New Joints More Hip? Learning in the Use of New Components*

Kamalini Ramdas†, Khaled Saleh‡, Steven Stern§, and Haiyan Liu¶

August 2012

Abstract

There is a vast literature on experience curves. The measure of experience is typically product or service procedure volume, at the level of an organization, a facility, a team, or an individual. Yet most products and services involve a variety of components and sub-processes. We examine learning in hip replacement surgery as a function of a surgeon's experience at the very granular level of specific surgical component versions in addition to total experience, using a unique hand-collected dataset. In our study, five surgeons used over 1200 unique stock keeping units (236 distinct versions ignoring size variations) of four main component types, in a three year period. Experience accrued at the level of specific component versions has a significant impact on duration of surgery, which is well known to impact both quality of outcomes and cost. A single prior use of certain component versions can reduce duration of surgery by 25%. In addition, learning accrues very gradually for some component types and rapidly for others. Our findings suggest that hospitals may benefit from requiring minimum volumes for surgeons at the level of specific component versions. Further, we provide evidence of important heretofore unnoticed benefits from adopting new technologies such as surgical simulators, derived from the tremendous proliferation in devices today. We also draw important implications for the medical devices industry.

*We are grateful to Marianne Corbishley, Hyoun Ahn, Whitney Deck, and Amanda Wilson for help with data collection. For useful comments, we thank Dr. Wael Barsoum, Gerard Cachon, Sanjay Jain , Serguei Netessine, Wendy Novicoff, Nicos Savva, Elizabeth Teisberg, Karl Ulrich, and seminar attendees at Emory Business School. Mike Guthrie at Zimmer, Jerry Kie at Smith & Nephew, Chris Petrie at Stryker, and Robert McGlothin at Depuy provided valuable industry insight and expertise.

†Management Science and Operations, London Business School, kramdas@london.edu.

‡Division of Orthopedic Surgery, Southern Illinois University School of Medicine, ksaleh@siumed.edu

§Department of Economics, University of Virginia, sns5r@eservices.virginia.edu.

¶Department of Economics, University of Virginia, hl4y@virginia.edu.

1 Introduction

A long history of research on learning curves has documented the relationship between production volumes and both unit cost and quality (e.g. Wright 1936, Baloff 1971, Lieberman 1984, Fine 1986, Argote and Epple 1990, Mukherjee, Hatch and Mowery 1998, Argote 1999, Terwiesch and Bohn 2001).

More recently, researchers have begun to focus on the underlying drivers of learning (Terwiesch and Bohn 2001). Several studies have examined organizational aspects of learning (e.g. Pisano et al. 2001, Edmondson et al. 2001, Tucker, et al. 2007). Reagans et al. (2005) distinguish organizational experience as a driver of learning from individual experience and experience working together in teams, in the context of hip replacement surgery. Huckman et al. (2009) and Huckman and Staats (2011) find evidence of team-based learning in the software industry. Clark et al. (2011) examine both individual and organizational learning in the context of outsourced radiology services and find evidence that customer-specific learning is beneficial in both cases.

While researchers have examined learning as a function of individual experience, team experience, and organizational experience, the measure of experience is most commonly product or service procedure volume, at the level of an organization, a production or service facility, a team, or an individual. Yet most products or service procedures involve a multitude of subprocesses. For example, assembled products such as automobiles are comprised of a wide variety of components, each of which is manufactured in a number of different versions. With innovation commonly based on core platforms (Robertson and Ulrich 1998, Gopal et al. 2011), new car models typically have some uniquely designed component versions and others that are carried over from previous models or designed to be shared with other models going forward. Even products that are manufactured through continuous processes involve a number of subprocesses, some shared and others unique. Similarly, new service products, such as new financial instruments, media products, or insurance policies, also typically involve a combination of common and unique subprocesses. In the field of medicine, innovative new forms of healthcare delivery such as group delivery of preventive care combine common and unique subprocesses (Ramdas et al. 2012). Most surgeries also involve the use of a variety of medical devices and instruments. Innovation in medical devices has lead to an exploding assortment of devices, which themselves come in an ever-increasing palate of distinct versions, some of which are more commonly used than others.

Despite this ubiquitous proliferation in components or subprocesses, very little research has examined how experience at the level of product components, production subprocesses, or phys-

ical items such as medical devices used in a service procedure impacts cost or quality. Ramdas and Randall (2008) have examined the impact of experience at the level of specific component versions on the reliability of automotive braking systems. These authors find evidence that component reliability is increasing in component volume, suggesting that, for manufactured products such as automobiles, experience at the level of component versions is an important driver of quality. In a medical industry study, Diwas and Staats (2011) find that, in minimally-invasive cardiac bypass surgery, in addition to the number of surgeries performed in this procedure and those performed in the traditional invasive procedure, the variety of subtasks experienced within each surgerical procedure also impacts outcomes.

Our research furthers the investigation of whether learning occurs at more granular levels than product or service volume. We focus on service processes and operationalize our thinking in the context of surgery. It is well known in the medical profession that more experienced surgeons deliver better surgical outcomes. When a patient needs to undergo surgery, she may often have some flexibility in picking a hospital and surgeon. Based on research to date on how experience affects outcomes, she would likely be advised to pick a hospital that does a high volume of the procedure in question and a surgeon who has done many surgeries of the procedure in question. For example, the Leapfrog Group, a consortium of large corporations and public agencies in the US that purchase healthcare, has advocated volume-based referral since 2000 (Birkmeyer et al. 2003, Finks et al. 2011).[1] Leapfrog uses minimum hospital volume standards for several hospital procedures. Based on their finding that better outcomes at high volume hospitals are driven largely by high volume surgeons, Birkmeyer et al. (2003) suggest that, in addition to volume-based referral at the level of hospital volume, standards based on surgeon volumes also should be used. Increasing surgeons' volumes on a specific type of surgery requires active management of the way in which patients needing that surgery are distributed to surgeons within a hospital – essentially, the surgery type in question needs to be restricted to a smaller number of surgeons. However, the conclusions reached by Birkmeyer et al. (2003) rely on the assumption that covariation of volume and quality does not reflect consumers' preference for high-quality surgeons; in other words, quality may drive volume instead. Later in this paper, we provide some evidence associated with this issue.

Due to device proliferation, surgeons today face a plethora of choices in medical devices and their variants, even within a particular surgery type. We examine in this study whether it is enough to consider a surgeon's experience in terms of total number of surgeries of the type in

[1] Private payers and professional organizations in the US are also establishing minimum hospital volume standards in order for a hospital to be accredited as a Center of Excellence, for a variety of operations (e.g., see http://www.acsbscn.org/docs/Program_Manual_v4.03-10-11.pdf)

question, such as total hip replacement surgery, or if what is needed is an even more granular approach, focusing on experience with specific device versions. We examine whether, to what extent, and how learning occurs at the level of what we call "component-specific experience," where a "component" refers to a medical device. If learning occurs at the level of specific component versions, then a surgeon who has done thousands of surgeries could still falter on his or her first usage of a particular component version. Also, the global characteristics of the learning curve would determine the number of surgeries necessary for a surgeon to become proficient with a particular component version. Depending on whether learning occurs quickly in the first usage or gradually with multiple usages of a particular component version, it would take a surgeon less or more experience to get up to speed with specific component versions.

To examine these research questions, we assembled together a unique and extremely detailed dataset that contains information on the specific component versions used on each hip replacement surgery performed at the University of Virginia Hospital over a three year period, 2006-2008. Our dataset combines inputs from different hospital databases with hand-collected data from multiple sources within the hospital. Although a great number of studies have examined learning in medical and other settings, to our knowledge we are the first to examine learning at the level of granularity that has been made possible by our unique dataset.

We find that for a specific type of surgery – total hip replacement – component-specific experience within certain component types is a significant driver of reduction in duration of surgery. Duration is a widely used outcome measure that is known to impact both cost (by affecting capacity – e.g., see Olivares et al. 2008, Saleh et al. 2009) and quality (by affecting infection rates, blood loss and other complications, e.g., see Peersman et al. 2006, Yasunaga 2009). Consistently across a variety of different specifications, we find that a single prior use of a component version within certain component types reduces the duration of surgery by about 40 minutes, at a statistical significance level of 1%. Since the average duration of surgery in our data is only 166 minutes, this represents an almost 25% increase in duration. On the other hand, other component types exhibit a more gradual learning curve, while experience with some components does not seem to impact learning at all.

Our research has important implications for hospitals seeking to organize surgery in ways that are well-suited to the high device variety and short device life cycle environment that characterizes medicine today. The implications we highlight in this paper have been informed by our discussions with a number of medical professionals and managers at device manufacturers, as well as the personal experience of one of our authors, who is a practising orthopedic surgeon.[2] In

[2] And to some extent that of another of our authors, who has experienced orthopedic surgery first hand.

organizing surgery to cope with high device variety, our findings suggest that surgeon volume per se is not adequate to ensure higher quality. For certain types of surgical components, minimum standards may need to be based on minimum volumes by component version for each surgeon.

Aside from managing the assignment of individual cases to surgeons, policy decisions and hospital management can influence the choice of devices and instruments available to surgeons. The US government has recently proposed the use of "gainsharing" arrangements through which physicians at a hospital can share in cost savings that result directly from productivity gains or increased efficiency (Ketcham and Furukawa 2008). Gainsharing essentially offers physicians a financial stake in controlling hospital costs. While physicians use hospital facilities, traditionally they have billed and been paid by insurance providers independent of the hospital based on volume of procedures performed. Hospitals, on the other hand, are reimbursed separately a fixed amount per procedure, depending on admitting diagnosis. This payment covers most costs including those that physicians control, such as the use of supplies and selection of medical devices. Thus traditionally physicians have had little incentive to control costs. With gainsharing, a surgeon is incentivized to substitute a cheaper device version for a costly one if there are no detrimental effects from doing so. Gainsharing therefore is expected to lower purchasing costs due to subsitution of cheaper devices, combined with volume discounts due to higher volumes concentrated on a smaller set of device stock keeping units (SKUs) (Buczko 2011). Our findings suggest that, aside from these savings, gainsharing also has the potential to reduce costs through freeing up expensive OR capacity, and to improve quality due to focusing learning on a smaller set of devices.

