Forthcoming in Management Science

Variety and Experience: Learning and Forgetting in the Use of Surgical Devices

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

August 2016

Abstract

We use a unique, hand-collected dataset to examine learning and forgetting in hip replacement surgery as a function of a surgeon's experience with specific surgical device versions and the time between their repeat uses. We also develop a generalizable method to correct for the left-censoring of device-version-specific experience variables that is a common problem in highly granular experience data, using Maximum Simulated Likelihood Estimation (MSLE) with simulation over unobservables conditional on observables. Even for experienced surgeons, the first usage of certain device versions can result in at least a 32.4% increase in surgery duration, hurting quality and productivity. Furthermore, with the passage of time, surgeons can forget knowledge gained about the use of particular devices. For certain devices, when the time gap between repeat uses increases from its median to its 75^{th} percentile, surgery duration increases by about 3.4%. The high productivity and quality costs associated with device variety suggest that the gain from a new device design needs to be large enough to compensate for the short term disadvantages of starting up on a new learning curve and of increasing the chances of knowledge depreciation over time.

Keywords: Product Variety**,** Learning and Forgetting, Experience Curves, Productivity, Health Care

1 Introduction

Employees in industries as diverse as IT support, auto servicing and repair, healthcare, home loans and mortgages, investment banking, and retail customer service work with a great variety of products and services on a daily basis. A car mechanic services a wide variety of makes and models and a variety of model years and trim levels even within a particular make and

*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@virginia.edu.
§Department of Economics, University of South Florida, hliu4@usf.edu.

model. An IKEA kitchen designer helps customers design kitchens that come in a wide variety of design suites, with a number of different cabinets, drawer systems, counters, and appliances within each suite. A server at an IT help desk may receive calls related to a wide variety of problems on a variety of different computer makes and models. Client work in law, consulting, architecture, and other professional services often entails combining elements from a variety of different knowledge bases, frameworks, or methodologies. Most surgeons perform several different kinds of surgery using a variety of medical devices that each come in many variants. In such environments, it is quite common for employees to encounter product or service variants with which they have nil or very limited prior experience, even if they have been working in the same job for many years.

A long stream of research has documented the benefits and costs of product variety (Macduffie et al. 1996, Ho and Tang 1998, Ramdas 2003). On the cost side, recent research in operations management and economics suggests that high product variety can slow down production or worsen quality due to limited learning spillover from one type of activity to the next (Benkard 2000, Boh et al. 2007, Ramdas and Randall 2008, KC and Staats 2012, Clark et al. 2012, Staats and Gino 2012). Another potential reason for production slowdown or quality degradation in high variety settings is that it often can be a while before a worker performs any particular type of activity again, despite operating at high volume. This could lead to forgetting the intricacies of specific tasks. Naturally, the costs of variety should be weighed against its benefits in deciding how much variety to offer.

Our focus is on the learning and forgetting-related costs of variety. Workers in high-variety settings often encounter new tasks. While it is widely known that learning occurs steeply at first and then flattens out with experience (e.g., Wright 1936, Lieberman 1984, Argote and Epple 1990, Argote 1999), in empirical estimation, emphasis is almost never placed on the first few exposures to a new task.[1] Also, the estimation of forgetting effects has received far less attention than the estimation of learning rates (Bailey 1989, Argote et al. 1990). In particular, to our knowledge, no one has examined how forgetting with the passage of time at the granular level of specific tasks – such as a car mechanic servicing a car of a specific make-model-year and trim level or a surgeon using a specific surgical device variant – impacts production. The impact of having had any prior exposure to a specific task and that of forgetting in relation to specific tasks is particularly relevant in high variety settings. It is the estimation of these effects – at the level of specific tasks – that we focus on, in the context of medical devices used

[1] Two exceptions are Narayanan et al. (2009) who estimate the impact of software programmers' first experience with a new software module, and Pisano, Bohmer and Edmondson (2001) who trace surgical teams' experience from the very first use of a new technique in cardiac bypass surgery.

in surgery.

The past few decades have seen an explosion in the variety in medical devices (Gelberman et al. 2010). In 2004, the US FDA[2] was regulating 500,000 models of 1700 different medical devices (Maisel 2004). In this environment, a surgeon's ease in using a device version that he has never previously used has important implications for productivity and quality. Further, high device variety increases the time gap between repeat uses of any particular device version by a surgeon. This can result in forgetting over time of device-version-specific knowledge. The impact of forgetting over time at the level of specific tasks has not been examined previously.

Our goal is to examine how device proliferation impacts production, in the context of surgery. A longer duration of surgery reduces productivity, eating into expensive Operating Room (OR) and surgeon capacity (Olivares et al. 2008, Saleh et al. 2009) and using up time in which additional surgeries could have been performed. All else equal, shorter duration is also preferable as the risks of infection, blood loss, and post-surgical complexities are well-known to increase with surgery duration (Peersman et al. 2006, Yasunaga 2009).

Investigating how device-specific learning and forgetting impacts the productivity of surgeons poses three major empirical challenges. First, doing this requires very detailed data about device usage at the individual surgeon level, as opposed to data on hospital or even surgeon volumes. We use a unique, hand-collected dataset from the University of Virginia Hospital to examine learning and forgetting at the level of specific surgical devices used in hip replacement surgery. We assembled data from multiple sources including OR records, patient charts, and hospital accounting databases. Our very detailed dataset enables us to consider a richer set of hypotheses than previously has been possible. Our dataset includes all hip replacement surgeries performed at the University of Virginia Hospital from August 2006 until November 2008. We obtained information on the specific version of each of four key devices used in surgeries performed during this period – the "stem" or femoral device, which is inserted into the patient's thigh bone; the "shell" or acetabular device, which is inserted into the patient's hip socket; and the "head" and "liner" devices, which together comprise the ball and socket joint (see Figure 1). There are many versions of each of these four devices which differ in shape, material, coatings, and other characteristics that are likely to affect a surgeon's ease in using them (see Figure 2 for two distinct stem versions). We limit our analysis to devices made by four vendors – Stryker, Depuy, Smith & Nephew, and Zimmer – that account for around 90% of the surgeries performed in our study period. When counting device versions made by only these four vendors, a total of 563 SKUs (or 121 unique device versions after accounting for devices

[2] Food & Drug Administration

that differ only in size) of these four key devices were used by just 4 surgeons in performing 671 hip replacement surgeries during our study period, indicating high variety in device versions (see Table 1).

Our dataset includes both first-time hip replacement surgeries as well as revision surgeries, which tend to take longer. Table 2 provides a simple description of our data, broken down into these two categories.[3] Each of these categories is further divided into three subcategories. The first contains surgeries in which the surgeon had prior experience in our sample with the specific versions of all four key devices used in the surgery. The second contains surgeries in which the surgeon used one device version (among the four key devices) that he had not used before in our sample, and the third contains the remaining surgeries in the category. A cursory glance at this table suggests that longer average duration of surgery is associated with a surgeon using a greater number of device versions that he has not used before, both for revision surgeries and first-time surgeries. We will examine whether this pattern withstands a rigorous analysis.

The second major empirical challenge we face is related to data censoring. Since our data have a fixed starting time with no surgeries observed prior to it, we cannot be certain that an observed first usage of a specific device version by a surgeon is indeed the true first usage. In fact, even if we had data from when a surgeon joined the hospital, it would still be impossible to know if an observed first usage of a specific device by a surgeon is a true first usage (as surgeons typically get their initial training and experience at one hospital and then move elsewhere to practise). In reality, given the tremendous and ever-changing variety in devices, the fragmented way in which device data are recorded, and the lack of attention that has been paid to this type of data in healthcare research and management to this date, it is very difficult to obtain current information on device usage at most hospitals, and even more so going back in time.[4] This type of left-censoring of device usage data, while widely prevalent in hospital data, introduces a challenging econometric problem. A similar problem would arise with the use of highly granular experience data in other contexts where the goal is to examine learning and forgetting at the level of subtasks or subprocesses of a service procedure. We present a generalizable approach to address this type of problem. In essence, we use observed information on the distribution of time between usages of specific devices to infer the probability that an observed first usage of a device by a surgeon is indeed a true first usage and incorporate this information into a maximum likelihood estimation procedure to estimate our coefficients of interest.

[3] This table is based on a sample of 483 surgeries for which we have complete information on all variables used in our study. Details on how we assembled this dataset are presented in Section 3.

[4] An alternative is to use only surgeries where all devices used were introduced to the market after the starting time of the sample period. This approach can reduce sample size and result in non-randomly missing data. Also, it is impractical because hospitals do not store data on when specific device versions were introduced.

The third major empirical challenge we face is that our learning and forgetting-related variables may be endogenous, for several reasons. First, unobserved factors may impact both a surgery's duration and the surgeon's choice of devices to use in it. For example, it is possible that certain "difficult-to-use" device versions are used only occasionally, when there is an unusual patient need. Even in the absence of learning, this might cause first use of a new device version and the time gap between uses to appear to impact surgery duration. Second, patients' choice of surgeons may also result in endogeneity. More able surgeons may be chosen more frequently and may also operate faster. Third, unobserved factors in the OR could result in omitted variables bias. We account for these sources of endogeneity in our empirical analysis.

We find that a single prior usage of a stem (shell) version reduces surgery duration by about 32.4% (27.6%) relative to the average duration of stem (shell) versions that have been used at least once. The time spent on first usages of stems and shells represents a potential 5% increase in the number of hip surgeries that can be performed annually with no increase in hospital and physician costs. For stems, which are by far the most tricky device to implant, we find some evidence of learning even on second, third, and fourth usage. Accounting for this would further increase available OR capacity.

Forgetting is also costly. By our estimates, the reduction in forgetting obtainable by halving the variety of stems and liners would result in a 1% increase in hip surgeries annually, under conservative assumptions. Also, keeping the time gap between repeat uses of a device constant, forgetting increases with the number of surgeries between repeat uses. With the number of joint replacements rising steeply year on year, increases in capacity that could be obtained by reducing device variety would be well-utilized.

One might well ask why there is so much variety in medical devices. One reason is that different devices are suitable for different patients. The incentives for variety in the medical devices industry also provide some insight into why variety is high, while they are not our focus. The current regulatory environment in many countries allows medical device vendors to sell devices that have no proven health benefit (Meier 2011). In the US, if a manufacturer can show that a new device is "substantially equivalent" to a legally marketed existing "predicate" device, it can bypass clinical trials and go through a relatively straightforward 510(K) FDA approval process that often takes less than three months. Only one percent of devices listed with the FDA in recent years have required the more stringent pre-market approval (PMA) process which requires clinical trials (GAO Report, 1999). A substantially equivalent device needs to be only at least as safe and effective as the predicate, and to have the same intended use. In this regulatory environment, manufacturers are well-known to "tweak old models and patent the

changes as new products" (Rosenthal 2013). At the same time, a number of devices approved through the 510(K) process have been recalled due to life-threatening adverse consequences for patients (Garber 2010). Curfman and Redberg (2011) caution against "putting defective medical devices onto the market where they cause harm to patients, waste health care dollars, and may kill jobs when they are withdrawn." With little evidence of any long-term benefits from device proliferation, accurate estimates of the short-term costs of such proliferation are crucially needed to inform policy targeted at improving patient outcomes and lowering healthcare provider costs. Our goal is to shed some light on these short term costs.

2 Hypotheses on Learning and Forgetting

The relationship between production volumes and both unit cost and quality is well documented (Wright 1936, Lieberman 1984, Argote and Epple 1990, Mukherjee et al. 1998, Argote 1999). In healthcare, researchers have documented the impact of medical procedure volume on outcomes in many settings including several types of surgery (e.g., Birkmeyer et al. 2003, Reagans et al. 2005, Huckman and Pisano 2006, Shwartz et al. 2008, KC, and Staats 2012 and Clark et al. 2012). Several studies also have examined the impact of hospital volume and surgeon volume on outcomes in hip replacement surgery specifically. In a review of research on learning in hip replacement surgery, Shervin et al. (2007) find that higher hospital volumes and surgeon volumes are associated with improved outcomes. They call for further research to identify the causal factors – such as new surgical technology – underlying these volume-outcome relationships. Our dataset enables us to examine both the classical volume-outcome relationship at the level of individual surgeon experience as well as learning and forgetting with respect to a critical dimension of surgical technology – the key devices used in surgery.

Researchers also have started to examine how product (or service) variety affects production. For example, Benkard (2000) finds that productivity suffers when switching from production of one commercial aircraft model to another. Ramdas and Randall (2008) find limited learning spillovers for carmakers who use the same brake components on different car models. Limited learning spillovers across tasks have also been documented for programmers who perform maintenance tasks on different software modules (Narayanan et al. 2009), bank employees who work on different stages in home loan application processing (Staats and Gino 2012), surgeons who perform different procedures for minimally invasive heart surgery (KC and Staats 2012) and remote radiologists who read scans for different body organs or from different hospitals Clark et al. (2012). At an organizational level, Clark and Huckman (2012) find that in multi-specialty hospitals, organizational focus improves outcomes. We build on this literature by estimating

the impact of experience at the level of specific device versions on surgeon productivity.

Naturally, one would expect the productivity or quality losses associated with product variety to be small if product variants are very similar to one another. One might expect the differences among versions of the same device, within the same type of surgical procedure, performed at a single hospital to be smaller than the differences among product variants examined in previous studies. Furthermore, in the case of orthopedic devices, it is widely acknowledged that most new devices are very minor variants of existing devices (Demske 2008), unlike the case of two different aircraft models, brake designs, software modules, loan applications, cardiac procedures, or body organs. The basic design of orthopedic devices has remained relatively stable for a few decades (Bauer 1992, Gelberman et al. 2010, Salemi 2011) with the vast majority of new devices having been deemed similar enough to existing devices to not require any clinical evaluation. In this environment, where new devices are often very similar to existing ones, we will examine whether device variety hurts surgeons' productivity.

Traditionally, surgery has been taught using the apprenticeship model best exemplified by the phrase "see one, do one, teach one" (Gorman et al. 2000). Naturally, in a high-variety environment, the likelihood of having seen or used a specific device variant before can be quite low. We therefore estimate the impact on surgery duration of the first few previous exposures to a specific device version, an econometrically challenging task.

In contrast to the vast literature on learning, little emphasis has been placed on knowledge depreciation (Argote and Epple 1990, Argote 1999). At the individual level, forgetting is a key cause of knowledge depreciation (Benkard, 2000). A critical determinant of the extent of forgetting is the amount of time between learning a task – such as how to use a specific device version – and recall of that learning the next time it is needed (Wixted, 2004). When a limited number of tasks are being performed, the time gaps between repeated performance of any one task are likely to be smaller, and therefore the role of forgetting may be less important. On the other hand, when there is a large variety of tasks, as in our context, forgetting is more likely to come into play as any one task is performed less frequently.

Much of the literature on individual knowledge depreciation comes from the field of psychology and consists of theory and laboratory experiments (Bailey 1989), with little estimation of individual knowledge depreciation rates outside of the laboratory. Shafer et al. (2001) use a simulation model to examine learning and forgetting on an assembly line. A few studies estimate forgetting at the level of overall production at a manufacturing or service facility (e.g., Argote, Beckman and Epple 1990, Thompson 2007, Boone et al. 2008, Agrawal and Muthulingam 2015). Similarly, Mincer and Ofek (1982), and Anderson et al. (2002) estimate depreciation in

general human capital. Keane and Wolpin (1997) estimate depreciation in occupation-specific human capital as a function of occupation changes. Others have examined forgetting in the context of specific models or tasks within a facility. Nembhard and Osothsilp (2001) compare a variety of forgetting models using data from an assembly line that produces car radio models. Nembhard and Uzumeri (2000) compare forgetting rates for a manual task and a procedural task. Nembhard and Osothsilp (2002) examine how task complexity impacts the variance of forgetting rates across workers. Yamaguchi (2012) models human capital as a vector of task complexity measures on cognitive, motor and other tasks, with depreciation of skills from one year to the next. In contrast to these studies, our dataset tracks the usage by individual surgeons of specific device versions over time, including the time between repeat usages of each version and measures of the intensity and variety of the other tasks performed in between. We are thus able to estimate knowledge depreciation at the level of individual subtasks as a function of time between repeat instances of a specific subtask and the type of work performed in between. This approach provides a natural way to think about knowledge depreciation that is also supported by research in psychology (Bailey 1989). We are able to measure the effect of performing a greater number of distinct tasks and the effect of performing a greater volume of tasks between two occurrences of the same task.

