# The Effect of Failure on Performance over Time:

# The Case of Cardiac Surgery Operations

Emmanouil Avgerinos

IE Business School (Instituto de Empresa), emmanouil.avgerinos@ie.edu

Bilal Gokpinar

UCL School of Management, University College London, b.gokpinar@ucl.ac.uk

Ioannis Fragkos

Rotterdam School of Management, Erasmus University fragkos@rsm.nl

Failure is a common occurrence in many operational contexts involving knowledge work. Concentrating on highly critical cardiac surgery operations, we investigate how failure affects subsequent performance over time. In addressing our research questions, we draw on the sensemaking perspective and incorporate behavioral aspects of failure that are often overlooked. We develop three hypotheses on the effects of failure (i.e., in-hospital mortality of a patient) and test them with a unique dataset of 4,306 cardiac surgery operations from a large European hospital, spanning five years. Our findings show that while failure promotes learning over time and improves task execution quality (as measured by patients' reduced length of stay) in the long term, its effect is the opposite in the short term. Our work also unravels how relational dynamics (i.e., familiarity) may reduce the short-term effects of failure. We find evidence that team familiarity mitigates the detrimental effects of recent failures. This implies that certain team assignment strategies after failure (e.g., putting individuals into familiar teams) may be preferable than others. We explore and illustrate this by conducting a policy simulation based on our data. Our paper provides new insights into how operations managers can support their employees in moving forward after failure.

*Key words*: healthcare operations, failure, sensemaking, quality, team familiarity.

## 1. Introduction

From academics submitting grant proposals or research papers to lawyers representing clients before a court, an unsuccessful outcome or failure is a common occurrence in many operational contexts involving knowledge work. This is more so in healthcare settings where individuals (e.g., surgeons) perform many critical tasks with often uncertain outcomes (e.g., surgeries). While failure may facilitate individual and organizational learning and improve future performance (Cannon & Edmondson, 2005; KC, Staats, & Gino, 2013) by stimulating the development of new ways of approaching existing problems and the adoption of new strategies (Baum & Dahlin, 2007), it may also induce negative reactions involving anxiety, anger or shame, and be detrimental to performance (Carmeli & Gittell, 2009; Zhao, 2011).

This study examines the behavioral and learning dynamics that are likely to take place after failure and the consequences this can have on individuals' subsequent execution of tasks, and seeks to answer the following two research questions: How does failure affect subsequent performance over time? Does team familiarity affect the relationship between recent failure and subsequent performance and if so, to what extent? While recent studies have investigated the effect of failure on individuals (KC et al., 2013), the mechanism by which this effect unfolds over time has received little attention. We focus on the consequences of failure based on temporal dynamics and investigate both its long-term and short-term impacts. Exploring the link between failure and subsequent performance is interesting from a theoretical perspective, but it is also of significant practical importance for operations managers who have to make team allocation decisions in the aftermath of a recent failure and also need to support their employees in moving forward after failure.

Learning from failure seldom takes place in isolation but involves interpersonal activities and dynamic interactions among organizational members (Elkjaer, 2003; Kozlowski & Ilgen, 2006). We consider the relational aspects associated with learning from failure and the recent work which highlights the importance of shared work experience (i.e., team familiarity) on learning and task performance (Reagans et al., 2005, Huckman et al., 2009; Avgerinos & Gokpinar, 2017). Consequently, our second research question addresses whether and to what extent team familiarity affects the relationship between recent failure and subsequent performance.

In addressing our research questions and examining the performance implications of failure in the short and long term, we draw on the rich and growing literature on sensemaking (Weick, 1995; Weick et al., 2005; Maitlis & Christianson, 2014; Schabram & Maitlis, 2017). Sensemaking is the process by which individuals work to understand and react to uncertain events or occurrences around them through continuous interpretation and action (Thomas et al., 1993; Christianson & Sutcliffe, 2009). A key feature of sensemaking is the active nature of understanding, as individuals play a role in creating the actual situations they try to understand in retrospect (Weick et al., 2005; Cornelissen, 2012; Sutcliffe, 2013). Research has shown how past failures can promote sensemaking among individuals, leading to learning and improved performance (Weick, 1988, 1990, 1993; Gephart, 1993; Thomas et al., 2001; Colville et al., 2013; Maitlis & Christianson, 2014). Similarly, we argue that sensemaking is a key process through which individuals learn from failure (Christianson et al., 2009; Catino & Patriotta, 2013) in our setting. Moreover, we propose that failure is likely to trigger negative emotions (Shepherd & Cardon, 2009; Zhao, 2011), which in turn may hinder the process of sensemaking and learning from the failure experience (Shepherd, 2009; Byrne & Shepherd, 2015). Emotions are increasingly considered to be part of the sensemaking process (Maitlis & Christianson, 2014), with emerging research connecting negative emotions with superficial sensemaking (Liu & Maitlis, 2014). Finally, we examine how past shared experiences interact with recent failure by arguing that highly familiar individuals share greater levels of information, which leads to the creation of shared mental models that help them to create cognitive maps of their situation (Weick, 1995) and therefore promote the process of sensemaking (Akgun et al., 2012).

We address our research questions using a highly detailed data set of all cardiac surgeries conducted in a private hospital in Europe over five years. Patients' in-hospital mortality has been commonly characterized as failure in the healthcare operations literature (KC & Terwiesch, 2011; KC et al., 2013) and has been found to be a highly negative experience for hospital employees (Gerber & Workman, 1995; Lenart et al., 1998). Throughout our study, we use failure as meaning "lack of success" (Oxford English Dictionary, 2018), which is different than an 'error'. Our setting involving cardiac operations is ideal to address our research questions for several reasons. First, the longitudinal nature of our sample allows us to observe the effects of failure over time. Second, in cardiac surgery, failure is well defined. In-hospital mortality is widely accepted as a measure of failure for cardiac operations (KC & Terwiesch, 2011; KC et al., 2013). Thirdly, epitomizing today's dynamic, high pressure and high uncertainty knowledge work environment, cardiac surgery operations have significant ramifications (Edmondson, 1999; Tucker & Edmondson, 2003; Nembhard & Edmondson, 2006), whereby behavioral dynamics are likely to be at play alongside learning. In a cardiac surgery setting, individuals are assembled to work on a specific operation in fluid teams. After the operation, the team is dissolved, and individuals become part of another team to perform a different operation. That is, individuals develop familiarity with each other over time through their shared work experience in the same teams. Our sample in this study includes 4,306 cardiac operations of the following types: Coronary artery bypass grafting (CABG), valve replacement/repair, tumor removal, heart failure, congenital heart surgery, routine cardiac surgery, other normal surgeries which do not belong in any of the aforementioned groups, double surgeries which include two cardiac surgeries of the previous groups and triple surgeries. Finally, our performance measure for a cardiac operation (i.e., task execution quality) is a patient's hospital length of stay (LOS) after an operation, a widely accepted quality measure for cardiac surgeries (Schneider & Epstein, 1996; Hannan et al., 1997; Guru et al., 2005; KC & Terwiesch, 2009; KC &Terwiesch, 2011).