Our research also has important implications for how surgery is taught. In medical school, surgeons-in-training practice on cadavers and synthetic plastic bones using surgical tools and implants. Students are often taught using only one or two variants of a medical device, resulting in an a priori high probability that a surgeon will be using a device for the first time on a patient. While on-the-job training is also an important part of surgical education, Gorman et al. (2000) lament that surgery education is still "largely mired in the 100-year-old apprenticeship model best exemplified by the phrase: see one, do one, teach one." Gallagher and Cates (2004) also criticize the apprenticeship model for its lack of structure as junior surgeons are trained based on the random arrival of particular cases to a hospital. Quite apart from this concern, given the ever-increasing variety of devices and device versions available today, it is practically impossible for a junior surgeon to have observed a surgery performed using every device version she is likely to use. For example, in the context that we examine, total hip replacement surgery, a total of just over 1200 unique SKUs (236 distinct versions ignoring size variations) of the four

main device types were in usage during the three year period of our study.

A relatively recent innovation in surgery training that can help address this concern is computer-based surgical simulation. Such simulation takes many forms including virtual reality simulation (Meier et al. 2001), and has been shown to reduce duration and improve surgical outcomes (Seymour et al. 2002, Noble et al. 2003, Saleh et al. 2009). Aggarwal and Darzi (2006) point out that, while simulation-based training is a prerequisite in industries such as airlines and nuclear power, where high reliability is critical, it is "a niche player in medical education." These authors suggest that, while surgical simulators are often criticized for not being lifelike, "the real problem has more to do with a lack of motivation or understanding on the part of educational leaders than with the eventual outcomes." Even if surgical simulators are less lifelike than cadavers or plastic bones, the true choice today is more often between practicing use of a new device version via simulation, or not at all. The wide assortment of available device versions, combined with our finding that learning in fact occurs at the level of device versions, suggests a much higher-than-anticipated payoff from investing in innovative approaches such as virtual reality simulation for surgery training. Ideally one could imagine a surgeon practicing on a device version that she has never used before via simulation prior to operating on a patient.

Of course, committing to surgical education via simulation tools requires significant investment (Saleh et al. 2009), therefore quantification of the potential benefits from doing so is important. We find that a single prior usage of a "stem" component version (the stem is a component is inserted into the patient's thigh bone in hip replacement surgery) can reduce duration of surgery by 25% on average. Our discussions with orthopedic surgeons[3] suggest that surgical simulation can reduce the extra time associated with the first usage of such component versions subtantially. Of the 763 total hip replacement surgeries performed at the UVA hospital over a three year period, 93 involved a new stem version, resulting in an estimated 3,720 minutes (7.75 days) of additional OR time due to new stem versions alone. Any reduction in this time would directly translate into increased OR capacity. Since operating theaters are used to do a variety of surgeries, such gains will very likely accrue accross a variety of different components, freeing up a significant amount of expensive OR suite capacity. Saleh et al. (2009) note that OR capacity is the most expensive resource in a hospital. To our knowledge, we are the first to estimate the potential savings in OR time that can be made possible by speeding up the learning curve for new device versions.

Our finding that learning accrues at different rates for different types of devices also has

[3] We also drew on the experience of one of our coauthors who is an orthopedic surgeon.

direct implications for how surgery should be taught. For devices with an initial steep learning curve, one exposure to a new device version may be enough to derive most of the benefits of experience. On the other hand, for devices with a gradual learning curve, surgeons in training may benefit from repeated practice on cadavers, plastic bones, or surgery simulators.

Unlike all prior research we know of on learning, which typically draws implications for the organization where the learning occurs or sometimes for its interactions with customers (e.g. Clark et al. 2011, Crawford and Shum 2005), our research has important implications for an upstream provider industry - in our case, the medical devices industry. The reason we are able to offer these insights is that we examine learning at a much more granular level that has not been done in prior research. Our finding that experience accrues at the level of specific component versions implies that product proliferation potentially can hurt outcomes, as documented by Ramdas and Randall (2008). On the other hand, Ramdas and Randall (2008) also find that components that are tailored to the products that they are used on improve reliability. Our research suggests that, in medical devices as well, device makers must trade off the benefits of a component version best tailored to specific patient needs against the cost of surgeons needing to move up the learning curve for a new component version. Our finding that learning varies drastically across component types also has several implications for device proliferation. Other things being equal, it makes sense to encourage innovation in device types that exhibit a steep learning curve over those that exhibit a shallow learning curve, as surgeons can come up to speed faster on the former. For device types that exhibit a gradual learning curve, it makes sense for device makers to investigate which features require more learning effort and standardize these features where possible. Note that Ramdas and Randall (2008) are unable to comment on where to focus component variety, as they examine only one type of component.

Given the significant learning costs associated with new devices, device makers should benefit from actions that enable surgeons to learn faster on new medical devices. In fact, device makers today are starting to set up cadaver labs where surgeons can practise their devices. They may also benefit from teaming up with surgical simulation companies to further enhance surgeons' learning.

In the next section, we formalize our hypotheses. Section 3 contains a description of the data. In Section 4, we discuss methodology. Sections 5 and 6 contain results and concluding remarks.

2 Hypotheses

The literature on learning curves provides empirical evidence that product cost decreases and quality increases in the cumulative production volume of a manufactured product. A similar pattern has been observed in services, including healthcare. Researchers in business, economics, and medicine have documented the impact of medical procedure volume on quality of outcomes in a number of settings including many types of surgery (e.g., Birkmeyer et al. 2003, Kelsey et al. 1984, Huckman and Pisano 2006, Shwartz et al. 2008, Reagans et al. 2005, Carty et al. 2009, Diwas and Staats 2011, Clark et al. 2011, and Finks et al. 2011). A number of recent studies have also examined the impact of hospital volume and surgeon volume on outcomes in hip replacement surgery. In a review of twenty-six medical research studies examining learning in hip replacement surgeries, Shervin et al. (2007) find that hospital volumes and surgeon volumes are important drivers of improved outcomes.

The literature on learning in surgery has examined several outcomes, including mortality, the need for revision surgery, and procedure completion time, or "duration of surgery." Duration of surgery (procedure completion time) is a widely used outcome measure for a variety of different types of surgeries. Duration is considered to be an important outcome measure for several reasons. First, in any surgery, the chances of infection are directly proportional to the duration of surgery. Thus, complications due to infection can be reduced by shortening the duration of surgery (Peersman et al. 2006). Second, blood loss and other post-operative complications are also associated with longer duration of surgery (Yasunaga 2009), resulting in longer convalescence and longer length of hospital stay. Third, other measures, such as mortality and the possibility of revision, would require a much larger sample to measure precise effects because these variables are somewhat rare events. Duration of surgery also directly impacts costs through it effect on OR suite capacity, which is the most expensive resource in a hospital (Olivares et al. 2008, Saleh et al. 2009).

Hypothesis 1: Greater individual experience in terms of total number of surgeries performed of a particular procedure type reduces procedure completion time.

The impact of experience on cost and quality is well-documented, and more recent research has focused on the underlying drivers of these relationships. A stream of research has examined to what extent learning accumulated in the course of production at one site transfers to other sites or to other products manufactured at the same site. Argote (1999) notes that learning transfer across facilities is incomplete. In a study of process innovation, Hatch and Mowery (1998) note losses in semi-conductor yield when production is moved from development facilities

to manufacturing facilities. Examining product volumes and costs, Benkard (2000) reports that, in aircraft production, only a partial transfer of knowledge occurs across different models built at a production site. Examining component volumes and quality, Ramdas and Randall (2008) find that, in the automotive industry, component-specific learning accumulates faster when component volume is attained on a single model rather than over multiple models. Similarly, in the healthcare context, Huckman and Pisano (2006) find that surgeons' performance is not fully portable across hospitals, suggesting that performance improvement due to experience is at least in part firm-specific, likely due to greater familiarity with organization-specific assets. Clark et al. (2011) find that outsourced radiologists' experience is not fully portable across customer hospitals. In the context of coronary artery bypass surgery, Diwas and Staats (2011) find that experience with related surgery types is less beneficial than experience with the focal surgery type. On the other hand, Eaton and Kortum (2002) find significant learning spillovers across national borders. Taking a much more granular view, we examine whether a surgeon's experience is portable across component variations within the *same* type of surgical procedure, at a single hospital.

In the surgical context, we anticipate that learning is a function not only of a surgeon's overall experience with a particular surgical procedure but also his or her component-specific experience at the level of variants of the medical devices used. We expect such component-specific experience to matter because there are significant differences across variants of the main devices used in most surgeries. For example, in the case of total hip replacement surgery, there are four main components: the stem or femoral component, which is inserted into the patient's thigh bone; the shell or acetabular component, which is inserted into the patient's hip socket; and the head and liner components, which together comprise the ball and socket joint (see Figure 1). There are many variants within each of these four components, which differ substantially in shape, material and coatings, and characteristics that are likely to affect a surgeon's ease in using them (Figure 2 shows two distinct stem versions).

The different components used in any surgery can differ substantially in terms of the degree of difficulty involved in their use. For example, of the four main components used in total hip replacement surgery , two components - stems and shells - come in direct contact with the patient's bone, while the other two components - heads and liners - do not. A stem component needs to be positioned properly and joined to the patient's thigh bone or femur, while the shell component needs to be positioned and joined to the patient's acetabulum or hip bone. The other two components are essentially popped on and do not require delicate maneuvering and attaching. Hip surgery is in some ways quite akin to highly skilled carpentry. The placement of

a stem or shell component is similar to getting ceiling moldings to fit together exactly right at the corners of a room, while the placement of a head or liner component is more like popping a towel bar into its brackets. One might expect that experience would be more significant for those components that require greater skill and dexterity to place properly.

One also might expect that the way in which learning accumulates might differ substantially across components. Some components might involve a learning curve that is initally very steep, with most of the learning occuring on the very first usage, while others might involve a gradual learning curve. What type of learning curve might be associated with a particular component is largely an empirical question and an important one with implications for innovation in how surgery is organized and taught, as well as for innovation in the medical device industry.

