Causality, Learning and Forgetting in Surgery

Gautam Gowrisankaran
Department of Economics
Universitty of Arizona
and National Bureau of Economic Research

Vivian Ho
Baker Institute, Rice University
and Department of Medicine, Baylor College of Medicine

Robert J. Town
School of Public Health
University of Minnesota
and National Bureau of Economic Research

Current Draft: April 2008

Abstract

In this paper we distinguish between two causal explanations for the volume-outcome relationship, learning-by-doing and selective referral. We use data on three surgical procedures for which a volume-outcome relationship has been documented, the Whipple, coronary artery bypass graft (CABG) and repair of abdominal aortic aneurysm (AAA). We distinguish between the competing explanations by estimating the relationship between mortality and volume in a non-linear system of equations where volume is endogenous and where predicted volume is used to "instrument" for volume. In this system, we also allow for the possibility of forgetting. This model will identify learning-by-doing by the extent to which differences in expected volume based on variation in the numbers of competitors and patients near the hospital affect mortality at the hospital. For AAA and CABG increased volume appears to cause lower mortality, while the direction of causality is less certain for the Whipple. Using the assumption that volume is exogenous, we find that a significant amount of the learning is retained from quarter to quarter for the Whipple and AAA. For CABG, the impact of an exogenous increase in contemporaneous volume on mortality depreciates from one quarter to the next.

[*] We thank Nancy Baxter, Bryan Dowd, Marty Gaynor, Willard Manning, Sean Nicholson, Ariel Pakes, Harold Seider, Bill Vogt and seminar participants at several institutions for helpful comments. We acknowledge funding from the Federal Trade Commission.

1. Introduction

Over the past quarter century, a large literature comprising over 135 studies has analyzed the relation between physician and hospital volumes for surgical procedures and patient outcomes, principally mortality. Roughly 70% of these studies find a significant positive correlation between volume and outcomes (Halm, et al. 2002). While the volume-outcome correlation is well known, the underlying reasons for this relationship are not well established. Two hypotheses, with opposite causal implications, have been offered as explanations.

One explanation is that the volume-outcome relationship is a consequence of some economies of scale in quality. Usually the posited specific mechanism underlying these scale economies is learning-by-doing. The idea is simple and has been in economic thought at least since Adam Smith: one gets better at a given task by performing it more often. This same principle underlies production processes that rely on specialization of tasks. In our context, high-volume hospitals perform more surgeries and perform them more frequently, which increases their skills and improves outcomes. An alternative explanation is selective referral, which postulates that high volume hospitals are on average better because patients, perhaps with the advice of a physician, prefer to be admitted to high quality hospitals. Opposite to learning-by-doing, selective referral implies that high quality causes high volume. Although both explanations are plausible, the medical literature has largely assumed that learning-by-doing is *the* correct interpretation for the observed volume-outcome relationship.[1]

In this paper, we attempt to understand the extent to which learning-by-doing and selective referral explain the volume-outcome relationship for three different surgical procedures, the Whipple, coronary artery bypass graft (CABG) and the repair of abdominal aortic aneurysm (AAA). Our results can shed light on the extent of learning-by-doing for an important part of the economy. The hospital sector is also one for which the implications of potential policies depend crucially on the mechanisms through which high volumes lead to good outcomes. If learning-by-doing is the right explanation for the volume-outcome relationship then consolidation of procedures may be in the public interest. If the cause is selective referral, then regionalizing procedures will serve to reduce competition and may also make it more difficult for patients to sort to good hospitals. The reduction in competition may increase market power and allow firms to raise prices (Ho, Town and Heslin, 2005). While the increase in market power is of significant concern in small markets, recent studies have found substantial market power for hospitals even in large metropolitan areas (Capps, Dranove and Satterthwaite, 2003, Town and Vistnes, 2001).

Our results also have implications beyond surgery. Understanding the extent of learning-by-doing in different industries has long been the focus of significant theoretical and empirical work.[2] More recently, a small literature has sought to understand the type of learning process, in particular, to evaluate the theoretical implications and empirical importance of *forgetting*.[3] The skills of a surgical team may deteriorate when they do not perform the procedure often, implying that forgetting is important. We are able to estimate the technology of learning for our industry because we use detailed data that may not exist for other industries. These data allow for the

[1] For example, a recent editorial in the *New England Journal of Medicine* advocated public policy and private efforts to reduce the number of procedures at low volume hospitals (Epstein 2002). A notable exception is Luft, Hunt and Maerki (1987) who first proposed the term selective referral.

[2] For example, the literature has identified learning in the following industries: airplane manufacturing, automobile manufacturing, chemical manufacturing, textiles, semiconductor manufacturing, ship building, pizza making, refined petroleum products, power plants, Kibbutz farming and nuclear power plants. See Argote (1999) for a thorough discussion of this literature.

[3] For instance, Doraszelski and Satterthwaite (2005) examine the theoretical implications of forgetting. Benkard (2000) provides finds significant forgetting in commercial aircraft production. Darr, Argote and Epple (1995) find that there is significant learning-by-doing and forgetting in the cost of pizza production. Argote, Beckman and Epple (1990) find significant knowledge depreciation in Liberty Shipbuilding during World War II.

identification of learning, selective referral and forgetting effects based on micro-level shocks. Unlike studies in the manufacturing sector, we measure learning in terms of better outcomes, and not lower costs. Yet, better outcomes surely translate into lower social costs, and so ultimately, these measures can be tied back. Moreover, data on mortality is likely more accurate than data for other industries, e.g. cost data for manufacturing industries.

Our basic model is that patient mortality depends on hospital volume, hospital characteristics and patient characteristics, via a probit specification. However, if selective referral is even partly correct, then high hospital volume is partly caused by good hospital outcomes. The error term in this equation will then be correlated with volume, in which case a simple probit alone will yield inconsistent results. We use simultaneous equation methods to control for the endogeneity, and ultimately to inform us about both the level of learning-by-doing and selective referral. Our methods rely on exogenous variables that are plausibly correlated with volume but uncorrelated with unobserved components of hospital quality. We create exogenous variables using the predicted hospital volume by procedure from a multinomial choice model. The key predictor in the multinomial choice model is distance to a hospital, which we allow to interact with hospital and patient characteristics. The location of patients relative to hospitals is a determinant of hospital volume that we think might reasonably be thought to be exogenous and hence that we use in a similar way to an instrument. Although our model is identified from similar forces as instrumental variables, our exact specification, which we detail below in Section 2, uses maximum likelihood methods that are consistent with the non-linearities inherent in our model.

The idea of our identification is that the different locations of hospitals provide a source of exogenous variation that should affect mortality for surgical procedures but that can properly be excluded as a regressor. For instance, a hospital which experiences a positive shock to the number of patients living nearby will on average accrue more volume and move down any learning curve more quickly, regardless of the extent of selective referral. In contrast, hospitals with many competitors relative to patients nearby will on average move slowly down the learning curve. The level of forgetting is identified by the relative extents to which lagged and current predicted volumes affect mortality. A sufficient condition for our source of exogenous variation to be valid is that the underlying distribution of *residual* hospital quality (which is the predictor of mortality that remains after controlling for hospital and patient characteristics including hospital fixed effects in some specifications) is identically distributed in the population. For example, our source of valid variation would be if the unobserved determinants of hospital quality are due to random variation in how smoothly different surgical teams function. Selective referral would be important to the extent that patients or their physicians are aware of the unobserved determinants of hospital quality, and choose hospitals on this basis. Distance can plausibly be excluded as a regressor in the mortality equation for the surgeries that we examine because time to the hospital is not going to significantly affect survival probabilities.

Because the identification is based on the amount of potential competition, we have to be careful in how we define potential competitors. For instance, a hospital which performs high quality Whipple surgeries may cause nearby hospitals to have no Whipple patients in a given quarter, thereby boosting its predicted volume if the nearby hospitals are omitted. For this reason, we define predicted volume based on the set of hospitals that are potential entrants to the procedure, instead of the set that are currently performing it.

Only a few other papers have attempted to distinguish between the competing explanations using an endogenous variables approach. Luft, Hunt and Maerki (1987) instrument for hospital volume using the number of appendicitis procedures, the size of the hospital, teaching affiliation and medical staff as instruments. These would only be valid instruments if the relative qualities of different procedures within a hospital are unrelated. Their identification did not rely on patient flow data at all, likely because of the lack of availability of necessary data or computational power at the time of their study. A recent paper, Gaynor, Seider and Vogt (2005), analyzes the impact of scale effects on CABG outcomes using an approach that is identified from forces similar to ours.[4] Their estimates suggest that hospital volume causes better outcomes. Some other papers have used fixed effects estimators to control for unobserved hospital heterogeneity that may be correlated with volume (Hamilton and Ho (1998), Hamilton and Hamilton (1997), and Farley and Ozminkowski (1992)).

There is even less work that estimates the extent of forgetting for surgical procedures. The medical literature almost always regresses contemporaneous volume on mortality, implicitly assuming perfect forgetting. The only other paper that estimates a model that allows forgetting is Gaynor, Seider and Vogt (2005). They estimate the level of forgetting by including lagged values of annual volume in their mortality regression. While they have difficulty identifying the impact of contemporaneous volume from lagged values because of multicolinearity, their coefficient estimates suggest that the volume effects are only contemporaneous suggesting significant forgetting. Our work is different from Gaynor, Seider and Vogt (2005) in that our empirical specification is consistent with the inherent nonlinearity in learning and forgetting processes. In addition, we examine two other procedures besides CABG and we measure hospital volume at a higher frequency (quarter versus year) which allows us to better understand the extent to which learning is retained over time.