An underlying issue in knowledge accumulation and depreciation is the transferability of what one has learned. For example, when one learns how to differentiate a polynomial, this learning transfers over if the next polynomial to be differentiated has different coefficients or different variable names. However, differentiating a different class of functions may involve some separate learning and/or depreciation. Are different device versions like different polynomials or are they like different function types, or are they like integration? If there really is something to learn (or forget), this would suggest that they are more than just different polynomials. Our analysis sheds light on this underlying question in the context of orthopedic devices.

In orthopedics, patient characteristics (e.g., anatomy and bone quality) and device characteristics (e.g., the geometry of the device, its material, and its type of coating) are key determinants of the difficulty of a surgery. In hip replacement surgery, placing the stem into the thigh bone requires preparing, shaping, and opening up the bone canal using instruments specific to each stem. Slight differences in the shape of a stem can alter how the stem is inserted because differently shaped stems can get caught up in different parts of the bone cavity. Also, unexpected variation - e.g., certain stems sit a little higher when placed in the thigh bone canal than the stem height specified on the box - can result in rework. All told, the high variation in patient and device characteristics may increase uncertainty and necessitate significant learning

for stems while also increasing the chances of forgetting over time. Preparing the hip socket for shell insertion is relatively straightforward and similar across all shells. Therefore, one might expect there to be less learning and forgetting for shells relative to stems, even though the shell, like the stem, touches the patient's bone.

The liner and head do not touch bone, reducing complexity of insertion. Inserting the liner into the shell takes little time, although it does require delicate maneuvering. The head is easy to insert and requires no instrumentation. One might expect that there is little to learn and forget for heads. Aside from whether a device touches bone, the extent of variation across device versions can impact learning and forgetting. We do not intend to distinguish between such effects.

3 Data

We obtained data from the University of Virginia hospital for all hip replacement surgeries performed from August 2006 to November 2008. Data on all devices used in each surgery were obtained from a hospital database that is used for operational and accounting purposes. Despite there being many studies of learning in surgery, to our knowledge, no other study has examined learning at the level of devices. This may be due in part to the difficulty in accessing detailed data on device usage. We use our data to develop measures of surgeon experience at the level of specific device 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 was a painstaking process. We hired three nurses to perform this task. 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, it was necessary to locate and read through the surgical note for every patient in order to identify reasons for surgery and complexities during surgery. 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.

During our sample period, 752 hip replacement surgeries were performed by four surgeons at the University of Virginia hospital.[5] Column (1) of Table 3 shows how these 752 surgeries are distributed across the four surgeons. These data are used to define the variable that measures the total experience for each surgeon during the sample period at the time of each surgery. Column (2) of Table 3 contains frequency by surgeon of surgeries in which at least one of the

[5] Two other surgeons performed 11 surgeries in total. We exclude these due to the low volumes.

main devices used (head, stem, liner, and shell) were made by one of the four major vendors (Stryker, Depuy, Smith & Nephew, and Zimmer). This sample of 671 surgeries accounts for almost 90% of our sample. We limit our sample in this way as we needed to interact closely with a vendor representative from each vendor to correctly classify device versions in our dataset.[6] We use this sample to define our surgeon-specific device-version experience variables at the time of each surgery. Due to missing values on surgery duration and some control variables, our sample shrinks to 555.[7] In order to include only those surgeries for which all of the major devices were from one of the four major vendors, our sample is further reduced to 483 surgeries for which we have complete data. Column (3) contains frequency of surgeries by surgeon for the final sample that we use for our empirical analysis. We discuss below how we use our data to define the variables used in our estimation procedure. Columns (4) to (7) provide additional information about the education and professional background of all surgeons in our data sample. All of our surgeons are highly experienced. We discuss how this impacts our results in Section 6.

Outcome Variable

Our outcome variable is *duration of surgery*, defined as the number of minutes from the start of a surgery, i.e. skin opening, until the end of the surgery, i.e. skin closing. Our measure of 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. We use the natural log of duration, resulting in the widely used log-linear experience curve (Reagans et al. 2005).

Although duration of surgery is commonly used as a measure of both productivity and health outcomes quality in the healthcare and OM literatures, mortality rate or follow-up complications are also common outcome measures. Death from hip replacement surgery is very rare, therefore mortality rate is not appropriate for our setting. Follow-up complications are of interest, but, given how rare they are, we would need a much larger multi-hospital dataset to identify effects. For instance, need for revision is a common follow-up measure of surgery quality. However, a patient may go to a different hospital for revision surgery many years after the first time surgery.

[6] For example, often the same device variant was recorded under slightly different names, needing an expert to identify the underlying variant.

[7] Although we lose around 20% of our data sample due to missing values of our outcome variable and control variables, this is mainly because the UVA hospital did not systematically record all information related to surgeries and patients during the sample period. When we check variables for which we have complete information, we do not find any systematic differences between the sample we drop and the one we keep. Therefore, we believe that the data is missing completely at random and does not bias our results.

Experience Variables

We define a surgeon's *total experience* as the number of hip replacement surgeries that the surgeon has performed during the study period prior to a particular surgery considered. We calculate total experience for each surgeon using all 752 surgeries completed by the four surgeons in our sample.

Aside from gaining overall experience over time, each surgeon also accumulates experience over time with specific device versions. We learned from our discussions with orthopedic experts including our orthopedic surgeon coauthor that the shell, stem, liner, and head devices are the primary drivers of the time taken to complete a surgery. These devices also are quite expensive. The prices for devices in our dataset range from $624 to $7,400 for shells, $1,525 to $6,955 for stems, $998 to $4,050 for liners, and $356 to $5,100 for heads. We focus on the possible learning and forgetting of these four main devices.

The most granular level at which device experience can be accumulated is the device SKU. A total of 114 unique shell SKUs, 162 unique stem SKUs, 122 unique liner SKUs, and 165 unique head SKUs were used in our sample period by just four surgeons to perform 671 hip replacement surgeries which had at least one device from one of our four main vendors, as listed in Table 1. Within each of the four key devices – shells, stems, liners, and heads – device SKUs differ in technology, shape, materials, surface, coatings, and size.

For our purposes, we group together SKUs whose labels differ only in size into a single device version for each of the four devices and for each of the four vendors included in our study.[8] In some cases, SKUs that differ only in size have slightly different item descriptions due to inconsistent use of abbreviations by the staff who originally recorded the data. Therefore, we enlisted the help of the hospital's orthopedic device vendor representative for each vendor, to accomplish this grouping.[9] Quite a bit of mixing and matching is possible over the versions of the four devices, both within and across vendors. Thus, it would be inappropriate to think of the appropriate unit of analysis as a fixed combination of specific device versions.

Through the procedure described above, the large number of SKUs was reduced to a much smaller number of device versions, as summarized in Table 1. In the sample of 671 surgeries, there are 20 shell versions, 35 stem versions, 28 liner versions, and 38 head versions, ignoring size variations. From now on, we use the term "device version" to denote all SKUs that vary only in size.[10]

[8] For example, during the sample period, in the Zimmer line, the Trilogy Multi-Holed Shell contained 14 different size variations, ranging from 44mm to 70mm in diameter, while the Trilogy Uni-Holed Shell contained 11 different size variations, ranging from 46mm to 68mm in diameter.

[9] These representatives attend surgery and have extensive knowledge of the devices used.

[10] The observed variety in device versions is not limited to specific surgery types (first time vs. revision),

We created two types of surgeon experience variables at the level of specific device versions for each one of the four main devices by using the data from the 671 surgeries summarized in Column (2) of Table 2 as well as in Table 1.

Variables related to device-specific learning:

For each surgeon, surgery, and device, the *first use* dummy takes on the value of one if and only if the surgeon in question is using the specific device version used in the surgery for the *first* time during our study period.

Additional variables that we examine for each surgeon, surgery, and device are *nth use* dummies which takes on the value of one if and only if the surgeon in question is using the specific device version used in the surgery for the *nth (2nd, 3rd, or 4th)* time during our study period. Also, for each surgeon, surgery, and device, we measure device-specific experience as a count of how many times the surgeon has used the specific device version used in the surgery at hand, since the start of our sample. Using this variable in addition to the dummy variables for the first few usages allows us to check whether learning occurs very quickly.

Variables related to device-specific forgetting:

For each surgeon, surgery, and device, *experience gap* is defined as the amount of calendar time (in days) since the last use by the surgeon of the specific device version used in the surgery. We use log values of "experience gap" variables to reduce the effect of outliers.[11]

We consider three additional variables related to forgetting. For each surgeon, surgery, and device, *surgeries-between* is a count of the number of surgeries that the surgeon has performed since last use of the specific device version used in the surgery at hand. Also at the level of surgeon, surgery, and device, the *device-switch* dummy indicates whether the surgeon has used other device versions since his last use of the specific device version used in the surgery at hand, whereas *switch-variety* is a count of how many other device versions have been used in-between two consecutive uses of a specific device version.

Control Variables

We use a number of variables to control for the impact of patient, surgery, device, and surgeon characteristics on duration of surgery. Patient characteristics include age, gender, body mass index (BMI), anesthetic severity assessment (ASA), and patient comorbidities. *BMI* is a standard measure of obesity of patients and is calculated as the ratio of weight to squared height. BMI directly affects duration of surgery because a more obese patient can take longer to

surgeons, or patient severity levels. We control for these factors in our empirical specifications.

[11] Note that we use log(experience gap+1) to avoid taking logs of zero which occurs when a surgeon performs two or more surgeries on the same day. Our results are insensitive to the choice of constant.

operate. *ASA* 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.[12] The *number of comorbidities* is coded as the sum of ten indicator variables which indicate the presence of each of the ten most common patient comorbidites in hip surgery.[13]

We include several controls for the surgery itself. *Both Legs* is a dummy indicating whether the surgery is performed on one or both hips. Surgeries that involve both hips generally are expected to take longer. *Unihead* is a dummy for the use of a unipolar head device, used for fractures and associated with longer duration. Through discussions with our orthopedic surgeon coauthor and other orthopedic experts**,** we learned that we can aggregate stem device versions from all vendors into two groups based on the method used for joining the device 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 device. *Cemented* is a dummy for use of a cemented stem. Revision surgeries do not always use all of the four main devices. Therefore, we also include a dummy called *Use_ device* for each of the four main devices. For example, *Use_ stem* is a dummy for use of a stem device.

We also control for the reasons for surgery. We include indicator variables for each of the most frequently cited reasons for surgery.[14] The reasons-for-surgery dummies are non-exclusive. For example, a surgery can be conducted because of arthritis and fracture. In the case of revision surgeries, we include an additional variable, "reasons for revision," which is the sum of indicator variables for each of the following reasons for revision surgeries: acetabular osteolysis, aseptic loosening, infection, pain, dislocation, and hematoma. In addition, we include manufacturer fixed effects and surgeon fixed effects.

Finally, we include a linear and quadratic time trend[15] 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.[16] Table 4 reports the pairwise correlation coefficients

[12] 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.

[13] 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 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.

[14] The most frequently cited reasons include Revision, Avascular Necrosis (AVN), Displasia, Arthritis, Severe Arthritis, End-stage Arthritis, Fracture. The "Other Reasons" category includes very infrequently cited reasons such as deformity, childhood disease, and post-traumatic bone conditions.

[15] Time trend is defined as the number of days since start of the sample period divided by 1000.

[16] Due to space constraints, some controls included in the regression are not reported in this table. Complete

of our experience variables. We do not find high correlation between our different experience measures.[17] Table 5 provides descriptive statistics of our main variables.

4 Empirical Specification

4.1 Baseline Specification

We first model device-specific learning through the *1st usage* dummy for each of the four main device versions used in a surgery and device-specific forgetting through the *experience gap* since prior usage of each of the four main device versions[18] as

$$y_{st} = \beta X_{st} + \gamma e_{st} + \alpha w_{st1} + \theta \log [w_{st2}] + \varepsilon_{st}. \tag{1}$$

Here, y_{st} is the log value of duration of the surgery performed by surgeon s at time t, e_{st} is the total experience of surgeon s at time t, X_{st} is a vector of control variables including fixed effect dummies, w_{st1} and w_{st2} are vectors of (observed) device-specific experience variables related to learning and forgetting for surgeon s at time t, as explained below, and ε_{st} is the error term. Log-linear or "exponential" total experience curves are widely used in the literature to capture the diminishing returns from additional units of experience (Argote 1999, Thornton and Thompson 2001).[19]

Define $k_{st} = (k_{s1t}, k_{s2t}, k_{s3t}, k_{s4t})$ where k_{sjt} is an index for the specific version of device j used by surgeon s in his t^{th} surgery, where $j = 1, 2, 3, 4$ indexes the four main devices – shell, stem, liner, and head. Next, we define $w_{st1} = (w_{s1t1}, w_{s2t1}, w_{s3t1}, w_{s4t1})$ where w_{sjt1} is a dummy equal to 1 if and only if surgery t is the first *observed* surgery using device k_{sjt}. Define $w_{st2} = (w_{s1t2}, w_{s2t2}, w_{s3t2}, w_{s4t2})$. For each sjt combination, w_{sjt2} is the *observed* time gap between the current surgery and the most recent prior surgery which used the device version k_{sjt}, if we observe a prior usage of device k_{sjt} (i.e., if $w_{sjt1} = 0$). For those cases where we observe no prior usage of device k_{sjt} (i.e., $w_{sjt1} = 1$), we set $w_{sjt2} = 0$ without loss of generality.

Since the error ε_{st} may have a different variance for each surgeon, we test for grouped heteroskedasticity using the test proposed by Levene (1960) and Brown and Forsythe (1974). The results shows that we cannot reject the null hypothesis; thus, we continue to use a homoskedasticity assumption.[20] Since we have an unbalanced panel with varying time gaps between ob-

tables may be requested from the authors.

[17] We also compute eigenvalues of the inner product of explanatory variables and variance inflation factors for each of our variables. Both eigenvalues and VIFs are within acceptable ranges. Thus multi-collinearity is not a concern.

[18] We will introduce other measures for device-specific learning and forgetting later.

[19] We also use duration instead of its log value in alternative specifications, with qualitatively similar results.

[20] We also run the OLS specifications using the cluster, robust command in Stata to model heteroskedasticity

servations 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 to detect any possible serial correlation of errors. We find that serial correlation in ε_{st} is not a concern. See online Appendix for details.

4.2 Correction for Left Censoring of Device-Specific Experience

A serious problem in the above specification is that our two key measures of device-specific experience, namely, whether a specific device version is being used for the first time, w_{st1}, and the amount of time since the last use of a specific device version by a surgeon, w_{st2}, suffer from left censoring. This censoring problem arises for the first observed usage of device version k_{sjt} by surgeon s: is it the true first, or was there a usage prior to the start of our sample? If there was a prior usage, then the true experience gap will be larger than the observed time gap between the start of our sample and the first observed usage of device k_{sjt}.[21] Clearly, left censoring of device-specific experience is a serious problem because it affects our two main sets of variables of interest. This type of left censoring is a pervasive problem in hospital data because there is little data available on surgical device usage patterns going back in time and also because surgeons typically move across hospitals over their careers. A similar concern arises in other contexts when using highly granular experience data. Below, we develop a generalizable estimation procedure that corrects for this problem.

We rewrite equation (10) as

$$y_{st} = \beta X_{st} + \gamma e_{st} + \alpha z_{st1} + \theta \log [z_{st2}] + \varepsilon_{st} \tag{2}$$

where both z_{st1} and z_{st2} are vectors of (unobserved) device-specific experience variables for surgeon s at time t. We define z_{sjt1} as a dummy equal to 1 if and only if surgery t is the *true* first surgery performed by surgeon s using device k_{sjt}, and z_{sjt2} as the *true* amount of calendar time since the last use of k_{sjt} by surgeon s.[22] For observations with $w_{sjt1} = 1$, $z_{sjt} = (z_{sjt1}, z_{sjt2})$ is not observed because surgeon s may or may not have used the same device k_{sjt} in a surgery prior to the beginning of the sample period. For notational simplicity, redefine w_{sjt2} as the time gap between the starting day of our sample period and the date of the surgery at hand when

by surgeon and arbitrary correlation of errors within each surgeon. Our results are qualitatively unchanged. Also, to check whether a random-effects model is preferable to the pooled OLS model with fixed effects, we run the Breusch and Pagan (1980) Lagrange Multiplier test. Based on the residuals from pooled OLS, $LM = 1.34$ which follows a χ^2 distribution under H_0; thus we fail to reject the null hypothesis that pooled OLS with fixed effect is appropriate for our data.

