\)

The literature on hospitals’ motives suggests a wide spectrum of profit orientation and the quality outcomes (Sloan 2000). In order to better identify the key trade-offs in studying hospital/physician interaction, in our main model, we assume hospital is profit-driven with the sole motivation to reduce costs, while, the physicians aim to both provide appropriate quality of care\(^3\) and increase their revenues. In Section 4.1, we extend this base model to consider the setting where in addition to physicians, hospital is also quality-considerate.\(^4\) As such, similar to the established literature in health economics and policy (Dor and Watson 1995, Jelovac and Macho-Stadler 2002, Crainich et al. 2008), we model the practiced care intensity, \(I\), to be jointly produced by the hospital and physicians as follows:

\[
I(i_h, i_p) := (1 - i_h)\Psi + (1 - \Psi)i_p, \tag{1}
\]

where intensity \(I\) represents the practiced level of care intensity and captures the fraction of patients assigned to the more expensive pathway, \(i_h \in [0,1]\) represents the physician influencing effort, \(i_p \in [0,1]\) is the hospital influencing effort, and \(\Psi \in (0,1)\) is the physician alignment coefficient. Physician alignment (also referred to as physician integration) is a well-established and widely studied concept in medical and health economics literature, and is defined as the degree to which physicians share the same mission, vision, and objectives with their hospital systems and work toward their success (Shortell et al. 2001, Ma 1994, Huang and McCarthy 2015). Empirical research has found that physicians within large group practices, those physicians that receive a stipend, and older physicians\(^5\) in general have higher alignment level with the hospital (Shortell et al. 2001).

3.1. FFS Payment Model and Analysis

In the base model, we assume the hospital’s objective is to reduce costs and increase its profits while the physicians weigh both their monetary benefits and the patient’s interest as captured by the quality

\(^2\) Such hospital-driven intensity reduction was observed during the transition to the DRG-based system in the 80s from the cost-based reimbursement, where hospitals were paid based on average cost of a patient per-diem and the length of stay. At that time, hospitals were able to reduce the average length of stay for non-surgical patients from 9.4 days to 7.2 days within only a few years (Altman 2012).

\(^3\) The physician controls quality of care by deciding on the intensity level.

\(^4\) While our overall results are qualitatively similar, the analysis in the base model better reveals the key trade-offs.

\(^5\) The authors of the conducted empirical work argue that this is perhaps due to lower level of competitiveness and fewer available alternatives older physicians have, given that they are at the later stages of their career.
of care. This dichotomy in physician’s behavior between monetary benefits and some measure of benevolence, altruism, or professionalism is well established and is used widely in the health economics literature (e.g., Ellis and McGuire 1986), and is captured by the physician benevolence coefficient \( w_b \) in our analysis. We model FFS case, the status quo, as a Nash equilibrium of a single-stage game where hospital and physician simultaneously choose their effort levels. The utility functions for the hospital and physician are characterized as:

\[
F_{h}^{\text{FFS}} = r_h - c_1 I(i_h, i_p) - c_2 (1 - I(i_h, i_p)) \\
F_{p}^{\text{FFS}} = -w_b (I(i_h, i_p) - I_0)^2 + r_{1,p} I(i_h, i_p) + r_{2,p} (1 - I(i_h, i_p))
\]  

(2)

We remark that \( F_{h}^{\text{FFS}} \) and \( F_{p}^{\text{FFS}} \) respectively represent the average per-patient utility for the hospital and physician. Under FFS, the hospital receives a single DRG-based payment, \( r_h \), and incurs costs due to patients receiving care via either the costly or cheaper pathway. In contrast, the physician receives pathway-dependent payments and incurs disutility due to any deviation from the optimal care intensity. The following lemma characterizes the physician preference if the care intensity is solely determined by her.

**Lemma 1.** Under FFS, the physician’s utility function is maximized when the level of care intensity \( I \) is equal to \( I_0 + \frac{\Delta r_p}{2w_b} \).

The above lemma shows that, under FFS, the physician prefers to increase the care intensity beyond the optimal care intensity, \( I_0 \), by the benevolence-adjusted financial motives, \( \frac{\Delta r_p}{2w_b} \), when the hospital has no influence on the care intensity. As expected, the financial motives, measured by the physician payment/revenue differential between the costly and cheaper pathways, \( \Delta r_p \), increases the extent of deviation while the benevolence factor, \( w_b \), decreases the extent of deviation from the optimal care intensity. Next, we present an intuitive but helpful result, which states that the optimal intensity under FFS (i.e., status quo) is inversely related to the physician alignment. That is, the higher the alignment, the lower the intensity (i.e., more patients will follow the cheaper pathway). Let \( I^* \) denote the equilibrium intensity under FFS, then the following lemma provides an upper bound for \( I^* \).

**Lemma 2.** The equilibrium intensity under FFS is bounded from above by \( 1 - \Psi \), that is

\[
I^* \leq 1 - \Psi.
\]  

(3)

Lemma 1 and 2 together imply that, under FFS, physicians would be inclined to set the optimal intensity level to \( I_0 + \frac{\Delta r_p}{2w_b} \), whenever they can; however, when physicians are highly aligned with the
hospital, the level of alignment dominates the physicians’ financial incentives to set the intensity at
the benevolence-adjusted level, hence the equilibrium intensity would be bounded by $1 - \Psi$.

In the following result, we fully characterize the optimal intensity under FFS as a function of the
physician alignment level, $\Psi$. In particular, we show that in hospitals with non-salaried physicians,
where physician alignment is lower than a certain threshold, $\bar{\Psi}$, the equilibrium intensity under FFS
is driven by the physicians only and is set to the maximizing value of the physician utility function,
$I_0 + \frac{\Delta r_p}{2w_b}$. On the other hand, in hospitals where physician alignment is higher than this threshold $\bar{\Psi}$,
physician utility maximizing intensity, $I_0 + \frac{\Delta r_p}{2w_b}$, becomes larger than the maximum intensity level,$1 - \Psi$ (by Lemma 2), and hence the intensity level is set to its maximum value, $1 - \Psi$. We define
this threshold value $\bar{\Psi}$ in the lemma below, and throughout the remainder of this paper, we refer to
those hospitals with $\Psi > \bar{\Psi}$ ($\Psi < \bar{\Psi}$) as hospitals with high (low) physician alignment.

**Lemma 3 (Status-quo intensity).** The equilibrium intensity under FFS $I^*$ is given by

$$I^* = \begin{cases} 
I_0 + \frac{\Delta r_p}{2w_b} & \text{if } \Psi \leq \bar{\Psi} \\
1 - \Psi & \text{otherwise}
\end{cases}$$

where

$$\bar{\Psi} := 1 - I_0 - \frac{\Delta r_p}{2w_b} \quad (4)$$

3.2. Bundled Payment Model and Analysis

In line with the commonly practiced retrospective bundled payments models, we assume that the
hospital is initially operating under FFS. The hospital is then offered bundled payments, under which
the hospital is allowed to reward the physicians for cooperating with the hospital for cost reductions
through a gainsharing contract. In studying the equilibrium under the bundled setting, we consider
a two-stage game, where the hospital first offers a gainsharing contract to the physicians and sets its
effort, and then, in response, the physicians choose their effort. We derive the subgame perfect Nash
equilibrium using backward induction.