Our paper offers a number of important contributions to operations and healthcare management literatures. First, it provides a more nuanced understanding of the effects of failure by distinguishing between its long term and its short term effects. Drawing on the sensemaking perspective and incorporating often overlooked emotional aspects associated with failure, we hypothesize and find that while failure promotes learning over time, which in turn improves task execution quality in the long term, it reduces task execution quality in the short term.

Secondly, we provide new insights into how operations managers can support employees in moving forward after failure. Our work unravels how relational dynamics (i.e., shared work experience) may reduce the short-term effects of failure. We find evidence that shared work experience mitigates the detrimental effects of recent failures. This result implies that certain team assignment strategies after failure (e.g., putting individuals into familiar teams) may be preferable than others as they lead to greater execution quality in subsequent tasks. Previous research highlighted a significant research gap relating sensemaking with key team processes (Maitlis and Christianson 2014). Our paper is an attempt to address this gap by linking team familiarity –an important team level property (Reagans et al. 2005)—with the sensemaking process in an environment with significant emotional dynamics.

Finally, we provide practical insights for hospital managers which may help mitigate the possible negative effects of failure. We build and analyze a simple policy simulation to examine team assignment policies based on our data. We devise a simulation policy and compare this with current assignments policies by quantifying corresponding effects based on our empirical model and the real data.

In the next section we develop our theoretical framework and motivate our hypotheses, describe our setting and data, and then present our empirical strategy, findings, robustness checks, and simulation policies. Finally, we discuss the conclusions of our study and acknowledge its limitations.

## 2. Hypotheses Development

Previous research has highlighted the significance of failure in the future success of organizations as well as individuals (Chuang & Baum, 2003; Edmondson, 2011). There is wide consensus that past failures can promote performance both at the organizational and at the individual level. Failure has been shown to be important for organizational learning (Haunschild & Sullivan, 2002; Haunschild & Rhee, 2004) as it can lead to the development of new company strategies (Baum & Dahlin, 2007). Failures can also trigger an organization to adopt new approaches and solutions to avoid similar results in the future and can increase creativity and innovation within organizations (Kim et al., 2009; Madsen & Desai, 2010). Similarly, past failures can be helpful for an individual because they may highlight what went wrong and lead to the adoption of new approaches that can address existing problems and difficulties (Sitkin, 1992; March & Simon, 1993). After a failure, an individual may be motivated to identify why it occurred, which could then result in a different behavior (Cyert & March,1963; Locke et al., 1981), including changing the way of processing and sharing available information (Sitkin, 1992; KC et al., 2013).

While the aforementioned literature makes a strong case for learning from failure, resulting in improved performance, it does not focus on how such learning develops over time. We explore this notion of failure-driven learning through the lens of the sensemaking framework (Weick, 1995; Weick et al., 2005), which can help us to understand the behavior of individuals when they are confronted with issues or events that are unexpected, ambiguous, or uncertain. That is, individuals make a retrospective sense of the occurrences, which then provokes their subsequent actions (Weick, 1993; Maitlis, 2005). In constructing our theoretical framework and developing our hypotheses, we build on the sensemaking perspective as the key explanatory mechanism for the learning process that takes place after a failure has occurred.

Sensemaking is essentially a process through which individuals deal with uncertainty by creating a rational account of events that enables action (Maitlis, 2005) and through which they continue to enact the environment (Gioia & Chittipeddi, 1991; Brown, 2000). The sensemaking process involves three stages (Gioia & Chittipeddi, 1991; Shepherd, 2009): i) scanning (i.e., information gathering and selective search), ii) interpretation (i.e., attending to and ascribing meaningful labels to incoming information for comprehension and action), and iii) learning (i.e., changes to existing practices and actions taken).  Research has established an important link between this three-stage sensemaking process and performance outcomes (Thomas et al., 1993; Haas, 2006)

Sensemaking has been suggested to be instrumental in teams' and individuals' learning from experience (Huy, 1999; Christianson et al., 2009). A significant body of work relates learning from failures through the process of sensemaking (Weick, 1988, 1990, 1993; Gephart, 1993; Thomas et al., 2001; Colville et al., 2013; Maitlis & Christianson, 2014). Kayes (2004) examines how sensemaking affects team learning for a group of climbers. Ron et al. (2006) shows the importance of sensemaking for flight crews to achieve psychological safety and team learning, and Haas (2006) studies how teams operating in knowledge-intensive environments fail to learn if they do not engage in sensemaking. At the individual level, Catino and Patriotta (2013) show the importance of sensemaking for learning from in-flight errors in the Italian Air Force, and Christianson et al. (2009) examine how leaders' sensemaking response to the collapse of a museum roof facilitated learning. Ravasi and Turati (2005) find that sensemaking plays a vital role in learning from past mistakes for entrepreneurs.

Building on this understanding of the sensemaking perspective, we argue that people learn from failure by continuously developing plausible retrospective accounts of previous failure occurrences (i.e, scanning) (Weick et al., 2005) and by converting these into meaningful knowledge (i.e., interpretation), which then informs actions (Thomas et al., 1993). Indeed, learning develops from the interaction of interpretation and action and over time (Schwandt, 2005), and each failure incidence reveals additional pieces of information (Weick & Sutcliffe, 2007). This newly generated knowledge should be put into action to assess its validity and use. Consequently, we posit that a greater number of failures leads to a better sensemaking and comprehension of the corresponding tasks. We therefore expect such sensemaking as a result of previous failures to improve task execution quality (Catino & Patriotta, 2013).