[21] Note that the left censoring of the total experience of each surgeon, e_{st}, does not pose a problem as prior experience of each surgeon is fully captured by his/her surgeon fixed effect.

[22] If $z_{sjt1} = 1$, then we can just set $z_{sjt2} = 0$ (or any other constant) with no loss of generality.

$w_{sjt} = 1$. In fact, if $z_{sjt1} = 0 |\ w_{sjt1} = 1$, then $z_{sjt2} > w_{sjt2}$ and therefore z_{sjt2} is censored.

For each device j, although we cannot observe the true values of z_{sjt1} and z_{sjt2} directly, if we can estimate the distribution of $z_{sjt} = (z_{sjt1}, z_{sjt2})$ conditional on observed w_{sjt1} and w_{sjt2}. Then we can correct for the censoring problem using simulation methods. The intuition for our method is simple and can be illustrated using Figure 3, which plots the usage over time of two devices, A and B. Suppose the time of the first observed use of device A coincides with that of device B, and that subsequently B is used much more frequently across surgeons in our sample than A. Here the observed first use is more likely to indicate a true first use for device B than for device A. Our methodology uses this logic to incorporate the "probability of first usage" into the likelihood. Fader et al. (2005) address a similar problem in a different way.[23]

Recall that z_{sjt2} is defined as the true amount of calendar time since the last use of the same version of device j by surgeon s, therefore, it is a typical "time to event" variable and we can use survival analysis to deal with the data censoring issue. Let $S_j(\cdot)$ denote the survivor function of z_{sjt2}, i.e., the probability that z_{sjt2} is larger than a certain value. By using the Kaplan-Meier estimator which takes into account the censoring problem, we can estimate $S_j(\cdot)$ nonparametrically from our data. Let $f_j(\cdot)$ be the density of z_{sjt2} with distribution $F_j(\cdot)$. Then the estimate of $f_j(\cdot)$ and $F_j(\cdot)$ can be derived from $S_j(\cdot)$. We estimate the distribution of z_{sjt2} for each device type separately, assuming that $z_{sjt2} \sim F_j(\cdot)$ for each version of device j.[24] Figure 4 shows a graph of the estimated $F_j(\cdot)$ for each device.

We next construct the likelihood. Define $h_j(z_{sjt} \mid w_{sjt})$ as the density of $z_{sjt} = (z_{sjt1}, z_{sjt2})$ conditional on observed $w_{sjt} = (w_{sjt1}, w_{sjt2})$. We construct $h_j(z_{sjt} \mid w_{sjt})$ as follows. First, we define

$$h_{j1}(z_{sjt1}, z_{sjt2} \mid w_{sjt1} = 1, w_{sjt2}) = \begin{cases} F_j(w_{sjt2}) & \text{if } z_{sjt1} = 1 \\ f_j(z_{sjt2}) & \text{if } z_{sjt1} = 0; z_{sjt2} > w_{sjt2} \end{cases} . \tag{3}$$

Intuitively, the first line of equation (3) indicates that, when the observed first usage of a particular device version is the true first usage, $z_{sjt1} = 1$. The probability of the true first usage occurring prior to the sample is $1 - F_j(w_{sjt2})$;[25] therefore the likelihood in the case where $z_{sjt1} = 1$ is $F_j(w_{sjt2})$. The second line indicates that, when the observed first usage of a particular device version is not the true first usage, then $z_{sjt1} = 0$ and z_{sjt2} must be larger than w_{sjt2}. By the same logic, we have a density for z_{sjt} if we observe that there are prior usages of

[23] Fader et al. (2005) model purchasing behavior in settings where customers may drop out over time.

[24] Both assumptions are made due to data size limitation. If we had a much larger sample, we could relax these two assumptions.

[25] If $z_{sjt1} = 0$, then $h_{j1}(z_{sjt1}, z_{sjt2} \mid w_{sjt1} = 1, w_{sjt2}) = \int_{w_{sjt2}}^{\infty} h_{j1}(0, z_{sjt2} \mid w_{sjt1} = 1, w_{sjt2}) dz_{sjt2} = 1 - F_j(w_{sjt2})$.

a particular device version by surgeon s before surgery t. In this case, define

$$h_{j0}(z_{sjt1}, z_{sjt2} \mid w_{sjt1} = 0, w_{sjt2})$$

with all of its mass at w_{sjt2}. Then we can write $h_j\left(z_{sjt}|w_{sjt}\right)$ as

$$h_j\left(z_{sjt1}, z_{sjt2}|w_{sjt1}, w_{sjt2}\right) = \begin{cases} h_{j1}\left(z_{sjt1}, z_{sjt2}|w_{sjt1}, w_{sjt2}\right) \ if \ w_{sjt1} = 1 \\ h_{j0}\left(z_{sjt1}, z_{sjt2}|w_{sjt1}, w_{sjt2}\right) \ if \ w_{sjt1} = 0 \end{cases},$$

and define $H_j\left(z_{sjt1}, z_{sjt2}|w_{sjt1}, w_{sjt2}\right)$ as the corresponding distribution function.

If we make a functional form assumption about the distribution of ε_{st}, then we can construct a likelihood term reflecting our imperfect knowledge of left-censored waiting time. In particular, we assume that $\varepsilon_{st} \sim iidN\left(0, \sigma_\varepsilon^2\right)$. Then, the likelihood contribution for $(y_{st} \mid X_{st}, e_{st}, z_{st}, \upsilon_s)$ is

$$g\left(y_{st} \mid X_{st}, e_{st}, z_{st}\right) = \frac{1}{\sigma_\varepsilon}\phi\left(\frac{y_{st} - \beta X_{st} - \gamma e_{st} - \alpha z_{st1} - \theta \log\left[z_{st2}\right]}{\sigma_\varepsilon}\right).$$

The likelihood contribution for surgeon s performing surgery t conditional on observed $w_{st} = (w_{st1}, w_{st2})$ can be obtained by integrating over unobserved variables, z_{st}, as[26]

$$L\left(y_{st} \mid X_{st}, e_{st}, w_{st}\right) = \int \frac{1}{\sigma_\varepsilon}\phi\left(\frac{y_{st} - \beta X_{st} - \gamma e_{st} - \alpha z_{st1} - \theta \log\left[z_{st2}\right]}{\sigma_\varepsilon}\right) \prod_{j=1}^{4} dH_j\left(z_{sjt}|w_{sjt}\right). \tag{4}$$

Then we can use simulation to approximate $L\left(y_{st} \mid X_{st}, e_{st}, w_{st}\right)$ (McFadden 1989, Stern 1997). The underlying intuition behind the simulation is to draw random values of z_{sjt} from its distribution $H_j\left(z_{sjt}|w_{sjt}\right)$ for each device j and use them to compute the sample mean of $\frac{1}{\sigma_\varepsilon}\phi\left(\frac{y_{st} - \beta X_{st} - \gamma e_{st} - \alpha z_{st1} - \theta \log[z_{st2}]}{\sigma_\varepsilon}\right)$. In particular, if z_{st}^r, $r = 1, 2, \cdots, R$ are R independent draws from the joint distribution, $H\left(z_{st}|w_{st}\right)$, then

$$\widetilde{L}\left(y_{st} \mid X_{st}, e_{st}, w_{st}\right) = \frac{1}{R}\sum_{r=1}^{R}\left[\frac{1}{\sigma_\varepsilon}\phi\left(\frac{y_{st} - \beta X_{st} - \gamma e_{st} - \alpha z_{st1}^r - \theta \log\left[z_{st2}^r\right]}{\sigma_\varepsilon}\right)\right] \tag{5}$$

is used to approximate equation (4).[27] Details of the simulation are provided in Appendix A.1.

[26] Since z_{sjt} are independent from each other, the joint conditional density function of $z_{st} = (z_{s1t}, z_{s2t}, z_{s3t}, z_{s4t})$, $h\left(z_{st}|w_{st}\right)$, can be written as $h\left(z_{st}|w_{st}\right) = \prod_{j=1}^{4} h_j\left(z_{sjt}|w_{sjt}\right)$.

[27] Note that some of the devices used by a particular surgeon are the same. One might think this causes z_{ijt} to be dependent over t if they share k. However, for this specification of the experience variables, there is no dependence because the randomness in z_{ijt} applies only to the first observed occurence of the use of type-j device k.

The likelihood function can be written as

$$L\left(y \mid X, e, w\right) = \prod_{s} \prod_{t} L\left(y_{st} \mid X_{st}, e_{st}, w_{st}\right). \tag{6}$$

So equation (6) can be simulated as

$$\widetilde{L}\left(y \mid X, e, w\right) = \prod_{s} \prod_{t} \widetilde{L}\left(y_{st} \mid X_{st}, e_{st}, w_{st}\right). \tag{7}$$

Instead of choosing parameters to maximize the likelihood value of equation (6), we choose parameters to maximize the simulated likelihood value of equation (7). The results from MSLE are presented in column (4) of Table 6.

Within healthcare, our approach is applicable to the usage of new medical devices and instruments and adoption of new surgical procedures for a wide variety of surgeries. This approach is also widely applicable outside healthcare. For example, architectural work for house renovation can include a variety of jobs such as attic conversion, basement conversion, porch extension, chimney removal or stairwell redesign. An architecture firm can use its own historic panel data to examine whether architects are much slower on their first few instances of a new type of job to decide whether to invest in ways to ramp up their learning curve. However, as the panel may cover a limited time period and also as architects move between firms, the data would be left-censored.

4.3 Endogeneity

We face three sources of potential endogeneity. The first is a surgeon's choice of devices. In selecting device versions for a specific surgery, a surgeon attempts to choose devices that provide the best match with the patient's specific needs. Therefore, factors related to patient, surgery, device, and surgeon characteristics may impact both the surgery's duration and the surgeon's choice of devices. To address potential endogeneity due to device selection, we attempt to fully capture patient, surgery, device, and surgeon characteristics through an extensive set of controls described in Section 3. To determine whether we still have this type of endogeneity, we then model a surgeons' decision to use a new device version using a probit specification and test whether there are common unobserved factors driving both duration of surgery and the decision to use a new device version. A surgeon's decision to use a new device version for a surgery, for each of the four key device types, can be represented in the probit model,

$$
\begin{aligned}
m^{*}_{sjt1} &= \gamma_j X_{st} + \nu_{sjt}, \\
\nu_{sjt} &\sim iidN\left(0,1\right), \\
m_{sjt1} &= 1\left(m^{*}_{sjt1} > 0\right).
\end{aligned} \tag{8}
$$

Here X_{st} is the vector of observed exogenous control variables and ν_{sjt} is the error term representing unobserved factors impacting the device choice. The issue of interest is whether the error term of the surgery duration equation (10), ε_{st}, is correlated with the errors $\nu_{st} = (v_{s1t}, v_{s2t}, v_{s3t}, v_{s4t})$, from the "new-device-choice" probit model for each device. More formally, we assume that, for surgeon s during surgery t,

$$
\begin{pmatrix} \varepsilon_{st} \\ \nu_{s1t} \\ \vdots \\ \nu_{sJt} \end{pmatrix} \sim iidN \left[0, \begin{pmatrix} \sigma^2_{\varepsilon} & \rho_{\varepsilon 1 \nu} & \cdots & \rho_{\varepsilon J \nu} \\ \rho_{\varepsilon 1 \nu} & 1 & \cdots & \rho_{\nu} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{\varepsilon J \nu} & \rho_{\nu} & \cdots & 1 \end{pmatrix} \right]
$$

where $\rho_{\varepsilon j \nu}$ is the covariance between ε_{st} and ν_{sjt}. $c_{\varepsilon j \nu} = \frac{\rho_{\varepsilon j \nu}}{\sigma_{\varepsilon}}$ is the corresponding correlation, and we construct the hypothesis of interest as $H_0 : c_{\varepsilon j \nu} = 0$ against $H_A : c_{\varepsilon j \nu} \neq 0$.

The intuition for this test is simple. If there are common unobserved factors driving both duration of surgery and the choice to use a new device version, ε_{st} and ν_{sjt} should be correlated, and $c_{\varepsilon j \nu}$ should be significantly greater than zero. Details of constructing generalized residuals (ε_{st}, ν_{sjt}) and correlation ($c_{\varepsilon j \nu}$), as well as the endogeneity test itself, are provided in Appendix A.2. The test results show no significant correlation between these error terms. We therefore conclude that there is no potential endogeneity of this type.

A patient's selection of surgeon may also be endogenous as a patient may seek out a surgeon with higher quality, causing high-quality surgeons to have more experience. This endogeneity problem is likely to cause a downward bias to the coefficient of total experience. However, if patients with more complex problems are more likely to seek out high quality surgeons, it is possible that the coefficient of total experience will be biased upward. Our large set of control variables for surgery complexity and patient characteristics and our surgeon-specific fixed effects control for endogeneity due to surgeon selection. Note that prior research on the impact of experience on outcomes in hip replacement surgery has not included surgeon fixed effects, resulting in potentially biased results (e.g., Reagans et al. 2005, Shervin et al. 2007, Yasanuga et al. 2009).

Finally, many unobserved factors in the OR influence duration of surgery. When using a new device for the first time, a surgeon might want to work with team members with whom he is most familiar. However, the effect of using new devices on productivity would prevail even if surgeons can mitigate the impact of using a new device by adding the right professionals to the team. A surgeon may also try to speed up if the OR is highly congested and he is running late. Also, the composition of the surgical team may influence the duration of surgery, both directly and due to team experience effects (e.g., Reagans et al. 2005, Huckman et al. 2009, Huckman and Staats, 2011). These factors are not a source of bias because surgeons decide which specific device versions to use in each surgery well in advance of (and without knowing) the exact surgery date.

5 Results

5.1 Baseline Results

The first four columns of Table 6 contain estimates for different versions of our baseline specification in equation (10), estimated with OLS. In columns (1) and (2), we examine the impact of total experience on surgery duration in the absence of any controls for device-specific experience, which is the main focus of prior research on learning in orthopedic surgery (e.g., Reagans et al. 2005, Shervin et al. 2007, Yasanuga et al. 2009). Column (1) shows that total experience significantly reduces the duration of surgery, which is consistent with previous research. However, on inclusion of surgeon dummies as in column (2), total experience is no longer significant. Thus, the negative coefficient of total experience in the specification in column (1) is likely due to variation in surgeon quality or unobservable patient characteristics across surgeons rather than due to within-surgeon learning with experience. Given the high experience level of our surgeons (see Table 3, columns (4)-(7)), it is not surprising to find that they appear to have reached the flat portion of their experience curves with regard to general learning about hip replacement surgery.

In the specifications in Columns (3)-(6), we consider device-specific experience at the highly granular level of device versions within each of the four key devices. Columns (3) and (4) present estimates for equation (10) using OLS. The results in column (3) suggest that the first observed use of a stem version by a surgeon results in an approximately 26.2% increase in duration of surgery, all else equal, relative to cases where the surgeon has been observed using the stem version before. The estimates on the forgetting variables in column (3) suggest there is knowledge depreciation over time in the case of both stems and liners. For these device

types, a 1% increase in the number of days since previous usage of a specific device version results in an approximately 0.03% increase in surgery duration. This implies that, when the experience gap for stems increases from its median (7 days) to its 75^{th} percentile (24 days), surgery duration increases by about 3.4%, all else equal. Similarly, when the experience gap for liners increases from its median (9 days) to its 75^{th} percentile (25 days), surgery duration increases around 2.9%, all else equal. In column (4), we control for three rare cases: usage of multiple new devices in a surgery, first usage of a shell and liner together, and first usage of a stem and head together.[28] Results in this column suggest that first usage of stems continues to result in a substantial (29.6%) and highly statistically significant increase in duration of surgery with an increase in the size of the effect relative to column (3). Furthermore, we find that first usage of shell versions also results in a statistically significant increase in duration of surgery, albeit a smaller size of effect (26.2% approximately) than that for the first usage of stems.[29] The results on forgetting variables are similar to those in column (3).[30]

Columns (5) and (6) contain results for our second baseline specification (equation (2)), which we estimate using our Maximum Simulated Likelihood Estimation procedure with simulation of unobservables conditional on observables. As described in Section 4.2, we develop this procedure to control for the left-censoring of our device-specific experience variables. Note that, while we cannot be sure whether an observed first usage of a device version by a surgeon is indeed his true first usage of this version, the number of observed first usages can only exceed or at best equal the number of true first usages. If true first usages take longer than repeat usages, then the positive effect of true first usages will be masked by including some later usages as first usages, which means OLS coefficients may be underestimates of the true coefficients. Although left censoring also affects the experience gap variable, one cannot sign the bias in this case and OLS coefficients cannot be considered as underestimates or overestimates of the true coefficients. The sign of the bias depends on the correlation between the censored variable(experience gap) and other control variables which are not left censored.[31] Results in columns (5) and (6) support the above predictions: compared with columns (3) and (4), the estimated coefficients for both stem and shell become larger and their significance is either improved or stays the same. The coefficients of experience gap since last use of stems and liners continue to be statistically

[28] We do the latter because certain shell and liner versions are constrained to be used as a pair, as are certain stem and head versions.