While under FFS, the hospital and physicians are paid separately, under bundled payments, the
hospital (or its parent health system) is responsible as the “convener” to receive and distribute the
payment. The total payment under BP, which is paid to the hospital and is shared with physicians,
is captured by $r_{BP}$. We note that $r_{BP}$ accounts for all the costs from the services included in the
bundle, in addition to hospital payment. In line with the retrospective bundled payment practice,
we assume that the hospital reimburses the physicians at the FFS rates, but the hospital can now
also pay an additional gainsharing amount that is proportional to the physicians’ influencing effort
toward the cheaper pathway, i.e, $(1 - i_p)T$, to the physicians to incentivize them to reduce costs by
reducing intensity. Hence, the decision variables are $i_h$ and $T$ (gainsharing) for the hospital and $i_p$ for the physicians. Then, the hospital’s and physician’s utility functions under the bundled payments become:

$$F_{h}^{BP} = r_{BP} - (c_1 + r_{1,p})(I(i_h, i_p) - (1 - I(i_h, i_p)) - (1 - i_p)T,$$

$$F_{p}^{BP} = -w_b(I(i_h, i_p) - I_0)^2 + r_{1,p}I(i_h, i_p) + r_{2,p}(1 - I(i_h, i_p)) + (1 - i_p)T. \hspace{1cm}(5)$$

Bundling will occur if each of the stakeholders—the physicians, the hospital, and the payer—has higher payoffs, compared with FFS. Clearly, in order for the hospital not to lose compared with the FFS, the total payments to the physicians under bundled payments should not be too high, and the overall payment to the hospital from the payer under bundled payments, $r_{BP}$, should not be too small. This means that in addition to $r_{BP}$ being less than $r_{FFS}$ (the total payment from the payer under FFS, including hospital and physician reimbursements), the difference should not be too high in absolute terms. As we study the effect of hospital-physician alignment on bundled payments in this paper, we state this payer-relevant condition\(^6\) as an assumption below and characterize the hospital and physician-related conditions for bundling in Theorem 1.

**Assumption 1.** $r_{FFS} - r_{BP}$ is not too large; specifically

$$0 < r_{FFS} - r_{BP} \leq \begin{cases} 
(\Delta c + \Delta r_p)(1 - I_0 - \Psi) - \frac{1}{2w_b} \cdot \left[ \frac{T^2}{(1 - \Psi)^2} \right] \hspace{1cm}, & \text{if } \Psi \geq \bar{\Psi} \\
\frac{\Delta r_p^2}{2w_b} - \frac{1}{2w_b} \cdot \left[ \frac{T^2}{(1 - \Psi)^2} - \frac{T}{(1 - \Psi)}(\Delta c + 2\Delta r_p - 2w_b(1 - I_0 - \Psi)) + \Delta r_p^2 \right] \hspace{1cm}, & \text{if } \Psi \leq \bar{\Psi} 
\end{cases}$$

where $T$ is quantified later in Proposition 1, $\Delta c$ is the cost differential, and $\Delta r_p$ is the revenue differential between the pathways.

The following theorem characterizes the incentives for bundling in the context of physicians’ alignment with hospital.

**Theorem 1 (When do they bundle?).** Suppose the condition in Assumption 1 holds. Then, hospitals and physicians will bundle if and only if:

$$|\Psi - \bar{\Psi}| < \frac{\Delta c + \Delta r_p}{2w_b} \hspace{1cm}(6)$$

\(^6\)Note that if this condition is not satisfied, that is, if the payment by the paper under BP is too small, then, clearly bundling will not be feasible.
Theorem 1 has several important implications. First, we observe that the following factors are critical for bundling to occur: i) hospital influence relative to physician in determining the care intensity (physician alignment), ii) the financial incentives for each party as determined by cost and revenue differentials, iii) optimal care intensity associated with the disease condition ($I_0$ as captured within $\bar{\Psi}$) relative to physician alignment ($\Psi - \bar{\Psi}$) and iv) physician’s level of care for quality (as measured by the benevolence factor $w_b$). Because bundling is a joint decision under conflicting incentives, it is intuitive that the relative influence matters in bundling decisions. The incentives of physicians and hospitals respectively manifest in the cost and revenue differentials, and therefore, the chance of bundling increases with increasing differentials. In some sense, the cost and revenue differentials represent the cost saving and revenue enhancement opportunities from hospital’s and physicians’ perspectives, respectively. Second, we observe that when the physician alignment level and the care intensity threshold ($\bar{\Psi}$) are within a range defined by the cost and revenue differentials, hospital and physician will bundle. Given the care intensity threshold, bundling will not occur in hospitals with too high physician alignment or with too low physician alignment. An intuitive explanation for this interesting finding is as follows. We know from Lemma 4 that as the physician alignment level increases, level of care intensity under FFS decreases. Thus, in hospitals where physician alignment is high, most of the potential cost savings would have been already realized under FFS, and there would be very little room for further cost reduction and hence savings through bundled payments. On the other hand, in hospitals with low physician alignment, level of care intensity under FFS would be higher than the preferred intensity from patient’s perspective ($I_0$), and hence there will be more opportunities for cost-reduction via bundling. However, in the bundling scenario, the relative revenue loss for physicians outweighs the revenue gain from cost reduction. Physicians, therefore, will not have enough incentives to cooperate with the hospital and engage in bundling activities, and given the loose level of alignment, hospital lacks the power to influence physicians and integrate them to the bundling initiatives. As such, when physician alignment level is low, although there is much room for cost-reduction, bundling will not occur.

Next, in Proposition 1, we characterize the optimal solution when bundling occurs, and highlight the role of gainsharing, an incentive mechanism for physicians that is not allowed under FFS.

**Proposition 1 (Optimal solution and the role of gainsharing).** If bundling occurs as outlined in Theorem 1, then the optimal solution becomes the following:

\[
i_h = 1,
\]

\[
T = 2w_b(1 - \Psi) \min \left\{ \frac{1}{2} (I_0 + \frac{\Delta c + 2\Delta r_p}{2w_b} - (1 - \Psi)), I_0 + \frac{\Delta r_p}{2w_b} \right\},
\]

\[
i_p = \frac{1}{1 - \Psi} \left( I_0 + \frac{\Delta r_p}{2w_b} - \frac{T}{2w_b(1 - \Psi)} \right).
\]

In this case, it always holds that $T > 0$. 
One notable finding in Proposition 1 is that the gainsharing amount, $T$, is always positive, which is in line with the expert opinions that gainsharing is critical in moving bundled payments forward (Froimson et al. 2013). Intuitively, this is because while the hospital would be inclined to minimize intensity, the physicians in return may resist and attempt to keep the intensity as high as it was in FFS. Therefore, in order for the hospital to incentivize physicians to reduce the level of care intensity, the hospital would have to compensate physicians through gainsharing.