[4] In contrast to our use of predicted volume based on potential competition, they use the number of CABG surgeries and number of hospitals performing CABGs within a fixed radius as exogenous shifters.

Recall that we examine three separate procedures. The advantage of examining multiple procedures is that we can understand the extent to which learning, forgetting and selective referral differ across the procedures. The three procedures that we examine have quite different technologies. The Whipple procedure, which is a treatment for early stage pancreatic cancer, involves removing cancerous tumors from parts of five organs. AAA repairs involve opening the aortic aneurysm and sewing a vascular graft in place of the weakened segment of the aorta. CABG is an open-heart procedure in which blocked coronary arteries are replaced with vessels taken from other parts of the body.

Our results indicate that for AAA and CABG the relationship between volume and mortality appears to be causal, while for the Whipple the results are somewhat ambiguous on this point. The point estimates from the static model indicate a 50% increase in the number of procedures would reduce the expected mortality rate by 20% (2.1 percentage points) for the Whipple procedure, 9.5% (.5 percentage points) for AAA and 12% (.5 percentage points) for CABG for a patient with the mean mortality rate and median volume levels.

Our parameter estimates from the model that allows learning and forgetting indicate that for the Whipple procedure the mortality value of an exogenous increase in volume deteriorates slowly while for AAA accumulated learning deteriorates by approximately 50% within a quarter. For CABG the mortality value of an exogenous increase in volume in a given quarter is fully depreciated by the following quarter. We posit that the pattern of retained learning across the procedures is related to the routinization and complexity of the surgery.

The rest of the paper has the following structure. The next section presents our empirical framework. Section 3 discusses the procedures we study and the data used in the analysis. Section 4 presents the findings and Section 5 concludes.

2. Empirical Framework

We develop a simple model of surgical learning that we use to estimate the levels of learning-by-doing and forgetting for Whipple, AAA repair and CABG surgery. In our model, hospital personnel can learn how to better care for patients undergoing a surgical procedure by performing the procedure repeatedly. This then allows them to increase the likelihood of post-operative survival.

Consider a patient i who requires surgery at time $t(i)$. Throughout this paper a period is a quarter of a year. The patient and/or her physician decide at which hospital the patient will obtain the surgery. Call this hospital $j(i)$. Mortality for the patient will depend on her illness severity and also on the quality of care provided by the hospital. Define a latent mortality index m_i^* and assume that the patient will die if and only if $m_i^* > 0$, so that mortality is an indicator function, $m_i = \mathbf{1}\left\{m_i^* > 0\right\}$.

We write the latent mortality index as:

$$(1) \qquad m_i^* = f\left(\beta_1, E_{j(i)t(i)}\left(q_{j(i)1}, \ldots, q_{j(i)t(i)}\right)\right) + x_i\beta_2 + \beta_3 z_{j(i)t(i)} + \varepsilon_i + \omega_{j(i)t(i)},$$

where $q_{j(i)t(i)}$ is the quantity or volume of surgical procedures performed in the quarter, $E_{j(i)t(i)}$ is the surgical experience of hospital $j(i)$ at quarter $t(i)$, which is a function of current and lagged quantities,[5] f() denotes the effect of experience on mortality, x_i are the observable risk factors of the patient, $z_{j(i)t(i)}$ are observable characteristics of the hospital that might influence mortality, β

are parameters to estimate, and ε_i and $\omega_{j(i)t(i)}$ are unobserved components of mortality, that reflect both patient severity of illness (ε_i) and hospital performance ($\omega_{j(i)t(i)}$). We let $v_{1i} = \omega_{j(i)t(i)} + \varepsilon_i$ and assume that v_{1i} is normally distributed with mean 0 and variance 1. Note that a key assumption implicit in (1) is that distance to a hospital does not directly affect mortality. Distance will then be used to identify predicted experience below.

We allow for learning-by-doing in (1). Specifically, as hospital personnel obtain more experience with a surgical procedure they obtain skills that may increase the survival rate. This may happen through several different mechanisms that may vary across procedures. For instance, volume may improve surgical skills, may help anesthesiologists, intensivists, nursing staffs, paramedical staff, and rehabilitation staff perform better and may also result in better preoperative care. We let

(2) $E_{j1} = q_{j1}$ and $E_{jt} = q_{jt} + \lambda E_{jt-1}$ for $t > 1$,

where λ is a parameter that represents the fraction of retained learning at a quarterly level. We choose a simple square root form for the effect of experience on quality,

(3) $f\left(E_{jt}, \beta_1\right) = \beta_1 \sqrt{E_{jt}}$,

because this appeared to fit the data the best.[6] We focus on two cases: perfect forgetting, $\lambda = 0$, and partial forgetting, where λ is estimated. The medical literature has largely assumed perfect forgetting.[7]

[5] Note that we are using volume only at the hospital level and not at the physician level. For the Whipple and AAA most hospitals will have only one team that can perform the surgery and so the two measures will be equivalent for most observations. For CABG surgery, investigating physician volume is an important avenue for future research.

[6] We tried a linear specification for (3). This specification yielded very similar, but less precise, parameter estimates and a higher mean squared error than the square root specification in (3).

[7] For example, see Birkmeyer, et al. (2002).

Our principal goal is to estimate β_1, the impact of experience on mortality, and, in some specifications, the forgetting parameter λ. One strategy would be to estimate (1) (and (2) in cases where λ is not fixed) using a probit type analysis. A potential problem with this strategy is the endogeneity of experience E_{jt}. Endogeneity will occur if the unobserved component of hospital skill, ω_{jt}, varies across hospitals, and if patients are aware of this variation. In this case, patients may disproportionately select hospitals with a low ω_{jt}, and so these hospitals will then obtain a high level of experience. This is the selective referral hypothesis. In this case, experience does not lead to high quality, but rather the reverse is true.[8]

We control for the endogeneity of experience by estimating a simultaneous equations system with two equations. The first equation is the mortality equation given above in (1). Experience is the dependent variable for the second equation. Our model is similar to instrumental variables in that this equation specifies experience as a function of exogenous characteristics, most notably the experience that a hospital would obtain on average given its location and other exogenous characteristics. Specifically, our second equation is:

$$(4) \qquad \sqrt{E_{jt}} = \alpha_1 \sqrt{\hat{E}_{jt}} + \alpha_2 \hat{E}_{jt} + \alpha_3 \hat{E}_{jt}^2 + \alpha_4 z_{jt} + v_{2jt},$$

where $\hat{E}_{jt}$ denotes the predicted experience for hospital j at time t and $v_{2,jt} \sim N\left(0,\sigma^2\right)$. We use the square root of experience as the dependent variable in (4) to be consistent with (3). Consistent with E_{jt} being endogenous, we allow for a correlation ρ between v_{1i} and $v_{2,j(i)t(i)}$. We assume that the covariance structure is such that the unobservable components of mortality

[8] Note that another source of endogeneity is due to ε_i, that patients with a high severity of illness may be more likely to select hospitals with high experience levels. This source of endogeneity, which was addressed by Geweke, Gowrisankaran and Town (2003), will also be controlled for by our methods.

for two individuals v_{1i_1} and v_{1i_2} who are treated at the same hospital in the same quarter are independent conditional on the error term in (4), which is $v_{2,j(i_1)t(i_1)}$ or equivalently $v_{2,j(i_2)t(i_2)}$.

Although our model is non-linear, it is helpful to think of as similar to instrumental variables in order to understand our identification strategy.[9] In particular, we would expect the model to be well identified only when there is some "instrument" that will cause exogenous variation in experience. Our idea is that the predicted experience of a hospital, which is a function of the distance between patients and hospitals, can serve that role. We construct predicted experience $\hat{E}_{jt}$ by estimating and aggregating a model of patient flows that predicts the choices of patients based on exogenous factors, principally distance.

Continuing with the idea that $\hat{E}_{jt}$ is similar to an instrument for E_{jt}, the principal identifying assumptions that we employ are that $\hat{E}_{jt}$ does not enter into (1), is orthogonal to ω_{jt}, and is correlated with E_{jt}. It seems reasonable to assume that distance and hence predicted volume (or experience) will not directly affect mortality conditional on the actual experience, since time to hospital is not crucial for the surgeries that we examine, and that predicted volume will be correlated with actual volume. Thus, the key assumption is that $\hat{E}_{jt}$ is uncorrelated with ω_{jt}, the unobservable component of hospital performance. A sufficient condition for this is that ω_{jt} is randomly drawn and hence uncorrelated with hospital locations.

To understand whether this key assumption is reasonable, note that we allow for hospital characteristics in $z_{j(i)}$ including size and teaching status, and hospital fixed effects in some

[9] Since our model is not instrumental variables, (4) needs to be fully specified and an incorrect functional form for (4) will lead to inconsistent parameter estimates in (1).

specifications. With fixed effects, a sufficient condition for this assumption to be valid is that hospitals of a given type cannot make specific investments to lower their mortality rate in response to demand shocks. Investments that improve the performance for these surgeries may include training in teamwork and group dynamics,[10] higher quality staff, and different hospital layouts,[11] all of which are difficult to change in the short-run. In contrast, our model would not be consistent if ω_{jt} can be chosen in response to $\hat{E}_{jt}$. Even in this case, our analysis will still predict how expected volume would affect outcomes and our results will be relevant to policies that change expected volume, such as mandates to consolidate services.