[29] If surgeons choose to use multiple new devices in simpler surgeries, not controlling for use of multiple new devices may mask the average effect of first use of a shell.

[30] It can be shown that our estimated impact of forgetting on productivity is considerably higher than that implied in the ship building study of Thompson (2007).

[31] If we assume the correlation between experience gap and other non-censored control variables is zero or small enough, then the bias has the same sign as the coefficient. A proof is provided in the online appendix.

significant and of very similar magnitude to the OLS results.[32]

As a robustness check, we interacted the surgeon dummies with total experience. Our results are consistent with the those presented in Table 6. We also ran a specification in which we allowed for a different coefficient for overall experience only for surgeon 4, who had much less US-based experience than the other surgeons, although he does have non-US experience. We still see no effect for learning with overall experience. In another specification, we interacted the junior surgeon (i.e., surgeon 4) dummy with our first usage dummies and experience gap variables, and our main results stay the same. Since the data sample for the junior surgeon is small, we do not use this as our main specification. The stability of our main results across specifications in terms of size of effects and significance is reassuring.

5.2 Alternative experience variables

With regard to device-specific learning, our baseline specifications focus mainly on the first usage of a device version. An alternative is to include dummies to control for the first few usages of a device version or to include usage count variables. Either of these approaches can help reveal the curvature of learning. Note, however, that, if device-specific learning is steep in the beginning and soon flattens out, then usage count variables will show little learning effect and insignificant results. Column (1) of Table 7 shows the results when we add usage count variables in addition to first usage variables. None of the usage count variables' coefficients are significant. We therefore focus on the steep start of the learning curve. Although including dummy variables for the first few usage of a device version can help map out the early learning trajectory, we still face a left censoring problem for the second usage, the third usage, etc. of a device version, since we cannot be sure as to whether an observed first usage is a true first usage. In the case of the first usage variables, our MSLE approach controls for left censoring. However, for further usage, left censoring remains, and controlling for it is significantly more difficult than for the first usage. Therefore, we present the results from the OLS estimation using the first four usages for each device in column (2), acknowledging that the related estimated coefficients can be biased. We find that the first four usages of a stem variant result in a statistically significant increase in duration. The coefficient of first usage is almost double that of the subsequent three usages, and these differences are statistically significant at the 1% level. However, the sum of the coefficients for the second through fourth usages is significantly greater than the coefficient of the first usage. This suggests that additional learning occurs in these subsequent usages. As

[32] In all specifications, the coefficients for control variables are consistent with intuition and qualitatively similar. To save space, we have omitted coefficient estimates for control variables from the table. A full set of results is posted in the online appendix.

these results are based on OLS, focusing on first usages provides a conservative lower bound on added time due to learning.

With regard to device-specific forgetting, our baseline specification focuses on *experience gap* which is defined as the time elapsed (in days) since the last use by the surgeon of the specific device version used in the surgery. A surgeon's knowledge about a specific device version can depreciate over time simply because he or she has not used that device version for a while; therefore this experience gap naturally becomes our best choice. Nonetheless, it is interesting to explore alternative factors which can cause forgetting. For example, distractions due to using devices other than the "focal" device may affect forgetting. We construct three additional sets of variables to capture forgetting due to such distractions: the number of surgeries in-between, a device switch dummy and device switch variety, as defined in Section 3. Results for these three sets of variables are presented in columns (3) - (6) of Table 7. Column (3) shows that, for shells and liners, the number of surgeries in-between increase surgery duration at the 10% significance level. On the other hand, in the results in columns (5) and (6), we find that the other two sets of variables (device switch dummies and device switch variety variables) do not have a statistically significant effect on duration. Thus, we do not find support for forgetting due to switching to one or more different device versions in-between repeat uses of a particular device version.[33]

Our finding of significantly higher surgery duration in the case of surgeries involving first use of stem and shell versions is consistent with the notion that experience would be more significant for those devices that require greater skill and dexterity to place properly, e.g., because they touch bone. While first use of a liner version does not significantly impact surgery duration, this may be due to the fact that the total time needed to insert a liner is a small part of the total surgery time, even for a difficult insertion instance. It is not surprising that there is little learning or forgetting in the case of heads, which are easy to insert.

Significant depreciation of knowledge over time in the case of stems and liners is also consistent with the difficulty associated with these devices. In the case of shells, we see no knowledge depreciation over time.

6 Discussion and Conclusions

We have found that first-time use of a new stem (shell) version increases duration of surgery by about 32.4% (27.6%) with a p-value of 0.01 (0.05). These increases in duration increase

[33] We acknowledge that all three sets of alternative forgetting variables also suffer from left censoring, which may lead to biased estimates.

the likelihood of infection, blood loss, and other complications.[34] In our sample period, about 10% of all surgeries included first observed usage of a stem version, and 5% of all surgeries included first observed usage of a shell version. The average surgery duration in our sample is 165 minutes. With about 330 hip replacement surgeries performed each year at the UVA Hospital in our sample period, this translates into 1764 additional minutes each year for surgeries involving first-time stems and 754 additional minutes each year for surgeries involving first-time shells. Hospital ORs increasingly face severe capacity constraints (Sokal et al. 2006). Using structural estimation, Olivares et al. (2006) have estimated that the implied cost of OR idle time far exceeds that of OR staff overtime. Freeing up OR time by reducing variety would allow hospitals to perform more surgeries. For example, at the UVA hospital, given average surgery duration of 165 minutes, fifteen additional hip replacement surgeries could have been performed per year in the additional time spent when operating with new stems or shells, a 5% increase. We would expect a 5% increase at the national level as well under similar assumptions as above. Accounting for the additional time associated with 2nd, 3rd, and 4th usage (predicted by our estimates in Table 7, column (2)) would further increase available OR capacity. Of course, reduced capacity is only a part of the total short-term cost of product variety. Infection and blood loss due to longer duration of surgery at the outset are other short term costs.

The above estimates of the productivity losses from first use of devices are conservative for three reasons. First, having access to data for only experienced surgeons has allowed us to examine the impact of high device proliferation on a highly experienced surgeon pool. For less experienced surgeons, who are likely to see more first time uses, this type of capacity loss would be even greater. Second, ORs are used for many other procedures that involve a variety of devices and instruments. Variety in stems and shells is only a small but illustrative slice of the plethora of variety in devices used in hospital ORs (Maisel 2004). Third, to be conservative we only use the coefficients of first usage for these calculations. If we were to use the coefficients of the first four usages, the effects would be doubled.

Our estimated costs of forgetting are also high. On average, surgeons in our sample perform a surgery using a stem once a week,[35] and they use the same stem version on average once a month. If all stems were identical (no stem variety), the time gap between surgeries using the same stem version would be the time gap between surgeries using a stem. We can calculate the hypothetical surgery duration in this case and compare it with the real duration of each

[34] Yasunaga et al. (2009) report that surgeon volume in excess of 500 cases is inversely related to operating time (odds ratio 0.20;$P < 0.01$), blood loss (odds ratio 0.54; $P = 0.02$), and postoperative complications (odds ratio 0.53; $P = 0.01$).

[35] Some revision surgeries may not use a stem at all.

surgery.[36] In our sample period at UVA, about 1587 minutes (623 minutes) in total are added to surgery duration due to the longer experience time gap associated with the high variety of stems (liners).[37] More conservatively, suppose the device-specific experience gap is reduced by half due to lower device variety. In this case, the time saved each year is 370 minutes for stems (312 minutes for liners). This represents 1% more hip surgeries that could be performed annually in the US if stem and liner variety is halved.

Hospitals can reduce the costs of device variety through better surgical education (Aggarwal and Darzi 2006). Our research highlights a specific need area – ways to adequately train surgeons on the wide variety of available device versions. In medical school, surgeons-in-training practice surgery on cadavers and synthetic plastic bones, often using only one or two versions of a medical device,[38] so graduates are very likely to encounter new device versions. A surgeon can prepare prior to using a new device version by carefully reading the documentation, examining the device itself beforehand, talking to a colleague who has used the device before, and by using surgical simulation software. Our discussions with surgeons suggest that most do not take these preparatory steps.

Given the extraordinarily high variety of device SKUs available today for most medical devices, our findings also have very significant implications for policy makers. The high productivity and quality costs associated with device variety suggest that the gain from a new device design needs to be large enough to compensate for the short term disadvantages of starting up on a new learning curve, and, also, of increasing the chances of knowledge depreciation over time. Better measures of the long term benefits and costs of device variety are needed to navigate this tradeoff. Such measurements would be facilitated by implementing nationwide medical device registries to gather information about devices that are in use, and by requiring greater price transparency in the medical device market.

Future research should examine the underlying reasons for the extremely high and seemingly inefficient level of variety in medical devices. A related issue is hospitals' and surgeons' incentives and disincentives to control costs through choice of medical devices.

We have focused on only one hospital and one type of surgery. Future research can examine other settings. Estimation of other costs of device variety (such as greater instrumentation

[36] We use the following equations to calculate the hypothetical duration for each surgery using a stem:

$$\begin{aligned} \log(duration\_\text{real}) &= x + 0.03 \times \log(gap\_stem + 1); \\ \log(duration\_hypothetical) &= x + 0.03 \times \log(gap + 1) \end{aligned}$$

where x is the contribution of the remaining terms in the ln(duration) equation (which remains the same under the assumption of no stem variety) and *gap* is the time gap between two surgeries using a stem.

[37] Including as an independent variable the amount of time since the last surgery performed by the surgeon does not impact our main results.

[38] This information was shared with one of the authors by the Chief Information Officer of the Cleveland Clinic.

costs and higher inventory costs), and its benefits (such as better fit to patient needs), are also fruitful research areas. Furthermore, future research can consider behavioral aspects of device choice in the spirit of emerging behavioral research in healthcare operations (e.g., Mennicken et al. 2014).

Acknowledgments

The authors thank Serguei Netessine, the associate editor, and three reviewers for extremely useful comments. The authors 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 Berkeley, Boston University, Chicago, Duke, Emory, Georgetown, HBS, Kellogg, the 5th LBS Operational Innovation Workshop, the Wharton Empirical Operations Conference, Maryland, MIT, NYU, UCLA and USF. Mike Guthrie at Zimmer, Jerry Kie at Smith & Nephew, Chris Petrie at Stryker, and Robert McGlothin at Depuy provided valuable industry insight and expertise.

References

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

Agrawal, A. and S. Muthulingam. 2015. Does Organizational Forgetting Affect Vendor Quality Performance? An Empirical Investigation. *Manufacturing & Service Operations Management.* 17(3):350-367.

Anderson, D., M. Binder, and K. Krause. 2002. The Motherhood Wage Penalty: Which Mothers Pay It and Why? *American Economic Review.* 92(2):354-358.

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

Argote, L., S.L. Beckman and D. Epple. 1990. The Persistence and Transfer of Learning in Industrial Settings. *Management Science.* 36(2): 140-154.

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

Bauer, G.C.H. 1992. What Price Progress? Failed Innovations of the Knee Prosthesis. *Acta Orthopaedica Scandinavica.* 63 (3): 245-246.

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

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

Boh, W.F., S.A. Slaughter and J.A. Espinosa. 2007. Learning from Experience in Software Development: A Multilevel Analysis. *Management Science.* 53(8): 1315-1331.

Boone, T., R. Ganeshan and R. L. Hicks. 2008. Learning and Knowledge Depreciation in Professional Services. *Management Science.* 54(7): 1231-1236.

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.

Clark, J.R. and R.S. Huckman. 2012. Broadening Focus: Spillovers and the Benefits of Specialization in the Hospital Industry. *Management Science*. 58(4): 708-722.

Clark, J.R., R.S. Huckman and B.R. Staats. 2013. Learning from Customers in Outsourcing: Individual and Organizational Effects. *Org. Sci.* 24(12): 1539-1557.

Demske, G.E. 2008. Examining the Relationship Between the Medical Device Industry and Physicians. Testimony provided to the Office of Inspector General. Department of Health and Human Services.

Fader, P.S., B.G.S. Hardie and K.L. Lee. 2005. "Counting your customers" the Easy Way: An Alternative to the Pareto/NBD model. *Marketing Science.* 24(2): 275-284.

Gelberman, R.H., D. Samson, S.K. Mirza, J.J. Callaghan and V.D. Pellegrini. 2010. Orthopaedic Surgeons and the Medical Device Industry: The Threat to Scientific Integrity and the Public Trust. *Journal of Bone and Joint Surgery.* 92:765-77.

Gorman, P.J., A.H. Meier, C. Rawn and T.M. Krummel. 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.

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.

KC, D. and B. R. Staats. 2012. Accumulating a Portfolio of Experience: The Effect of Focal and Related Experience on Surgeon Performance. *Manufacturing & Service Operations Management.* 14(4): 618-633.

Keane, M.P. and K.I. Wolpin. 1997. Career Decisions of Young Men. *Journal of Political Economy.* 105(3): 473-522.

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

MacDuffie, J.P., K. Sethuraman and M.L. Fisher. 1996. Product Variety and Manufacturing Performance: Evidence from the International Automotive Assembly Plant Study. *Management Science.* 42(3): 350-369.

McFadden D. 1989 A Method of Simulated Moments for Estimation of Discrete Response Models Without Numerical Integration. *Econometrica*. 57(5): 995-1026.

Meier, B. In Medicine, New Isn't Always Improved. New York Times. June 25, 2011.

Mennicken, R., L. Kuntz and S. Scholtes. 2015. Stress on the Ward - An Empirical Study of the Nonlinear Relationship between Organizational Workload and Service Quality. *Management Science* 61(4):754-771.

Mincer, J. and H. Ofek. 1982. Interrupted Work Careers: Depreciation and Restoration of Human Capital. *Journal of Human Resource*s. 82(17): 3-24.

Narayanan , S., S. Balasubramanian and J.M. Swaminathan. 2009. A Matter of Balance: Specialization, Task Variety, and Individual Learning in a Software Maintenance Environment. *Management Science*. 55(11): 1861-1876.

Nembhard, D. A. and N. Osothsilp. 2001. An Empirical Comparison of Forgetting Models. IEEE Transactions on Engg. Mgmt. 48(3): 283-291.

Nembhard, D. A. and N. Osothsilp. 2002. Task Complexity Effects on Between-Individual Learning/Forgetting Variability. Intl. J. of Industrial Ergonomics. 29: 297–306.

Nembhard, D. A. and M. V. Uzumeri. 2000. Experiential learning and forgetting for manual and cognitive tasks. Intl. J. of Industrial Ergonomics. 25(4): 315-326.

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, R. Laskin, J. Davis, M.G. Peterson and T. Richart. 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, K. 2003. Managing Product Variety: An Integrative Review and Research Directions. *Production and Operations Management*. 12(1): 79-101.

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

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.

Rosenthal, E. In Need of A New Hip, but Priced Out of the US. New York Times. August 3, 2013

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.

Salemi, T. 2011 Are Device VCs Becoming Spineless? *IN VIVO*. 29(5).

Schilling, M. A., P. Vidal, R. E. Ployhart and A. Marangoni. 2003. Learning by Doing Something else: Variation, Relatedness, and the Learning Curve. *Management Science*. 49(1): 39-56.

Shafer, S. M., D.A. Nembhard and M.V. Uzumeri. 2001. The Effects of Worker Learning, Forgetting, and Heterogeneity on Assembly Line Productivity. *Management Science* 47(12): 1639-1653.

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.