Hypothesis 2: Greater component-specific experience with the components used in a surgery reduces procedure completion time.

While each surgical component often comes in a wide variety of versions, at times there can be similarities across subsets of component versions, due, for example, to similarity in shape, material, or method of insertion. For example, of the components used in total hip replacement surgery, stem component versions can be classified in two broad families based on the method used to insert the stem component into the thigh bone. One might expect learning spillovers associated with component versions that are within the same family, due to the closer similarity in the tasks involved in using such components. We refer to this type of experience across closely related component versions as component-family experience.

Hypothesis 3: Greater component-family experience with the components used in a surgery reduces procedure completion time.

In the next section we describe data and variables.

3 Data and Variables

We obtained data from the University of Virginia hospital for all total hip replacement surgeries performed during the time period starting August 2006 and ending November 2008. A total of 763 total hip replacements were performed by 6 surgeons during this period.[4] We obtained data on all of the components used in each of these surgeries from a hospital database that is used for operational and accounting purposes. This database is not maintained by the surgeons and

[4] We include 5 surgeons out of those 6 surgeons in our final sample, since one of the surgeons performed only 2 surgeries and has missing values for some variables.

others involved in actually performing the surgeries. We were able to obtain access to this database through the permission of the Heath Sciences Information Systems Department at the University of Virginia. Perhaps partly due to the difficulty in accessing this type of detailed data on components usage, we are aware of no other study that has examined learning at the level of components, despite there being a plethora of studies of learning in the surgical context. We use our data to develop measures of surgeon experience at the level of specific component versions. We supplement this data with data on outcome and control variables from multiple sources including hand-collected data from individual patient records, other hospital databases, and hand-collected data from records kept in the operating theaters. Hand-collection of data from individual patient records was a painstaking process. We hired three nurses to identify and read through the relevant sections of each patient's paper medical record binder, and gather the needed data. Since a patient's medical record was often a thick binder covering all visits to the hospital and its associated clinics, finding and correctly interpreting the relevant data required trained medical expertise. As an example, one of our nurse research assistants needed to locate and read through the surgical note for every patient in order to identify reasons for surgery and any complexities during surgery, which were then coded and used in our analysis. Similarly, obtaining information from the records kept at operating theaters required our nurse research assistants to access these paper documents through the operating theater nurses. Due to missing data on outcome and control variables, we have complete data for 554 of the 763 surgeries performed in our study period.

We chose to limit our study to surgeries performed using components from one of four major vendors that account for over ninety percent of our sample - Stryker, Depuy, Smith & Nephew and Zimmer - resulting in an eventual sample size of 503 surgeries. Due to missing values on the reasons for surgery, which we include as a control variable, our sample size is further reduced to 488 surgeries. Sample comparison tests detect no significant differences between the initial sample of 763 surgeries and the subsample of 488 surgeries that we use in our analysis.

Outcome

Duration: The duration of surgery, $Duration_{ist}$, is the amount of time in minutes from the start of surgery, i.e. skin opening, until the end of the surgery, i.e. skin closing, performed on patient i by surgeon s on day t. Duration does not include the time taken to anesthetize the patient or the time that the patient may remain in the operating theater to "wake up" before being taken to the post-anesthesia care unit. Figure 3 is a histogram of the duration of surgeries in our sample.

Experience

A surgeon's overall experience, Exp_{ist} is the number of total hip replacement surgeries that surgeon s has performed during the study period prior to the surgery performed by s on patient i at time t. We generate overall experience for each surgeon using all 763 surgeries completed in our study period.[5]

Aside from gaining overall experience over time, surgeons also accumulate experience over time with specific components. While a variety of minor components including screws and springs are used in each surgery, we learned from our discussions with orthopedic experts that stem, head, liner, and shell components are the primary drivers of the time taken to complete a surgery. These components also are quite expensive. The prices for components in our dataset ranged from \$1,525 to \$6,955 for stems, \$624 to \$7,400 for shells, \$356 to \$5,100 for heads, and \$998 to \$4,050 for liners.

In order to determine whether experience accrues at more granular levels within a particular surgery type, we consider different levels of aggregation of component-specific experience within total hip replacement surgery. The most granular level at which component experience can be accrued is the component SKU. In the period of our study, a total of 369 unique stem SKUs, 262 unique shell SKUs, 275 unique head SKUs, and 303 unique liner SKUs were used, as listed in Table 1. Within each of these four categories, component SKUs differ in technology, shape, materials, surface, coatings, and size. For our purposes, we group together SKUs that differ only in minor size variations. For each of the four component types and for each of the four vendors included in our study, all SKUs whose labels differed only in size were aggregated into a single component version. Accomplishing this was complicated by the fact that, in some cases, SKUs that differed only in size had slightly different item descriptions due to inconsistent use of abbreviations. We therefore enlisted the help of four of the hospital's surgical representatives for orthopedic surgery, one from each of the four vendors included in our study, to help us group all of the SKUs for each of the four component types into subgroups of component versions such that SKUs within each subgroup varied only in size. These surgical representatives used their extensive knowledge of the specific component versions used for hip replacement surgeries at the University of Virginia Hospital to categorize SKUs. [6] Through this procedure, the large number of SKUs was reduced to a much smaller number of component versions in Table 1. For

[5] For each surgeon, experience accrued prior to our study period does not vary from one surgery to another within the study period, and is fully captured via a surgeon dummy in the case of regressions including surgeon fixed effects.

[6] For example, in the Zimmer line, the Trilogy Multi-Holed Shell component version contained 14 different size variations, ranging from 44mm to 70mm in diameter, that were used at the hospital in the study period, while the Trilogy Uni-Holed Shell component version contained 11 different size variations, ranging from 46mm to 68mm in diameter.

example, for Stryker, we obtained 15 stem, 19 head, 17 liner, and 8 shell component versions.

We also consider whether learning is driven by a surgeon's experience at intermediate levels of aggregation between component versions at the most granular level and surgeries at the most aggregate level. Through discussions with orthopaedic experts, we learned that, for stem components, there is a natural way to think about intermediate levels of aggregation, which we describe below. For shell, head, and liner components, on the other hand, we learned that there were no natural intermediate levels of aggregation.

At an intermediate level of aggregation for stems, we aggregate stem component versions from all vendors into two groups based on the method used for joining the component to the femur. *Cemented* stems have a smooth surface, and a cement-based adhesive is used to attach the stem to the femur. *Uncemented* stems, on the other hand, have a rough surface such that a proper joining of device and bone occurs when the bone grows around the implant. The dummy variable $Cemented_{ist}$ takes on a value of 1 if a cemented stem is used for the surgery performed on patient i by surgeon s on day t and is included as a control variable.

We next create surgeon experience variables at the level of each component version and at the intermediate level of aggregation for stem components.

Component–Version-Experience: $Expstem_{ist}$ is the number of times surgeon s has used a stem SKU from the same component version as the stem SKU used in the surgery performed by s on patient i at time t. $Expshell_{ist}$, $Exphead_{ist}$, and $Expliner_{ist}$ are defined simlarly.

Component-Version-Used-Before: $Stem\_before_{ist}$ takes on the value one if surgeon s who performed surgery i at time t had used a stem SKU belonging to the same component version as the stem SKU used in this surgery, prior to time t[7]. $Shell\_before_{ist}$, $Head\_before_{ist}$, and $Liner\_before_{ist}$ are defined similarly. It allows us to distinguish between learning occuring in the first use of a component and subsequent learning.

Joining-Experience: The variable $Expstem\_join_{ist}$ is the number of times surgeon s has used the joining method - cemented or uncemented joining - appropriate for the stem SKU on patient i at time t prior to that surgery, during the study period. $Stem\_before\_join_{ist}$ takes on value one if surgeon s who performed surgery i at time t had used the joining method appropriate for the stem SKU used in that surgery prior to time t and in the duration of our study period.

Other variables

We control for a number of patient characteristics that may affect both duration of surgery

[7] A value of zero indicates that the surgeon had no prior experience with the stem version, or that the surgery did not use a stem component. We control separately for the number of stems used in each surgery to account for the latter.

and the components chosen by a surgeon, including age, gender, body mass index, anesthetic severity index, and patient comorbidities, as detailed below.

Body Mass Index: Body mass index (BMI_{it}) is a standard measure of obesity calculated as the ratio of weight to squared height. BMI directly affects duration of surgery as more obese patients can take longer to operate on, and it also affects the choice of components used for surgery.

Anesthetic Severity Assessment: ASA_{ist} is another standard variable used in the medical literature that takes on integer values between 1 and 4 and is a rating of the overall fitness of the patient prior to surgery based on a system developed by the American Society of Anesthesiologists.[8]

$Nconfound_{it}$ is the number of comorbidities, coded as the sum of ten indicator variables, which indicate the presence of each of the ten most common patient comorbidites in hip surgery.[9]

In addition, we control for a number of variables related to the surgery itself. Since revision surgeries are generally much more complex than first-time surgeries, we include a revision dummy. A "bilateral" dummy indicates whether surgery is performed on one or both hips. Surgeries that involve both hips are generally expected to take longer. We include a dummy for the use of a unipolar head component.[10] We include counts of the number of component SKUs of each type used in a surgery. For example, n_stem_{ist} is the number of different stem SKUs used in the surgery performed by surgeon s on patient i at time t.[11] In an overwhelming majority of cases, at most one component SKU of each type is used. In some cases, all four component types are not used, as only malfunctioning components may be replaced. Also, occasionally, more than one component SKU are used due to, for example, dropping a component SKU or poor fit of the first component SKU used. Such cases are typically correlated with a longer duration of surgery.[12]

We also control for the reasons for surgery. We include indicator variables for each of

[8] At the UVA hospital, an ASA score for each patient is provided by both the anesthesiologist and a surgical team member. The two scores are highly correlated, and we use the average of the two scores. Our results are robust to the use of each of the individual scores instead.

[9] The ten most common patient comorbidites are diabetes, kidney disease, liver disease, respiratory disorder, COPD, immune deficiency, prior venous thromboembolism, substance dependence, cardiovascular disease, high blood pressure, and bleeding disorders. In other specifications we also used the Charlson comorbidity index (Charlson 1987), a also a sum of indicators for presence of each five digit ICD9 code description for a patient condition, and a sum of indicators for presence of each three digit ICD9 code description (these are slightly more aggregate descriptions). We also estimated specifications in which we interacted complexity measures with the revision dummy, and with time trend.