We derive predicted experience $\hat{E}_{jt}$ by estimating a conditional logit model of patient choices. We specify patient utility as:

$$(5) \qquad u_{ij} = \gamma_1 d_{ij} + \gamma_2 d_{ij}^2 + \gamma_3 d_{ij} x_i + \gamma_4 d_{ij} z_j + v_{3,ij},$$

where d_{ij} indicates distance from patient i to hospital j, x_i are patient characteristics such as age, z_j are hospital characteristics such as teaching hospital status, and $v_{3,ij}$ is distributed type I extreme value. We do not include an outside alternative, since it is unlikely that not obtaining the surgery is one of the relevant choices for patients in our sample.[12] We estimate (5) using maximum likelihood. We create predicted experience $\hat{E}_{jt}$ by adding the predicted flows using (5) to create predicted volume,[13] and then converting predicted volume to predicted experience using the analogous expression to (2) .

[10] See Edmondson, Bohmer, and Pisano (2001).

[11] See Herzlinger and Stavros (2002).

[12] Since there is no outside alternative, there is also no reference group with utility normalized to zero, which implies that a choice-specific or overall constant term in (5) could not be identified.

[13] See Gowrisankaran and Town (2003) for details of this type of computation.

One significant issue is the set of hospitals to specify as potential competitors when calculating predicted volume. As Whipple is an extremely rare procedure and AAA repair is somewhat uncommon, there are several hospitals that perform one or two procedures in one quarter, none the following quarter, and one the quarter after that, etc. Yet, these hospitals are likely in the market in the intermediate quarters. Excluding them will boost $\hat{E}_{jt}$ for nearby hospitals and hence may create an endogeneity problem since the fact that they performed no procedures in a given quarter is likely correlated with a nearby hospital having a low ω_{jt}. To avoid this endogeneity problem, we keep hospitals as potential competitors in quarters where they have no patients.

A related issue is that a hospital with a high ω_{jt} may deter other hospitals from performing the surgery at all, which would cause a similar endogeneity problem to that noted above. This potential issue is problematic in particular for CABG, as this procedure has the smallest set of hospitals performing the procedure, in spite of having the most patients by far. We control for this problem by including the set of potential CABG entrants as the set of hospitals with predicted volume. We let the set of potential CABG entrants be the set that perform cardiac catheterizations.

Note that we are estimating our model in two stages. In the first stage, we estimate the α coefficients and use them to derive $\hat{E}_{jt}$, and in the second stage, we use these predicted experience variables as exogenous shifters for experience.[14] Another potential way of estimating our model would be to perform these two procedures jointly. However, there is little efficiency gain from that since the coefficients in (5) are estimated very precisely, given the large number

[14] This is similar to Dubin and McFadden (1984).

of patients relative to hospitals. Moreover, estimating these two equations jointly would require having to specify the correlation structure between each $v_{3,ij}$ and v_{1i}, which is very complex. Instead, we need only specify the correlation structure between two variables, v_{1i} and $v_{2,j(i)t(i)}$ which is much simpler. Indeed, since the point of estimating the patient flow model is simply to provide exogenous shifters of experience, an alternate way of deriving $\hat{E}_{jt}$ would have been to pick reasonable values of α based on travel times and costs.

Our base estimation method for our mortality model specified by (1) and (4) is maximum likelihood. As derived in Wooldridge (2002), the likelihood function for this type of model can be split up into two parts, the density of observing a given hospital experience level given predicted experience and the conditional probabilities of observing a given mortality outcome for a patient at that hospital. Specifically, we can write the likelihood for hospital j at time t as:

$$\begin{aligned}
\ln L_{jt}\left(\lambda,\beta,\rho,\sigma,\alpha\right) = {} & \ln\left(\phi\left(\left(\sqrt{E_{jt}}-\bar{E}\right)\Big/\sigma\right)\Big/\sigma\right)+ \\
& \sum_{i|j(i)=j,t(i)=t} m_i \ln\Phi\left(\frac{1}{\sqrt{1-\rho^2}}\left(\beta_1\sqrt{E_{jt}}+\beta_2 x_i+\beta_3 z_{jt}+\frac{\rho}{\sigma}\left(\sqrt{E_{jt}}-\bar{E}\right)\right)\right)+ \\
& \sum_{i|j(i)=j,t(i)=t} \left(1-m_i\right)\ln\Phi\left(-\frac{1}{\sqrt{1-\rho^2}}\left(\beta_1\sqrt{E_{jt}}+\beta_2 x_i+\beta_3 z_{jt}+\frac{\rho}{\sigma}\left(\sqrt{E_{jt}}-\bar{E}\right)\right)\right), \\
& \text{for } \bar{E}=\alpha_1\sqrt{\hat{E}_{jt}}+\alpha_2\hat{E}_{jt}+\alpha_3\hat{E}_{jt}^2+\alpha_4 z_{jt},
\end{aligned} \tag{6}$$

where ϕ is the standard normal density and Φ the standard normal distribution. Estimation was performed in Stata using the maximum likelihood command. We obtain standard errors using the robust sandwich formula for the likelihood function. The standard errors are calculated treating one hospital over time as one observation, in order to not overstate the significance of results due to serial correlation.

We actually use a slightly different specification for the results, principally for robustness reasons. To explain, note that the functional form of (4) matters for the consistency of our results, since we are estimating a non-linear system of equations. Yet, we have specified (4) with a simple linear functional form, rather than trying to guess at the "true" functional form. A predictor with more relevant regressors is likely to better approximate the true underlying functional form. Our base model does not allow the predictor to depend on patient characteristics because it is at the hospital level. Thus, for the reported specifications, we use a patient-level variant of (4):

$$(7) \qquad \sqrt{E_{j(i)t(i)}} = \alpha_1\sqrt{\hat{E}_{j(i)t(i)}} + \alpha_2\hat{E}_{j(i)t(i)} + \alpha_3\hat{E}^2_{j(i)t(i)} + \alpha_4 z_{j(i)t(i)} + \alpha_5 x_i + v_{2j(i)t(i)}\,.$$

As (7) allows for patient characteristics, it can provide a better fit of the error term. For this variant, we estimate a similar likelihood function to (6), but where the first line has one observation for each patient:

$$(8) \qquad \begin{aligned} \ln L_{jt}\left(\lambda,\beta,\rho,\sigma,\alpha\right) = & \sum_{i|j(i)=j,t(i)=t} \ln\left(\phi\left(\left(\sqrt{E_{jt}} - \bar{E}\right)\Big/\sigma\right)\Big/\sigma\right) \\ & + \sum_{i|j(i)=j,t(i)=t} m_i \ln \Phi\left(\frac{1}{\sqrt{1-\rho^2}}\left(\beta_1\sqrt{E_{jt}} + \beta_2 x_i + \beta_3 z_{jt} + \frac{\rho}{\sigma}\left(\sqrt{E_{jt}} - \bar{E}\right)\right)\right) \\ & + \sum_{i|j(i)=j,t(i)=t} \left(1-m_i\right)\ln \Phi\left(-\frac{1}{\sqrt{1-\rho^2}}\left(\beta_1\sqrt{E_{jt}} + \beta_2 x_i + \beta_3 z_{jt} + \frac{\rho}{\sigma}\left(\sqrt{E_{jt}} - \bar{E}\right)\right)\right), \\ & \text{for } \bar{E} = \alpha_1\sqrt{\hat{E}_{jt}} + \alpha_2\hat{E}_{jt} + \alpha_3\hat{E}^2_{jt} + \alpha_4 z_{jt} + \alpha_5 x_i. \end{aligned}$$

All of our reported estimates use the model specified by (7).

For most specifications, we estimated the model specified by (7) with maximum likelihood. For the specifications with fixed effects and with perfect forgetting, we used a twostep estimator specified by Rivers and Vuong (1988). This entailed first estimating (7) using OLS and then estimating (1) with the predicted residual from (7), which we can call $\hat{v}_{2j(i)t(i)}$, included as a regressor.[15] The disadvantage of this method is that it is not efficient. This method can also can be used to construct a test of the endogeneity of $\hat{E}_{jt}$ that is robust to the functional form of $v_{2j(i)t(i)}$.

3. Whipple, CABG and AAA Repair—The Procedures and the Data

The Whipple Procedure

The Whipple procedure, or pancreaticoduodenectomy, is used primarily in patients with early stage, localized pancreatic cancer. During the procedure, the head and neck of the pancreas and parts of the stomach, common bile duct, gall bladder and small intestine are removed. The procedure is extremely complicated because it requires operating in an area of the body with several vital organs and important blood vessels. Many surgeons require 8 to 9 hours to perform the surgery, although some surgeons can complete the operation in half that time. The complexity of the surgery and its relative rarity (roughly 330 Whipple procedures are performed in the entire state of Florida per quarter), make it a strong candidate for a procedure in which learning-by-doing would reveal itself in the data.