Gainsharing aims to get physicians to cooperate in cost reduction efforts, which in turn creates value for hospitals. Under the bundled payments, physicians would also be interested in how much gainsharing, as a “value-based” part of their compensation, they actually receive. We characterize the gainsharing amount $T$ as a function of the physician alignment factor $\Psi$. The amount that is gainshared with physicians is maximal at $\Psi = \bar{\Psi} + \frac{1}{2}(I_0 - \Delta c/2\omega_b)$, first increasing with physician alignment but later decreasing, as showed in Figure 1. For lower physician alignment level values, $\Psi < \bar{\Psi} + \frac{1}{2}(I_0 - \Delta c/2\omega_b)$, an increase in the level of physician alignment requires larger gainsharing to compensate physicians’ higher financial loss from bundling. For higher physician alignment level values, $\Psi > \bar{\Psi} + \frac{1}{2}(I_0 - \Delta c/2\omega_b)$, an increase in the physician alignment requires smaller gainsharing because reduced cost saving opportunities along with better cooperating physicians facilitate a reduction.

![Figure 1 Gainsharing amount by physician alignment.](image)

Notes. Notice that the maximum is not attained at $\bar{\Psi}$ but rather at $\bar{\Psi} + \frac{1}{2}(I_0 - \Delta c/2\omega_b)$, as explained in the main text.

### 3.2.1. Savings under bundled payments

In the previous section, we have characterized the conditions under which bundling is preferred and the corresponding optimal solution. In this subsection, we analyze the extent of savings achieved under bundled payments as the alignment level $\Psi$ changes. Let $\Sigma := r^{FFS} - r^{BP}$, the difference between the total reimbursement under FFS and the minimal acceptable reimbursement under bundled payments, to represent the overall savings from
bundled payments. Then, we have the following result characterizing the overall savings from bundled payments as a function of the physician alignment level, $\Psi$:

**Proposition 2.** When bundling is feasible and preferred, in hospitals with relatively low physician alignment (i.e., $\Psi \leq \bar{\Psi}$), increasing alignment level leads to higher savings. In contrast, in hospitals with relatively higher physician alignment (i.e., $\Psi \geq \bar{\Psi}$), increasing alignment level further leads to lower savings. More specifically, for $i_p > 0$, the savings are given by:

$$\Sigma = \begin{cases} 
\frac{(\Delta c + 2\Delta r_p - 2w_b(1 - I_0 - \Psi))^2}{8w_b}, & \text{if } \Psi \leq \bar{\Psi} \\
\frac{(\Delta c + 2w_b(1 - I_0 - \Psi))^2}{8w_b}, & \text{if } \Psi \geq \bar{\Psi}
\end{cases}$$

and the savings $\Sigma$ are maximized when $\Psi = \bar{\Psi}$, as also illustrated in Figure 2.

Figure 2 visualizes Proposition 2. As seen from this figure, savings initially increase, with the highest savings occurring when $\Psi = \bar{\Psi}$, and then decrease as the alignment level further increases. Interestingly, this result implies that hospitals that have the opportunity for highest savings from bundling are not the ones with very high or low physician level alignment, but instead are the ones with moderately high level of physician alignment. A conclusive matching of alignment levels and specific hospital types is a difficult task. However, some good examples to hospitals with high, low, and moderate physician alignment levels could be integrated healthcare systems, stand-alone hospitals in competitive markets, and stand-alone community hospitals in less competitive markets, respectively. Based on these examples, the result in Proposition 2 implies that when bundling is feasible, integrated networks or stand-alone hospitals in competitive markets are expected to achieve relatively lower savings, compared with stand-alone community hospitals in less competitive markets. Although such a finding may appear counterintuitive at first, it has an intuitive explanation: inefficiencies, and hence the potential for savings, are highest in hospitals with low physician alignment. As the alignment level increases, the proportion of these savings that is realized, increases. Then, at a certain threshold alignment level, all potential savings are realized, and as the alignment level further increases, the room for savings tends to decrease.

### 3.2.2. Intensity and quality

In this subsection, we analyze how care intensity and quality will be influenced by bundled payments.

**Corollary 1.** The optimal care intensity under bundled payments, $I^t$, is less than that under the FFS, where $I^*$ is given by

$$I^t = I_0 + \frac{\Delta r_p}{2w_b} - \frac{T}{2w_b(1 - \Psi)} \leq I^*.$$  

(8)
Figure 2  Savings as a function of $\Psi$ under bundled payments.

Note: The y-scales of $r^{BP}$ and $r^{FFS} - r^{BP}$ are not comparable. For other cases of parameters, the figure does not differ dramatically, even though the bounds of $\Psi$ where bundling occurs may vary (even down to 0 or up to 1). Furthermore, the right, linear part of $r^{BP}$ may be decreasing rather than increasing. Details are given in Appendix ??.

Corollary 1 corroborates experts' intuition that, compared with FFS, bundled payments are expected to decrease intensity, and hence utilization and costs, which underlies the motivation of CMS to implement bundled payments (Mechanic and Altman 2009). However, it is unclear whether this decreased intensity would lead to a decrease or an increase in quality, which we investigate next.

Let $\Delta Q$ be the extent of deviation from the optimal care intensity under FFS and bundled payments, representing the quality difference between the two regimes. Specifically, let $\Delta Q := |I^* - I_0| - |I^d - I_0|$, where $I^* > I^d > 0$. $\Delta Q > 0$ implies less deviation from the optimal care intensity, $I_0$, under bundled payments, which may be interpreted as quality improvement under bundled payments as compared with the FFS. On the other hand, the higher the deviation from the optimal care intensity under bundled payments is (i.e., $|I^d - I_0|$), the lower $\Delta Q$ becomes, which may eventually become negative, implying a reduction in quality.