Sensemaking is suggested to be even more critical in settings where there is inherent uncertainty (Maitlis & Chirstianson, 2014), such as in healthcare (Jordan et al., 2009). Cardiac surgeries are a good example of such settings with uncertainty arising first and foremost from the condition of the patients (Argote, 1982; Gittel, 2002). We expect past failures to promote the process of sensemaking among surgical team members. After experiencing failure, surgical staff members are more likely to seek outside information (KC et al., 2013) and to explore new approaches and refinements to existing operating techniques. Especially in high-risk surgical operations such as cardiac surgery, team-level learning plays a significant role (Avgerinos & Gokpinar, 2018) in developing a deeper understanding and knowledge of the surgical tasks (KC & Staats, 2012). Past failures may encourage individuals to consult their colleagues to identify differences in their approaches in an attempt to detect the root cause of their failure (KC et al., 2013). Specifically, in our context, consultation among surgical staff members of the same hospital commonly takes place through informal and formal interactions or "grand rounds" (Parrino &White, 1990; Manian & Jannsen, 1996; Clark & Huckman, 2012). We therefore expect individuals to seek greater engagement and interaction after a failure.

Indeed, the stages of sensemaking and the subsequent learning process can be illustrated by several case studies of patient failures in the medical literature. For example, Thomson (2018) examines a case where a trauma patient experienced very low blood pressure after an operation, which eventually led to her death. The author explains how he first wrote the mandatory death note, which outlined the events that led to the patient's death (i.e., the information gathering or scanning stage). Later, he describes a self-reflecting process (i.e., the interpretation stage), which involved reviewing possible actions that could have led to a different outcome. She then mentions a team meeting where the case was reviewed, the potential causes of death were hypothesized (e.g., myocardial infarction), and alternative actions were outlined (e.g., taking a preoperative electrocardiogram). After many hours of reviewing and while being sued by the patient's relatives, the doctor concluded that the most likely cause of death was different than what they had supposed (anaphylactic reaction). Although the patient had shown some symptoms, such as nausea and vomiting, these had not been interpreted correctly because they commonly occur in various situations (i.e., the learning stage). The surgeon published this case to inform and raise awareness within the surgical community[1]. In another case, Michalsky et al. (2013) describe the case of an obese 19-year-old who died during a gastric bypass surgery. The surgeons felt that the mortality risk of obese young adults undergoing such operations was not well understood and therefore the decision to operate should be revised carefully. In a recent study, Raffensperger reports a structured approach to sensemaking (2019:p.70): "*if a patient died after an operation, it was important to find surgical errors [...]. We discussed these cases during the weekly surgical pathology conference, not so much as a rebuke to the guilty surgeon, but as a lesson so that the error would not be repeated.*"

As a result, we believe that the process of sensemaking is promoted among surgical team members over time through the acquisition of new information and will eventually result in new and improved

[1] Certain journals, such as the *Journal of Pediatric Surgery CASE REPORTS*, publish example cases that outline important lessons learned and often offer suggestions for improving or extending existing regulations.

approaches that can have a positive effect on future performance (Piaget, 1963; KC et al., 2013). Hence, we expect that:

Hypothesis 1: *Individuals' cumulative prior experience of in-hospital mortality of patients significantly decreases a patient's length of stay.*

**Recent Failure and Performance**

Although we expect learning to take place after failure, this process may not be simple and straightforward due to the uncertain nature of tasks in most knowledge-intensive environments (Haas, 2006). In such settings, the root cause of failure and the specific learning points may not be immediately apparent, as individuals need time to make sense of the failure by scanning relevant information, interpreting it, and learning from it (Daft & Weick, 1984; Gioia & Chittipeddi, 1991). Failure can quickly trigger negative emotions (Huy, 2002; Kiefer, 2005; Shepherd & Cardon, 2009) such as sadness, fear, anxiety, and guilt (Zhao, 2011), which can limit individuals' initiative taking (Urda & Loch, 2013) or make them less confident the next time they perform the same or similar task (Locke et al., 1981), both of which can have a negative effect on individual performance (Fugate et al., 2008).

Emotions are a key part of the sensemaking process (Bartunek et al., 2006; Dougherty & Drumheller, 2006; Liu & Maitlis, 2014). Rafaeli et al. (2009) argue that positive emotions within a team promotes the process of sensemaking by developing high-quality shared mental models shared more broadly by the team members. Liu and Maitlis (2014) empirically show how positive emotions can lead to deeper sensemaking, whereas mixed or negative emotions can result in superficial sensemaking. Cornelissen et al. (2014) show how emotion arousal and contagion negatively shape sensemaking and can lead to errors.

Negative emotions stirred quickly after a failure can significantly interfere with the sensemaking process as individuals' information processing abilities can be hindered by negative emotions (Mogg et al., 1990). Individuals' limited information gathering and processing capacity might also be preoccupied with the negative emotions generated by the failure event (Nolen-Hoeksema & Morrow, 1991). Such negative emotions are common among hospital staff members (Gerber & Workman, 1995; Lenart et al., 1998) particularly after the death of a patient, which is a highly stressful experience and can cause grief and burnouts to surgical staff (Hipwell et al., 1989; Gerber & Workman, 1995; Lenart et al., 1998). We expect that the death of a surgery patient will have a temporary negative effect on the task execution quality of the surgical team (Wilson & Kirshbaum, 2011) as a result of negative emotions and superficial sense-making (Liu & Maitlis, 2014).

Evidence that substantiates our theory can also be found in the medical literature. Goldstone et al. (2004) find qualitative evidence of increased surgical complications for patients who are operated by a surgeon who experienced death in another surgery up to 48 hours before the focal operation, and question whether a surgical team should continue working soon after experiencing a patient death during surgery. Whitehead (2012) reports that physicians may doubt their abilities after a patient's death, as it is often ambiguous whether they did everything they could to prevent such an outcome. Waterman et al. (2007) report that physicians are more anxious about future errors, less confident about their professional capacity, and have reduced job satisfaction and quality of sleep after serious medical errors. Although follow-up reviews of failures, such as mortality and morbidity meetings, aim to support individuals and help productive sensemaking, Whitehead (2012) reports that they often cultivate doubts about a physician's competence[2]. Such negative emotions hinder sensemaking (Walsh & Bartunek, 2011), and this implies that individuals with recent failure experience may not benefit from them immediately.