Sokal, S.M., D. L. Craft, Y. Chang, W.S. Sandberg and D.L. Berger. 2006. Maximizing Operating Room and Recovery Room Capacity in an Era of Constrained Resources. JAMA Surgery. 141: 389 – 396.

Staats, B.R. and F. Gino 2012. Specialization and Variety in Repetitive Tasks: Evidence from a Japanese Bank. *Management Science*. 58(6): 1141-1159.

Stern, S. 1997 Simulation-Based Estimation. *Journal of Economic Literature*. 35(4): 2006-2039.

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.

Thompson, P. 2007. How Much Did the Liberty Shipbuilders Forget? *Management Science.* 53(6): 908-918.

Thornton, R.A. and P. Thompson. 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.

Wixted, J.T. 2004. The Psychology and Neuroscience of Forgetting. *Annual Review of Psychology* 55: 235–69.

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

Yasunaga, H., K. Tsuchiya, Y. Matsuyama and K. Ohe. 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.

Appendix

A.1: Details of Simulation Algorithm of z_{st}^r

1. Estimate $\widehat{F}_j(\cdot)$ for device j using the Kaplan-Meier estimator;
2. Draw R random values l^r from uniform distribution and then use them to find corresponding values $z^r = \widehat{F}_j^{-1}(l^r)$;
3. For observations with $w_{sjt1} = 1$, compare each z^r with w_{sjt2}: if $z^r \leq w_{sjt2}$, then $\left(z_{sjt1}^r = 1, z_{sjt2}^r = 0\right)$, if $z^r > w_{sjt2}$, then $\left(z_{sjt1}^r = 1, z_{sjt2}^r = z^r\right)$. After the comparison, there should be a matrix of z_{sjt}^r with $2R$ elements;[39]
4. Use the density $h_{j0}(z_{sjt1}, z_{sjt2} \mid w_{sjt1} = 0, w_{sjt2})$ and $\left(z_{sjt1}^r = 0, z_{sjt2}^r = w_{sjt1}\right)$ for observations with $w_{sjt1} = 0$;
5. Do step 1-4 for each device to get $z_{st}^r = (z_{s1t}^r, z_{s2t}^r, z_{s3t}^r, z_{s4t}^r)$ used in equation (5).[40]

A.2 Endogeneity Test

In this appendix, we use the specification with OLS estimation to demonstrate our endogeneity test method. The full version including both OLS and MSLE, as well as endogeneity test results, is provided in our online appendix.

As mentioned in Section 4.3, we first construct generalized residuals of duration equation and device choice equation, ε_{st} and ν_{sjt}. Then we calculate the correlation coefficient, $c_{\varepsilon j\nu}$, of those two random variables, use it as the test statistic and see whether it is significantly different from zero. Particularly, we define $\widehat{\varepsilon}_{st}$ as the residual for equation (10) and

$$\begin{aligned}\widehat{\nu}_{sjt} &= E\left[\nu_{sjt} \mid w_{sjt1}, X_{st}\right] \\ &= \begin{cases} \frac{\phi\left(\widehat{\gamma}_j X_{st}\right)}{\Phi\left(\widehat{\gamma}_j X_{st}\right)} & \text{if } w_{sjt1} = 1 \\ \frac{-\phi\left(\widehat{\gamma}_j X_{st}\right)}{1-\Phi\left(\widehat{\gamma}_j X_{st}\right)} & \text{if } w_{sjt1} = 0 \end{cases}\end{aligned}$$

[39] There are 2R elements, since we need to use antithetic acceleration to reduce the variance of our simulators;
[40] z_{st}^r's are simulated prior to optimization of the likelihood function and never changed (McFadden 1989).

as the generalized residual for equation (14) (e.g., Gourieroux et al., 1987; Dean et al., 2015)

Next we can construct a correlation term either for each device j or for all devices together. The estimate of the device-specific correlation term is

$$\widehat{c}_j = \frac{n_j^{-1}\sum_{st}\widehat{\varepsilon}_{st}\widehat{\nu}_{sjt}}{\sqrt{\left(n_j^{-1}\sum_{st}\widehat{\varepsilon}_{st}^2\right)\left(n_j^{-1}\sum_{st}\widehat{\nu}_{sjt}^2\right)}},$$

where n_j is the total number of surgeries using device j and the correlation term for all devices together is

$$\widehat{c} = \frac{n^{-1}\sum_{st}\widehat{\varepsilon}_{st}\overline{\nu}_{st}}{\sqrt{\left(n^{-1}\sum_{st}\widehat{\varepsilon}_{st}^2\right)\left(n^{-1}\sum_{st}\overline{\nu}_{st}^2\right)}}$$

where n is the total number of surgeries in the sample and $\overline{\nu}_{st} = J^{-1}\sum_j \widehat{\nu}_{sjt}$.Under null hypothesis,

$$plim\widehat{c}_j \propto plim\left(n_j^{-1}\sum_{st}\widehat{\varepsilon}_{st}\widehat{\nu}_{sjt}\right) = 0 \tag{9}$$

where the proportionality factor is the plim of the denominator.

In order to actually use the test statistic, one must know something about the sample distribution of the test statistic. Instead of deriving the asymptotic distribution for our test statistic analytically, it is more straightforward to simulate the small sample distribution of the test statistic and then use simulated critical values to perform the test. In particular, define $\widetilde{\varepsilon}$ as the sample vector of $\widehat{\varepsilon}_{st}$ and $\widetilde{\nu}_j$ analogously for device j. Define $\widetilde{\nu}_j^r$ as the rth random reordering of $\widetilde{\nu}_j$.[41] If $\widetilde{\varepsilon}_{st} \sim iidF_\varepsilon$, $\widetilde{\nu}_{sjt} \sim iidF_{\nu j}$, and $\widetilde{\varepsilon} \perp \widetilde{\nu}_j$, then $\widetilde{\nu}_j^r \sim iidF_{\nu j}$ and $\widetilde{\varepsilon} \perp \widetilde{\nu}_j^r$ as well. Define

$$\widehat{c}_j^r = \frac{n_j^{-1}\sum_{st}\widetilde{\varepsilon}_{st}\widetilde{\nu}_{sjt}^r}{\sqrt{\left(n_j^{-1}\sum_{st}\widetilde{\varepsilon}_{st}^2\right)\left(n_j^{-1}\sum_{st}\left(\widetilde{\nu}_{sjt}^r\right)^2\right)}}$$

as a single draw of $\widehat{c}_j$ and repeat R independent times. Then find the 2.5% and 97.5% percentiles of $\left\{\widehat{c}_j^r\right\}_{r=1}^R$. These are the 5% critical values for the test statistic; reject H_0 iff $\widehat{c}_j$ falls outside the two critical values.

[41] Consider a vector of variables $\nu = (\nu_1, \nu_2, .., \nu_n)'$. Simulate $\xi^r = (\xi_1^r, \xi_2^r, .., \xi_n^r)'$ as a vector of random numbers where $\xi_k^r \sim iidU(0,1)$, and construct ν^r as ν reordered in the same way as ξ^r if sorted from smallest to largest; i.e., $\nu_m^r = \nu_k$ iff ξ_k^r is the m'th smallest element of ξ^r. ν_m^r is a random permutation of ν and independent across $r = 1, 2, .., R$.

A.3: Tables and Figures

Figure 1: Four Key Devices Used in Hip Replacement Surgery

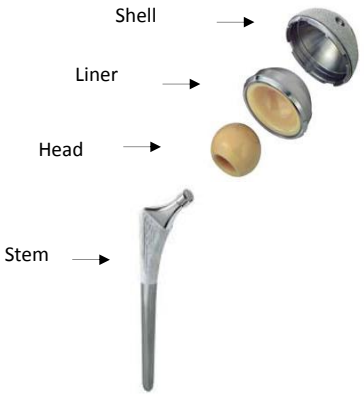


Figure 2: Two Distinct Stem Device Versions

Figure 3: Data Censoring Problem for Surgeries with First Observed Use of a Device Variant

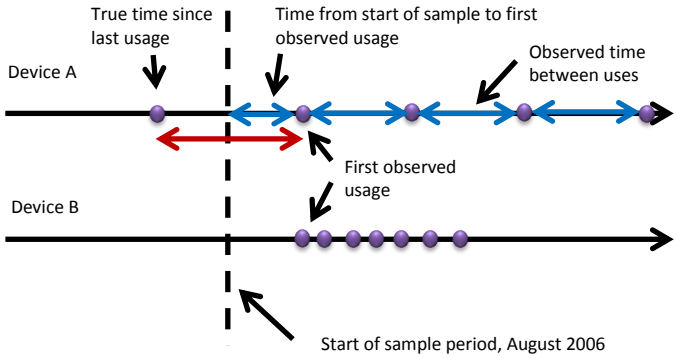


Figure 4: Kaplan-Meier Estimates of Distribution Functions for Experience Gap

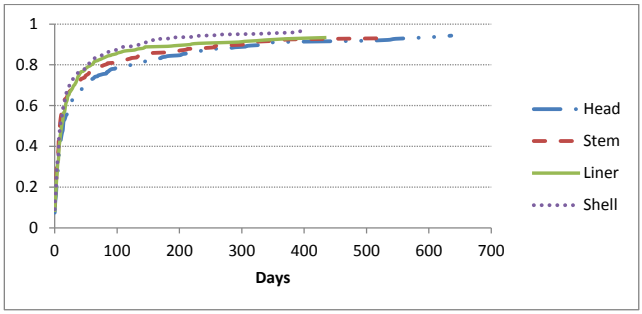


Table 1: Groupings of the Four Main Devices Used in Hip Replacement Surgeries

Company	Shell		Stem		Liner		Head		Total	
	# of SKUs	# of Device Variants	# of SKUs	# of Device Variants	# of SKUs	# of Device Variants	# of SKUs	# of Device Variants	# of SKUs	# of Device Variants
Zimmer	16	4	10	9	19	6	20	4	65	23
Depuy	45	8	60	10	61	10	63	11	229	39
Stryker	31	5	48	12	32	11	48	15	159	43
Smith	22	3	44	4	10	1	34	8	110	16
Total	114	20	162	35	122	28	165	38	563	121

Note: based on the 671 surgeries that are used to create device version variables.

Table 2: Surgery Duration across Sub-Samples of Surgeries

	All Device Variants Observed in Use Before			1 Device Variant Not Observed in Use Before			More than 1 Device Variant Not Observed in Use Before		
	# of obs.	Mean	Std. Dev.	# of obs.	Mean	Std. Dev.	# of obs.	Mean	Std. Dev.
Revision	76	202.47	69.59	27	218.6	102.94	14	243	105.95
First-time	312	146.89	50.16	42	169.5	91.98	12	170.75	98.3

Note: Surgery duration is measured in minutes.

Table 3: Information about Surgeons and Data Structure

Surgeon Number	(1) Number of Surgeries in Sample Period	(2) Number of Surgeries Using at Least One Device from 4 Main Vendors	(3) Estimation on Sample	(4) MD Completion Year	(5) Residency Completion Year	(6) Orthopedic Fellowship Completion Year	(7) Approximate Number of Hip Replacement Surgeries Pre-2006
1	365	350	268	1991	1996	1999	1000
2	189	146	94	1991	1997	1999	1000
3	119	110	80	1993	1999	2000	500
4	79	65	41	1984	2005	2006	–
Total	752	671	483	NA	NA	NA	NA

Table 4: Correlation Matrix of Experience Variables

		Total Exp.	First Time Use Dummy				Ln(Exp. Gap)			
			Shell	Stem	Liner	Head	Shell	Stem	Liner	Head
	Total Exp.	1								
First Time Use Dummy	Shell	-0.13	1							
	Stem	-0.15	0.26	1						
	Liner	-0.11	0.22	0.09	1					
	Head	-0.17	0.26	0.15	0.22	1				
Ln(Exp. Gap)	Shell	-0.12	-0.25	-0.12	-0.22	-0.07	1			
	Stem	-0.04	-0.05	-0.31	-0.10	-0.02	0.31	1		
	Liner	0.11	-0.04	-0.07	-0.25	0.03	0.20	0.11	1	
	Head	-0.04	-0.09	-0.02	-0.11	-0.43	0.28	0.19	0.11	1

Table 5: Descriptive Statistics

Variable	# of Obs.	Mean	Std. Dev.
Duration (minutes)	483	164.98	70.47
Total Experience	483	141.97	104.56
Variable: First Use Dummy			
Shell	414	0.05	0.21
Stem	408	0.09	0.29
Liner	349	0.08	0.28
Head	468	0.10	0.30
Variable: Experience Gap(days)			
Shell	394	24.44	47.87
Stem	371	30.08	64.57
Liner	320	26.02	52.08
Head	423	42.46	85.45
Dummies for New Device Combinations			
Shell and Liner	483	0.01	0.11
Stem and Head	483	0.02	0.14
Dummies for # of New Devices			
2 New Devices	483	0.04	0.19
3 New Devices	483	0.01	0.11
4 New Devices	483	0.00	0.06

Table 6: Estimation Results of Main Specification - Dependent Variable is Ln(Duration)

Explanatory Variable	Column (1) OLS: Total Experience		Column (2) OLS: Add Surgeon FE		Column (3) OLS: Add Device Experience		Column (4) OLS: Add Dummies for New Device Combinations and # of New Devices		Column (5) MLE: Add Device Experience		Column (6) MLE: Add Dummies for New Device Combinations and # of New Devices	
	Coef.	(Std. Err.)	Coef.	(Std. Err.)	Coef.	(Std. Err.)	Coef.	(Std. Err.)	Coef.	(Std. Err.)	Coef.	(Std. Err.)
Total Experience/100	-0.110***	(0.050)	0.027	(0.050)	0.027	(0.050)	0.030	(0.050)	0.026	(0.061)	0.030	(0.064)
First Time Use Dummy												
Shell					0.131	(0.091)	0.262**	(0.111)	0.155*	(0.087)	0.276**	(0.115)
Stem					0.262***	(0.073)	0.296***	(0.078)	0.286***	(0.060)	0.324***	(0.064)
Liner					0.078	(0.076)	0.129	(0.090)	0.089	(0.076)	0.125	(0.087)
Head					-0.029	(0.068)	0.007	(0.075)	-0.031	(0.076)	0.014	(0.084)
Log(Experience Gap)												
Shell					-0.009	(0.015)	-0.009	(0.015)	-0.009	(0.017)	-0.010	(0.016)
Stem					0.030**	(0.013)	0.030**	(0.013)	0.029**	(0.013)	0.029**	(0.013)
Liner					0.031**	(0.015)	0.030**	(0.015)	0.030*	(0.016)	0.030*	(0.016)
Head					0.001	(0.012)	0.002	(0.012)	0.001	(0.014)	0.002	(0.014)
Dummies for New Device Combinations												
Shell and Liner							-0.266	(0.232)			-0.266	(0.859)
Stem and Head							-0.116	(0.191)			-0.182	(0.190)
Dummies for # of New Devices												
2 New Devices							0.016	(0.122)			-0.012	(0.117)
3 New Devices							-0.203	(0.261)			-0.327	(1.013)
4 New Devices							-0.069	(0.456)			0.180	(1.783)
Surgeon FE	No		Yes		Yes		Yes		Yes		Yes	
Quadratic Time Trend	Yes		Yes		Yes		Yes		Yes		Yes	
# of Observations	483		483		483		483		483		483	
Adj. R-squared	0.381		0.397		0.429		0.432		-		-	

Note: Time Trend is defined as the number of days since start of the sample period divided by 100. Standard errors in parentheses. * p<0.10, ** p<0.05, *** p<0.01. All regressions include controls for patient characteristics, surgery characteristics and device characteristics defined in Section 3.