[10] Unipolar heads are used in the treatment of hip fractures, which often involve a distorted anatomy and more bloody surgical field, resulting in longer duration.

[11] Note that a component SKU is recorded as having been used as soon as it is taken out of its original casing.

[12] The use of more than one component SKU in the surgery can explain the longer duration if it is because of poor fit of the first component SKU used. However, it can also be the result of the learning process if it is because of dropping. It is more common for surgeons to make mistakes if they are not familiar with a component.

the most frequently cited reasons for surgery: arthritis, severe arthritis, end-stage arthritis, avascular necrosis, dysplasia, fracture, deformity, childhood disease, and post-traumatic bone conditions, as well as an "other reasons" category that includes very infrequently cited reasons. The reasons for a revision surgery can be any of the above reasons or other reasons that pertain only to revisions, for example loosening of components in the first surgery or infection. In the case of revision surgeries, we include an additional variable, $reasons\_for\_revision_{it}$, which is the sum of indicator variables for each of the following reasons that are specific to revision surgeries: acetabular osteolysis, aseptic loosening, infection, pain, dislocation, and hematoma.

Finally, we include a linear time trend variable to control for technological advances and other trends over time and surgeon-specific fixed effects to control for surgeon unobservables such as education and prior experience.[13] Table 2 provides descriptive statistics of all variables.

4 Empirical Analysis

4.1 Benchmark Specifications

Prior research in business, economics, and medicine has found that surgeon-specific volume can improve surgery outcomes and reduce surgery duration (e.g., Birkmeyer et al. 2003, Finks et al. 2011). We examine the impact of overall experience on surgery duration using the benchmark specification,

$$Duration_{ist} = \alpha_0 + \alpha_1 Exp_{ist} + \alpha_2 controls + \varepsilon_{it}.$$

Next, we examine the effect of component-specific experience over and above overall volume of surgeries performed at the lowest level of aggregation of component-specific experience, namely component versions. We first introduce our continuous measures of component-specific experience:

$$\begin{aligned} Duration_{ist} &= \beta_0 + \beta_1 Exp_{ist} + \beta_2 Exphead_{ist} + \beta_3 Expstem_{ist} \\ &\quad + \beta_4 Expliner_{ist} + \beta_5 Expshell_{ist} + \beta_6 controls + \xi_{it}. \end{aligned}$$

Next, we add our discrete measures of component-specific experience to the above formulation:

$$\begin{aligned} Duration_{ist} &= \gamma_0 + \gamma_1 Exp_{ist} + \gamma_2 Exphead_{ist} + \gamma_3 Expstem_{ist} + \gamma_4 Expliner_{ist} \\ &\quad + \gamma_5 Expshell_{ist} + \gamma_6 Stem\_before_{ist} + \gamma_7 Head\_before_{ist} \\ &\quad + \gamma_8 Liner\_before_{ist} + \gamma_9 Shell\_before_{ist} + \gamma_{10} controls + \omega_{it}. \end{aligned}$$

Finally, we examine the effect of component-specific experience at intermediate levels of aggregation. To determine whether a surgeon's learning is driven by experience accrued at the

[13] Note that a surgeon's prior experience is completely captured by a surgeon fixed effect.

level of number of surgeries she has performed, at the level of specific component versions, or at an intermediate level of aggregation, all of these three levels of aggregation need to be included at once as explanatory variables. Thus, we include two variables that measure intermediate experience in the joining method used to insert stem components:

$$\begin{aligned} Duration_{ist} &= \delta_0 + \delta_1 Exp_{ist} + \delta_2 Exphead_{ist} + \delta_3 Expstem_{ist} + \delta_4 Expliner_{ist} \\ &+\delta_5 Expshell_{ist} + \delta_6 Stem\_before_{ist} + \delta_7 Head\_before_{ist} \\ &+\delta_8 Liner\_before_{ist} + \delta_9 Shell\_before_{ist} + \delta_{10} Expjoin_{ist} \\ &+\delta_{11} Join\_before_{ist} + \delta_{12} controls + \upsilon_{it}. \end{aligned}$$

Experience curves traditionally have been modeled using log, log linear, polynomial, or piecewise linear specifications to capture the diminishing returns from additional units of experience (Argote 1999, Thornton and Thompson 2001). The above specifications assume constant returns from additional units of experience aside from an initial nonlinear term associated with whether a component version has been used before. We explore other non-linear specifications in section 5.3 below.

4.2 Empirical Results

Table 3 contains the results from the benchmark specifications discussed above.

Impact of Overall Experience

In Column 1 of Table 3, we see that overall experience has a significant negative coefficient in the absence of any measures of component-specific experience. Other things equal, an increase of 100 surgeries in a surgeon's overall experience in total hip replacement surgery will reduce the duration of surgery by 19 minutes on average. This result is consistent with past research on learning in hip replacement surgery which has pointed out that the overall experience of individual surgeons or hospitals can improve surgery performance and reduce surgery duration (e.g., Shervin et al. 2007, Yasanuga et al. 2009, Reagans et al. 2005). However, the magnitude of the coefficient is not very large. In fact, we see in Column 2 that, after we add measures of component-specific experience, it is no longer significant. We also see that surgery duration decreases significantly with age, likely due to deterioration of muscle mass with age of the patient, which reduces the time taken to cut through muscle tissue.

Surgeries on male patients take about 19 minutes longer than those on female patients, on average, likely due to greater body mass. Not surprisingly, revision surgeries and surgeries in which both the left and right hip joints are operated on display a significantly longer duration. Surgeries that involve more than one stem SKU result in a longer duration. In such surgeries, the initially used stem SKU is deemed inadequate, resulting in a longer surgery as the original

stem SKU may need to be removed and a second one inserted. An increase in the number of comorbidities, which is a measure of the complexity of surgery, significantly increases duration of surgery, with a coefficient of 7.65 minutes. Even after controlling for complexity of surgery in a variety of different ways,[14] we observe a significant positive trend over time. Our discussions with surgeons at the UVA hospital indicate that over time there has been a trend at this hospital towards using specialized nurses outside their main area of focus, in order to reduce the idle time costs associated with dedicated nurses. In fact, in a study of operating room efficiency conducted at the UVA hospital in the same timeframe as our study, McGowan et all (2007) note that "nurses with specialized skills had drifted to various parts of the hospital, sometimes in places where the skills were not maximally used." The increased duration of surgery we observe over time is most likely an unintended negative consequence of this trend. Increased duration of surgery may also be due to greater usage over time of residents to complete more stages in the surgery, which typically adds to the amount of time involved. Based on discussion with surgeons, neither of these factors is likely to affect the choice of components used in a surgery.

Impact of Component-Specific Experience

We next consider component-specific experience at the most granular level, i.e. component versions within each of the four key component types. Column 2 includes continuous measures of component-specific experience for the four component types, while Column 3 includes both continuous and discrete measures of component-specific experience. In Columns 2 and 3, the coefficients of our continuous measures of component-specific experience for stems and shells are negative and significant. For example, other things equal, an increase in a surgeon's experience within the same shell version by 10 surgeries reduces the duration of surgery by 1.9 minutes on average. For heads and liners, the coefficients of our continuous measures of component-specific experience are statistically insignificant.

In Column 3, we find that a single previous usage of a component version has a negative and statistically significant effect for stems and liners. If a surgeon has used a stem or liner component version before, surgery duration is reduced by about 36 minutes for stems and 18 minutes for liners, on average, other things equal. This suggests that, at the most granular level, the learning process is slow and cumulative for shells but very fast for stems and liners. As noted earlier, we find that total experience in terms of number of surgeries performed is no longer significant once component-specific experience is introduced.

[14] Recall that we included indicators for the ten most common patient comorbidites in THR surgery, and also used a variety of other measures as well as alternative specifications for complexity of surgery, in unreported regressions.

While just looking at the number of surgeries a surgeon has performed is not a good indicator of how he or she will perform on a specific surgery, it remains to be seen whether experience accrues only at the lowest or most granular level of aggregation. It is possible that, while experience with specific component versions within a component type matters, the experience accrued on specific component versions may have a positive spillover at least on related component versions within the same component type. To test if this is the case, we introduce the surgeon's experience with the joining method (cemented or uncemented) as an intermediate level of aggregation for stem components, described earlier. In Column 4, neither Exp_join nor $Join\_before$ significantly impact duration, suggesting that there are no learning spillovers at the intermediate level.

In the next section, we consider a number of robustness tests. While the significance of our results improves, we continue to find that stems and liners exhibit a very steep learning curve, while shells exhibit a shallow learning curve. Also, intermediate experience does not seem to matter in our setting.

5 Robustness Analysis

5.1 Surgeon-Specific Effects

Since our data is a panel with multiple observations on each surgeon over time, we decompose the error term into an individual-specific component and an idiosyncratic component. Our model becomes

$$\begin{aligned} y_{it} &= X_{it}\beta + u_i + \varepsilon_{it}, i = 1, \cdots, N \text{ and } t = 1, 2, .., T_i, \\ u_i &\sim iidN\left(0, \sigma_u^2\right), \varepsilon_{it} \sim iidN\left(0, \sigma_\varepsilon^2\right) \end{aligned}$$

where y_{it} is the dependent variable $Duration_{ist}$, X_{it} are all observed explanatory variables, u_i is an unobserved surgeon-specific individual effect, and ε_{it} is the idiosyncratic error. If $\sigma_u^2 > 0$, pooled OLS is inappropriate and a random effects model is preferable. We use the Breusch and Pagan (1980) Lagrange Multiplier test (LM) to test if $\sigma_u^2 > 0$. Based on the residuals from pooled OLS, $LM = 1.34$ which follows a X_1^2 distribution under H_0; thus we fail to reject the null hypothesis that $\sigma_u^2 = 0$.