**Proposition 3.** Compared with FFS, quality of care under bundled payments may decrease or increase, depending on the physician alignment level, $\Psi$. In particular:

1. For $\Psi \geq \hat{\Psi}$, we have

$$\Delta Q = \begin{cases} \frac{1}{2}(1 - \Psi - I_0 - \frac{\Delta c}{2w_b}) > 0 & \text{if } I_0 + \frac{\Delta c}{2w_b} < 1 - \Psi \\ \frac{1}{2}(3(1 - \Psi - I_0) - \frac{\Delta c}{2w_b}) \leq 0 & \text{if } I_0 < 1 - \Psi < I_0 + \frac{\Delta c}{2w_b} \\ \frac{1}{2}(I_0 + \Psi - 1 - \frac{\Delta c}{2w_b}) < 0 & \text{if } 1 - \Psi < I_0. \end{cases}$$

(9)

2. For $\Psi < \hat{\Psi}$,

$$\Delta Q = \begin{cases} -\frac{1}{2}(I_0 + \frac{\Delta c - 2\Delta r_p}{2w_b} - (1 - \Psi)) \text{ can be } < \text{ or } > 0 & \text{if } I_0 + \frac{\Delta c}{2w_b} \geq 1 - \Psi \\ \frac{1}{2}(I_0 + \frac{\Delta c + 2\Delta r_p}{2w_b} - (1 - \Psi)) = 2w_b(1 - \Psi)T > 0 & \text{if } I_0 + \frac{\Delta c}{2w_b} < 1 - \Psi \end{cases}$$

(10)

Proposition 3 presents a somewhat surprising result, which suggests that the care quality under bundled payment may increase or decrease, depending on the hospital and physician alignment level.
3.2.3. When the Payer Adjusts Payments to Achieve Certain Quality

Our analysis in 3.2.2 suggests that under bundled payments, intensity decreases and the associated quality may decrease or increase when a hospital is allowed to operate as an unconstrained profit maximizer. Expecting the intensity reduction with a concern for quality, it is plausible that under BP, the payer may set quality guarantees. Namely, he may set a lower bound on $I^\#$ by paying just enough for the hospital to raise the quality at or above $I^\#$. The analysis then corresponds to the characterization of the efficiency frontier for payments vs. the achievable quality. Our main findings are summarized in the following theorem.

**Theorem 2.** When the payer sets minimum quality requirements, three scenarios are possible:

1. If $I^* > I_0$, when bundling occurs, the resulting quality under bundled payments may be lower or higher than that under the FFS.
2. If $I^* < I_0$ and $\Delta r_p > \Delta c$, then a higher quality level can always be achieved for cheaper with bundled payments than FFS.
3. If $I^* < I_0$ and $\Delta r_p < \Delta c$, then a higher quality level may be achieved for cheaper with FFS, compared with bundled payments.

Figures 3 and 4 illustrate Cases 2 and 3 in Theorem 2, respectively. Perhaps the more interesting case is Case 3 presented in Figure 4, which implies that the payer may reach high quality more easily under FFS than under bundled payments. In other words, when the physician reimbursement differential between the two pathways is lower than the hospital cost differential, FFS may offer higher quality for lower cost. While this is a somewhat counterintuitive result, there is a reasonable explanation: under FFS, increasing intensity (hence in this case quality) increases physician reimbursement but not hospital reimbursement. Hence, if $\Delta r_p$ is relatively low compared to $\Delta c$, the payer can easily and cheaply motivate physicians, and FFS is then more desirable if high quality is required.

A typical example for $\Delta r_p < \Delta c$ may be the common problem faced by the hospitals when physicians’ preference of implants determine the cost of knee replacement to the hospital. When the choice is between a cheaper and an expensive implant, the two pathways is determined by physician’s implant choice. The procedure conducted by the physician may be similar in both cases (i.e., $\Delta r_p$ is small), but the hospital incurs a higher cost when the expensive implant is used ($\Delta c$ is large). In contrast, $\Delta r_p > \Delta c$ may hold, for instance, if the physician can order additional laboratory or radiological testing in the costlier pathway which generates relatively higher revenues for physicians, even though the cost differential for the hospital may be lower as compared with revenue differential for the physician.
4. Extensions

Hospitals are businesses that provide health services to make profits and rely on the physicians to provide high quality. However, some hospitals may value high-quality care in addition to the quality arising from physicians’ altruism. Indeed, health economics literature suggests that non-profit hospitals tend to value quality more than for-profit hospitals (Chang and Jacobson 2012). In addition, reimbursement mechanisms that tie the hospital payments to value (i.e., quality) is becoming more common, as in the case of performance-based payment models. Under a performance-based payment model, hospital payments are adjusted based on quality performance (Rosenthal and Dudley 2007), as in the readmission penalties discussed earlier (Andritsos and Tang 2015, Zhang et al. 2016). Consistent with this theme, Section 4.1 extends the base model to capture the bundling decisions of a hospital whose utility function incorporates the resulting quality into their decision making beyond what physicians decide on. Then, in 4.2, we extend the quality model to analyze a setting where physicians are salaried employees of the hospital, physicians and hospitals contractually agree on a compensation with a pre-established performance requirements from the physicians. This latter analysis is relevant and important because several qualitative studies have discussed that different physician compensation models lead to differences in incentives and that hospital-physician integration through physician employment is a promising direction for success under new payment models, including bundled payments (Lee et al. 2012, O’Malley et al. 2011).

4.1. Quality-Aware Hospital

As we discussed earlier, hospitals differ with respect to quality valuations. In this section, we consider a quality-aware hospital model (henceforth, referred to as the quality model), which extends the base model by considering utility reduction that quality-conscious hospitals experience as the practiced intensity deviates from the optimal care intensity. We scale the deviation from the optimal intensity with a factor of \( w^\text{FFS}_q \) (respectively \( w^\text{BP}_q \)) under FFS (respectively under BP) which denotes the dollar

![Figure 3](comparison_of_efficiency_frontiers_for_Dc<Drp.png)

**Figure 3** Comparison of efficiency frontiers for \( \Delta c < \Delta r_p \).

![Figure 4](comparison_of_efficiency_frontiers_for_Dc>Drp.png)

**Figure 4** Comparison of efficiency frontiers for \( \Delta c > \Delta r_p \).
weight that a hospital places on per unit quality. In the quality model, the hospital’s and physician’s utility functions under FFS and BP are given by:

\[
\begin{align*}
F_{FFS}^h &= r_h - c_1 I(i_h, i_p) - c_2 (1 - I(i_h, i_p)) - w_{FFS}^q (I(i_h, i_p) - I_0)^2, \\
F_{FFS}^p &= -w_b (I(i_h, i_p) - I_0)^2 + r_{1,p} I(i_h, i_p) + r_{2,p} (1 - I(i_h, i_p)), \\
F_{BP}^h &= r_{BP} - (c_1 + r_{1,p}) I(i_h, i_p) - (c_2 + r_{2,p}) (1 - I(i_h, i_p)) - w_{BP}^q (I(i_h, i_p) - I_0)^2 - (1 - i_p) T, \\
F_{BP}^p &= -w_b (I(i_h, i_p) - I_0)^2 + r_{1,p} I(i_h, i_p) + r_{2,p} (1 - I(i_h, i_p)) + (1 - i_p) T.
\end{align*}
\]

Our analysis in this section considers the case where \( w_{BP}^q = w_{FFS}^q =: w_q \), namely the hospital’s valuation of quality does not change in a bundled payment setup as compared with the FFS. Alternatively, in the Appendix, we also consider another practical case where \( w_{BP}^q > w_{FFS}^q \), which indicates a lower disutility from lower quality under the FFS. This second case captures the possibility that hospitals could make a profit from the readmission stay under the FFS, resulting in a lower care provision during the index admission (while however patients are typically seen by different providers in readmissions, compared with index admissions). In contrast, under a bundled payment setup, the initial admission and readmissions will be bundled into one as both are part of a single episode.