It is rare for a patient undergoing the Whipple procedure to die in the operating room. In-hospital death most often results from post-operative complications, including bleeding, infection, delay in the function of the stomach, pneumonia, and intra-abdominal abscesses. Other patients suffer nutritional compromise after the surgery. Experienced surgeons may be more likely to prevent patient complications such as bleeding and delay in the function of the stomach,

[15] See also Newey, 1987 and Wooldridge, 2002.

and their shorter operating time may reduce the probability of infection. In addition to the surgeon, many other clinicians in the hospitals are likely to affect the probability of post-operative complications that can lead to patient harm or death.

Several authors have found correlations between hospital volume and mortality rates for the Whipple procedure. Ho, Town and Heslin (2003) find a correlation between volume and mortality for the Whipple. They do not find much evidence of learning spillovers from nearby hospitals nor do they find any impact of competition on outcomes. Birkmeyer, et al. (1999a), Birkmeyer, et al. (1999b), Gordon et al. (1998) and Glasgow and Mulvihill (1996) all find significant correlations between hospital volume and mortality rates (measured by in-hospital or 3-year mortality rates). None of these studies attempt to distinguish learning-by-doing from selective referral as the causal mechanism underlying the correlations between volume and mortality for the Whipple.

Our outcome measure is in-hospital mortality. This is an appropriate measure for a few reasons. First, in-hospital death is a relatively common outcome. In our sample, patients undergoing the Whipple procedure have a 10% chance of dying in the hospital. Second, using a wider mortality window can begin to blur the quality of the Whipple procedure with the progression of the cancer that may be unrelated to the surgery. While these patients have pancreatic cancer, it is in an early stage and it is very unlikely that the cancer is the cause of in-hospital death. Finally, in-hospital death is a common outcome measure that is widely used to analyze the volume–outcome relationship for Whipple and other procedures.

It is possible that there may be two quality dimensions to the Whipple procedure: the probability of surviving the procedure itself and the impact of the surgery on surviving the cancer. We will only measure the first dimension of quality. Measuring the second dimension is rather difficult and, as is clear from the probability of in-hospital death, the first dimension is, at the very least, an important dimension of Whipple quality.

We include patient-specific case-mix variables to control for disease severity. These variables specify patient age and gender and indicators for the presence of a myocardial infarction, renal failure, liver disease, diabetes and whether the admission was an emergency. They are intended to pick up factors that correlate with the ability of the patient to withstand complex surgery and post-surgery complications.

AAA

A weakening of the aorta, which is the main blood vessel that carries blood from the heart to the rest of the body, causes abdominal aortic aneurysms. As the aorta weakens, the vessel balloons. If left untreated, the aneurysm will generally grow larger and eventually rupture. Ruptured aortas are medically important: they are the 13th leading cause of death in the U.S., accounting for 15,000 deaths each year. Surgical treatment of AAA has been performed for more than 50 years. The treatment, which is major surgery requiring hospital stays of 7 to 10 days, is to replace the diseased part of the aorta with a graft. In 1999, the FDA approved an endovascular grafting technology that allows surgeons to repair the AAA by delivery of a bypass graft through a small incision in the groin. This procedure generally reduced hospital stays to a single night. The widespread use of endovascular grafting occurs after the period we study here.

Death is also a common outcome for AAA surgery. The 30-day, in-hospital mortality in our data is 5.6%. We control for many of the factors that are correlated with AAA mortality including patient demographics, myocardial infarction, renal failure, liver failure, stroke, and patient demographics (Chen, et al, 1996). Consistent with the previous literature, we limit our analysis to unruptured aneurisms. Once an aneurism ruptures the patient is at grave risk and requires emergency surgery. The mortality rates for ruptured aneurisms approaches 50%—substantially higher mortality than the 6.5% 30-day, in-hospital mortality for unruptured AAAs. In general, patients with ruptured aneurisms should be taken to the closest hospital and are not subject to selective referral.

There is a large literature (approximately 20 papers) studying the relationship between hospital volume and mortality for AAA using data from the US, Canada and Europe.[16] This literature is almost unanimous in finding an inverse relationship between hospital volume and mortality, both inpatient and overall. However, none of these studies effectively controls for the potential endogeneity of volume.

CABG

Heart disease is the leading cause of death in the U.S.[17] CABG surgery is an operation designed to detour blood around a narrowed segment of a heart artery in an effort to restore blood flow to the heart muscle. The surgery involves the removing a "clean" vessel (graft) from the leg, chest, or arm and attaching it to the areas around the blocked artery in order to restore blood flow. In traditional bypass surgery, the heart is stopped and a heart-lung machine (cardiopulmonary bypass machine) is used to pump blood and perform the duties of the lungs. A recently developed technique – called "beating heart bypass" or "minimally invasive bypass" – allow surgeons to perform the surgery without stopping the heart. This method uses a special device that stabilizes the part of the heart on which the surgeon is operating. The rest of the heart continues to beat while the surgeon operates.

[16] The more recent US studies include Goodney, Lucas and Birkmeyer, 2003; Cowan et al., 2003; Dimick, et al., 2002; Dimick, et al, 2002; Sollano et al. 1999; Pearce, et al. 1999; Manheim et al. 1998; and Hannen et al. 1992. The Canadian studies are Urbach, Bell, and Austin, 2003 and Chen et al., 1996. The European studies are Kantonen et al (1997) and Amundsen et al. (1990).

[17] National Center for Health Statistics at http://www.cdc.gov/nchs.

Dr. David C. Sabiston Jr. performed the first coronary bypass surgery in a human in 1962. CABG was a relatively rare procedure until the 1980s. Since then the procedure has become widespread. In 1999, there were approximately 571,000 CABG procedures performed in the U.S.[18] The procedure is common, complicated and expensive and has proven to yield significant health improvements for those patients suffering from severe angina.[19]

We are aware of nine studies of the relationship between volume and mortality for CABG. Halm, et al. (2002) reviews eight papers examining the relationship between in-hospital mortality and volume for CABG. Six of the eight studies find a negative and significant relationship between volume and mortality. More recently, Birkmeyer et al. (2002) studied the mortality rate of Medicare patients undergoing CABG and finds the same relationship for overall (not necessarily in–hospital) 30–day mortality.

We use 30-day, in-hospital mortality as our primary measure of quality. There are many measures of CABG quality in the literature including 30-day, 90-day, 180-day mortality rates and hospital readmission. Mortality is widely considered to be an important dimension of CABG quality. We focus on the in-hospital mortality rate, as that is the measure used by most of the studies of the volume-outcome relationship for CABG, as noted above. We have also performed the analysis using the 180–day mortality rate for death at *any* hospital and the conclusions are identical to using the in-hospital mortality rate. For our sample, 30-day, in–hospital mortality is approximately 3.7% suggesting that mortality is a relatively common outcome for those undergoing CABG, and hence a useful measure of quality.

As with Whipple, we include patient-specific case-mix variables to control for disease severity. We chose measures based on the consensus statement prepared by a panel of

[18] National Center for Health Statistics at http://www.cdc.gov/nchs.

researchers from the major CABG reporting programs (Block et al., 1998), using those variables which can be constructed from discharge data. Our case–mix variables are age, sex, race, renal disease, diabetes, ventricular arrhythmia, stroke, AMI, number of vessels bypassed, angioplasty, congestive health failure, and liver failure. In our limited experimentation, our conclusions are insensitive to the exact set of risk adjusters.

Data

The data used in this study are hospital discharge data from two sources: the Florida Agency for Health Care Administration (AHCA) from 1988 to 1999 and the California Office of Statewide Health Planning and Development (OSHPD) from 1993 to 1997. Following previous studies in the literature, we defined Whipple surgery by an ICD-9-CM procedure code of 52.7 (radical pancreaticoduodenectomy) from the AHCA and OSHPD data.[20] We excluded patients who were undergoing the Whipple procedure due to a trauma accident as indicated by the ICD-9 codes rather than cancer from the analysis. The raw Florida data had 3,182 observations. We dropped those observations with missing values, leaving us with 2,894 observations.[21] For the California data, we started with 1,640 observations and by removing observations with missing values left us with 1,582 observations. In our analysis we merged the California and Florida data giving us a total of 4,455 observations.

We designate AAA by procedure codes of 38.34, 38.44 or 38.64 (regardless of diagnosis codes) and procedure codes 39.25, 39.51 or 39.52 together with diagnosis codes 441.0, 441.02, 441.03, 441.4, 441.7 or 441.9. We started with 14,778 observations from the California data and 39,056 observations from the Florida data. After removing missing values we have 14,207

[19] In California, average hospital charges for CABG in 1997 are approximately $74,000 (investigators' calculations using OSHPD data).

[20] We base all of our procedure indications on whether *any* of the codes, not just the primary code, matches the listed value.

California observations and 36,384 Florida observations. This leaves us with a total of 50,520 observations to estimate the coefficients.

The CABG analysis uses hospital discharge abstract data from AHCA and OSHPD. We define CABG surgery by ICD-9-CM procedure codes of 36.1x, for any x. CABG is a common procedure. We initially extracted 131,155 observations from California and 289,146 from Florida. After removing missing values we have a total of 404,565 CABG observations.

Both the Florida and the California data contain the home zip code of the patient. We use this information to construct approximate distances from the patient home to the hospital. We assign patient latitude and longitude by matching the zip code to the Census Bureau's Tiger database that lists the latitude and longitude for the center of each US zip code. We also used zip code information to merge in per-capita median zip code income from the Census Bureau. We obtain hospital latitudes and longitudes from the American Hospital Association. With this information it is straightforward to calculate the distance from the center of each zip code to the hospital. We also use information on the teaching status and the number of beds for each hospital. For both the Whipple procedure and CABG we obtained information on the beds size from the same agency that provided the patient discharge data. Teaching status is determined by membership in the Council of Teaching Hospitals.