Next, we treat the surgeon effects as fixed effects. Column 1 to 4 in Table 4 are similar to those in Table 3, with the addition of surgeon fixed effects. The results are very similar to those in Table 3. The main difference is the impact of overall experience in Column 1: once we use surgeon fixed effects, this coefficient is no longer significant, suggesting that the negative and significant coefficient in the pooled OLS regression in Column 1 of Table 3 was largely driven by

variation in experience (or any other omitted surgeon-specific covariates that are correlated with experience) across surgeons, with more experienced surgeons finishing surgeries faster. When surgeon fixed effects are included, the coefficients of our experience variables are being identified by variation in experience over time within each surgeon. Once we focus on only within-surgeon variation, we find that learning accumulates at the most granular level of component versions, rather than at the level of overall or intermediate experience. Note that prior studies on learning in hip replacement surgery that have reported learning at the level of surgeon volumes have not controlled for individual surgeon fixed effects (e.g., Shervin et al. 2007, Yasanuga et al. 2009, Reagans et al. 2005). Of course, whether learning occurs at the level of overall surgery volume or at a more granular level is an empirical question, and the answer will vary depending on the nature of the surgery, service procedure, or manufacturing process in question.

The error component may have a different variance for each surgeon, so we test for group heteroskedasticity using the Breusch-Pagan test. $LM = 1.76$ which follows a X_4^2 distribution under H_0; thus we can not reject the null and we continue to use a homoskedasticity assumption.

5.2 Endogeneity and IV Approach

The duration of a surgery is chiefly determined by two factors: the complexity of the surgery itself and the "quality" of the lead surgeon who performs the surgery. The complexity of a surgery is determined by many factors including the reasons for the surgery, the physical condition of the patient, and whether it is a revision surgery. The "quality" of the surgeon is determined by her overall experience, her component-specific experience, and other unobserved surgeon-specific attributes such as her training and innate dexterity or ability. Intuitively, all else equal, more complex surgeries should take longer to perform, while higher "quality" surgeons should finish faster. In practice, at the UVA hospital, patients call into a calling center and either ask for a particular surgeon, or if not, are randomly assigned to an available surgeon. Thus patients may seek out surgeons with higher quality, causing high-quality surgeons to have more experience. This endogeneity problem is likely to cause a downward bias in our results. On the other hand, if patients with more complex problems are more likely to seek out high quality surgeons, it is possible that an upward bias will occur.

While we have included a large set of variables that are related to complexity of surgery, it is impossible to completely control for the true complexity of a surgery. The fixed effects model already controls for unobserved surgeon-specific heterogeneity to some degree. An alternative approach is to use instrumental variables for our overall experience variable. To control for the above source of endogeneity, we follow Altonji and Shakotko (1987) and use the deviation of a

surgeon's experience from its average sample value as an instrument. The proposed instrument is uncorrelated with surgeon-specific quality because, by construction, its mean is differenced away. However, it is still correlated with experience to the degree that surgeon experience is time-varying. The use of this instrument significantly reduces the value of surgeon-specific effects because only the variation of experience within a surgeon is used to estimate the effect of experience on duration. Berkovec and Stern (1991) find similar results in a labor market setting after controlling for endogeneity of specific experience. With the inclusion of this instrument, we therefore drop surgeon fixed effects. We do not instrument for experience with specific component versions. Given the current state of knowledge regarding how experience impacts outcomes, we believe it is highly unlikely that a patient would select a surgeon based on the surgeon's experience with the specific component versions to be used in his or her impending surgery.

With a large set of control variables for the complexity of a surgery and the proposed instrument for surgeon-specific quality, we can partially reduce the bias caused by these possible endogeneity problems. Our model may still have some problems. For example, surgeons may prefer to use component versions that they are more familiar with for more complex surgeries. However, with the upward bias caused by this type of endogeneity, our results can be taken as "conservative" results.

Column V in Table 4 shows the results from OLS instrumenting for overall experience. The results are very similar to those obtained in the other specifications.

5.3 Nonlinearity

It is well-known that the effect of learning is usually nonlinear (Argote 1999 and any Bayesian learning models). In the benchmark specifications, we assume that, aside from first usage, the marginal impact of component-specific experience on duration is constant. Now, we investigate the possibility that component-specific experience has nonlinear effects on duration.

In order to identify component types that may exhibit a nonlinear learning effect on surgery duration, we use residuals from our benchmark regressions, excluding one component-specific experience variable at a time. For example, to detect any potential nonlinearity in *Exphead*, we estimate our benchmark specification excluding *Exphead* and save the residuals from this regression. Next, we use the residuals as dependent variables and include a third order polynomial in *Exphead* in a new regression; this shows us any possible nonlinear relationship between the residual from the previous regression and *Exphead*. After repeating the same procedure for each of the component-specific experience variables, we find that only *Expstem* deviates

significantly from a linear relationship. Based on the pattern of the *Expstem* scatterplot, we estimate a regression including a third degree polynomial in *Expstem*, keeping the other variables the same as before. The results in Column 6 of Table 4 show that the linear, quadratic, and cubic terms of component-specific experience for stems are all statistically significant at the 5% level. Increasing experience with stems reduces surgery duration to start with, and then starts to increase duration at some point, but finally reduces duration again.

In order to gain a more intuitive understanding of the nonlinear learning effect of stem-specific experience, we identifiy spline kink points for *Expstem* and estimate a piecewise linear regression equation. We find that a piecewise linear model with kink points at 50 and 150 surgeries approximates the polynomial regression very well. The results in Column 7 of Table 4 show that, for the first 50 surgeries, the slope of stem-specific experience is -0.43 and for the next 100 surgeries the change in the slope from the preceding interval is the 0.65. After 150 surgeries, the change in the slope from the preceding interval becomes -0.56. Although the change in the slope of stem-specific experience is positive for the second interval of this variable, when we check the slope (0.65-0.43 = 0.22) of that interval, it is positive but not significant. So it appears that the learning process mainly occurs in the first 50 surgeries.

With regard to our other variables, the polynomial regression and the piecewise linear regression give us very similar results to our linear OLS regression specifications. Our results are also robust to other non-linear specifications such as log-linear OLS.[15]

5.4 Serial Correlation

Since we have an unbalanced panel with varying time gaps between observations for each surgeon (for example, a surgeon may do three surgeries on one day, none the next, and two the day after that), we construct a nonparametric estimator of the correlation of two errors from the same surgeon following Stern et al. (2010). In particular, consider writing any of the models discussed above as

$$\begin{aligned} y_{it} &= X_{it}\beta + \upsilon_{it},\ t = 1, 2, .., T_i; \\ \upsilon_{it} &= \rho\left(d_{t,t-1}\right)\upsilon_{it-1} + \eta_{it}; \eta_{it} \sim iid\left(0, \sigma_\eta^2\right). \end{aligned}$$

The online appendix shows how to estimate a correlation function $\rho\left(d_{ts}\right)$ for two surgeries t and s by the same surgeon as a function of the time gap between them $d_{ts} = |t - s|$. Figure 4 displays the estimated correlation function. We see that $\widehat{\rho}\left(0\right) \approx 0.28$, implying that even

[15] In our data set, component-specific experience can take on a value of zero if a component has never been used before. Using a log-linear specification requires an ad hoc adjustment such as adding 1 to every instance of the experience variable. While we obtain very similar results, we prefer the piecewise linear and polynomial specifications because they avoid using an ad hoc adjustment and are also more flexible.

for surgeries performed by a surgeon on the same day, the estimated correlation coefficient is relatively small. Also, the estimated correlation function dies out pretty quickly. Therefore, serial correlation in ε_{it} is not a concern even though we have long panel for some surgeons.

6 Discussion and Conclusions

In this paper, we have examined the impact of learning on outcomes by examining experience accrued at the level of component versions of four major component types used in total hip replacement surgery at a single hospital. Although there is a vast literature on learning curves in surgery, to our knowledge, we are the first to examine learning as a function of experience accrued on specific component versions within a component type. We were able to examine learning at this very granular level by virtue of assembling together a unique dataset combining hospital data from different existing data sources with hand collected data, using the in-depth assistance of three nurses and four highly specialized surgical representatives. Using this data, we find that learning takes place at a highly granular level of component versions that differ only in that they are minor size variations of one another. Across a variety of specifications, we find that first time use of a new stem version increases the duration of surgery by about 40 minutes, at a statistical significance level of 1%. This almost 25% increase in duration proportionately increases the likelihood of infection, blood loss, and other complications. Also, based on our estimates, 3,720 minutes (7.75 days) of OR time could be attributed to the use of new stem versions alone, over the three year period of our study. On the other hand, shell components exhibit a more gradual learning curve, while experience with heads does not seem to impact learning at all. Given the extraordinarily high variety of component SKUs available and in use today for most medical devices,[16] these findings have very significant implications for how surgery is organized and taught, and also for device manufacturers, in today's high device variety environment.

The fact that learning occurs at the level of component versions for certain component types suggests that hospitals may need to think about how to reorganize the allocation of patient cases to surgeons, and how to incentivize surgeons to use certain component versions, so as to increase the usage of some component versions while possibly eliminating others. Our results suggest that schemes such as gainsharing, which incentivize surgeons to use a smaller variety of device versions, can result in improved quality and reduced cost due to learning benefits, aside from previously identified benefits. On the other hand, some infrequently used component versions

[16] For example, as mentioned earlier, over 1200 SKUs (236 variants ignoring size variations) of just four component types were used at the UVA Hospital in the three year period of our study.

may bring particular benefits because they are custom-tailored to specific patient needs. Being able to have the best of both worlds, i.e. to use custom-tailored components without compromising outcomes or cost due to lack of experience, calls for serious innovation in how surgery is taught. Many recent articles in the most prestigious medical journals have decried the current state of surgical education for its lack of structure and lack of innovation (e.g. Aggarwal and Darzi 2006, Reznick et al. 2006, Epstein 2007). Our research highlights an unrecognized advantage of new techniques such as surgical simulation, which can enable surgeons to practice on a wide variety of component versions prior to performing surgery. Prior research has shown that surgical residents trained using virtual reality simulation operate significantly faster and better than others trained using standard methods, when doing the same standard procedure using identical instruments (Saleh et al. 2009, Seymour et al. 2002). Seymour et al. (2002) mention the need for simulations that train "very, very specific procedures rather than focus on acquisition of basic skills". Our quantitative estimates of the substantially longer duration associated with first time use of certain components highlight a key source of variation in duration even within a specific procedure, and underscore the need for on-the-job training whenever a surgeon encounters a new device version, for certain device types. Ensuing reductions in duration of surgery should improve quality of outcomes and at the same time reduce cost by freeing up OR capacity.