Overall, our results resemble those of the base model with the exception that under the quality model, bundling is more difficult and the implied quality is no lower than that of the base model. In particular, under the quality-aware hospital setup, some FFS arrangements are a priori strictly superior to any bundling scenario (Corollary 2) where physicians’ and payer’s incentives are no longer fully aligned (Proposition 4). Although bundling is less likely under the quality model, when realized, the quality may be higher than that under the base model (Corollary 4). We proceed with presenting our main findings and start with a result which is the analogue of Lemma 3 in the base model.

**Lemma 4 (Status-quo intensity under FFS).** The equilibrium intensity under FFS \( I^* \) is given by

\[
I^* = \begin{cases} 
I_0 + \frac{\Delta r_p}{2 w_b} & \text{if } \Psi \leq \bar{\Psi} \\
1 - \Psi & \text{if } \bar{\Psi} \leq \Psi \leq \tilde{\Psi} \\
I_0 - \frac{\Delta c}{2 w_q} & \text{if } \Psi \geq \tilde{\Psi}
\end{cases}
\]

where

\[ \bar{\Psi} := 1 - I_0 - \frac{\Delta c}{2 w_q} \]

Note that the first two cases correspond to the cases in Lemma 3. The third case occurs in hospitals with highly aligned physicians, where setting intensity as \( 1 - \Psi \) as in Lemma 3 would imply much deviation from the optimal care intensity \( I_0 \), which is not preferred when both physician and hospital
are quality-conscious. As such, in this case, the intensity is bounded below by $I_0 - \frac{\Delta c}{2w_q}$. Next we present a result analogous to Theorem 1, assuming that the difference between $r^{FFS}$ and $r^{BP}$ will be small enough to allow for hospital profitability, as in Assumption 1.

**Assumption 2.** $r^{FFS} - r^{BP}$ is not too large; specifically:

$$0 < r^{FFS} - r^{BP} \leq \begin{cases} 
\frac{1}{4w_q(2w_q + w_b)}(w_q \Delta r_p + w_q(\Delta c + 2\Delta r_p) - 2w_b^2(1 - I_0 - \Psi))^2 & \text{if } T \leq T_{\text{max}}, \Psi \leq \bar{\Psi} \\
\frac{1}{4w_b^2}(2I_0w_b + \Delta r_p)(w_q \Delta r_p + 2w_b(-I_0w_q + \Delta c + \Delta r_p) - 4w_q^2(1 - \Psi)) & \text{if } T > T_{\text{max}}, \Psi \leq \bar{\Psi} \\
\frac{1}{4(2w_q + w_b)}(\Delta c + 2w_b(1 - I_0 - \Psi) + 2w_q(1 - I_0 - \Psi))^2 & \text{if } T \leq T_{\text{max}}, \Psi \geq \bar{\Psi} \\
(-2I_0(w_b + w_q) + \Delta c + w_q(1 - \Psi))(1 - \Psi) & \text{if } T > T_{\text{max}}, \Psi \geq \bar{\Psi} 
\end{cases}$$

Under the quality model, hospital’s and physicians’ simultaneous interest in bundling does not necessarily translate into payer benefiting from bundling, which in turn results in many special cases in the solution space. For clarity of the presentation, in the following result, we present only one case, which is analogous to Theorem 1. The interpretation is also very similar to that of Theorem 1, but is applicable in a more limited setting. The full set of solutions is included in the Appendix, in Proposition 2.

**Proposition 4 (When do they bundle?).** Suppose the condition in Assumption 2 holds. Then, hospitals and physicians will bundle if:

$$\Delta c \geq 2I_0w_q + 2w_b(1 - I_0 - \Psi), \Psi \leq 1 - I_0 - \frac{\Delta r_p}{2w_b},$$

$$\frac{\Delta r_p}{2w_b} \cdot \left(\frac{w_q}{w_b} + 1\right) + \frac{\Delta c}{2w_b} > \bar{\Psi} - \Psi, \text{ and}$$

$$\frac{\Delta r_p}{2w_b} + \frac{\Delta c}{2(w_b + w_q)} > -(\bar{\Psi} - \Psi).$$

In the base model, getting physicians on board was sufficient to align hospital-physician-payer trio. Under the quality model, the payer will bundle only in certain cases, of which one is presented in Proposition 4: when the opportunities for savings are high (implied by $T = T_{\text{max}}$) and the physicians are not aligned well with the hospitals (implied by $\Psi \leq 1 - I_0 - \frac{\Delta r_p}{2w_b}$), bundling would occur as long as Condition (13) holds. This condition is an almost immediate equivalent of Condition (6) in Theorem 1 except i) the lower and upper bounds for the distance from the critical threshold for the physician alignment level, $|\Psi - \Psi|$, are shifted, ii) the bounds are no longer symmetric around the critical threshold $\bar{\Psi}$, iii) the alignment range where bundling is feasible may be smaller or larger as compared with the base model depending on the parameter values.

**Corollary 2.** When $\Psi \geq 1 - I_0 + \frac{\Delta c}{2w_q}$, bundling is not attractive for hospitals and physicians.

Hospitals can gain from bundled payments by decreasing intensity. However, this will often also decrease quality as well. When a hospital is concerned about the quality (as in the quality model), an
increase in cost savings due to decreasing intensity may or may not outweigh the utility losses from decreasing quality. Many new payment mechanisms tie payments to quality, which in our framework essentially is the value of the hospital quality objective (specifically, it would increase $w_q$). Corollary 2 characterizes the region where bundling is not a profitable proposition for the hospitals and the physicians where an increase in $w_q$ suggests a wider range of $\Psi$ values. Hence, increasing $w_q$ makes bundling more difficult. This observation warns CMS and other payers that they must be cautious when combining bundled payments with other quality-improving payment mechanisms (which would make the hospitals value the quality more). Indeed, we already see that hospitals participating in bundled payments complain about such conflicting payment mechanisms (Dummit et al. 2015). An uncareful launch of bundling and performance-based programs can discourage hospitals from bundling or it can lead them to incur losses, if bundling was mandatory. In the following result, we compare how a quality-aware hospital achieves different quality outcomes under bundled payments than a quality-blind hospital as in the base model. The quality model can be considered as the FFS vs. the bundled payment mechanism where the payer simultaneously offers performance-based payment programs to hospitals.