Table 1 presents some summary statistics of our data. As mentioned above the mortality rate is approximately 10% for the Whipple procedure, 6.5% for AAA repair and 4.3% for CABG For all three procedures, Florida hospitals have higher mortality rates despite performing roughly two times as many procedures per hospital. CABGs are performed much more frequently than the Whipple procedure or AAA repair. The typical patient receives treatment at a hospital performs 58 times as many CABGs as Whipple procedures (117 versus 2) and 15 times as many

[21] The majority of the missing observations are a consequence of the failure to merge in per capita zip code income.

CABGs as AAAs (117 versus 8). More hospitals perform AAA (331) in California than either CABG (189) or the Whipple procedure (265).

Floridians, on average, travel a little further than Californians to receive their care. The mean distance traveled is 21.3, 26.0 and 17.2 km for the Whipple, CABG and AAA repair in Florida, while the Californian in our data traveled an average of 18.6, 16.5 and 22.6 kilometers for the Whipple procedure, CABG and AAA repair, respectively. All of the patient populations are elderly—the mean age for the three procedures is over 64 years of age.

While the majority of the variance in quarterly volume occurs between hospitals, nevertheless there is significant within hospital variance in volume. OLS, patient weighted regressions of quarterly volume on hospitals and year fixed effects yield R^2s and (Root MSE) of .75 (1.6), .80 (3.8) and .86 (54.6) for the Whipple procedure, AAA repair and CABG, respectively.

4. Results

A. Hospital Choice Model

Table 2 presents the results of our multinomial hospital choice model for both Whipple and CABG. In addition to distance and distance squared (measured in kilometers) we include an indicator for the closest hospital, the number of available/staffed beds, the number of beds interacted with distance, age interacted with distance (age dummies for the Whipple), an indicator for a teaching hospital, and teaching hospital interacted with distance.

Columns (1), (2) and (3) present the results for the Whipple, AAA repair and CABG, respectively. The CABG estimates are based on a 25% random sample. The coefficients are in line with our expectations. For all procedures, distance is inversely related to the probability of being admitted to the hospital. The coefficient is negative and precisely estimated for all procedures. The impact of distance is slightly concave as the coefficient on distance squared is positive. However, the impact of distance squared is modest given the distances between patients and hospitals in Florida and California. In all cases, patients prefer to go to the closest hospital. Patients also prefer to go to larger hospitals. The coefficients on the age interacted with distance are all negative. Half of these coefficients are insignificant at traditional levels of confidence. All else equal, teaching hospitals are more desirable than their non-teaching counterparts.

The predicted volumes generated from the multinomial choice model are, as expected, highly correlated with the actual volumes. Predicted volume will only be a "good instrument" if it is correlated with actual volume. To explore the predictive quality of these variables we perform the standard F-test of the joint significance of predicted volume (and its square and cube) regressed on the square root of actual volume with the relevant risk adjusters included as regressors. For the Whipple procedure the p-value of the F-stat is 345 (p-value= .000001). For AAA the F-test is 52.6 (p-value = .00001), and for CABG the F-test is 8.59 (p-value = .0001). With the exception of CABG, the F-tests are well above typical thresholds used to investigate instrument validity. For CABG, while the F-test is highly significant, it is on the borderline of the typical instrument validation criteria.

B. Volume and Outcomes

Whipple Analysis

The first column of Table 3 presents the coefficients from the maximum likelihood probit mortality-volume regression. The coefficient on the square root volume is negative and significant—increasing volume is associated with decreased mortality.[22] The implied magnitude of the impact of volume on mortality is large. An individual with a 10% chance of dying (approximately the mean) at a hospital performing two Whipple procedures a quarter (the median, patient weighted, hospital volume) has an approximately 8.2% (20% less) chance of dying at a hospital performing 4 Whipple procedures per quarter. The coefficients on the severity variables are generally sensible. The presence of renal and liver disease significantly reduces the chance of survival. There does not seem to be a relationship between either the size or the teaching status of the hospital and the likelihood of death. Finally, conditional on the other right hand side variables, hospitals in Florida have higher mortality rates.

In column (2) of Table 3 we present the coefficients from the hospital fixed effects specification. The coefficient on the square root of volume is a third as large in magnitude as in the specification where volume is treated as exogenous. However, the parameter estimates on the square root of volume are very imprecise. The confidence interval of the marginal impact on mortality of an increase in volume in this specification overlaps the point estimate of the marginal impact implied by the simple probit. The Wald test of the joint significance of the hospital fixed effect resoundingly rejects the hypothesis that they are equal to zero. This suggests that there are important differences between hospitals in the quality of care they provide beyond the volume of patients they treat. However, as we discuss below, the results below suggest that these hospital differences in quality are uncorrelated with hospital volume.

In column (3) we present the maximum likelihood results treating volume as endogenous. The coefficient on volume in this specification is modest in magnitude and not significantly different from zero at traditional levels of confidence. Importantly, the coefficient estimate of ρ, the parameter that captures the cross-equation correlation in the error terms, is negative but not significant at traditional levels of confidence. That is, the results in this specification fail to reject the learning-by-doing hypothesis for the Whipple procedure. The fixed effects, Rivers and

[22] Standard errors are corrected for hospital level clustering.

Vuong estimates treating volume as endogenous are presented in column (4). The coefficient estimates on the square root of volume is negative but imprecisely estimated. The coefficient estimate of $\hat{v}_{2j(i)t(i)}$ is positive and is also very imprecisely estimated. In sum, the results for the Whipple procedure are somewhat mixed but we interpret them to weakly support the volume causes mortality hypothesis.

AAA Repair Analysis

The results for AAA repair are presented in Table 4. Treating volume as exogenous (column (1)), the coefficient estimates imply a negative relationship between mortality and contemporaneous procedure experience. The coefficient on volume is negative and precisely estimated. The magnitudes of the volume effects implied by the coefficients are nontrivial. Starting at a base mortality probability of .063 at a hospital performing 6 procedures in a quarter, the coefficient estimates implies that a 50% increase in the number of procedures will reduce mortality by approximately .5 percentage points or 9.5%. The coefficients on the severity variables are sensible. Mortality is increasing and convex in age, higher for blacks and women, and more likely in the presence of myocardial infarction, renal disease, liver disease and an emergency admission. Mortality is increasing in the number of co-morbid conditions. Higher per-capita zip code income implies lower mortality. Again, conditional on the other right hand side variables, patients in Florida have significantly higher mortality rates.

In the second column of Table 4 we present fixed effects estimates of the volume-mortality relationship. The coefficient on volume modestly declines relative to the estimate in column (1), but it is still significantly different from zero at the 1% level of confidence. Like the case of the Whipple procedure, the joint test that the hospital fixed effects are all equal convincingly rejects that hypothesis.

Column (3) presents the maximum likelihood estimates treating volume as endogenous. The coefficient estimate on volume is negative and similar in magnitude to the estimate in column (1), but it is significantly different from zero. The coefficient on ρ is small and is not significant at a traditional level of confidence. We do not reject the learning-by-doing hypothesis in this model.

Column (4) presents the fixed effects, Rivers and Vuong estimates treating volume as endogenous. The coefficient estimate from this regression is smaller in magnitude that the corresponding coefficients in columns (1)-(3). However, the coefficient is very imprecisely estimated and the coefficient on $\hat{v}_{2j(i)t(i)}$ is small and is not significant at a traditional level of confidence.

In sum, the results in Table 4 suggest that, increases in volume results in reduced mortality and this appears to be the best explanation for the correlation between volume and outcomes for the repair of AAA.

However, it is important to note that the estimates imply two important aspects of the data. First, individual hospital heterogeneity is more important than differential volume in explaining differential mortality rates across hospitals. Second, better hospitals do not attract more patients, conditional on the control variables. These findings suggest that while consolidation of facilities may lead to a decrease in mortality—larger decreases may be had by simply directing patients to better hospitals.

CABG Analysis

In Table 5 column (1), we present the probit model estimates for CABG treating volume as exogenous. The parameter estimate on volume is negative and precisely estimated (p-value = .001). The implied magnitude of the coefficient estimates is quite large. The coefficients imply that hospitals that a person with a 3.6% chance of dying at a hospital performing 117 CABGs a quarter, has a 12% (.5 percentage points) less) chance of dying at a hospital performing 175 CABGs per quarter. The coefficients on the severity variables are sensible and follow a similar pattern to those present for AAA repair. However, there are three notable differences between the control variable coefficients between the two estimates. Being black, the size of the hospital and zip code income does not impact expected mortality.

In the second column of Table 4 we present the fixed effects estimates of volume-mortality relationship. The coefficient is very similar in magnitude to the volume effects estimated without the fixed effects. The parameter is precisely estimated and significantly different from zero at the 1% level of confidence. The hospital fixed effects parameters are also jointly significant.

The maximum-likelihood estimates allowing for endogenous volume is presented in column (3). The estimated effect of volume is actually larger in magnitude (.040 versus .022) that the estimates in column (1). The coefficient on ρ is positive and significant (t-statistic = 2.02) indicating that there some modest unobserved selection, but in an unexpected direction. Hospitals with higher mortality rates have higher volumes. However, given that this particular result is somewhat sensitive to model specification we do not place significant weight on it.