To investigate why the first usage of a new stem version can take so much longer, we talked to orthopedic surgeons. We learned that each new implant variant inolves the use of new instruments. The placement of a stem into the thigh bone is the hardest part of a hip replacement surgery, and involves "broaching", i.e. using a mold to prepare the bone canal and shape it, and "reaming", i.e. using a long thin drill to open up the bone canal. The peculiarities of a new instrument are often only understood by using it, on a patient or cadaver, or through plastic bones or simulation. Aside from needing to learn a new instrument, there can also be particular idiosyncracies associated with specific device versions. For example, we learned from one orthopedic surgeon that certain stems tend to sit a couple of millimeters higher when placed in the thigh bone canal, than the stem height specified on the outside of the box. A surgeon who does not know this would only realize after reaming and broaching and inserting the stem that a deeper opening is needed - causing all of these steps to be redone. This type of problem could be easily mitigated by flagging such particular idiosyncracies to surgeons new to a device.

Our findings also have important implications for the medical devices industry. Our finding that experience accrues at the level of specific component versions implies that product proliferation potentially can hurt outcomes, as documented by Ramdas and Randall (2008). On

the other hand, Ramdas and Randall (2008) also find that components that are tailored to the products that they are used on improve reliability. Our research suggests that, in medical devices as well, device makers must trade off the benefits of a component version best tailored to specific patient needs against the cost of surgeons needing to move up the learning curve for a new component version. Our finding that learning varies drastically across component types also has several implications for device proliferation. Other things being equal, it makes sense to encourage innovation in device types that exhibit a steep learning curve over those that exhibit a shallow learning curve, as surgeons can come up to speed faster on the former. For device types that exhibit a gradual learning curve, it makes sense for device makers to investigate which features require more learning effort and standardize these features where possible. Note that Ramdas and Randall (2008) are unable to comment on where to focus component variety, as they examine only one type of component.

Our findings also that learning occurs at the level of device versions suggests that device makers should think carefully about the costs and benefits of introducing a new device version. The gain from the new design needs to be large enough to compensate for the disadvantage of starting up on a new learning curve. Our finding that learning accrues at very different rates and in different ways for different types of components also has direct implications for where device manufacturers should focus variety. Other things equal, device makers may want to offer more variety in devices that exhibit rapid learning curves, or those where learning is irrelevant. Future research that examines the underlying causes of these differential learning rates may be useful. Device makers should also reap benefits from actions that enable surgeons to learn faster on new medical devices.

While we have taken an empirical approach in this study, the tradeoffs we have unearthed open up new areas for analytically modelling the tradeoffs in organizing surgery, or in determining the focus of innovation efforts in medical devices. Also, we have taken a reduced form approach in our empirical analysis. Future empirical work can use structural estimation to model surgeons' choice of what device version to use in each surgery that could formalize the selection and learning effects that we have found some evidence for. Future research can also consider behavioral aspects of component choice, in the spirit of emerging behavioral research in healthcare operations (e.g., Mennicken et al. 2011).

Aside from the implications for healthcare, our work has more general implications for any industry where products or services are comprised of numberous subprocesses. Our findings suggest that it is important to study learning at the level of subprocesses, as learning at this very granular level has significant implications for both cost and quality.

References

Altonji, J., and R. Shakotko. 1987. Do Wages Rise with Job Seniority? *Review of Economic Studies*. 54(3): 437-459.

Aggarwal, R. and A. Darzi. 2006. Technical-Skills Training in the 21st Century. *The New England Journal of Medicine*. 355: 2695-2696

Argote, L. 1999. Organizational Learning: Creating, Retaining and transferring Knowledge. Kluwer Academic Publishers.

Argote, L., D. Epple. 1990. Learning Curves in Manufacturing, *Science*. 247: 920–924.

Baldwin, C. Y. and K. B. Clark. 2000. Design Rules: The Power of Modularity. MIT Press.

Baloff, N. 1971. The Extension of the Learning Curve. *Operational Research Quarterly*. 22: 329-340.

Benkard, C. L. 2000. Learning and Forgetting: The Dynamics of Aircraft Production. *The American Economic Review*, 90(4): 1034-1054.

Berkovec, J. and S. Stern. 1991. Job Exit Behavior of Older Men. *Econometrica*. 59(1): 189-210.

Birkmeyer, J. D., T. A. Stukel, A. E. Siewers, P. P. Goodney, D. E. Wennberg and F. Lucas. 2003 Surgeon Volume and Operative Mortalityin the United States. *The New England Journal of Medicine*. 2117 - 2127.

Breusch, T. S. and Pagan, A. R. 1980. The Lagrange Multiplier Test and Its Applications to Model Specification in Econometrics. *Review of Economic Studies*. 47(1): 239-53.

Buczko, William. 2011. Medicare Gainsharing Demonstration: Report to Congress on Quality Improvement and Savings.

Carty, M. J., R. Chan, R. Huckman, D. Snow and D. P. Orgill. 2009. A Detailed Analysis of the Reduction Mammaplasty Learning Curve: A Statistical Process Model for Approaching Surgical Performance Improvement. *Plastic & Reconstructive Surgery*. 124(3): 706-714.

Charlson, M. E., P. Pompei, K. L. Ales and C. R. MacKenzie. 1987. A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation. *Journal of Chronic Diseases*. 40(5): 373-383.

Clark, Jonathan R., Huckman, Robert S. and Staats, Bradley R. 2011. Learning from Customers in Outsourcing: Individual and Organizational Effects. Harvard Business School. Working Paper.

Crawford, Gregory S. and Shum, Matthew. 2005. Uncertainty and Learning in Pharmaceutical Demand, *Econometrica*, 73(4): 1137-1173.

Diwas, K. C. and B. R. Staats. 2011. Accumulating a Portfolio of Experience: The Effect of Focal and Related Experience on Surgeon Performance. Working Paper.

Eaton, J. and S. Kortum. 2002. Technology, Geography, and Trade. *Econometrica*, 70(5): 1741-1779.

Edmondson, A. C., R. M. Bohmer and G. P. Pisano. 2001. Disrupted Routines: Team Learning and New Technology Implementation in Hospitals, *Administrative Science Quarterly*, 46(4): 685-716.

Epstein, R. M. 2007. Assessment in Medical Education. *The New England Journal of Medicine*. 356: 387-396.

Fine, C. 1986. Quality Improvement and Learning in Productive Systems. *Management Science*. 32(10): 1301-1315.

Finks, Jonathan F., Osborne, Nicholas H. and Birkmeyer, John D. 2011. Trends in Hospital Volume and Operative Mortality for High-Risk Surgery. The New England Journal of Medicine. 364: 2128-2137.

Gallagher, A. G and C. U. Cates. 2004. Virtual Reality Training for the Operating Room and Cardiac Catheterisation Laboratory. *The Lancet.* 364(9444): 1538-1540.

The impact of New Product Introduction on Plant

Gopal, A. M. Goyal, S. Netessine and M. Reindorp, 2011. Productivity in the North American Automotive Industry. Working Paper.

Gorman, Paul J., Meier, Andreas H., Rawn, Chantal and Krummel, Thomas M.. 2000. The Future of Medical Education Is No Longer Blood and Guts, It Is Bits and Bytes. *The American Journal of Surgery.* 180(5): 353-356.

Hatch, N.W. and D. C. Mowery. 1998. Process Innovation and Learning by Doing in Semiconductor Manufacturing. *Management Science.* 44(11): 1461-77.

Huckman, R.S. and Gary P. Pisano. 2006. The Firm Specificity of Individual Performance: Evidence from Cardiac Surgery. *Management Science.* 52(4): 473-488.

Huckman, R. S., B. R. Staats and D. M. Upton. 2009. Team Familiarity, Role Experience, and Performance: Evidence from Indian Software Services. *Management Science.* 55(1): 85-100

Huckman, R. S. and B. R. Staats. 2011. Fluid Tasks and Fluid Teams: The Impact of Diversity in Experience and Team Familiarity on Team Performance. *Manufacturing & Service Operations Management.* 13(3): 310-328.

Kelsey, S. F., S. M. Mullin, K. M. Detre, H. Mitchell, M. J. Cowley, A. R. Gruentzig and K. M. Kent. 1984. Effect of Investigator Experience on Percutaneous Transluminal Coronary Angioplasty. *The American Journal of Cardiology* 53(12): 56–64.

Ketcham, Jonathan D. and Furukawa, Michael F. 2008. Hospital-Physician Gainsharing In Cardiology, *Health Affairs.* 27(3): 803-812.

Lieberman, M. 1984. The Learning Curve and Pricing in the Chemical Processing Industries. *The RAND Journal of Economics.* 15(2): 213-228.

Meier, A. H, Rawn, C. L and Krummel, T. M. 2001. Virtual Reality: Surgical Application–Challenge for the New Millennium. *Journal of the American College of Surgeons.* 192(3): 372-384

Mennicken, R., L. Kuntz and S. Scholtes. 2011. Stress on the Ward - An Empirical Study of the Nonlinear Relationship between Organizational Workload and Service Quality. Working Paper. Ruhr Economic Papers.

Noble, P.C., N. Sugano, J.D. Johnston, M. T. Thompson, M.A. Conditt, C.A. Engh. and K.B. Mathis. 2003. Computer Simulation: How Can it Help the Surgeon Optimize Implant Position? *Clinical Orthopaedics & Related Research.* 417: 242-252.

Olivares, M., C. Terwiesch, L. Cassorla. 2008. Structural Estimation of the Newsvendor Model: An Application to Reserving Operating Room Time. *Management Science.* 54, 1, 41–55.

Peersman G, Laskin R, Davis J, Peterson MG and Richart T. 2006. Prolonged Operative Time Correlates with Increased Infection Rate after Total Knee Arthroplasty. *HSS Journal.* 2(2): 70-72.