In all, we argue that sensemaking after a failure is a process that involves multiple stages (i.e., scanning, interpretation, and learning) over time, and in the immediate aftermath of a failure, negative emotions are likely to dominate and hinder the sensemaking process or lead to superficial sensemaking. In turn,

[2] As one physician reports (Whitehead 2012, p.272): "You're exposed, and... people are able to see you're not perfect, and you have flaws... You always want that ability to say you are competent—and that detracts from your professional identity"

individuals' processing and decision-making will be hampered, leading to possible suboptimal actions, which in turn result in lower task execution quality.

Given our setting, we define as recent any death of any patient that occurred within a week (seven days) prior to the focal operation. Similar definitions have been used by other scholars depending on their specific context. For example, in a bank setting Staats and Gino (2012) define recent (i.e., short term) as the same day, but they also noted: "Short- and long-term time periods would differ in other operational contexts. For instance, in a context for which tasks generally last only several seconds, a short time period might be comparatively smaller (e.g., an hour), whereas a setting with more tasks at least five hours long would have a comparatively longer short time period (e.g., perhaps a week)." (Staats & Gino, 2012: p. 143). With similar reasoning (i.e., the average duration of an operation in our sample is about five hours) and also after consulting with healthcare professionals in our setting, we use a seven-day window to define recent operations.

*Hypothesis 2: Individuals' recent (same week) experience of in-hospital mortality of patients significantly increases a patient's length of stay.*

### Recent Failure and Shared Work Experience

We next focus on how recent failures interact with past shared work experience. An emerging stream of research has highlighted the significant role of shared work experience (e.g., team familiarity) on performance (Reagans et al., 2005; Huckman et al., 2007; Avgerinos & Gokpinar, 2017). As individuals become more familiar with each other, teams become better coordinated and individuals become more motivated in performing their tasks. As a result of acquiring essential information about the abilities and areas of expertise of each other, teams with shared work experience tend to perform better than teams with limited or no shared work experience. (Reagans et al., 2005). Moreover, team familiarity can enhance the relationships among team members (Edmondson, 1999; Reagans et al., 2005; Easton & Rosenzweig, 2012), which in turn makes individuals more motivated and willing to work harder to meet the team’s standards and goals. Huckman, Staats, and Upton (2009) also show that team familiarity increases quality performance of project teams in the software industry.

As we argued above, the sensemaking process after a failure involves significant psychological dynamics stemming from negative emotions especially shortly after failure (Walsh & Bartunek, 2011; Holt & Cornelissen, 2014). Negative emotions as a result of failure not only make individuals more risk averse (Lerner & Keltner, 2001), but also affect other attitudes and behaviours that can lead to decreased trust and commitment among members and to lower work performance (Patterson & Cary, 2002; Kiefer, 2005). In contrast, shared work experience has been suggested to promote psychological safety among team members (Edmondson, 1999), especially in the context of cardiac surgeries (Edmondson et al., 2003). Shared work experience also provides greater levels of trust among members (Faraj & Sproull, 2000) and helps to develop cohesion (Evans & Dion, 1991; Mullen & Copper, 1994; Gully et al., 1995), which in turn improves performance (George & Bettenhausen, 1990; Vinokur-Kaplan, 1995). As such, we propose that the emotion-driven detrimental effects of a recent failure are mitigated by increased levels of shared work experience due to increased trust, psychological safety, and cohesion among team members, all of which can facilitate the sensemaking process (Akgun et al., 2012).

A specific type of sensemaking (known as resourceful sensemaking) refers to the ability of individuals to appreciate the perspectives of others through interactions, and in turn use this comprehension to act accordingly (Wright et al., 2000). Highly familiar individuals tend to share more information, resulting in the creation of shared mental models, which can help team members develop cognitive maps of their environment (Weick, 1995). Familiar individuals may also feel more comfortable in expressing their disagreement and discussing their concerns, which improves the sensemaking process (Ashmos & Nathan, 2002).

Although negative emotions as a result of a failure make it difficult to discuss and learn from the experience (Shepherd et al. 2009) and engage in deeper sensemaking through discussions (Liu & Maitlis, 2004), shared work experience can alleviate such difficulties as it increases the willingness of individuals to engage in a relationship with others (Reagans et al. 2005), which in turn can help to identify the appropriate course of action. Since sensemaking is a social act that is grounded both at the individual level and at interactions with others (Weick, 1995; Maitlis, 2005), past shared work experience can promote the sensemaking process within team members. Due to the above reasons, we expect that:

Hypothesis 3: *Team familiarity interacts with individuals' recent (same week) experience of in-hospital mortality of patients so that the negative effect of recent failure on patient's length of stay is significantly decreased.*

## 3. Setting, Data, and Variables

As the focus of this study is knowledge-intensive healthcare operations, similar to recent work in the operations management literature (KC & Staats, 2012; Avgerinos & Gokpinar, 2018), we test our hypotheses using data from cardiac surgery operations. Our sample is drawn from the cardiac unit of a European private hospital, which includes 300 beds and serves more than 2,000 patients every year. Our dataset includes all cardiac operations that were conducted in the hospital between April 1, 2011 and April 30, 2016.

A typical cardiac surgery team in our archival dataset consists of two to eight individuals. Specifically, each team includes one lead surgeon, up to four assistant surgeons, up to one anesthesiologist, up to one perfusionist, and up to three scrub nurses. Our sample comprises 102 individuals: 15 lead surgeons, 31 assistant surgeons, 10 anesthesiologists (six since the beginning of our dataset), 10 perfusionists (six since the beginning of our dataset), and 36 nurses (35 since the beginning of our dataset). For every case, our sample contains data regarding the specific surgery type, the patient's characteristics and severity of condition, the members of the surgical team, the length of stay after the operation, and the in-hospital mortality, which we define as failure. Our sample has a total of 169 in-hospital deaths, 160 of which occur within 60 days of the corresponding operation. It is worth noting that we define failure as the time of the patient's death and not the time of the patient's operation that led to death. After admission and before the operation, each patient is allocated into one of the three categories with the following tags: "severe", "medium", and "mild". Finally, we also carried out several informal interviews with healthcare professionals in our hospital. This helped us form a better view of cardiac surgery in general and gave us information about hospital policies and cardiac unit practises.