Finally, column (4) presents the fixed effects, Rivers and Vuong estimates treating volume as endogenous. The coefficient estimate on volume is very close to the estimate in column (1) and but not nearly as precisely estimated (t-statistic = 2.25). The coefficient on $\hat{v}_{2j(i)t(i)}$ is positive but insignificant.

Thus, the CABG estimates indicate that like AAA repair, learning-by-doing is the probable explanation for the correlation between volume and outcomes. Again, like AAA repair, better outcomes are possible through consolidation on procedures. However, unlike the AAA repair results, individual, time invariant, hospital heterogeneity does not seem important in explaining mortality rates. The contribution of the hospital fixed effects to the likelihood is modest. These results suggest that efforts to steer patients to high volume hospitals for CABG is a reasonable strategy for those patients as hospitals do not appear to differ in the quality of care they provide conditional on their volume. This strategy will lead to better outcomes, not only for those patients, but also for the other patients that are already likely to receive their surgery in that hospital.

Robustness

We examined the robustness of our parameter estimates to different specifications and samples of the data. Our first robustness check is to estimate the parameters for each procedure separately for each state. In all cases, volume did not appear to be endogenous and the magnitudes to the impact of volume on mortality were similar between California and Florida. Our results are also insensitive to the exclusion of outliers. We re-estimated the models dropping both the lowest 5% volume hospitals and the highest 5% volume hospitals from the sample and the parameter estimates and the precision was not greatly impacted.

Learning-by-Doing and Forgetting

In Table 6 we present our maximum likelihood results on the amount of learning/forgetting in surgery for the three procedures. In these specifications we are allowing λ in equation (2) to be a free, estimated parameter.[23] For all of the procedures, any evidence suggesting endogeneity was not overwhelming of volume and therefore we estimate the

[23] We examined the robustness of our results to models that use the first year's worth of data to calculate the experience and the estimated retained learning parameters are similar to the ones presented in Table 6.

parameters treating volume as exogenous.[24] Column (1) presents the coefficient estimates for the Whipple procedure. The coefficient on the square root of retained experience is negative and significant and the coefficient of degree of retained learning is .93 with a modest standard error.[25] These estimates indicate that the effects of an exogenous increase in volume reduce mortality and the mortality benefits persist well into the future.

Column (2) presents the estimates for AAA repair. Again, the coefficient on the square root of retained experience is negative and significant, but for this procedure, there is significant but imperfect retained learning. Fifty-one percent of the impact of an exogenous increase in volume in a quarter is carried over to the following quarter. Interestingly, for the CABG procedure, the parameter estimates indicate that the impact of an increase in volume is fully depreciated by the following quarter.

In order to better understand the dynamics of learning and forgetting implied by our coefficient estimates we perform the following simple simulation. We track the impact of a transitory and exogenous doubling of volume for the Whipple procedure and AAA repair. We do not examine the CABG procedure as the coefficient estimates imply that changes in volume do not have any dynamic consequences for mortality. For each procedure we assume a steady-state level of experience at the median quarterly volume and the base mortality rate is roughly the mean mortality rate for each procedure. In the experiment we then double the volume in period 1 and volume is assume to return back the median level for the other periods.

In Figure 1 presents the results of this simulation. For the Whipple procedure, the one-time increase in volume yields modest one period reductions in mortality but those gains are spread out over a significant period. In the period of the increase in volume, mortality decreases

[24] We also estimated the forgetting parameters treating volume as endogenous but the results were not sensible.

by approximately 3% (.32 percentage points). However, mortality is meaningfully lower 6 quarters into the future. For the AAA repair, the one-time increase in volume leads to a significant immediate drop in mortality of 6% (.4 percentage points). The impact of this increase in volume is essentially dissipated by the third quarter.

It is instructive to compare out estimates of knowledge depreciation rates to those from other industries. Argote, Beckman and Epple (1990) find a quarterly retained learning rate of .42 in Liberty Shipbuilding. Darr, Argote and Epple (1995) estimate a retained learning rate of .07 after one quarter in franchise pizza making. Benkard (2000) finds relatively low knowledge depreciation within one airline manufacturing line. He estimates that after one quarter 85% of the accumulated knowledge is retained. Like our estimates for different procedures, the few estimates from the literature span the possible values for retained learning rates.

The effect of retained learning differs distinctly across conditions, with history mattering most for the Whipple, for AAA somewhat, and not at all for the CABG. Two factors may cause this pattern: the routinization of the operation and the frequency of a surgeon performing the operation. The Whipple is a non-routine and complex procedure, with outcomes primarily depending on physician skill, and surgeons perform the Whipple intermittently while primarily performing other types of abdominal surgery. At the other end of the spectrum is CABG which is highly routinized, and most surgeons who perform CABGs perform only CABGs and valve procedures. While the variation in Whipple outcomes are driven by surgeon skill and practice variation, the CABG outcomes are plausibly driven by differences in pre- and post- surgical care. This care is provided by teams of nursing and other hospital staff. The composition of these teams changes frequently due to scheduling and, to a lesser degree, employee turnover. Our

[25] Not surprisingly the estimates on the forgetting parameters are sensitive to the frequency of the data. Using data at an annual frequency yields forgetting parameters near zero for all procedures.

results are consistent with the notion that the knowledge accumulation is team specific and team composition changes relatively frequently. AAA repair likely lies somewhere between the Whipple procedure and CABG in the relative roles of the physician and the rest of the hospital staff in determining patient outcomes. Vascular surgeons perform many different procedures; however, AAA repair likely comprises a large percentage of their surgical portfolio. There is less physician specialization than CABG but more than the Whipple procedure.

An obvious question is: are we measuring learning-by-doing or some other scale effect for quality? The most obvious potential causes of these scale effects are better physicians are attracted by higher volume hospitals, and, related, higher volume hospitals may be able to afford higher quality surgical staff and equipment.[26] In this case, the retained learning coefficient would be related to the retention rate of surgeons as a function of accumulated experience. The coefficients for AAA and CABG imply a turnover that is too high relative to actual surgical staff turnover. Thus, our view is that the parameter estimates are inconsistent with other volume causing mortality explanations.

The results on forgetting parameter estimates raise several issues—both for understanding the nature of knowledge accumulation within an organization and the impact of competition policy and market dynamics on surgical quality. An obvious question is why the retained learning rate varies across procedures. These procedures differ in their difficulty and volume, each may influence the rate of knowledge depreciation. This is a topic we plan to explore in the future. Finally, while it has been known for some time that if there is learning-by-doing in surgery, there may be some significant welfare gains by consolidating procedures. Our results hint at the possibility that in some circumstances a forced uncoupling of merged hospitals

[26] In general, surgeons are not employees of the hospital.

may result in a significant loss of organizational knowledge (at least for some procedures) and the de-merger may lead to welfare loss.

5. Conclusions

It has been known for some time that there is a positive relationship between the number of procedures and the expected outcome for many surgical procedures. In this paper we examine the data for three procedures (the Whipple, CABG, and the repair of aortic aneurysms) in two states (Florida and California) in an attempt to sort out the causal relation between volume and outcomes. We seek to determine if volume causes improved outcomes or whether hospitals that are better at performing a procedure attract more patients.

Our results indicate that for at least two of the three procedures learning-by-doing plays an important role in explaining the difference across hospitals in their risk-adjusted outcomes. That is, for AAA repair and CABG, our results support the presumption of most of the medical literature—volume causes mortality reductions. This may also hold for the Whipple procedure. Furthermore, our findings suggest that the consolidation of these procedures has the potential to improve welfare through a reduction of mortality. Our results also suggest that the degree of organizational forgetting differs significantly across procedures, with the Whipple procedure having the slowest forgetting rate and CABG having the highest.

References

Amundsen, S. et al. (1990) “Abdominal aortic aneurysms. Is there an association between surgical volume, surgical experience, hospital type and operative mortality? Members of the Norwegian Abdominal Aortic Aneurysm Trial,” *Acta Chir Scandanvian*, 156(4) 323-327.

Argote, L. (1999) *Organizational Learning: Creating, Retaining and Transferring Knowledge*, New York, NY: Springer.

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

Benkard, L. (2000) “Learning and Forgetting: The Dynamics of Aircraft Production,” *American Economics Review*, 90(4): 1034-54

Birkmeyer, J. et al. (1999a) “Effect of hospital volume on in-hospital mortality with pancreaticoduodenectomy,” *Surgery*, 125(3):250-256.

Birkmeyer, J. et al. (2002) “Hospital volume and surgical mortality in the United States,” *The New England Journal of Medicine*,” 346:1128-37.

Birkmeyer, J. et al. (1999a) “Relationship between hospital volume and late survival after pancreaticoduodenectormy,” *Surgery*, 126176-183.

Block, P.C., et al., (1998) Identification of variables needed to risk adjust outcomes of coronary interventions: evidence-based guidelines for efficient data collection. *Journal of the American College of Cardiology*, 32(1): p. 275-82.

Capps, Cory, David Dranove, and Mark Satterthwaite (2003) “Competition and Market Power in Option Demand Markets.” *RAND Journal of Economics* 34: 737-63.