Table 7: Estimation Results using Alternative Experience Variables: Dependent Variable is Ln(Duration)

Explanatory Variable	Column (1) OLS: Add nth Usage Counts		Column (2) OLS: Add 2nd to 4th Usage Dummies		Column (3) OLS: Add # of Surgeries in-between		Column (4) OLS: Add Switch Dummy		Column (5) OLS: Add Switch Variety	
	Coef.	(Std. Err.)	Coef.	(Std. Err.)	Coef.	(Std. Err.)	Coef.	(Std. Err.)	Coef.	(Std. Err.)
Total Experience/100	0.091	(0.059)	0.031	(0.051)	0.024	(0.050)	0.024	(0.051)	0.031	(0.052)
First Use Dummy										
Shell	0.190*	(0.118)	0.247**	(0.112)	0.242**	(0.113)	0.267**	(0.112)	0.246**	(0.112)
Stem	0.294***	(0.085)	0.308***	(0.080)	0.316***	(0.079)	0.308***	(0.079)	0.302***	(0.079)
Liner	0.112	(0.093)	0.098	(0.091)	0.089	(0.092)	0.135	(0.092)	0.110	(0.091)
Head	-0.018	(0.084)	-0.048	(0.079)	-0.013	(0.078)	-0.033	(0.080)	0.018	(0.078)
Second Use Dummy										
Shell			-0.008	(0.095)						
Stem			0.147*	(0.080)						
Liner			0.116	(0.091)						
Head			-0.026	(0.068)						
Third Use Dummy										
Shell			0.126	(0.090)						
Stem			0.179**	(0.082)						
Liner			-0.104	(0.089)						
Head			-0.024	(0.065)						
Fourth Use Dummy										
Shell			0.103	(0.086)						
Stem			0.130	(0.081)						
Liner			0.033	(0.093)						
Head			-0.001	(0.077)						
Nth Usage Counts/100										
Shell	-0.101	(0.065)								
Stem	0.000	(0.053)								
Liner	-0.031	(0.065)								
Head	-0.046	(0.058)								
Log(Experience Gap)										
Shell	-0.018	(0.017)	-0.014	(0.016)	-0.022	(0.019)	-0.008	(0.017)	-0.026	(0.020)
Stem	0.030**	(0.015)	0.012	(0.015)	0.043**	(0.017)	0.024*	(0.014)	0.036*	(0.019)
Liner	0.033**	(0.015)	0.025*	(0.015)	0.013	(0.018)	0.026*	(0.016)	0.020	(0.020)
Head	-0.002	(0.014)	0.005	(0.014)	-0.004	(0.016)	0.008	(0.013)	0.004	(0.018)
# of Surgeries in-between/10										
Shell					0.024*	(0.014)				
Stem					-0.009	(0.010)				
Liner					0.027*	(0.016)				
Head					0.003	(0.008)				
Device Switch Dummies										
Shell							0.011	(0.039)		
Stem							0.038	(0.039)		
Liner							0.027	(0.044)		
Head							-0.075	(0.046)		
Device Switch Variety										
Shell									0.018	(0.012)
Stem									-0.003	(0.008)
Liner									0.010	(0.011)
Head									-0.001	(0.005)
Dummies for New Device Combinations										
Shell and Liner	-0.280	(0.232)	-0.273	(0.236)	-0.275	(0.231)	-0.265	(0.232)	-0.279	(0.232)
Stem and Head	-0.127	(0.191)	-0.078	(0.194)	-0.130	(0.191)	-0.107	(0.192)	-0.130	(0.192)
Dummies for # of New Devices										
2 New Devices	0.054	(0.123)	0.053	(0.125)	0.030	(0.123)	0.013	(0.122)	0.023	(0.123)
3 New Devices	-0.133	(0.263)	-0.183	(0.267)	-0.183	(0.262)	-0.205	(0.262)	-0.187	(0.262)
4 New Devices	0.041	(0.459)	0.000	(0.462)	0.002	(0.458)	-0.083	(0.457)	0.000	(0.459)
Surgeon FE	Yes		Yes		Yes		Yes		Yes	
Quadratic Time Trend	Yes		Yes		Yes		Yes		Yes	
# of Observations	483		483		483		483		483	
Adj. R-squared	0.433		0.438		0.436		0.431		0.432	

Note: Time Trend is defined as the number of days since start of the sample period divided by 100. Standard errors are in parentheses. * p<0.10, **p<0.05, *** p<0.01. All regressions include controls for patient characteristics, surgery characteristics, and device characteristics defined in Section 3.

Online Appendix

OA.1: Test for Serial Correlation of Error Term

In section 4.1 of our paper, the baseline specification is

$$y_{st} = \beta X_{st} + \gamma e_{st} + \alpha w_{st1} + \theta \log [w_{st2}] + \varepsilon_{st}. \tag{10}$$

Here, y_{st} is the log value of duration of the surgery performed by surgeon s at time t, e_{st} is the total experience of surgeon s at time t, X_{st} is a vector of control variables, w_{st1} and w_{st2} are vectors of (observed) device-specific experience variables related to learning and forgetting for surgeon s at time t, as explained below, and ε_{st} is the error term.

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

$$\varepsilon_{st} = \rho (d_{t,t-1}) \varepsilon_{st-1} + \eta_{st}; \eta_{st} \sim iid \left(0, \sigma_\eta^2\right). \tag{11}$$

In this appendix, we show how to estimate a correlation function $\rho (d_{t,v})$ for two surgeries t and v by the same surgeon as a function of the time gap $d_{t,v} = |t - v|$ (measured in days) between them.

Define $\widetilde{\epsilon}_{st}$ as the residual from the pooled OLS presented by specification (10), $\widetilde{\epsilon}_{s\cdot}$ as the sample mean of $\widetilde{\epsilon}_{st}$ for each individual, and $\widehat{\widetilde{\epsilon}}_{st}$ as the standardized residual which is

$$\widehat{\widetilde{\epsilon}}_{st} = \frac{\widetilde{\epsilon}_{st} - \widetilde{\epsilon}_{s\cdot}}{\sigma_{\widetilde{\epsilon}}}.$$

where $\sigma_{\widetilde{\epsilon}}$ is the standard deviation of $\widetilde{\epsilon}_{st}$. A kernel-based estimate of the correlation function $\rho (d)$ is

$$\widehat{\rho} (d) = \frac{\sum_s \sum_{t,v} K (d_{t,v} - d) \widehat{\widetilde{\epsilon}}_{st} \widehat{\widetilde{\epsilon}}_{sv}}{\sum_s \sum_{t,v} K (d_{t,v} - d)} \tag{12}$$

where $K (\cdot)$ is a kernel function. We use

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

and set $b = \sigma_d$, the sample standard deviation of distances. Note that the model which can be described by both equations (10) and (11) assumes a balanced panel, but the estimator for $\rho (d)$ in equation (12) does not require a balanced panel.

Figure A1 displays the estimated correlation function $\rho (d_{t,v})$. We see that $\widehat{\rho} (0) \approx 0.25$,

implying that, even 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_{st} is not a concern even though we observe surgeons over a long time period.

OA.2: Possible Biases Caused by Left-Censored Experience Variables

We use two examples to discuss the possible biases of our estimates if we don't consider the data censoring issue in our estimation. First, consider the simple case which has only a left-censored explanatory variable:

$$y_i = \beta z_i + u_i \tag{13}$$

with $\beta > 0$. Instead of observing z_i, we observe $w_i = \min(w_i^c, z_i)$ which means we can only observe a threshold value w_i^c whenever $z_i > w_i^c$. If we run the regression,

$$y_i = bw_i + e_i,$$

then

$$\begin{aligned}
\widehat{b} &= \frac{n^{-1}\sum_i w_i y_i}{n^{-1}\sum_i w_i^2} \\
&= \frac{n^{-1}\sum_i w_i (\beta z_i + u_i)}{n^{-1}\sum_i w_i^2} \\
&= \beta\frac{n^{-1}\sum_i w_i z_i}{n^{-1}\sum_i w_i^2} + \frac{n^{-1}\sum_i w_i u_i}{n^{-1}\sum_i w_i^2} \\
&= \beta\frac{n^{-1}\sum_i w_i (z_i + w_i - w_i)}{n^{-1}\sum_i w_i^2} + \frac{n^{-1}\sum_i w_i u_i}{n^{-1}\sum_i w_i^2} \\
&= \beta\left[\frac{n^{-1}\sum_i w_i (z_i - w_i)}{n^{-1}\sum_i w_i^2} + \frac{n^{-1}\sum_i w_i^2}{n^{-1}\sum_i w_i^2}\right] + \frac{n^{-1}\sum_i w_i u_i}{n^{-1}\sum_i w_i^2} \\
&= \beta\left[\frac{n^{-1}\sum_{i:z_i > w_i^c} w_i (z_i - w_i^c)}{n^{-1}\sum_i w_i^2} + 1\right] + \frac{n^{-1}\sum_i w_i u_i}{n^{-1}\sum_i w_i^2};
\end{aligned}$$

$$\begin{aligned}
plim\widehat{b} &= \beta + \beta\frac{plim\left(n^{-1}\sum_{i:x_i > w_i^c} w_i (z_i - w_i^c)\right)}{plim\left(n^{-1}\sum_i w_i^2\right)} + \frac{plim\left(n^{-1}\sum_i w_i u_i\right)}{plim\left(n^{-1}\sum_i w_i^2\right)} \\
&= \beta + \beta\frac{plim\left(n^{-1}\sum_{i:x_i > w_i^c} w_i (z_i - w_i^c)\right)}{plim\left(n^{-1}\sum_i w_i^2\right)} > \beta.
\end{aligned}$$

For this simple case, we know that, if we do not take the data left-censoring issue into account, then the estimate of the parameter has bias in the same direction as the true sign of the coefficient. However, if we have other covariates which have no censoring issues, we may

not be able to sign the bias. Now, we rewrite the equation (13) as

$$y_i = X_i\alpha + \beta z_i + u_i$$

where X_i includes variables which are not left-censored and z_i is the left-censored variable with threshold value w_i^c. Like the first example, we can observe only $w_i = \min(w_i^c, z_i)$, and we run the regression,

$$y_i = X_i a + b w_i + e_i.$$

By the same logic and after some algebra, we have

$$\begin{aligned}
\begin{pmatrix} \widehat{a} \\ \widehat{b} \end{pmatrix} &= \begin{pmatrix} n^{-1}\sum_i X_i'X_i & n^{-1}\sum_i X_i'w_i \\ n^{-1}\sum_i w_i'X_i & n^{-1}\sum_i w_i^2 \end{pmatrix}^{-1} \begin{pmatrix} n^{-1}\sum_i X_i'y_i \\ n^{-1}\sum_i w_i'y_i \end{pmatrix} \\
&= \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} n^{-1}\sum_i X_i'X_i & n^{-1}\sum_i X_i'w_i \\ n^{-1}\sum_i w_i'X_i & n^{-1}\sum_i w_i^2 \end{pmatrix}^{-1} \\
&\quad \begin{pmatrix} \beta n^{-1}\sum_i X_i'(z_i - w_i) + n^{-1}\sum_i X_i'u_i \\ \beta n^{-1}\sum_i w_i'(z_i - w_i) + n^{-1}\sum_i w_i'u_i \end{pmatrix};
\end{aligned}$$

$$\begin{aligned}
plim\begin{pmatrix} \widehat{a} \\ \widehat{b} \end{pmatrix} &= \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + plim\begin{pmatrix} n^{-1}\sum_i X_i'X_i & n^{-1}\sum_i X_i'w_i \\ n^{-1}\sum_i w_i'X_i & n^{-1}\sum_i w_i^2 \end{pmatrix}^{-1} \\
&\quad plim\begin{pmatrix} \beta n^{-1}\sum_i X_i'(z_i - w_i) + n^{-1}\sum_i X_i'u_i \\ \beta n^{-1}\sum_i w_i'(z_i - w_i) + n^{-1}\sum_i w_i'u_i \end{pmatrix} \\
&= \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \beta plim\begin{pmatrix} n^{-1}\sum_i X_i'X_i & n^{-1}\sum_i X_i'w_i \\ n^{-1}\sum_i w_i'X_i & n^{-1}\sum_i w_i^2 \end{pmatrix}^{-1} \\
&\quad plim\begin{pmatrix} n^{-1}\sum_{i:z_i>w_i^c} X_i'(z_i - w_i^c) \\ n^{-1}\sum_{i:z_i>w_i^c} w_i'(z_i - w_i^c) \end{pmatrix}.
\end{aligned}$$

In general, all we can say is that

$$plim\left(n^{-1}\sum_{i:z_i>w_i^c} w_i'(z_i - w_i^c)\right) > 0.$$

But this is not enough to sign any of the biases. However, if we can assume that

$$plim\left[n^{-1}\sum_{i:z_i>w_i^c} X_i'(z_i - w_i^c)\right] = 0$$

or is small and

$$plim\left[n^{-1}\sum_i X_i' w_i\right] = 0$$

or is small, then,

$$\begin{aligned} AsyBias\begin{pmatrix}\widehat{a}\\ \widehat{b}\end{pmatrix} &\approx \beta plim\begin{pmatrix} n^{-1}\sum_i X_i'X_i & 0\\ 0' & n^{-1}\sum_i w_i^2\end{pmatrix}^{-1} \\ & \quad plim\begin{pmatrix}0\\ n^{-1}\sum_{i:z_i>w_i^c} w_i'\left(z_i - w_i^c\right)\end{pmatrix} \\ &= \beta plim\frac{n^{-1}\sum_{i:z_i>w_i^c} w_i'\left(z_i - w_i^c\right)}{n^{-1}\sum_i w_i^2} \end{aligned}$$

which has the same sign as β.

OA.3 Endogeneity Test

To determine whether our models havepotential endogeneity issues, we model a surgeon's decision to use a new device version using a probit specification and test whether there are common unobserved factors driving both duration of surgery and the decision to use a new device version. A surgeon's decision to use a new device version for a surgery for each of the four key device types can be represented as

$$\begin{aligned} m^*_{sjt1} &= \gamma_j X_{st} + \nu_{sjt}, \\ \nu_{sjt} &\sim iidN\left(0,1\right), \\ m_{sjt1} &= 1\left(m^*_{sjt1} > 0\right) \end{aligned} \tag{14}$$

where X_{st} is the vector of observed exogenous control variables, and ν_{sjt} is the error term representing unobserved factors impacting new device choice decision. The issue of interest is whether the error term of the duration equation (10), ε_{st}, is correlated with any of the errors $\nu_{st} = (v_{s1t}, v_{s2t}, v_{s3t}, v_{s4t})$, from the "new-device-choice" probit model presented in equation (14) for each device. More formally, we assume that, for surgeon s during surgery t,

$$\begin{pmatrix}\varepsilon_{st}\\ \nu_{s1t}\\ \vdots\\ \nu_{sJt}\end{pmatrix} \sim iidN\left[0, \begin{pmatrix}\sigma_\varepsilon^2 & \rho_{\varepsilon 1\nu} & \cdots & \rho_{\varepsilon J\nu}\\ \rho_{\varepsilon 1\nu} & 1 & \cdots & \rho_\nu\\ \vdots & \vdots & \ddots & \vdots\\ \rho_{\varepsilon J\nu} & \rho_\nu & \cdots & 1\end{pmatrix}\right]$$

where $\rho_{\varepsilon j\nu}$ is the covariance between ε_{st} and ν_{sjt}. $c_{\varepsilon j\nu} = \frac{\rho_{\varepsilon j\nu}}{\sigma_\varepsilon}$ is the corresponding correlation coefficient, and we construct the hypothesis of interest as $H_0 : c_{\varepsilon j\nu} = 0$ against $H_A : c_{\varepsilon j\nu} \neq 0$.