Our initial dataset included 4,306 operations. After removing observations with missing data, we ended up with 4,272 operations. After consulting with a number of healthcare professionals in our setting, we divided our operations into the following types: Coronary artery bypass grafting (CABG), valve replacement/repair,  tumor removal, heart failure, congenital heart surgery, routine cardiac surgery, other normal surgery (namely operations that do not belong in any of the aforementioned groups), double surgery (namely operations that include two cardiac surgeries of the previous groups) and triple surgery (namely operations that include two cardiac surgeries of the previous groups). All operations in the last two categories are characterized as "severe".

## Variables

**Dependent Variable.** Our dependent variable is a patient's length of stay (i.e., the number of days) after the operation. This variable captures the execution quality of the operation and is commonly used as a quality measure for cardiac surgeries (Schneider & Epstein, 1996; Hannan et al., 1997; Guru et al., 2005; KC & Terwiesch, 2009, 2011). Shorter lengths of stay have been associated with better clinical outcomes (Gaynes et al., 2001; Guru et al., 2005; Gibbons et al., 2011), a fact that was also confirmed during our staff interviews in the hospital.

### Independent Variables

**Average Individual Failure.** We count the number of times each team member experienced the death of their patients (in-hospital mortality) in the past (excluding the current operation). We then divide this number by the number of team members. We exclude recent failures when calculating this variable.[3]

**Recent Average Individual Failure.** We count the number times each team member experienced a death of their patients during the past week[4] prior to the current operation). We then take the sum of all team members and divide it by the number of team members. Unlike other studies that use a standardized 30-day or 60-day mortality as a performance measure (dependent variable), we are interested in the aftermath of experiencing an in-hospital mortality (i.e., failure) and its subsequent performance implications. Thus, in-hospital mortality in our study is the primary independent variable, and as long as we observe in-hospital mortality during or after a surgery (regardless of when it happens), we consider it an instance of our independent variable. That is, we are not interested in when in-hospital mortality occurs, but rather what happens after in-hospital mortality takes place (i.e., its learning and subsequent performance implications).

**Shared work experience.** We operationalize this in two ways and then include as it a key interaction term in our models (see below). In our first set of analysis (Table 4), we use **Team Familiarity.** We first calculate how many times every pair of the team has collaborated in a previous operation (excluding the focal one). Next, we sum this up for every pair of the team and then divide by all possible pairs within the team. Our approach follows past literature (Reagans et al., 2005; Huckman et al., 2009; Avgerinos & Gokpinar, 2017) that uses average familiarity of the team. In our second set of analysis (Table 5), we use **Leader Familiarity.** We count how many times the lead surgeon worked together with the other team members in the past, take the sum of all these pairs and then divide this by all possible pairs within the team that include the lead surgeons.

[3] Our results remain the same qualitatively (i.e., in terms of hypotheses support) after including recent failures too.

[4] We re-ran our analyses using five and ten days instead and obtained results with the same level of significance and hypotheses support.

**Recent Average Individual Failure x Team (Leader) Familiarity.** We multiply these two variables to test our third hypothesis.

## Control Variables

**Team Size.** This variable is equal to the number of team members. Team size can affect team performance (Gladstein, 1984; McGrath, 1984), as larger teams may have increased resources, but they may also face coordination challenges (Hackman, 2002; Reagans et al., 2005; Huckman et al., 2009). Team size also varies with cardiac surgery types (see Table 1).

**Average Individual Direct Experience.** This variable captures the experience of each team member on each surgery type. To ensure that failure experience drives our results and not general experience, we control for overall experience. As described above, each operation falls into one of the nine categories. We calculate how many times each team member takes part in a specific operation category before the current one (excluding the focal one). Next, we sum it up for all members and divide this number by the number of team members.

**Time Fixed Effect.** We include monthly dummies indicating the month since the beginning of our dataset to capture organizational experience and any potential changes in hospital policy or technological development that can affect a patient’s length of stay. We also consider an alternative operationalization where we replace our monthly dummy variables with a continuous variable, which is equal to the number of months since the beginning of our dataset. We obtain the same results with this alternative specification.

**Lead Surgeon Fixed Effect.** We capture the experience and skill level of the lead surgeon, which may affect the length of stay of a patient after an operation.

**Male.** We take into account the gender of the patient with this variable, which is equal to one when the case involves a male patient and zero if the patient is female.

**Age.** We also control for the age of the patient since it can have an effect on the length of stay after an operation.

**Indicators for Severity.** During the first hours of admission, each patient is classified into one of three categories (mild, medium, or severe) by the physician examining the patient based on his/her characteristics that may influence the risk level of the operation such as past and current health problems, sex and age. Severe operations are more critical and challenging tasks than mild and medium ones. We therefore include the dummy variables "medium" and "severe" to control for the patient's condition. The variable "medium" is set equal to one when the patient's condition is considered medium. Similarly, the variable "severe" is set equal to one if the patient's condition is considered severe.

**Indicators for Procedure Type.** Each operation is classified into one of the nine categories. We consider CABG, which is the most frequently occurring operation type in our sample, as the reference category and we include eight dummy variables for the eight indicators we have. We set them equal to one if the operation is characterized as the respective type and zero otherwise.

## 4. Empirical Strategy and Analysis

Table 2 presents basic statistics for all our variables in their raw form. For our main analysis we use ordinary least squares regression in which each surgery is an observation. We also check for normality of residuals, heteroscedasticity, and serial autocorrelation. The Breusch-Pagan test (Breusch & Pagan, 1979) indicated potential heteroscedasticity, which led us to run all models using robust standard errors (Huber, 1967; White, 1980). We also use the commands *gladder* and *ladder* in Stata to check the distribution of our variables. We take the logarithm of our continuous variables *Length of Stay*, *Average*

*Individual Failure*, *Team Size*, *Average Individual Direct Experience* and *Team Familiarity* according to the traditional learning-curve model[5]. Hence, our final model is the following:

$$
\begin{aligned}
ln(los_t) = \beta_0 + \beta_1\, & ln(Average\ Individual\ Failure_t) \\
& \beta_2\ Recent\ Average\ Individual\ Failure_t + \\
& \beta_3\ Recent\ Average\ Individual\ Failure_t \ x\ ln(Team\ Familiarity_t) + \\
& \beta_4\, ln(Team\ Familiarity_t) + \\
& \beta_5\ ln(Team\ Size_t) + \\
& \beta_6\ ln(Average\ Individual\ Direct\ Experience_t) + \\
& \beta_7\ Severe_t + \\
& \beta_8\, Medium_t + \\
& \beta_9\ Male_t + \\
& \beta_{10}\ Age_t + \\
& \beta_{11}\ Other\ Controls_t + \\
& \beta_{12}\ u_t
\end{aligned}
$$

Table 3 shows the descriptive statistics and correlations among our variables for all team members. Note that the numbers are the logged values for all variables except Recent Average Individual Failure, Severe, Medium, Male, and Age. Table 4 shows the results for all our Hypotheses. In model 1, we include only the control variables. In model 2, we add *Average Individual Failure* and *Recent Average Individual Failure* and the adjusted R$^2$ increases by 8.74%. Next, we conduct an F-test and find that model 2 is superior than model 1 at 1%. *Average Individual Failure* is negative and significant at 1%, providing support for our first hypothesis. An increase of one standard deviation in this variable decreases length of stay by 22.69% (2.38 days). *Recent Average Individual Failure* is positive and significant at 5%, providing support for H2. An increase of one standard deviation in *Recent Average Individual Failure* increases length of stay by 24.58% (2.58 days).