Chen, et al. (1996) “Predictors of death in nonruptured and ruptured abdominal aortic aneurysms,” *Journal of Vascular Surgery*, 24(4): 614-20.

Cowan, J. et al. (2003) “Surgical treatment of intact thoracoabdominal aortic aneurysms in the United States: hospital and surgeon volume-related outcomes,” *Journal of Vascular Surgery*, 37(6):1169-1174.

Darr, E., Argote, L. and Epple, D. (1995) “The Acquisition, Transfers and Depreciation of Knowledge in Service Organizations: Productivity in Franchises,” *Management Science*, 41(11): 1750-62.

Dubin, J., and McFadden, D. (1984) “Econometric Analysis of Residential Electric Appliance Holdings and Consumption,” Econometrica, 52 (2): 345-62.

Dimick, J. et al. (2002a) "The volume-outcome effect for abdominal aortic surgery: differences in case-mix or complications?" *Archives of Surgery*, 137(7): 828-832.

Dimick, J. et al. (2002b) "Variation in death rate after abdominal aortic aneurysmectomy in the United States: impact of hospital volume, gender, and age," *Annals of Surgery,* 235(4): 579-85.

Edmondson, Amy, Bohmer Richard and Pisano, Gary (2001). "Speeding Up Team Learning." *Harvard Business Review*.

Epstein, A. (2002) "Volume and outcome — It is time to move ahead," *The New England Journal of Medicine*, Editorial, 346: 1161-1164.

Farley, D.E., and Ozminkowski, RJ. (1992) "Volume-outcome relationships and in-hospital mortality: the effect of changes in volume over time," *Medical Care*, 30:77-94.

Gaynor, M., Seider, H. and Vogt, W. (2004) "Volume-Outcome and Antitrust in US Health Care Markets," Mimeo.

Glasgow, R.E., Mulvihill, S.J. (1996) "Hospital volume influences outcome in patients undergoing pancreatic resection for cancer," *Western Journal of Medicine*, 165(5):294-30.

Goodney, P., Lucas, F., and Birkmeyer, J. (2003) " Should volume standards for cardiovascular surgery focus only on high-risk patients?" *Circulation*, 107(3): 384-387.

Gordon, et al. (1999) "Complex gastrointestinal surgery: Impact of provider experience on clinical and economic outcomes," *Journal of the American College of Surgery*, 189(1): 46-56.

Gordon, et al. (1998) "Statewide regionalization of pancreaticoduodenectormy and its effect on in-hospital mortality," *Annals of Surgery*, 228(1):71-78.

Gowrisankaran, G., Town, R.J. (2003). "Competition, Payers and Hospital Quality." *Health Services Research* 38: 1403-22.

Halm, E., et al. (2002) "Is volume related to outcome in heath care? A systematic review and methodologic critique of the literature," *Annals of Internal Medicine*, 137: 511-52.

Hannan, E. (1992) "A longitudinal analysis of the relationship between in-hospital mortality in New York State and the volume of abdominal aortic aneurysm surgeries performed," *Health Services Research*, 27(4): 517-542.

Hamilton, B.H., and Hamilton V. (1997) "Estimating surgical volume-outcome relationships applying survival models: account for frailty and hospital fixed effects, *Health Economics*, 6:383-95.

Hamilton, B.H., and Ho V. (1998) "Does practice make perfect? Examining the relationship between hospital surgical volume and outcomes for hip fracture patients in Quebec," *Medical Care*, 36: 892-903.

Herzlinger, Regina E. and Peter Stavros (2002). *MedCath Corporation A*. Harvard Business School Case 303041.

Ho, V. Town R.J. and Heslin, M., (2005) Regionalization versus Competition in Complex Cancer Surgery," Mimeo. Rice University.

Kantonen, I. et al. (1997) "Mortality in ruptured abdominal aortic aneurysms. The Finnvasc Study Group," *European Journal of Vascular and Endovascular Surgery*, 14(5):375-379.

Luft HS, Bunker JP, Enthoven AC. (1979) "Should operations be regionalized? The empirical relation between surgical volume and mortality," *The New England Journal of Medicine*, 301:1364-1369.

Luft, H.S., Hunt S.S., Maerki S.C., (1987) The volume-outcome relationship: practice makes perfect or selective referral patterns? *Health Services Research*, 22:157-82.

Manheim, L. et al. (1998) "Hospital vascular surgery volume and procedure mortality rates in California, 1982-1994," *Journal of Vascular Surgery*, 28(1): 45-56.

Newey, W. (1987) "Efficient estimation of limited dependent variables models with endogenous explanatory variables," *Journal of Econometrics*, 36: 231-250.

Pearce, W. (1999) "The importance of surgeon volume and training in outcomes for vascular surgical procedures," *Journal of Vascular Surgery*, 29(5): 768-776.

Rivers, D. and Vuong, Q. (1988) "Limited Information Estimators and Exogeneity Tests for Simultaneous Probit Models," Journal of Econometrics, 39(3): 347-66.

Sollano, J. et al. (1999) "Volume-outcome relationships in cardiovascular operations: New York State, 1990-1995," *Journal of Thoracic and Cardiovascular Surgery*, 117(3) 419-428.

Town, R. and Vistnes, G. (2001) "Hospital Competition in HMO Networks*," Journal of Health Economics*, 20(4): 733-53.

Ubach, D., Bell, C. and Austin, P. (2003) "Differences in operative mortality between high- and low-volume hospitals in Ontario for 5 major surgical procedures: estimating the number of lives potentially saved through regionalization," *Canadian Medical Association Journal*, 168(11): 1409-1414.

Wen, S.W. et al. (1996) “Hospital volume, calendar age, and short term outcomes in patients undergoing repair of abdominal aortic aneurysms: the Ontario experience, 1988-92,” *Journal of Epidemiology and Community Health*, 50(2): 207-213.

Wooldridge, C. (2002). *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

Table 1
Summary Statistics of Estimation Samples
(standard deviations in parentheses)

	Whipple		AAA		CABG	
Variable	FL	CA	FL	CA	FL	CA
	(1)	(2)	(3)	(4)	(5)	(6)
In-hospital mortality	10.4%	9.2%	6.6%	6.1%	4.4%	4.1%
Mean quarterly volume (patient weighted)	3.2 (3.5)	2.1 (1.6)	11.7 (9.2)	6.6 (5.0)	192.2 (163.2)	98.0 (74.2)
Number of hospitals	167	265	192	331	132	189
Mean distance (km) traveled to admitting hospital[1]	21.3 (25.0)	18.6 (21.5)	17.2 (19.7)	16.5 (20.6)	22.4 (22.1)	22.6 (24.6)
Percent female	47.2	47.7	20.3	20.9	28.3	27.6
Mean age	65.5 (11.7)	64.1 (10.7)	70.1 (12.5)	71.4 (7.94)	67.2 (10.2)	66.5 (10.6)
Mean per capita zip code income	$22,391	$25,216	$22,575	$24,585	$22,298	$24,417
Mean number of beds of hospital (patient weighted)	663.9 (523.6)	323.6 (197.0)	526.3 (404.9)	285.0 (149.6)	644.6 (446.7)	322.2 (150.9)
Percent teaching (patient weighted)	39.4	26.4	19.9	13.5	29.6	15.2
N	2,894	1,582	36,384	14,207	275,948	128,716

[1] Conditional on the distance being less than 120 km.

Table 2
Multinomial Hospital Choice Model Parameters

Variable	Whipple	AAA	CABG
Distance	-.12 (.0089)	-.11 (.0018)	-.11 (.0025)
$(\text{Distance}/100)^2$	.00044 (.0000098)	.00075 (.0000098)	.00035 (.0000076)
Closest	.60 (.082)	.047 (.016)	.47 (.013)
Beds/100	.0011 (.00011)	.00016 (.000030)	.000046 (.000016)
Distance×Beds/1000	.00054 (.00024)	.00010 (.000083)	.0025 (.000048)
Distance × Age	-.024 (.011)	-.065 (.0020)	-.0096 (.0028)
Teaching Hospital	-.66 (.13)	-.36 (.032)	1.14 (.018)
Distance×Teaching Hospital	.044 (.0042)	.011 (.00094)	-.0061 (.0056)
N Log Likelihood	4,455 -2,917	49,020 -77,466	96,047 -125,470

Note: standard errors in parentheses.