The intuition for this test is straightforward. If there are common unobserved factors driving both duration of surgery and the choice to use a new device version, then ε_{st} and ν_{sjt} should be correlated and $c_{\varepsilon j\nu}$ should be greater than zero. Therefore, we first construct generalized residuals of duration equation and device choice equation, ε_{st} and ν_{sjt}. Then we calculate the correlation coefficient, $c_{\varepsilon j\nu}$, of those two random variables, use it as the test statistic, and determine whether it is statistically significantly greater than zero. Particularly, for baseline specification with OLS estimation, we define $\widehat{\varepsilon}_{st}$ as the generalized residual for equation (10) and

$$\begin{aligned}\widehat{\nu}_{sjt} &= E\left[\nu_{sjt} \mid w_{sjt1}, X_{st}\right] \\ &= \begin{cases} \frac{\phi\left(\widehat{\gamma}_j X_{st}\right)}{\Phi\left(\widehat{\gamma}_j X_{st}\right)} & \text{if } w_{sjt1}=1 \\ \frac{-\phi\left(\widehat{\gamma}_j X_{st}\right)}{1-\Phi\left(\widehat{\gamma}_j X_{st}\right)} & \text{if } w_{sjt1}=0 \end{cases}\end{aligned}$$

as the generalized residual for equation (14) (e.g., Gourieroux et al., 1987; Dean et al., 2015)

For MSLE, the probit model in (14) can be written as

$$\begin{aligned} m^*_{sjt1} &= \gamma_j X_{st} + \nu_{sjt}, \\ \nu_{sjt} &\sim iidN\left(0,1\right), \\ m_{sjt1} &= 1 \Rightarrow \left(m^*_{sjt1} > 0 \text{ and } d_{sjt}=1\right), \text{or } \left(m^*_{sjt1} < 0 \text{ and } d_{sjt}=0\right), \\ m_{sjt1} &= 0 \Rightarrow m^*_{sjt1} < 0. \end{aligned} \tag{15}$$

where $d_{sjt} \sim Bernoulli\left(p_{sjt}\right)$ is the event that the observed first usage is the true first usage (i.e., $w_{sjt1} = z_{sjt1} = 1$) and p_{sjt} the probability of the observed first usage being the true first usage. p_{sjt} is calculated as the proportion of simulated $z^r_{sjt1} = 1$ when $w_{sjt1} = 1$.[42] Note that above, if $w_{sjt1} = 0$, then it cannot be the case that $z_{sjt1} = 1$, as a prior usage has been observed. On the other hand, if $w_{sjt1} = 1$, then either $z_{sjt1} = 1$ and therefore $d_{sjt} = 1$ (this happens if the utility surgeon s gets from using a new device version for device-type j exceeds zero on his t^{th} surgery), or his utility is less than zero and $d_{sjt} = 0$.

The MSLE probit estimator of γ_j, $\widehat{\gamma_j}$, maximizes

$$\begin{aligned} L_j = \sum_{s,t} \quad & \left\{ w_{sjt1} \log\left[p_{sjt} \Phi\left(\gamma_j X_{st}\right) + \left(1-p_{sjt}\right)\left(1-\Phi\left(\gamma_j X_{st}\right)\right)\right] \right. \\ & \left. + \left(1-w_{sjt1}\right) \log\left[1-\Phi\left(\gamma_j X_{st}\right)\right] \right\}. \end{aligned}$$

[42] As in the description of simulation details in Section A.2, for surgeries using observed new devices, we compare the simulated experience gap with the observed time gap between the sample start day and the surgery day. If the simulated value is smaller than the observed value, the simulated first usage dummy, $z^r_{sjt1} = 1$. Otherwise, $z^r_{sjt1} = 0$.

Next, the generalized residual $\widehat{\nu}_{sjt}$ when $w_{sjt1} = 1$ is

$$\begin{aligned}
\widehat{\nu}_{sjt} &= E\left[\nu_{sjt} \mid w_{sjt1} = 1, X_{st}\right] \\
&= \frac{p_{sjt} E\left(\nu_{sjt} \mid m^*_{sjt1} > 0\right) \Phi\left(\widehat{\gamma}_j X_{st}\right) + \left(1 - p_{sjt}\right) E\left(\nu_{sjt} \mid m^*_{sjt1} < 0\right) \left[1 - \Phi\left(\widehat{\gamma}_j X_{st}\right)\right]}{\Pr\left[w_{sjt1} = 1 \mid X_{st}\right]} \\
&= \frac{p_{sjt} \frac{\phi\left(\widehat{\gamma}_j X_{st}\right)}{\Phi\left(\widehat{\gamma}_j X_{st}\right)} \Phi\left(\widehat{\gamma}_j X_{st}\right) - \left(1 - p_{sjt}\right) \frac{\phi\left(\widehat{\gamma}_j X_{st}\right)}{1 - \Phi\left(\widehat{\gamma}_j X_{st}\right)} \left[1 - \Phi\left(\widehat{\gamma}_j X_{st}\right)\right]}{p_{sjt} \Phi\left(\widehat{\gamma}_j X_{st}\right) + \left(1 - p_{sjt}\right) \left[1 - \Phi\left(\widehat{\gamma}_j X_{st}\right)\right]} \\
&= \frac{p_{sjt} \phi\left(\widehat{\gamma}_j X_{st}\right) - \left(1 - p_{sjt}\right) \phi\left(\widehat{\gamma}_j X_{st}\right)}{p_{sjt} \Phi\left(\widehat{\gamma}_j X_{st}\right) + \left(1 - p_{sjt}\right) \left[1 - \Phi\left(\widehat{\gamma}_j X_{st}\right)\right]} \\
&= \frac{\left(2 p_{sjt} - 1\right) \phi\left(\widehat{\gamma}_j X_{st}\right)}{p_{sjt} \Phi\left(\widehat{\gamma}_j X_{st}\right) + \left(1 - p_{sjt}\right) \left[1 - \Phi\left(\widehat{\gamma}_j X_{st}\right)\right]}.
\end{aligned}$$

Note that, if an observed first usage of a device version occurs towards the end of the sample period, then it is more likely that this observed first usage represents a true first usage. In this case, $p_{sjt} \to 1$, and the generalized residual converges to

$$\widehat{\nu}_{sjt} \to \frac{\phi\left(\widehat{\gamma}_j X_{st}\right)}{\Phi\left(\widehat{\gamma}_j X_{st}\right)}.$$

On the other hand, if an observed first usage of a device version by a surgeon occurs towards the beginning of the sample period, then it is less likely that the observed first usage is a true first usage. In this case, $p_{sjt} \to 0$, and the generalized residual converges to

$$\widehat{\nu}_{sjt} \to \frac{-\phi\left(\widehat{\gamma}_j X_{st}\right)}{\left[1 - \Phi\left(\widehat{\gamma}_j X_{st}\right)\right]}.$$

When $w_{sjt1} = 0$, then $\widehat{\nu}_{sjt}$ is

$$\begin{aligned}
\widehat{\nu}_{sjt} &= E\left[\nu_{sjt} \mid w_{sjt1} = 0, X_{st}\right] \\
&= E\left(\nu_{sjt} \mid m^*_{sjt1} < 0\right) = -\frac{\phi\left(\widehat{\gamma}_j X_{st}\right)}{1 - \Phi\left(\widehat{\gamma}_j X_{st}\right)}.
\end{aligned}$$

Next we can construct a correlation term either for each device j or for all devices together. The device-specific correlation term is

$$\widehat{c}_j = \frac{n_j^{-1} \sum_{st} \widehat{\varepsilon}_{st} \widehat{\nu}_{sjt}}{\sqrt{\left(n_j^{-1} \sum_{st} \widehat{\varepsilon}^2_{st}\right) \left(n_j^{-1} \sum_{st} \widehat{\nu}^2_{sjt}\right)}}$$

where n_j is the total number of surgeries using device j, and the correlation term for all devices

together is

$$\widehat{c} = \frac{n^{-1}\sum_{st}\widehat{\varepsilon}_{st}\overline{\nu}_{st}}{\sqrt{\left(n^{-1}\sum_{st}\widehat{\varepsilon}_{st}^{2}\right)\left(n^{-1}\sum_{st}\overline{\nu}_{st}^{2}\right)}}$$

where n is the total number of surgeries in the sample and $\overline{\nu}_{st} = J^{-1}\sum_{j}\widehat{\nu}_{sjt}$. Under the null hypothesis,

$$plim\widehat{c}_{j} \propto plim\left(n_{j}^{-1}\sum_{st}\widehat{\varepsilon}_{st}\widehat{\nu}_{sjt}\right) = 0 \tag{16}$$

where the proportionality factor is the plim of the denominator.

In order to actually use the test statistic, one must know something about the sample distribution of the test statistic. Instead of deriving the asymptotic distribution for our test statistic officially, we simulate the small sample distribution of the test statistic and then use simulated critical values to perform the test. In particular, define $\widetilde{\varepsilon}$ as the sample vector of $\widehat{\varepsilon}_{st}$ and $\widetilde{\nu}_{j}$ analogously for device j. Define $\widetilde{\nu}_{j}^{r}$ as the rth random reordering of $\widetilde{\nu}_{j}$.[43] If $\widetilde{\varepsilon}_{st} \sim iidF_{\varepsilon}$, $\widetilde{\nu}_{sjt} \sim iidF_{\nu j}$, and $\widetilde{\varepsilon} \perp \widetilde{\nu}_{j}$, then $\widetilde{\nu}_{j}^{r} \sim iidF_{\nu j}$ and $\widetilde{\varepsilon} \perp \widetilde{\nu}_{j}^{r}$ as well. Define

$$\widehat{c}_{j}^{r} = \frac{n_{j}^{-1}\sum_{st}\widetilde{\varepsilon}_{st}\widetilde{\nu}_{sjt}^{r}}{\sqrt{\left(n_{j}^{-1}\sum_{st}\widetilde{\varepsilon}_{st}^{2}\right)\left(n_{j}^{-1}\sum_{st}\left(\widetilde{\nu}_{sjt}^{r}\right)^{2}\right)}}$$

as a single draw of $\widehat{c}_{j}$ and repeat R independent times. Then find the 2.5% and 97.5% percentiles of $\left\{\widehat{c}_{j}^{r}\right\}_{r=1}^{R}$. These are the 5% critical values for the test statistic; reject H_0 iff $\widehat{c}_{j}$ falls outside the two critical values.

Table A1-1 and Table A1-2 list the test results which show no significant correlation between these error terms. We therefore believe that our main results are robust to this type of potential endogeneity.

[43] Consider a vector of variables $\nu = (\nu_1, \nu_2, .., \nu_n)'$. Simulate $\xi^r = (\xi_1^r, \xi_2^r, .., \xi_n^r)'$ as a vector of random numbers where $\xi_k^r \sim iidU(0,1)$, and construct ν^r as ν reordered in the same way as ξ^r if sorted from smallest to largest; i.e., $\nu_m^r = \nu_k$ iff ξ_k^r is the m'th smallest element of ξ^r. ν_m^r is a random permutation of ν and independent across $r = 1, 2, .., R$.

OA.4: Tables and Figures

Table A1-1: Endogeneity Test for Fist Use Dummies – OLS

Device		Shell	Stem	Liner	Head	New Device
# of Observations		414	408	349	468	483
Correlation Term		-0.005	0.040	-0.022	0.010	0.044
α=1%	CV-Lower	-0.141	-0.125	-0.131	-0.114	-0.107
	CV-Upper	0.142	0.132	0.136	0.109	0.124
	Rejection	NO	NO	NO	NO	NO
α =5%	CV-Lower	-0.092	-0.097	-0.101	-0.092	-0.089
	CV-Upper	0.096	0.097	0.110	0.084	0.093
	Rejection	NO	NO	NO	NO	NO
α =10%	CV-Lower	-0.084	-0.083	-0.091	-0.075	-0.073
	CV-Upper	0.083	0.078	0.085	0.074	0.077
	Rejection	NO	NO	NO	NO	NO

Note: CV -- Critical Value; Null Hypothesis: Correlation Term =0

Table A1-2: Endogeneity Test for Fist Use Dummies - MSLE

Device		Shell	Stem	Liner	Head	New Device
# of Observations		414	408	349	468	483
Correlation Term		-0.015	0.045	-0.018	0.012	0.049
α =1%	CV-Lower	-0.125	-0.128	-0.141	-0.110	-0.129
	CV-Upper	0.134	0.129	0.136	0.117	0.118
	Rejection	NO	NO	NO	NO	NO
α =5%	CV-Lower	-0.094	-0.093	-0.102	-0.089	-0.091
	CV-Upper	0.094	0.102	0.117	0.094	0.090
	Rejection	NO	NO	NO	NO	NO
α =10%	CV-Lower	-0.077	-0.081	-0.087	-0.073	-0.072
	CV-Upper	0.079	0.084	0.090	0.077	0.077
	Rejection	NO	NO	NO	NO	NO

Note: CV -- Critical Value; Null Hypothesis: Correlation Term =0

Table A2: Estimation Results for the Baseline Specification: Dependent Variable is Ln(Duration)

Explanatory Variable	Column (3) OLS: Add Device Experience		Column (4) OLS: Add Dummies for New Device Combinations and # of New Devices		Column (5) MSLE: Add Device Experience		Column (6) MSLE: Add Dummies for New Device Combinations and # of New Devices	
	Coef.	(Std. Err.)	Coef.	(Std. Err.)	Coef.	(Std. Err.)	Coef.	(Std. Err.)
Total Experience/100	0.027	(0.050)	0.030	(0.050)	0.026	(0.061)	0.030	(0.064)
First Use Dummy								
Shell	0.131	(0.091)	0.262**	(0.111)	0.155*	(0.087)	0.276**	(0.115)
Stem	0.262***	(0.073)	0.296***	(0.078)	0.286***	(0.060)	0.324***	(0.064)
Liner	0.078	(0.076)	0.129	(0.090)	0.089	(0.076)	0.125	(0.087)
Head	-0.029	(0.068)	0.007	(0.075)	-0.031	(0.076)	0.014	(0.084)
Log(Experience Gap)								
Shell	-0.009	(0.015)	-0.009	(0.015)	-0.009	(0.017)	-0.010	(0.016)
Stem	0.030**	(0.013)	0.030**	(0.013)	0.029**	(0.013)	0.029**	(0.013)
Liner	0.031**	(0.015)	0.030**	(0.015)	0.030*	(0.016)	0.030*	(0.016)
Head	0.001	(0.012)	0.002	(0.012)	0.001	(0.014)	0.002	(0.014)
Dummies for New Device Combinations								
Shell and Liner			-0.266	(0.232)			-0.266	(0.859)
Stem and Head			-0.116	(0.191)			-0.182	(0.190)
Dummies for # of New Devices								
2 New Devices			0.016	(0.122)			-0.012	(0.117)
3 New Devices			-0.203	(0.261)			-0.327	(1.013)
4 New Devices			-0.069	(0.456)			0.180	(1.783)
Patient Characteristics								
Male	0.107***	(0.030)	0.105***	(0.030)	0.106***	(0.032)	0.105***	(0.032)
BMI/100	0.381*	(0.214)	0.469**	(0.218)	0.387*	(0.190)	0.466**	(0.220)
Age/100	-0.445***	(0.125)	-0.401***	(0.126)	-0.440***	(0.125)	-0.408***	(0.131)
ASA Average	0.002	(0.033)	-0.003	(0.033)	0.003	(0.035)	0.002	(0.036)
# of Comorbidities	0.030**	(0.012)	0.029**	(0.012)	0.029**	(0.013)	0.028**	(0.013)
Surgery Characteristics								
Both Legs	0.706***	(0.155)	0.719***	(0.155)	0.710	(0.559)	0.719	(0.545)
Reason: Revision	0.286***	(0.077)	0.292***	(0.078)	0.285***	(0.067)	0.293***	(0.069)
Reason: Avascular Necrosis	-0.073	(0.066)	-0.076	(0.066)	-0.072	(0.082)	-0.075	(0.082)
Reason: Displasia	0.068	(0.073)	0.077	(0.073)	0.068	(0.089)	0.081	(0.093)
Reason: Arthritis	-0.044	(0.071)	-0.054	(0.071)	-0.041	(0.081)	-0.043	(0.083)
Reason: Severe Arthritis	0.108	(0.098)	0.085	(0.098)	0.107	(0.118)	0.091	(0.117)
Reason: End Stage Arthritis	0.033	(0.092)	0.019	(0.092)	0.035	(0.117)	0.025	(0.119)
Reason: Fracture	-0.036	(0.077)	-0.040	(0.078)	-0.038	(0.081)	-0.041	(0.081)
Reason: Other	-0.019	(0.105)	-0.055	(0.106)	-0.022	(0.106)	-0.047	(0.106)
Reasons for revision	-0.026	(0.082)	-0.042	(0.082)	-0.026	(0.082)	-0.035	(0.082)
Use: shell	0.023	(0.078)	0.038	(0.078)	0.027	(0.076)	0.036	(0.076)
Use: stem	-0.074	(0.079)	-0.080	(0.080)	-0.077	(0.076)	-0.079	(0.075)
Use: liner	-0.057	(0.061)	-0.062	(0.061)	-0.057	(0.063)	-0.059	(0.062)
Use: head	-0.072	(0.099)	-0.062	(0.099)	-0.069	(0.101)	-0.059	(0.098)
Time Trend	0.072**	(0.029)	0.065**	(0.030)	0.070*	(0.035)	0.066*	(0.035)
Quadratic Time Trend	-0.003	(0.003)	-0.003	(0.003)	-0.003	(0.003)	-0.003	(0.003)
Unihead	-0.064	(0.075)	-0.061	(0.076)	-0.067	(0.078)	-0.056	(0.079)
Cemented	0.008	(0.052)	-0.003	(0.053)	0.010	(0.054)	0.005	(0.055)
Company 2	0.107	(0.076)	0.109	(0.076)	0.104	(0.068)	0.102	(0.069)
Company 3	0.078	(0.092)	0.087	(0.093)	0.076	(0.080)	0.063	(0.083)
Company: 4	0.210**	(0.097)	0.219**	(0.097)	0.210**	(0.083)	0.203**	(0.085)
Multiple Companies Indicator	0.146**	(0.071)	0.140*	(0.072)	0.150**	(0.061)	0.142**	(0.062)
Surgeon: 2	0.142	(0.087)	0.140	(0.089)	0.139	(0.090)	0.153*	(0.093)
Surgeon: 3	0.147*	(0.080)	0.144*	(0.081)	0.143*	(0.085)	0.159*	(0.086)
Surgeon: 4	0.193	(0.119)	0.201*	(0.122)	0.189	(0.133)	0.216	(0.141)
Surgeon FE	Yes		Yes		Yes		Yes	
Quadratic Time Trend	Yes		Yes		Yes		Yes	
# of Observations	483		483		483		483	
Adj. R-squared	0.429		0.432		-		-	

Note: Time Trend is defined as the number of days since start of the sample period divided by 100. Standard errors are in parentheses.
* p<0.10, ** p<0.05, *** p<0.01. All regressions include controls for patient characteristics, surgery characteristics, and device characteristics defined in Section 3.