[5] We also repeated our analysis using the linear form of all variables and obtained the same results in terms of significance and hypotheses support.

In model 3 we test our third hypothesis by adding the interaction term *Recent Average Individual Failure x Team Familiarity* and the adjusted $R^2$ further increases by 1.90%. We also conduct an F-test and find that model 3 is superior to model 2 at 1%. The interaction term is negative and significant at 1%, which provides full support for our third hypothesis.

Finally, we follow past literature (Aiken and West 1991, Dawson and Richter 2006) and perform post-hoc analysis for our third hypothesis. Specifically, we take two subsets of our sample: One with the values above the mean plus one standard deviation for Team Familiarity and one with the values below the mean minus one standard deviation for Team Familiarity. Next, we plot Recent Individual Average Failure and Length of Stay for these two subsets (Figure 1). The figure shows that the detrimental effect of failing recently indeed decreases in the presence of a highly familiar team.

**Lead Surgeons**

Next, we focus only on the lead surgeons and conduct our analysis by investigating only their failures. We include both the main effect of Leader Familiarity and as part of the interaction term to test our third hypothesis (i.e., we calculate only the familiarity of the lead surgeon with the rest of the team members). We also count the number of total failures and recent failures only for the lead surgeon for every team, while controlling for the experience of the other team members. Table 5 shows our results using only the lead surgeons. In model 1, we include only the control variables. In model 2, we add the variables *Individual Failure* and *Recent Individual Failure* and the adjusted $R^2$ increases by 8.78%. Next, we conduct an F-test and find that model 2 is superior than model 1 at 1%. *Individual Failure* is negative and significant at 1%, providing support for our first hypothesis. An increase of one standard deviation in this variable decreases Length of Stay by 5.49% (0.58 days). *Recent Individual Failure* is positive and significant at 5%, providing support for H2. An increase of one standard deviation in this variable increases length of stay by 27.77% (2.91 days).

In model 3 we test our third hypothesis by adding the interaction term *Recent Individual Failure x Leader Familiarity* and the adjusted $R^2$ further increases by 1.42%. We also conduct and F-test and find that model 3 is superior to model 2 at 1%. The interaction term is negative and significant at 5%, which provides full support for our third hypothesis.

## 5. Robustness Checks

Our results are significant both from a statistical and from an operational perspective. We run several checks to strengthen the robustness of our findings. First, we consider alternative specifications of our control variables. Namely, we replace our control variable *Average Individual Direct Experience* with another variable called *Average Individual Experience*. This new variable is calculated the same way as the former one, with the only difference being it counts all operations that each member has conducted prior to the current operation, as opposed to the direct version, which considers only the focal operation type in the past. We repeat our analysis and obtain similar results (i.e., we get the same support for all hypotheses).

A potential issue with our results is that we have no information before the starting date of our sample, which could affect our findings. We deal with this concern by repeating our main analysis after dropping a number of time intervals since the starting date of our sample. Specifically, we drop the first three, six and, nine months of our dataset, calculate all our variables using the remaining observations, and repeat our analysis. When we remove the first 9 months, for example, we only use the observations between months 9 and 61 to calculate *Team Familiarity*. We get full support for our three hypotheses. Table 6 shows the results after dropping the first nine months of our sample. We therefore believe that missing data does not constitute an issue for our analysis since our results remain the same both significantly and organizationally.

We also perform an analysis to check how sensitive our findings are to the removal of different time intervals (up to 12 months) since the starting date of our sample. In this case, we drop the first 12 months and repeat our analysis without recalculating our variables of interest. The results remain the same qualitatively (Table 7).

We also repeat our analysis in another subsample of our dataset which involves surgeons with more than average failure experiences. Specifically, we focus on the 11 lead surgeons (i.e., we drop the bottom 25th percentile of lead surgeons with the smallest percentage of failure) with the highest rate of failures in our sample and conduct our analysis with them. Table 8 presents the results which provide support for all our hypotheses.

In addition, we compare the average number of failures during a week after a failure experience by any of the team members and the weeks after an operation that did not result in a failure. We find that the former number is significantly higher at 5% than the latter, which provides further support for our second hypothesis.

Another concern for our second hypothesis could be that the lead surgeon who faces a recent failure may become more cautious and keep patients in the hospital for observation longer. To alleviate this concern, we repeat our analysis by only concentrating on the cases where the lead surgeon has not experienced a recent failure, but at least one another team member has experienced a failure within the past week. We therefore conduct a more conservative analysis in which the variable 'recent failures' excludes the lead surgeon's personal failures. The effect is positive (i.e., increasing Length of Stay) and significant, supporting hypothesis 2. We also compare the average duration of the operations within a week after a failure experience by the lead surgeons and the weeks after operations that did not result in a failure. Our rationale is that if lead surgeons were becoming more cautious after failure, then we would expect them to take more time to complete an operation after experiencing failures. However, our results suggest the opposite: The average duration of the operations during a week after a failure experience by the lead surgeon is actually shorter (p-value = 0.0216) than those that were conducted within a week that did not include a failure for the lead surgeon.

Another issue for our hypothesis 1 is that the variance of past individual failures within a team may affect our dependent variable. That is, a team in which some team members have very few or no failure experiences whereas others have many failure experiences would have high variance of individual failure experiences. This may then result in differences in risk-taking attitudes among team members and may cause interpersonal conflicts during surgery. To address this concern, we repeat our analysis by also accounting for failure dispersion. We define it as the standard deviation of all individual failures within a team following Harrison and Klein (2007)'s characterization to capture the difference between two individuals that differ in their position along a continuous attribute (i.e., number of past failures). After running our analysis, we find that variance of individual failures (i.e., variance dispersion) is insignificant, and our results do not change in terms of hypothesis support.