Pisano, G. P., R. M. Bohmer and A. C. Edmondson. 2001. Organizational Differences in Rates of Learning: Evidence from the Adoption of Minimally Invasive Cardiac Surgery. *Management Science.* 47(6): 752-768.

Ramdas, Kamalini. 2003. Managing Product Variety: An Integrative Review and Research Directions. *Production and Operations Management.* 12(1): 79-101.

Ramdas, Kamalini and Taylor Randall. 2008. Does Component Sharing Help or Hurt Reliability? An Empirical Study in the Automotive Industry. *Management Science.* 54(5): 922-938.

Ramdas, Kamalini, E.O. Teisberg, and A. Tucker. 2012. Redefining Service Delivery. *Harvard Business Review. forthcoming.*

Reagans, R., L. Argote and D. Brooks. 2005. Individual Experience and Experience Working Together: Predicting Learning Rates from Knowing Who Knows What and Knowing How to Work Together. *Management Science*. 51(6): 869-881.

Reznick, Richard K. and MacRae, Helen. 2006. Teaching Surgical Skills–Changes in the Wind. *The New England Journal of Medicine.* 355: 2664-2669.

Robertson, D. and K.T. Ulrich. 1998. Planning for Product Platforms. *Sloan Management Review*. 39: 19-31.

Saleh, K. J., W. M. Novicoff, D. Rion, L. H. MacCracken, and R. Siegrist. 2009. Operating-Room Throughput: Strategies for Improvement. *Journal of Bone and Joint Surgery*. 91: 2028-39.

Seymour, N. E., A. G. Gallagher, S. A. Roman, M. K. O'Brien, V. K. Bansal, D. K. Andersen, R. M. Satava. 2002. Virtual Reality Training Improves Operating Room Performance. *Annals of Surgery*. 236(4): 458–464.

Shwartz, M., J. Ren, E.A. Peko¨z, X. Wang, A.B. Cohen and J.D. Restuccia. 2008. Estimating a Composite Measure of Hospital Quality From the Hospital Compare Database: Differences When Using a Bayesian Hierarchical Latent Variable Model Versus Denominator-Based Weights. *Medical Care*. 46: 778–785.

Shervin, N, H. E. Rubash and J. N. Katz. 2007. Orthopaedic Procedure Volume and Patient Outcomes: A Systematic Literature Review. *Clinical Orthopedics and Related Research*. 457: 35–41.

Stern, S., E. Merwin, E. Hauenstein, I. Hinton, V. Rovnyak, M. Wilson, I. Williams and I. Mahone. 2010. The Effects of Rurality on Mental and Physical Health. *Health Services and Outcomes Research Methodology*. 10(1): 33-66.

Thornton, Rebecca Achee and Thompson, Peter. 2001. Association Learning from Experience and Learning from Others: An Exploration of Learning and Spillovers in Wartime Shipbuilding. *The American Economic Review*. 91(5): 1350-1368.

Tucker, A.L., I. M. Nembhard and A. C. Edmondson. 2007. Implementing New Practices: An Empirical Study of Organizational Learning in Hospital Intensive Care Units. *Management Science*. 53(6): 894-907.

Terwiesch, C. and R. E. Bohn. 2001. Learning and Process Improvement During Production Ramp-up. Internat. J. Production Econom. 70(1): 1–19.

Wright, T. 1936. Factors Affecting the Cost of Airplanes *Journal of Aeronautical Science* 3(4): 122-128.

Yasunaga, Hideo, Tsuchiya, Kazuaki, Matsuyama, Yutaka and Ohe, Kazuhiko. 2009. High-volume Surgeons in Regard to Reductions in Operating Time, Blood Loss and Postoperative Complications for Total Hip Arthroplasty. *Journal of Orthopaedic Science.* 14(1): 3-9.

Figure 1: Four Key Components Used in Hip Replacement Surgery

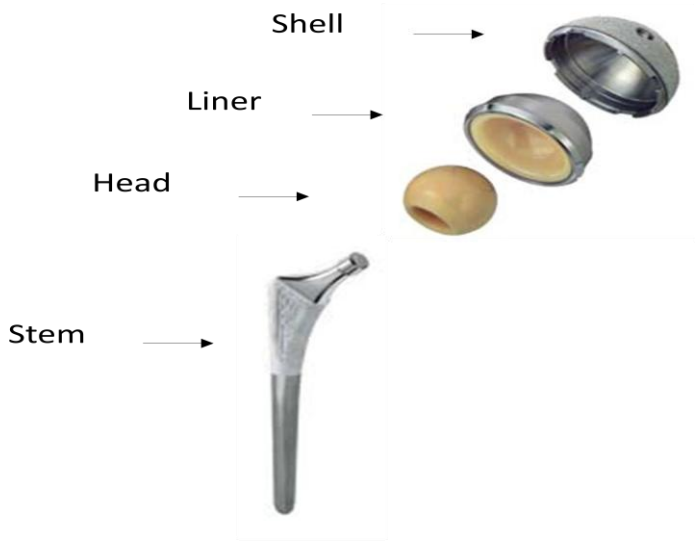

Figure 2: Two Distinct Stem Component Versions

Figure 3: Histogram of Duration of Surgery

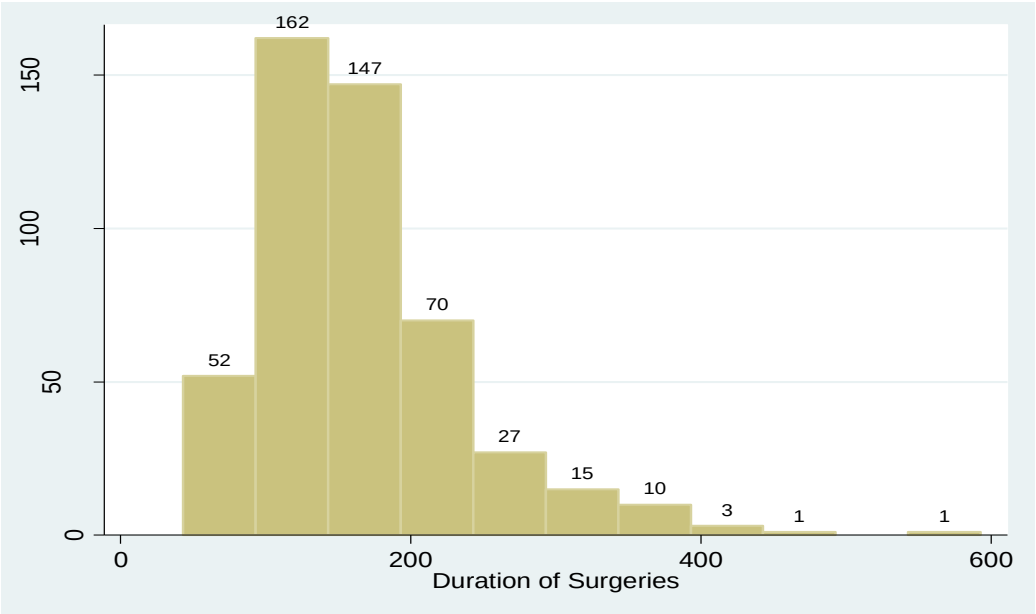

Figure 4: Serial Correlation Function $\rho(d)$

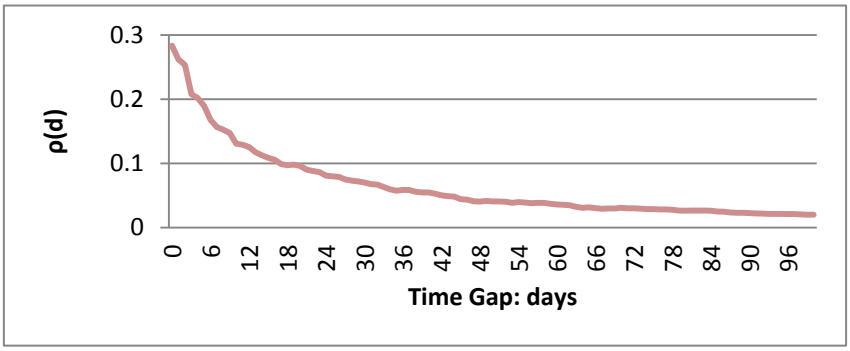

Table 1: Groupings of the Four Main Component Types

	Stem		Shell		Head		Liner	
Company	# of SKUs	# of Component Variants	# of SKUs	# of Component Variant	# of SKUs	# of Component Variants	# of SKUs	# of Component Variants
Zimmer	160	43	107	12	60	6	139	18
Depuy	100	17	75	16	93	14	95	16
Stryker	58	15	48	8	74	19	51	17
Smith	49	5	32	7	48	9	18	4
Total	369	80	262	43	275	48	303	55

Table 2: Descriptive Statistics

Variable	# of OBS	Mean	Std. Dev.	Min	Max
Duration	488	164.760	70.326	47	557
Exp	488	140.553	104.947	0	364
Exp_stem	488	43.381	58.463	0	199
Exp_join	488	70.500	75.989	0	251
Exp_shell	488	34.324	45.915	0	168
Exp_head	488	32.516	46.458	0	170
Exp_liner	488	20.619	30.187	0	124
Stem_before	488	0.756	0.430	0	1
Join _before	488	0.809	0.393	0	1
Shell_before	488	0.805	0.396	0	1
Head_before	488	0.848	0.359	0	1
Liner_before	488	0.648	0.478	0	1
Age	488	60.332	13.598	22	91
Male	488	0.492	0.500	0	1
Bilateral	488	0.008	0.090	0	1
BMI	488	29.887	6.997	15.438	62.125
Revision	488	0.242	0.429	0	1
ASA_average	488	2.428	0.517	1	4
AVN	488	0.113	0.317	0	1
Dysplasia	488	0.043	0.203	0	1
Arthritis	488	0.578	0.494	0	1
Severe arthritis	488	0.039	0.194	0	1
End stage arthritis	488	0.053	0.225	0	1
Fracture	488	0.072	0.258	0	1
Others reasons	488	0.049	0.216	0	1
Reasons of revision	488	0.166	0.399	0	1
Cemented	488	0.158	0.365	0	1
# of confounds	488	1.982	1.450	0	7