**Corollary 3.** Under bundling, the intensity decreases as compared with the FFS.

Corollary 3 compares care intensity under the FFS vs. the bundled payments for quality-aware hospitals, which is in line with the findings from the base model (see Corollary 1).

**Corollary 4.**

1. If $\Psi \geq \bar{\Psi}$, then the quality under bundling will be the same under the base model and the quality model.

2. If $\Psi < \bar{\Psi}$, then the quality under bundling will be higher in the quality model if

$$\frac{\Delta c - \Delta r_p}{2w_b} < \bar{\Psi} - \Psi.$$  \hspace{1cm} (14)

and higher in the base model otherwise.

Corollary 4 has important implications for the payer in designing the payment mechanism and setting the quality expectations from it. When combining the bundled payments with performance-based programs (e.g., $w_q > 0$), a payer should be careful in the implementation and should account for i) the hospital-physician alignment levels, and ii) cost and revenue differentials for the disease condition. For hospitals with highly aligned physicians ($\Psi \geq \bar{\Psi}$), the quality will remain same when performance-based programs are jointly offered with bundling. However, for hospitals that work with less-aligned physicians ($\Psi < \bar{\Psi}$), combining performance-based programs with bundling may increase the quality as intended, or may lead to an unintended reduction of quality. Corollary 2 and 4 together
imply that a payer should carefully balance the expected savings, incentives for bundling, and the quality implications when deciding to jointly offer the bundled and performance-based payment programs.

4.2. Physicians as Salaried Employees

In this section, we consider a salaried model setup, which represents a smaller but sizable portion of hospitals. Under the salaried setup, we find that the equilibrium intensity under FFS may be lower or higher in comparison to non-salaried setups (Lemma 5). However, even when the equilibrium FFS intensity is lower in comparison to non-salaried setups, it is still possible to bundle under some conditions. Further, among hospitals with salaried physicians and similar cost and revenue differentials, those that value quality highly are more likely to bundle; which is in contrast to the finding that quality-aware hospitals that do not employ physicians are less likely to bundle if they value quality highly (Proposition 5). Comparing the salaried model with the corresponding non-salaried analog based on the quality model, we find that bundling will be more feasible when cost savings between the pathways are low to moderate (Theorem 3).

When physicians are salaried employees, they are not active decision makers for bundling; instead the hospital pays physicians a fixed salary and collects the physician’s part of the reimbursement itself. Therefore, the modeling is decision-theoretic (henceforth referred to as the salaried model), rather than a game-theoretical model, where $\Psi = 1$. Similar to Section 4.1, we consider a setting where the hospital is quality-conscious (i.e. $w_q > 0$), and at the end of this section, we compare the quality and the salaried models (Theorem 3). Given this setup, the hospital’s utility function under FFS is given by

$$F_h^{FFS} = r_h - c_1 - c_1 p I(i_h) - c_2 - c_2 p (1 - I(i_h)) - w_q (I(i_h) - I_0)^2,$$  \hspace{1cm} (15)

where $r_S$ is the salary paid to physicians, prorated to a single patient visit. On the other hand, under bundled payments, the utility function is given by

$$F_h^{BP} = r^{BP} - c_1 I(i_h) - c_2 (1 - I(i_h)) - w_q (I(i_h) - I_0)^2.$$ \hspace{1cm} (16)

We also assume that the hospital’s concern for quality, $w_q$, and also the salary to the physicians, $r_S$, is the same under FFS and bundled payments. This is reasonable because hospitals generally do not provide additional financial motivation to physicians, beyond possible gainsharing (Dummit et al. 2015).

**Lemma 5 (Status-quo intensity).** The status quo FFS intensity $I^*$ is given by

$$I^* = \min[I_0 - \frac{\Delta c - \Delta r_p}{2w_q}, 1]^+. $$ \hspace{1cm} (17)
Lemma 5 suggests that the optimal care intensity under the salaried model is lower than the optimal care intensity under the base model and it can be lower or higher than the same under the quality model. Next, we characterize when hospitals with salaried physicians would be better off with bundling.\footnote{When bundled payments and FFS lead to same payments to hospitals and cost to the payer, we break the tie in favor of the bundled payments as coordination and less administrative costs are more desirable for the payer.}

**Proposition 5 (When do they bundle?).** The hospital will bundle iff

\[
\Delta c < \Delta r_p + 2w_q I_0
\]  

and \(r^{BP}\) is high enough.

Proposition 5 is a simple yet insightful finding. Based on the inequality (18), under the salaried model, the decision to bundle depends on the cost differential between pathways relative to potential gains from bundling. Specifically, the left-hand side of the inequality in (18) represents the opportunities for cost savings, and the right-hand side represents potential gains from i) payer’s reimbursement of physician services that are paid to the hospital and ii) the value derived from the optimal quality. One important implication of this result is that among hospitals with salaried physicians and similar cost and revenue differentials, those that value quality highly are more likely to bundle. Note that this result is in contrast to the finding that quality-aware hospitals that do not employ physicians are less likely to bundle if they value quality highly. The key intuition for this difference in these two results is the following. In the case of salaried physicians, since the hospital receives a higher physician payment for the more expensive pathway from the payer but pays the physicians a fixed salary, the hospital has an incentive to increase the intensity of care under FFS. However, under bundled payments, the hospital does not have such an incentive and hence tends to reduce the intensity. The increase in hospital utility through this reduction in intensity is higher if the hospital values quality highly, and hence a hospital that values quality highly is more likely to bundle if it has salaried physicians. On the other hand, when the physicians are not salaried employees of the hospital, since the physicians have an incentive to choose a high intensity and the hospital will have to compensate the physicians to reduce the intensity, a hospital that values quality highly will have to compensate the physicians more (or equivalently limit the reduction in intensity consistent with the valuation of the quality). This causes bundling to be less profitable compared to FFS for a hospital that values quality highly in case of non-salaried physicians. One example of such hospitals are academic medical centers, which typically salary the physicians and at the same time value quality highly. Indeed, in line with our findings, observational studies show that academic medical centers are more eager to bundle (Tsai et al. 2015).
Proposition 6 (Optimal solution). If bundling occurs as outlined in Proposition 5, then the optimal solution is as follows:

$$P^2 = [I_0 - \frac{\Delta c}{2w_q}]^+. \quad (19)$$

Moreover, the care intensity decreases under bundled payments.

The optimal care intensity under bundling of the salaried model is lower than that of the FFS, and not higher than the optimal care intensity under bundling of the quality model (Lemma 4). The comparatively lower care intensity in settings where hospitals and physicians are highly aligned, as in the salaried model, emphasizes the importance of performance-based payment models. To achieve good quality, payers should offer bundling and performance-based payment models simultaneously (i.e., $w_q \gg 0$) to ensure good quality (i.e., $\lim_{w_q \to \infty}[I_0 - \frac{\Delta c}{2w_q}] = I_0$) in salaried settings or settings where hospitals and physicians are highly aligned. The next theorem compares the salary model with the quality model. Although physicians do not manage costs and do not assume any risk for low-quality care in salaried settings, they are highly aligned with hospitals.