Moreover, anesthesiologists, perfusionists, and scrub nurses are assigned according to their shifts, and assistant surgeons are randomly assigned by the hospital. Nonetheless, one could assume that lead surgeons might choose collaborating with particular assistants. Hence, we repeat our analysis after excluding assistant surgeons when calculating our variables *Team Familiarity* and *Leader Familiarity*. Hereby we exclude any potential preferences of lead surgeons for working with specific assistant surgeons. In addition, we calculate all the percentages of the cases that each lead surgeon collaborated with each anesthesiologist, perfusionist, and head nurse in our sample. Our results indicate that these percentages are not significantly different for different lead surgeons. Specifically, the highest percentage for anesthesiologists is 47.83% (the lead surgeon collaborated with five of the ten anesthesiologists of our sample), the highest percentage for perfusionists is 19.57% (the lead surgeon collaborated with six of the ten perfusionists in our sample), and the highest percentage for head nurses is 22.22% (the lead surgeon  collaborated with eight of the 36 head nurses in our sample). Considering these percentages, we don't find any evidence for possible systematic selection of  anesthesiologists, perfusionists, or head nurses by the lead surgeons. We then repeat our analysis after excluding the observations that included lead surgeon-anesthesiologist couples, lead surgeon-perfusionist couples, and lead surgeon-head nurse couples with the highest percentages of collaborations and obtained the same results for all our hypotheses.

Our setting is a private hospital with an excellent reputation in Europe, where surgeons and other surgical team members have considerable experience and comparable skills and are asked to be able to conduct every cardiac surgery type. Our informal interviews with hospital management confirm that surgical teams are indeed formed based on a rotating schedule and there is no practice of certain lead surgeons working with particular team members (e.g., a lead surgeon cannot choose to work with a specific head nurse during surgery). Also, in our empirical models, we control for the severity of the operation as well as the cardiac surgery type to account for any inherent complexity differences across surgeries. Nevertheless, we conduct additional empirical analyses to further address concerns on team selection and possible endogenous team formation which may bias our findings on the interaction effect in hypothesis 3 (e.g., certain teams may be assembled more frequently and for more difficult cases, which may also experience a greater number of recent failures and greater lengths of stay).

To address these concerns, we first count the number of individuals from our sample who were members of a team in a case where the patient died. We get the following results: 39 surgeons (out of a total of 46), ten anesthesiologists (out of ten), ten perfusionists (out of ten) and 30 nurses (out of 36). Next, we compare the team familiarity between the "severe" and non-severe (i.e., "medium" and "mild") cases, and find that they are not significantly different (p-value>0.10). We also compare team familiarity between the cases where the patient died and the cases that did not, and again find that they are not significantly different (p-value>0.10). Second, we repeat the same two comparisons for the *Average Individual Direct Experience* of the team and find no significant differences (p-value>0.10 in both comparisons). Third, we repeat our analysis after dropping the observations in which team familiarity is higher than the 75th percentile (Table 9). Model 1 in Table 9 provides partial support for our first hypothesis and full support for our second hypotheses, and model 2 in Table 9 provides full support for our first hypothesis and for our third hypothesis.

Another concern could be that some surgeons may avoid cases involving higher risk and prefer accepting only medium and mild cases, which can bias our results. Similarly, the hospital may assign more severe cases to more experienced lead surgeons, which can also create a selection bias in our results. First, we examine the allocation of mild, medium, and severe cases among the lead surgeons in our sample with a chi-square test. We reject the null hypothesis and conclude that there is no indication of unevenly spread of cases among lead surgeons (p-value>0.10 in all three cases). Second, we investigate the allocation of patients' deaths among lead surgeons and again the null hypothesis is rejected, indicating that they are evenly spread among them (p-value>0.10). Third, we investigate the allocation of double and triple cardiac surgeries among lead surgeons, and again the null hypothesis is rejected, indicating that they are evenly spread among them (p-value>0.10). Fourth, we perform a chi-square test in order to test the average age of all patients of every lead surgeon. Again, the null hypothesis is rejected (p-value>0.10). Finally, we conduct limited interviews with hospital staff and managers. They all confirm that hospital policy dictates that surgeons are allocated according to shifts and that no surgeon is allowed to "cherry pick" patients or operations.

Furthermore, a potential bias for our dependent variable might be created by the fact that, according to our interviews, some patients who have undergone simple procedures may choose to stay in hospital longer than necessary to claim full reimbursement from their insurance company. Specifically, even if they can be discharged earlier, they prefer to stay in hospital longer because their insurance will cover their expenses only if they spend a minimum number of days in hospital. To eliminate this bias, we first repeat our analysis after excluding all operations that fall within the category of Routine Cardiac Surgery that includes the simplest procedures in our sample like replacement of a cardiac pacemaker or hemostasis. Our results remain the same qualitatively. In addition, we exclude observations within the 10th percentile for our dependent variable from our full sample and conduct our analysis. We hereby exclude patients who stayed in hospital up to seven days, who will most probably be the ones staying longer than usual, since seven days is a common minimum length of stay required by insurance companies according to our interviews. Again, we obtain the same results qualitatively.

Surgery duration could also affect the length of stay of a patient as longer surgeries could result in patients spending more days in hospital to recover. To eliminate this issue, we repeat our analysis for all our team members after controlling for surgery duration. Our results in Table 10 provide full support for all our hypotheses with surgery duration being partially significant at 10% in all models.