Table 3: Pooled OLS Regressions

Predictors	(1)	(2)	(3)	(4)
Exp	-0.19***	-0.06	-0.07	-0.05
	(0.04)	(0.06)	(0.06)	(0.07)
Expstem		-0.15*	-0.08	-0.06
		(0.08)	(0.08)	(0.09)
Exp_join				-0.08
				(0.10)
Expshell		-0.19*	-0.19*	-0.16
		(0.11)	(0.11)	(0.11)
Exphead		-0.05	-0.02	-0.02
		(0.09)	(0.09)	(0.09)
Expliner		0.04	0.07	0.06
		(0.11)	(0.11)	(0.11)
Stem_before			-35.99***	-42.12***
			(9.50)	(12.45)
Join_before				13.26
				(17.20)
Shell_before			-0.22	-0.83
			(10.79)	(10.84)
Head_before			-4.08	-4.24
			(8.33)	(8.34)
Liner_before			-18.10*	-17.42*
			(9.96)	(10.00)
Age	-0.77***	-0.79***	-0.76***	-0.77***
	(0.23)	(0.23)	(0.23)	(0.23)
Male	19.09***	21.22***	19.89***	20.25***
	(5.36)	(5.47)	(5.40)	(5.43)
Bilateral	124.28***	122.70***	109.97***	108.05***
	(30.896)	(30.72)	(31.11)	(31.34)
Revision	69.42***	64.59***	55.34***	55.69***
	(13.166)	(13.06)	(13.22)	(13.25)
# of Confounds	7.65***	7.49***	6.84***	7.01***
	(2.25)	(2.20)	(2.21)	(2.21)
Time Trend	0.14***	0.13***	0.13***	0.13***
	(0.02)	(0.02)	(0.02)	(0.02)
Surgeon dummy	NO	NO	NO	NO
# of Observations	488	488	488	488
Adj. R-squared	0.36	0.38	0.40	0.39

Notes:
1. Dependent variable: duration of surgery
2. Standard errors in parentheses ; *** p<0.01, ** p<0.05, * p<0.1
3. All regressions include other control variables described in the paper.

Table 4: Robustness Checks

Predictors	Fixed Effects (1)	Fixed Effects (2)	Fixed Effects (3)	Fixed Effects (4)	Instrument Variables (5)	Polynomial Regression (6)	Spline Regression (7)
Exp	-0.06	0.11	0.07	0.10	0.27	0.05	0.05
	(0.09)	(0.10)	(0.10)	(0.11)	(0.21)	(0.11)	(0.11)
Expstem		-0.15*	-0.08	-0.06	-0.05	-0.90**	-0.43**
		(0.08)	(0.08)	(0.09)	(0.09)	(0.41)	(0.23)
Eexpstem^2						0.012**	
						(0.01)	
Eexpstem^3						-0.00004**	
						(0.00)	
Expstem(51-150)							0.65**
							(0.35)
Expstem(150-)							-0.56
							(0.49)
Exp_join				-0.07	-0.31*	-0.08	-0.08
				(0.10)	(0.18)	(0.10)	(0.10)
Expshell		-0.21*	-0.20*	-0.18	-0.17	-0.20*	-0.21*
		(0.11)	(0.11)	(0.11)	(0.11)	(0.11)	(0.11)
Exphead		-0.05	-0.02	-0.02	-0.13	-0.02	-0.02
		(0.09)	(0.09)	(0.09)	(0.12)	(0.09)	(0.09)
Expliner		0.05	0.08	0.07	-0.08	0.11	0.11
		(0.11)	(0.11)	(0.11)	(0.15)	(0.12)	(0.12)
Stem_before			-35.74***	-42.26***	-42.03***	-37.03***	-38.58***
			(9.55)	(12.51)	(12.73)	(12.76)	(12.67)
Join_before				14.19	20.03	11.75	12.42
				(17.32)	(18.14)	(17.32)	(17.32)
Shell_before			2.16	1.39	3.59	3.84	3.56
			(11.00)	(11.06)	(11.40)	(11.14)	(11.14)
Head_before			-2.26	-2.52	-7.19	-0.84	-1.24
			(8.45)	(8.46)	(8.74)	(8.46)	(8.47)
Liner_before			-17.88	-17.19*	-18.62*	-17.37*	-17.69*
			(9.96)	(10.00)	(10.26)	(9.97)	(9.99)
Age	-0.72***	-0.74***	-0.72***	-0.73***	-0.85***	-0.73***	-0.73***
	(0.23)	(0.23)	(0.23)	(0.23)	(0.24)	(0.23)	(0.23)
Male	18.86***	21.28***	20.00**	20.36***	23.11***	21.01***	21.23***
	(5.36)	(5.49)	(5.44)	(5.46)	(5.85)	(5.46)	(5.47)
Bilateral	127.15***	125.13***	114.18***	113.50***	99.87***	118.66***	118.81***
	(30.95)	(30.73)	(31.27)	(31.47)	(32.59)	(31.79)	(31.79)
Revision	71.19***	67.69***	58.20***	58.55***	55.89***	55.24***	55.73***
	(13.36)	(13.25)	(13.47)	(13.50)	(13.55)	(13.58)	(13.57)
# of Confounds	7.05***	7.05***	6.40***	6.56***	7.52***	6.67***	6.46***
	(2.28)	(2.23)	(2.23)	(2.24)	(2.29)	(2.24)	(2.24)
Time Trend	0.10***	0.08**	0.08**	0.08**	0.07*	0.09***	0.09***
	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)	(0.03)	(0.03)
Surgeon dummy	YES	YES	YES	YES	NO	YES	YES
# of Observations	488	488	488	488	488	488	488
Adj. R-squared	0.36	0.38	0.40	0.40	0.37	0.40	0.40

Notes:

1. Dependent variable: duration of surgery.
2. Standard errors in parentheses ; *** p<0.01, ** p<0.05, * p<0.1.
3. All regressions include other control variables described in the paper.

Online Appendix

Surgeon-Specific Effects Test

Let

$$\begin{aligned} y_{it} &= X_{it}\beta + u_i + \varepsilon_{it},\ t = 1, 2, .., T_i, \\ u_i &\sim iidN\left(0, \sigma_u^2\right), \varepsilon_{it} \sim iidN\left(0, \sigma_\varepsilon^2\right). \end{aligned}$$

Test

$$H_0 : \sigma_u^2 = 0 \text{ vs } H_A : \sigma_u^2 > 0.$$

LM test is

$$LM = \left(\frac{1}{2N}\sum_i T_i\right)\left\{\sum_i \frac{1}{T_i - 1}\left[\frac{\sum_{i=1}^N T_i\left(T_i^{-1}\sum_{t=1}^{T_i} e_{it}\right)^2}{\sum_{i=1}^N T_i^{-1}\sum_{t=1}^{T_i}\left(e_{it}\right)^2} - 1\right]^2\right\} \sim \chi_1^2$$

$$e_{it} = u_i + \varepsilon_{it}$$

$$\begin{aligned} E\sum_{i=1}^N\left(T_i^{-1}\sum_{t=1}^{T_i} e_{it}\right)^2 &= \sum_{i=1}^N\left(\sigma_u^2 + T_i^{-1}\sigma_\varepsilon^2\right) \\ E\sum_{i=1}^N T_i^{-1}\sum_{t=1}^{T_i}\left(e_{it}\right)^2 &= \sum_{i=1}^N\left(\sigma_u^2 + \sigma_\varepsilon^2\right) \end{aligned}$$

$$\left[\frac{\sum_{i=1}^N T_i\left(T_i^{-1}\sum_{t=1}^{T_i} e_{it}\right)^2}{\sum_{i=1}^N T_i^{-1}\sum_{t=1}^{T_i}\left(e_{it}\right)^2} - 1\right] = \left[\frac{\sum_{i=1}^N\left(T_i\sigma_u^2 + \sigma_\varepsilon^2\right)}{\sum_{i=1}^N\left(\sigma_u^2 + \sigma_\varepsilon^2\right)} - 1\right] = 0$$

Serial Correlation

Let

$$\begin{aligned} y_{it} &= X_{it}\beta + \upsilon_{it},\ t = 1, 2, .., T_i; \qquad (1) \\ \upsilon_{it} &= \rho\left(d_{t,t-1}\right)\upsilon_{it-1} + \eta_{it}; \eta_{it} \sim iid\left(0, \sigma_\eta^2\right). \end{aligned}$$

Define $\widetilde{\upsilon}_{it}$ as the residual from pooled OLS, $\widetilde{\upsilon}_{i\cdot}$ as the sample mean of $\widetilde{\upsilon}_{it}$ for each individual, and $\widehat{\widetilde{\upsilon}}_{it}$ as the standardized residual,

$$\widehat{\widetilde{\upsilon}}_{it} = \frac{\widetilde{\upsilon}_{it} - \widetilde{\upsilon}_{i\cdot}}{\sigma_{\widetilde{\upsilon}}}. \qquad (2)$$

where $\sigma_{\widetilde{\upsilon}}$ is the standard deviation of $\widetilde{\upsilon}_{i\cdot}$. We need to define a correlation function $\rho\left(d_{ts}\right)$ for two surgeries, s and t, from the same surgeon as a function of the time gap d_{ts} (measured in days) between them. A kernel-based estimate of the correlation function is

$$\widehat{\rho}(d) = \frac{\sum_i \sum_{t,s} K(d_{t,s} - d)\, \widehat{\widehat{v}}_{it} \widehat{\widehat{v}}_{is}}{\sum_i \sum_{t,s} K(d_{t,s} - d)} \tag{3}$$

where $K(\cdot)$ is a kernel function and b is its corresponding bandwidth. We use

$$K(z) = \begin{cases} \frac{b^{-1}}{\sqrt{2\pi}} \exp\left\{-.5\frac{z^2}{b^2}\right\} & \text{if } |z| \leq 4 \\ 0 & \text{if } |z| > 4 \end{cases} \tag{4}$$

and set $b = \sigma_d$. Note that the model described in equation (1) assumes a balanced panel, but the estimator for $\rho(d)$ in equation (3) does not require a balanced panel.