Table A3: Estimation Results using Alternative Experience Variables: Dependent Variable is Ln(Duration)

Explanatory Variable	Column (1) OLS: Add nth Usage Counts		Column (2) OLS: Add 2nd to 4th Usage Dummies		Column (3) OLS: Add # of Surgeries in-between		Column (4) OLS: Add Switch Dummy		Column (5) OLS: Add Switch Variety	
	Coef.	(Std. Err.)	Coef.	(Std. Err.)	Coef.	(Std. Err.)	Coef.	(Std. Err.)	Coef.	(Std. Err.)
Total Experience/100	0.091	(0.059)	0.031	(0.051)	0.024	(0.050)	0.024	(0.051)	0.031	(0.052)
Shell 1st Use Dummy	0.190*	(0.118)	0.247**	(0.112)	0.242**	(0.113)	0.267**	(0.112)	0.246**	(0.112)
Stem 1st Use Dummy	0.294***	(0.085)	0.308***	(0.080)	0.316***	(0.079)	0.308***	(0.079)	0.302***	(0.079)
Liner 1st Use Dummy	0.112	(0.093)	0.098	(0.091)	0.089	(0.092)	0.135	(0.092)	0.110	(0.091)
Head 1st Use Dummy	-0.018	(0.084)	-0.048	(0.079)	-0.013	(0.078)	-0.033	(0.080)	0.018	(0.078)
Shell 2nd Use Dummy			-0.008	(0.095)						
Stem 2nd Use Dummy			0.147*	(0.080)						
Liner 2nd Use Dummy			0.116	(0.091)						
Head 2nd Use Dummy			-0.026	(0.068)						
Shell 3rd Use Dummy			0.126	(0.090)						
Stem 3rd Use Dummy			0.179**	(0.082)						
Liner 3rd Use Dummy			-0.104	(0.089)						
Head 3rd Use Dummy			-0.024	(0.065)						
Shell 4th Use Dummy			0.103	(0.086)						
Stem 4th Use Dummy			0.130	(0.081)						
Liner 4th Use Dummy			0.033	(0.093)						
Head 4th Use Dummy			-0.001	(0.077)						
Shell nth Usage Counts/100	-0.101	(0.065)								
Stem nth Usage Counts/100	0.000	(0.053)								
Liner nth Usage Counts/100	-0.031	(0.065)								
Head nth Usage Counts/100	-0.046	(0.058)								
Shell Log(Experience Gap)	-0.018	(0.017)	-0.014	(0.016)	-0.022	(0.019)	-0.008	(0.017)	-0.026	(0.020)
Stem Log(Experience Gap)	0.030**	(0.015)	0.012	(0.015)	0.043**	(0.017)	0.024*	(0.014)	0.036*	(0.019)
Liner Log(Experience Gap)	0.033**	(0.015)	0.025*	(0.015)	0.013	(0.018)	0.026*	(0.016)	0.020	(0.020)
Head Log(Experience Gap)	-0.002	(0.014)	0.005	(0.014)	-0.004	(0.016)	0.008	(0.013)	0.004	(0.018)
Shell # of Surgeries in-between/10					0.024*	(0.014)				
Stem # of Surgeries in-between/10					-0.009	(0.010)				
Liner # of Surgeries in-between/10					0.027*	(0.016)				
Head # of Surgeries in-between/10					0.003	(0.008)				
Shell Device Switch Dummies							0.011	(0.039)		
Stem Device Switch Dummies							0.038	(0.039)		
Liner Device Switch Dummies							0.027	(0.044)		
Head Device Switch Dummies							-0.075	(0.046)		
Shell Device Switch Variety									0.018	(0.012)
Stem Device Switch Variety									-0.003	(0.008)
Liner Device Switch Variety									0.010	(0.011)
Head Device Switch Variety									-0.001	(0.005)
Shell and Liner Comb. Dummy	-0.280	(0.232)	-0.273	(0.236)	-0.275	(0.231)	-0.265	(0.232)	-0.279	(0.232)
Stem and Head Comb. Dummy	-0.127	(0.191)	-0.078	(0.194)	-0.130	(0.191)	-0.107	(0.192)	-0.130	(0.192)
2 New Devices Dummy	0.054	(0.123)	0.053	(0.125)	0.030	(0.123)	0.013	(0.122)	0.023	(0.123)
3 New Devices Dummy	-0.133	(0.263)	-0.183	(0.267)	-0.183	(0.262)	-0.205	(0.262)	-0.187	(0.262)
4 New Devices Dummy	0.041	(0.459)	0.000	(0.462)	0.002	(0.458)	-0.083	(0.457)	0.000	(0.459)
Male	0.111***	(0.030)	0.109***	(0.030)	0.103***	(0.030)	0.107***	(0.030)	0.108***	(0.030)
BMI/100	0.456**	(0.219)	0.457**	(0.218)	0.481**	(0.218)	0.461**	(0.219)	0.480**	(0.219)
Age/100	-0.378***	(0.128)	-0.397***	(0.127)	-0.385***	(0.126)	-0.414***	(0.127)	-0.392***	(0.127)
ASA Average	-0.005	(0.033)	-0.005	(0.033)	-0.005	(0.033)	-0.001	(0.033)	-0.005	(0.033)
# of Comorbidities	0.031**	(0.012)	0.030**	(0.012)	0.027**	(0.012)	0.028**	(0.012)	0.028**	(0.012)
Both Legs	0.728***	(0.155)	0.751***	(0.155)	0.716***	(0.154)	0.691***	(0.156)	0.714***	(0.155)
Reason: Revision	0.298***	(0.078)	0.264***	(0.079)	0.299***	(0.078)	0.291***	(0.078)	0.292***	(0.078)
Reason: Avascular Necrosis	-0.064	(0.066)	-0.075	(0.066)	-0.080	(0.066)	-0.082	(0.066)	-0.083	(0.066)
Reason: Displasia	0.068	(0.074)	0.097	(0.073)	0.069	(0.073)	0.080	(0.073)	0.070	(0.073)
Reason: Arthritis	-0.042	(0.072)	-0.054	(0.072)	-0.056	(0.072)	-0.064	(0.072)	-0.060	(0.072)
Reason: Severe Arthritis	0.086	(0.099)	0.083	(0.099)	0.083	(0.098)	0.083	(0.099)	0.080	(0.099)
Reason: End Stage Arthritis	0.028	(0.092)	0.025	(0.092)	0.015	(0.092)	0.009	(0.092)	0.014	(0.092)
Reason: Fracture	-0.052	(0.079)	-0.060	(0.079)	-0.063	(0.079)	-0.047	(0.078)	-0.051	(0.078)
Reason: Other	-0.069	(0.107)	-0.050	(0.107)	-0.083	(0.107)	-0.061	(0.107)	-0.073	(0.107)
Reasons for revision	-0.059	(0.083)	-0.052	(0.083)	-0.066	(0.082)	-0.051	(0.083)	-0.053	(0.082)
Use: shell	0.097	(0.086)	0.038	(0.080)	0.030	(0.079)	0.025	(0.079)	0.032	(0.079)
Use: stem	-0.082	(0.088)	-0.095	(0.080)	-0.090	(0.080)	-0.086	(0.081)	-0.078	(0.080)
Use: liner	-0.049	(0.067)	-0.045	(0.062)	-0.033	(0.063)	-0.059	(0.063)	-0.047	(0.062)
Use: head	-0.021	(0.107)	-0.072	(0.099)	-0.046	(0.099)	-0.020	(0.102)	-0.063	(0.099)
Time Trend	0.057*	(0.030)	0.079**	(0.031)	0.062**	(0.029)	0.063**	(0.030)	0.064**	(0.030)
Quadratic Time Trend	-0.003	(0.003)	-0.004	(0.003)	-0.002	(0.003)	-0.002	(0.003)	-0.003	(0.003)
Unihead	-0.079	(0.076)	-0.065	(0.076)	-0.082	(0.076)	-0.065	(0.076)	-0.074	(0.076)
Cemented	-0.016	(0.053)	0.018	(0.054)	-0.015	(0.053)	-0.001	(0.053)	-0.006	(0.053)
Company 2	0.130*	(0.077)	0.116	(0.077)	0.140*	(0.077)	0.118	(0.077)	0.135*	(0.077)
Company 3	0.086	(0.093)	0.127	(0.094)	0.098	(0.093)	0.091	(0.093)	0.098	(0.093)
Company 4	0.218**	(0.097)	0.282***	(0.101)	0.219**	(0.097)	0.233**	(0.098)	0.232**	(0.098)
Multiple Companies Indicator	0.132*	(0.072)	0.153**	(0.073)	0.134*	(0.073)	0.146**	(0.072)	0.145**	(0.072)
Surgeon: 2	0.174*	(0.091)	0.102	(0.091)	0.175*	(0.091)	0.124	(0.090)	0.171*	(0.092)
Surgeon: 3	0.168**	(0.082)	0.101	(0.084)	0.174**	(0.083)	0.135*	(0.082)	0.179**	(0.083)
Surgeon: 4	0.268**	(0.127)	0.151	(0.124)	0.242**	(0.123)	0.192	(0.123)	0.246*	(0.127)
Surgeon FE	Yes		Yes		Yes		Yes		Yes	
Quadratic Time Trend	Yes		Yes		Yes		Yes		Yes	
# of Observations	483		483		483		483		483	
Adj. R-squared	0.433		0.438		0.436		0.431		0.432	

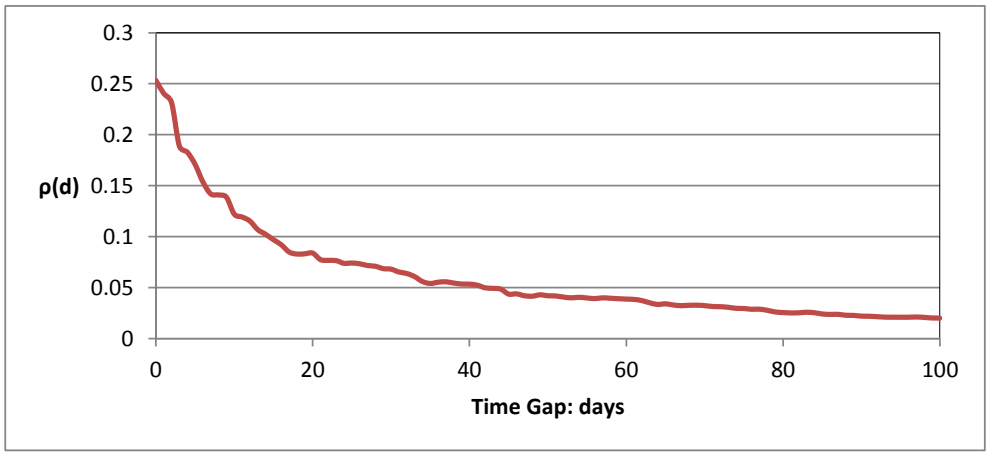
Note: Time Trend is defined as the number of days since start of the sample period divided by 100. Standard errors are in parentheses. * p<0.10, **p<0.05, *** p<0.01. All regressions include controls for patient characteristics, surgery characteristics and device characteristics defined in Section 3.

Table A4: Descriptive Statistics

Variable	# of Obs.	Mean	Std. Dev.
Duration (minutes)	483	164.98	70.47
Total Experience	483	141.97	104.56
Shell			
1st Use Dummy	414	0.05	0.21
2nd Use Dummy	414	0.04	0.19
3rd Use Dummy	414	0.04	0.19
4th Use Dummy	414	0.04	0.20
nth Usage Count	414	41.36	47.37
Stem			
1st Use Dummy	408	0.09	0.29
2nd Use Dummy	408	0.06	0.24
3rd Use Dummy	408	0.05	0.22
4th Use Dummy	408	0.04	0.21
nth Usage Count	408	52.56	60.62
Liner			
1st Use Dummy	349	0.08	0.28
2nd Use Dummy	349	0.04	0.20
3rd Use Dummy	349	0.04	0.20
4th Use Dummy	349	0.04	0.19
nth Usage Count	349	29.22	32.68
Head			
1st Use Dummy	468	0.10	0.30
2nd Use Dummy	468	0.08	0.27
3rd Use Dummy	468	0.07	0.26
4th Use Dummy	468	0.04	0.20
nth Usage Count	468	34.75	47.04
Shell			
Experience Gap	394	24.44	47.87
# of Surgeries inbetween	394	6.27	15.12
Device Switch Dummy	394	0.65	0.48
Device Switch Variety	394	1.61	2.07
Stem			
Experience Gap	371	30.08	64.57
# of Surgeries inbetween	371	8.44	23.03
Device Switch Dummy	371	0.59	0.49
Device Switch Variety	371	1.94	3.34
Liner			
Experience Gap	320	26.02	52.08
# of Surgeries inbetween	320	7.06	14.81
Device Switch Dummy	320	0.72	0.45
Device Switch Variety	320	1.93	2.33
Head			
Experience Gap	423	42.46	85.45
# of Surgeries inbetween	423	11.25	26.20
Device Switch Dummy	423	0.83	0.38
Device Switch Variety	423	3.86	5.41
Shell and Liner	483	0.01	0.11
Stem and Head	483	0.02	0.14
2 New Devices	483	0.04	0.19
3 New Devices	483	0.01	0.11
4 New Devices	483	0.00	0.06
Male	483	0.49	0.50
BMI	483	29.90	7.02
Age	483	60.34	13.65
ASA Average	483	2.43	0.51
# of Comorbidities	483	1.99	1.45
Both Legs	483	0.01	0.09
Reason: Revision	483	0.24	0.43
Reason: Avascular Necrosis	483	0.11	0.31
Reason: Displasia	483	0.04	0.20
Reason: Arthritis	483	0.58	0.49
Reason: Severe Arthritis	483	0.04	0.19
Reason: End Stage Arthritis	483	0.05	0.23
Reason: Fracture	483	0.07	0.26
Reason: Other	483	0.05	0.21
Reasons for Revision	483	0.17	0.40
Use: Shell	483	0.86	0.35
Use: Stem	483	0.84	0.36
Use: Liner	483	0.72	0.45
Use: Head	483	0.97	0.17
Unihead	483	0.07	0.25
Cemented	483	0.16	0.37
Company 2	483	0.52	0.50
Company 3	483	0.19	0.39
Company 4	483	0.24	0.43
Multiple Companies Indicator	483	0.05	0.22
Surgeon 1	483	0.55	0.50
Surgeon 2	483	0.17	0.37
Surgeon 3	483	0.19	0.40
Surgeon 4	483	0.08	0.28

Notes: # of surgeons = 4; # of observations = 483.

Figure A1: Serial Correlation Function $\rho(\mathbf{d})$



References

Dean, David, John Pepper, Robert Schmidt, and Steven Stern. 2015. The Effects of Vocational Rehabilitation for People with Mental Illness.

Gourieroux, C., A. Monfort, E. Renault, and A. Trognon. 1987. Generalized Residuals. *Journal of Econometrics*. 34: 5-32.