Table 3
Whipple Procedure: Impact of Volume on Mortality

Variable	ML Probit: Volume Exogenous	ML Probit: Hospital FE, Volume Exogenous	ML Probit: Volume Endogenous	Rivers and Vuong Probit: Hospital FE, Volume Endogenous
	(1)	(2)	(3)	(4)
Sqrt volume	-.21** (.065)	-.064 (.096)	-.089 (.13)	-.12 (.19)
Age	-.011 (.020)	-.024 (.029)	-.011 (.020)	-.028 (.032)
Age^2	.00026 (.00016)	.00041 (.00023)	.00026 (.00016)	.00045 (.00021)
Black	.042 (.12)	-.0030 (.17)	.056 (.12)	-.028 (.17)
Female	-.047 (.058)	-.11 (.072)	-.045 (.058)	-.093 (.073)
Myocardial infarction	-.21 (.17)	-.70* (.30)	-.23 (.17)	-.66* (.33)
Renal failure	1.56** (.14)	2.01** (.23)	1.56** (.14)	2.26** (.27)
Liver disease	.79** (.17)	.72** (.23)	.79** (.17)	.80** (.24)
Emergency admit	.27** (.090)	.20** (.11)	.27** (.089)	.21** (.098)
Number of comorbid conditions	.12** (.015)	.18** (.023)	.12** (.015)	.18** (.026)
Logarithm of zip code income	-.16 (.083)	-.10 (.12)	-.17 (.084)	-.099 (.13)
Teaching hospital	-.20 (.13)	---	-.24 (.14)	---
Logarithm of number of beds	-.010 (.057)	---	-.045 (.079)	---
Florida	.36** (.089)	---	.37** (.092)	---
ρ (for ML) or $\hat{v}_{2j(i)t(i)}$ (for Rivers and Vuong)	---	---	-.067 (.058)	.072 (.22)
Likelihood	-1,183	-908	-3,689	-786
N	4,475	3,337	4,475	3,263

Note: Standard errors in parentheses. Annual dummies also included as regressors.
**Significant at 1% level
*Significant at 5% level

Table 4

AAA: Impact of Volume on Mortality

Variable	ML Probit: Volume Exogenous	ML Probit: Hospital FE, Volume Exogenous	ML Probit: Volume Endogenous	Rivers and Vuong Probit: Hospital FE, Volume Endogenous
	(1)	(2)	(3)	(4)
Sqrt volume	-.078** (.014)	-.045* (.018)	-.056 (.033)	-.038 (.066)
Age	-.021** (.0033)	-.021** (.0052)	-.021** (.0033)	-.024** (.0051)
Age^2	.00030** (.00030)	.00031** (.00041)	.00030** (.000029)	.00033** (.00041)
Black	.17** (.063)	.14** (.068)	.18** (.064)	.14** (.071)
Female	.18** (.022)	.18** (.022)	.18** (.022)	.19** (.024)
Myocardial infarction	.98** (.046)	1.00** (.048)	.98** (.046)	.96** (.048)
Renal failure	1.05** (.032)	1.06** (.034)	1.05** (.032)	1.08** (.035)
Liver disease	1.29** (.19)	1.30** (.20)	1.29** (.19)	1.25** (.21)
Emergency admit	.63** (.035)	.62** (.036)	.63** (.035)	.63** (.040)
Number of comorbid conditions	.12** (.0052)	.14** (.0056)	.12** (.0052)	.14** (.0062)
Logarithm of zip code income	-.12** (.029)	-.081** (.033)	-.12** (.031)	-.084** (.034)
Teaching hospital	.10 (.042)	---	.099 (.043)	---
Logarithm of number of beds	.040 (.028)	---	.017 (.042)	---
Florida	.099* (.036)	---	.095* (.038)	---
ρ (for ML) or $\hat{v}_{2j(i)t(i)}$ (for Rivers and Vuong)	---	---	-.022 (.030)	-.00031 (.078)
Likelihood	-9,648	-9,235	-69,058	-9,237
N	50,520	50,520	50,520	50,520

Note: Standard errors in parentheses. Annual dummies also included as regressors.
**Significant at 1% level
*Significant at 5% level

Table 5
CABG: Impact of Volume on Mortality

Variable	ML Probit: Volume Exogenous	ML Probit: Hospital FE, Volume Exogenous	ML Probit: Volume Endogenous	Rivers and Vuong Probit: Hospital FE, Volume Endogenous
	(1)	(2)	(3)	(4)
Sqrt volume	-.022** (.0046)	-.014* (.0057)	-.040** (.0088)	-.027* (.012)
Age	-.021** (.0051)	-.018** (.0050)	-.021** (.0051)	-.018** (.0050)
Age^{2}	.00027** (.000039)	.00025** (.000038)	.00027** (.000040)	.00025** (.000038)
Black	-.0059 (.031)	-.038 (.033)	-.016 (.032)	-.038 (.033)
Female	.14** (.0097)	.13** (.0096)	.14** (.0096)	.13** (.0096)
Myocardial infarction	.27** (.016)	.26** (.015)	.27** (.015)	.26** (.015)
Renal failure	.94** (.023)	.93** (.023)	.93** (.024)	.93** (.023)
Liver disease	.99** (.074)	.98** (.073)	.98** (.075)	.98** (.073)
Emergency admit	.095** (.014)	.11** (.012)	.070** (.020)	.11** (.013)
Number of comorbid conditions	.13** (.0047)	.14** (.0041)	.13** (.0047)	.14** (.0041)
Logarithm of zip code income	-.030 (.023)	-.041** (.015)	-.016 (.026)	-.041** (.015)
Teaching hospital	.11* (.049)	---	.096 (.056)	---
Logarithm of number of beds	.054 (.037)	---	.11 (.053)	---
Florida	.17** (.041)	---	.19** (.040)	---
ρ (for ML) or $\hat{v}_{2j(i)t(i)}$ (for Rivers and Vuong)	---	---	.073* (.036)	.014 (.011)
Likelihood	-59,225	-58,276	-1,102,228	-58,275
N	404,575	404,565	404,565	404,565

Note: Standard errors in parentheses. Annual dummies, indicators for the presence of ventricular arrhythmia, performance of an angioplasty, and indicators for the number of vessels bypassed also included as regressors.
**Significant at 1% level
*Significant at 5% level

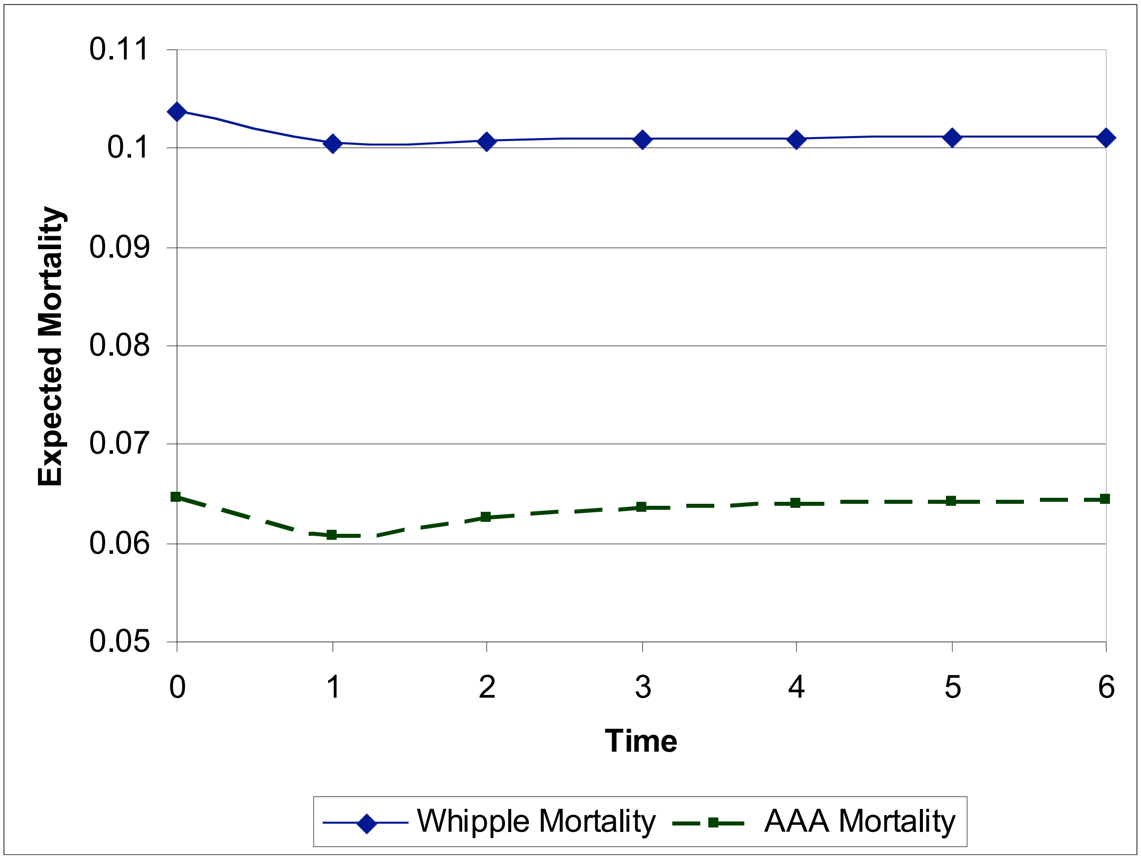
Table 6
Impact of Experience with Forgetting on Mortality

	Whipple	AAA	CABG
Sqrt experience	-.098**	-.062**	-.022**
	(.038)	(.016)	(.0053)
Retained learning	.93**	.51**	-.0093**
	(.080)	(.16)	(.000013)
Logarithm of zip code	-.15*	-.11*	-.030
income	(.082)	(.029)	(.023)
Logarithm of number of	.023	.054	.054
beds	(.058)	(.028)	(.037)
Teaching	-.16	.064	.11*
	(.12)	(.045)	(.049)
Log Likelihood	-1,180	-9,645	-59,224
N	4,475	50,520	404,565

Note: Standard errors in parentheses. The specifications also include patient severity and annual dummies as explanatory variables.
**Significant at 1% level
*Significant at 5% level

Figure 1

Dynamic Impact of Transitory Doubling of Volume in Period 1
for the Whipple Procedure and AAA
(Base volume: Whipple = 2; AAA = 8)


Expected Mortality
Time
Whipple Mortality
AAA Mortality