Theorem 3. Consider a hospital with highly-aligned physicians ($\Psi \to 1$). Then the following provides a characterization of bundling under the quality and the salary models:

1. There are cases when the hospital would bundle under the salary model but not under the quality model, namely, when $\Delta c < 2w_qI_0$ (relatively low savings).
2. There are cases when the hospital would bundle under the quality model but not under the salary model, namely, when $\Delta c - \Delta r_p > 2w_qI_0$ (relatively high savings).
3. Bundling may occur under both models when $2w_qI_0 + 2w_qI_0 < \Delta c < \Delta r_p + 2w_qI_0$ (relatively moderate savings).

Theorem 3 emphasizes the role of salaried physicians and the resulting bundling incentives for hospitals, physicians, and the payers when physicians are highly aligned with the hospitals. The race between the hospital’s valuation of quality and the savings opportunities with intensity reduction due to bundling determine the difference between the salaried and non-salaried settings. When hospital’s valuation of quality is higher than the cost savings as in part (1.) of the theorem, salary model indicates presence of sufficient incentives for bundling whereas the independence of physicians under the non-salaried setting suggests FFS as a better option for hospitals and physicians. When the hospital’s cost savings relative to physicians’ revenue reduction is larger than hospital’s valuation of quality as in part (2.) of the theorem, the non-salaried setting allows for hospital and physician alignment in bundling decisions whereas the hospitals operating in the salaried setting will not have the incentives to bundle because they would be already operating efficiently under the FFS. When savings are moderate as in part (3.) of the theorem, bundling occurs under both models with highly aligned physicians. Neither high nor low savings opportunities provide a balanced incentive environment for both the physicians and the hospital, thus facilitating bundling.
5. A Machine-learning Approach to Identify Common Service Bundles and Clinical Pathways

We demonstrated in Sections 3 and 4 that the costs associated with different clinical pathways play a key role in determining whether the current efforts related to bundling are likely to be successful. Consequently, it is imperative for hospitals and physicians to better understand and manage their combined costs resulting from different clinical pathways before proposing bundled payments to a payer or accepting a payer’s proposal for bundled payments. However, identifying the common clinical pathways may be quite complicated (Curran et al. 2005). In particular, although specific pathways are well established and understood by some hospitals, especially the ones with strong information technology capabilities, they are less obvious to most hospitals. Therefore, an important issue to address when considering bundling decisions is how to identify these “naturally-occurring” pathways with different costs via a data-driven approach.

Our purpose in this section is to illustrate a practical machine learning method to identify such pathways using historical data. We remark that the “pathways” that inform bundling decisions may not necessarily represent actual “physical” pathways, but rather different ways of delivering care, resulting in different costs. For instance, a more expensive pathway may mean using a more expensive implant, ordering several unnecessary tests, or prescribing special drugs.

The machine learning method we propose uses an Institutional Review Board-exempted dataset obtained from a hospital specializing in orthopedic surgery. Standard and reproducible orthopedic surgeries such as knee/hip replacement are considered as ideal for bundling. Therefore, to illustrate our proposed approach, we use cost data for DRG 470, corresponding to knee replacement. Our dataset is obtained from a hospital specializing in orthopedic surgery and includes 364 visits of DRG 470 made over a span of about two years, which is a volume comparable with the hospitals that are actually engaged in bundled payments for this DRG under the BPCI program. In addition to detailed information on costs and a breakdown of charges, our dataset included demographic information, insurance type, additional diagnoses, BMI, and a code for attending and assisting physicians. The machine learning method we propose consists of two stages.

(Stage 1): In this stage, we characterize “service bundles” from patient charges using Latent Dirichlet Allocation (LDA) model. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics, namely, a set of tests and services provided (Blei et al. 2003). LDA is commonly used in text mining, where a number of documents is available, and each document comprises several (unobserved) topics with different frequencies. The words in the documents are randomly drawn according to the topics. The goal is to determine how words relate to different topics and how different topics are represented in each document. Analogously, in our problem, each word corresponds to a charged service, each topic
corresponds to bundles of services (“service bundles”, e.g., a blood draw and laboratory tests) that typically occur together, and each document corresponds to a patient.

(Stage 2): In this stage, we identify the cost clusters using Gaussian mixture regression for a given DRG. The mixing probabilities depend on the service bundles (or “topics”) identified in Stage I and the costs are controlled for other relevant factors such as patient demographics. These final cost clusters then represent different pathways resulting in different costs for the same DRG.

Results. In our analysis, we assume normally distributed costs within each cluster. We fitted the clusters and costs jointly using the expectation-maximization algorithm and determined the number of clusters using the Bayesian Information Criterion (Leisch 2004). The analysis of actual data yielded two cost clusters with the following statistics:

- \( c_1 \approx 19400, \sigma_1 \approx 5000, I \approx 0.18, \)
- \( c_2 \approx 8700, \sigma_2 \approx 260, (1 - I) \approx 0.82, \)

| Table 2 |
|------------------|------|-------|------------------|
|                | Estimate | Std. Error | Pr(>|z|) |
| (Intercept)     | -2.35    | 0.67      | 0.0004 |
| topic 1         | 1.24     | 0.86      | 0.15   |
| topic 2         | 1.94     | 0.85      | 0.02   |
| topic 3         | -1.10    | 1.42      | 0.44   |
| topic 4         | 0.10     | 0.96      | 0.92   |
| topic 5         | 1.14     | 0.85      | 0.18   |
| topic 6         | 2.62     | 0.84      | 0.002  |
| topic 7         | 3.19     | 1.03      | 0.002  |
| topic 8         | -1.20    | 1.36      | 0.38   |
| topic 9         | -0.38    | 1.11      | 0.73   |

Table 3  Logistic regression estimates for a model predicting the probability of the expensive pathway using care topics.

where \( c_i \) is the estimated average cost for pathway \( i \) and \( \sigma_i \) represents the estimated standard deviation of cost in pathway \( i \). Similarly as in our theoretical models, intensity \( I \) represents the fraction of patients following the more expensive pathways and corresponds to the mixing probability of the more expensive pathway.

Our LDA analysis revealed that these cost clusters are partially explained by a set of service bundles topics as shown in Table 3. In the logistic regression estimates, a more positive coefficient for a topic indicates that the patients receiving the related service bundle are more likely to end up in the more expensive pathway. Similarly, a more negative coefficient for the topic indicates that the patients receiving the related service bundle are more likely to end up in the cheaper pathway.
Altogether, the degree of presence of service bundles corresponding to the topics determines the overall probability of patient care being in the expensive.

Discussing these results with the executive team of the collaborating hospital, we found out that some of these service bundles do play a substantial role in total cost, and that some of them could be avoidable. For instance, one of the topics with a positive coefficient corresponds to a particular brand of an implant, presumably an expensive one. Another topic (not significantly associated with the expensive pathway) corresponds to lab testing, including services such as sodium testing, potassium testing, chloride sera, hemoglobin testing, hematocrit testing, and a charge for a blood analyzer that the hospital was using. Other topics include drug combinations that are often administered together. The topics in the table that are not significant correspond to treatment patterns that are not correlated with either of the pathways; they instead correspond to treatment variations that on average incur similar costs.

The numerical exercise using real data demonstrated that the machine-learning approach could identify different inpatient clinical pathways. We further emphasize that such identification can be a starting point for hospital management and does not necessarily validate the insights we obtained from the analytical modeling. Our demonstration using a small number of observations suggests that the proposed method can i) help in better understanding of the costs associated with various pathways, ii) can encourage wider implementation of the optimal pathways once hospitals identify proper patient subgroups for each pathway, and iii) can be a primer for an empirical approach when evaluating whether switching to bundled payments is profitable based on the existing care patterns as defined by clinical pathways. Future research can extend our analysis to larger and richer datasets to provide a deeper understanding of cost clusters.

6. Discussion
The emerging payment models offer a new business model for healthcare organizations and is set to change the way healthcare is delivered. In this work, we studied alignment of hospitals and physicians in the face of bundled payments. While a few previous studies have also studied bundled payments, no prior operations management study has considered care coordination and physician-hospital power struggle, which is acknowledged to be a key factor in this space (Goldsmith et al. 2016). Our study also adds to the understanding of the characteristics of medical conditions that are ideal for bundling. Different medical conditions are characterized by different quality requirements, treatment intensities, and coordination needs (Sood et al. 2011, Tsai et al. 2015); for example, while bundled payments are widely touted for use in knee replacement or hip replacement, they are not brought up as much in the

\[8\] We disguise the service bundles and refer to them as topics in order not to reveal the cost structure of the collaborating hospital.
context of stroke, where appropriate quality measures are critical but not yet well developed, costs along the entire episode of care are not well understood, and new payment models and gainsharing are less familiar to physicians-neurologists (McClellan et al. 2014). Our analysis could be a starting point for hospitals in determining required features of a medical condition for structuring the discussions around bundling.

We found that: i) hospitals with very low or very high levels of physician alignment are not ideal for bundling, and they may be worse off under bundled payments compared with FFS; ii) to engage physicians, hospitals need to gainshare, a mechanism that was not available in traditional FFS-based payment models; iii) bundled payments will decrease care intensity and, unless carefully regulated, bundling may also lead to a reduction in care quality, and iv) in an environment where hospitals are also held accountable for quality, the incentives for bundling will differ in hospitals employing salaried physicians than those where physicians are independent contractors.

6.1. Managerial and Policy Implications

While the finding that bundling is not ideal for hospitals with very low physician alignment is more intuitive, the finding that hospitals with very high physician alignment are not ideal for bundling is counter-intuitive. Physicians that are not well aligned with hospitals are unlikely to give up on their power, and hence it is more difficulty to coordinate care and reduce system inefficiencies (such as redundant medical exams). On the other hand, while highly aligned hospital and physicians are able to further coordinate care, the extent of inefficiencies, and hence the room for improvement is small in such hospitals. Because the total historical cost under FFS is taken as the benchmark in determining the rates under bundled payments, highly-aligned hospital and physicians may tend not to lower their margins by engaging in bundling initiatives.

There are important implications of our findings around the impact of the extent of physician alignment on hospitals’ tendency to bundling. In general, physicians tend to be less aligned in competitive hospital marketplaces with access to alternatives (Wholey and Burns 1991, Burns et al. 2001). In contrast, physicians would expected to be more aligned in markets where health systems dominate, e.g. Intermountain Healthcare dominates in the state of Utah. Because the physician labor market is currently mostly undersupplied (Daly 2016), we anticipate physician presence or absence in competitive or monopolistic markets will determine the alignment spectrum. First, under the voluntary bundled payment models, the predominant form which is likely to grow further, hospitals with highly aligned physicians are unlikely to adopt bundled payments. However, this may not be a concern from a societal perspective, as such hospitals are already performing cost-efficiently and less of a burden in the overall healthcare costs. Second, as hospital with lowly aligned physicians lack the power to coordinate care and induce efficient care, the voluntary bundled payments will be
ineffective in reducing inefficiencies and hence unnecessary costs in such hospitals. Third, because hospitals with a moderate-level physician alignment are more likely to embrace bundled payments, such hospital and physician groups are set to gain from the voluntary model. This would further hurt the low-alignment, inefficient hospitals that compete with the moderate-alignment hospitals. This could lead to mergers (which we expect to increase hospital power and thus alignment) or service-line closures, thereby further concentrating the market.

There have been opposing views about the role of gainsharing in bundled payments, which is used by hospitals as a device to coordinate with physicians. On the one hand, gainsharing is recognized to play a crucial role in bundled payments (Froimson et al. 2013), but on the other hand, the importance of gainsharing is underplayed (American Hospital Association 2013). Our findings reinforces the former view that the hospitals and other conveners should take gainsharing into account from the outset, when considering bundled payments. Our study is a first attempt on paving the way for determining how gainsharing proceeds can be quantified (as represented by \( T \) in our modeling) and in doing so, how the role of care intensity, quality, and pathways can be assessed.

Third, we showed that the intensity will decrease under bundled payments, but this may sometimes lead to a lower quality. The payer can use two standard approaches to avoid this. First, it may set quality thresholds, and second, it may set stop-gain limits to avoid excessive cost savings. Imposing a quality threshold may be challenging because quality is often multidimensional and defining the right quality measures and risk adjusting based on patient severity may be difficult to achieve (Dranove and Jin 2010). As for the second approach, we have showed that the quality decreases, particularly the case when the opportunities for savings are large. The second approach of stop-gain provisions seem more promising through which hospital and physicians may choose to reduce the truly unneeded care that are also a source of cost. However, stop-gain provisions may further shift the risk to providers and, in return, providers’ incentives for bundling may reduce.

Finally, an organizational choice made by hospitals—whether or not physicians are employees—could have an impact on incentives for bundling. Our analysis suggests that, when hospitals value quality highly, a salaried physician setup may be more accommodating for bundling in contrast to a setup where physicians are independent contractors to hospitals. The implications for the hospital management is that, when engaging in bundled payments, the existing physician employment structure should be considered. From a policy-making perspective, organizational structures should be taken into account when offering bundled payments, especially in an era of value-based payment incentives.

References


Grady, D., R. F. Redberg. 2010. Less is more: how less health care can result in better health. *Archives of Internal Medicine* **170**(9) 749–750.


Mechanic, Robert, Stuart Altman. 2009. Payment reform options: episode payment is a good place to start. *Health Affairs* 28(2) w262–w271.