Finally, there might be an unobserved reason that can cause decreased performance for a lead surgeon in our sample, which can contribute to both failure (i.e., death of the patient) and the decreased quality (longer length of stay) in subsequent operations. To investigate this, we employ instrumental variables methods. We seek an instrument for our key independent variable recent failure (i.e., patient death) such that this instrument is highly correlated with recent failure, but does not have any direct effect on our dependent variable (i.e., length of stay). We identify such an instrument as follows: we use a dummy variable equal to one if the lead surgeon has performed an operation characterized as severe prior to the failure event (patient death during the past week), and call this variable "Recent Severe Dummy". We believe this new variable is a strong and valid instrument satisfying both the relevance and exclusion restriction conditions. The instrument clearly affects a lead surgeon's likelihood of experiencing patient death during the past week (all the deaths in our sample come from the severe group), but does not directly affect the lead surgeon's task quality (length of stay) after the experienced failure. That is, if a lead surgeon performs a severe operation on patient A on Monday, for instance, this should not affect the length of stay of the operation she performs on patient B on Thursday (i.e., surgeon's subsequent task execution quality). In addition, in our main model we already control for the experience of the lead surgeon. Hence, we believe that this dummy variable satisfies both the relevance and the exclusion condition, therefore constituting a valid instrument, which allows us to eliminate any potential bias associated with our variable Recent Individual Failure (Antonakis et al., 2010). Table 11 provides our results for the instrumental variable approach. Our instrument is significant at 1% and positive at the first stage (model 1), while the F-value is equal to 28.72, well above the common threshold of 10 for weak instruments (Staiger & Stock, 1997). *Recent Average Individual Failures* is significant at 5% and positive at the second stage (model 2), providing full support for our second hypothesis.

While the instrumental variables approach presented above addresses endogeneity concerns for our main variables, there may be remaining concerns on endogenous team formation, which in turn could bias our findings for the moderating effect of team familiarity. For this, we include an alternative instrumental variables analysis. We identify and use a plausible instrument for team familiarity based on the availability of other individuals that are not part of the team for each focal operation. Specifically, for each member of every focal operation, we count the number of days other individuals of the same role do not appear in our sample since the day that specific member joined the hospital. The higher this number (i.e., the higher the absence of other individuals of the same role), the greater the familiarity (or the opportunity to be familiar) of our focal member with others in the hospital.

For example, for the perfusionist of the focal operation, we count the number of days the other perfusionists in our sample (that have appeared at least once prior to the focal operation) are not assigned in an operation of our dataset between the time the perfusionist of the focal operation first appeared in our dataset and the focal operation. We repeat this approach for every individual of the focal operation and take the average for all team members. We name this variable *Team Availability* and suggest that it is a suitable instrument for team familiarity.

First, we expect *Team Availability* to be positively correlated with *Team Familiarity* because when the other individuals of the same role are not available, the focal individual has a higher chance of collaborating with other team members, therefore satisfying the relevance condition. Second, unavailability of the other team members in the past does not directly affect the length of stay of that individual's surgery, other than increasing her familiarity with the focal operation's team members. We therefore use *Team Availability* as an instrument for *Team Familiarity*. However, we also need a second instrument for the interaction term of *Team Familiarity* and *Recent Failures*. For this, we adopt Wooldridge (2002) as follows: We first run the first stage for *Team Familiarit*y using *Team Availability* as an instrument. Then we take the interaction term of the linear prediction from that model with recent failure and use it as our second instrument for the interaction term of team familiarity and recent failure. We can thus create and use two proper instruments to examine our third hypothesis (Wooldridge, 2002).

It is important to point out that our approach is different from the forbidden regression, in which the interaction term is used as a regressor in the second stage and can lead to biased results (i.e., we use fitted values as an instrument for the interaction term).

We first run the first stage (Table 12) and confirm that our instrument is indeed relevant. Both instruments are positive and significant at 1% for both endogenous variables, with F-values well above 10 (equal to 282.94 and 72.73, respectively). Both the underidentification test (Kleibergen LM Statistic) and the Kleibergen-Paap Wald test statistic to check weak instruments verify that our instruments satisfy the relevance condition. Our second-stage results in Table 13 provide full support for our third hypothesis with the interaction term Recent Average Individual Failure x Team Familiarity being negative and significant at 1%.

## 6. Policy Implications

To further assess the managerial implications of our study, we use our models to devise and simulate a lead surgeon-to-surgery assignment policy, and then compare its corresponding performance with the current dataset. Typically, lead surgeons are assigned to operations using a rotation policy, i.e., after the first surgeon on the list is assigned to a current operation, she is moved to the end of the list, and the next surgeon on the list is assigned to the next operation. This process is repeated for all upcoming operations. This is a well-known policy followed by many hospitals because it ensures a fair allocation and exposure of all surgeons to the full range of operations, thereby enhancing their learning experience. Moreover, it reduces the long-term dependency of the hospital unit on specific surgeons, as every surgeon can carry out a diverse set of operations. Using this common allocation rule as the basis, we devise a policy based on the insight gained from our theory and results. In particular, we propose *Policy Familiarity*, a policy that takes into account the interaction effect of recent failure and team familiarity to improve short-term task productivity. While more sophisticated policies can be designed, our main goal with the present analysis is to demonstrate how managers can design *simple* and *easy-to-implement* policies by making use of the insights generated by our study.

We run this policy using our sample starting from the beginning of the second year to the end of our data period (i.e., we use the first year as a warm-up period to initialize the familiarity score)[6]. During the time period of the policy simulation, we follow a corresponding rule, assign the lead surgeon accordingly, and then calculate the impact of the allocation policy on the length of stay based on our models after recalculating our variables of interest. Finally, we assume weekly shifts for the basis of our operations.

*Policy_Familiarity*: We calculate the Leader Familiarity score for the surgeon who is assigned according to the rotation policy. Note that this familiarity score depends on the number of times that the surgeon has collaborated with other team members. The composition and allocation of other team members are predetermined for the coming week and cannot be changed. If the surgeon experienced a recent failure, *Policy_Familiarity* compares her familiarity score with the prospective team (the first available operation) to her historic average familiarity score. If it is lower, then the surgeon is not deemed familiar with this specific team and is not assigned to the present operation. Instead, we assign her to the next operation, where the same process repeats. Her familiarity score with the team members of a prospective operation is calculated and compared to her historic average. The process is repeated until a suitably ‘familiar’ operation and corresponding team is identified. In the rare case that all possible operations in the coming week result in less-than-average familiarity scores for the surgeon, she is not assigned to any operations in that week. Overall, considering our insights on the role of familiarity in alleviating the negative effects of recent failure, the objective of this policy is to assign each lead surgeon to relatively familiar teams after having experienced a recent failure.

[6] We repeated our analysis in different time intervals of our dataset and the conclusions remain the same.

We conduct a simulation experiment to evaluate the performance of our policy. We implement a procedure that takes a sample of surgery operations, a warm-up period, and a policy name as inputs. It then outputs another data sample which is identical to the original one for surgeries that take place within the warm-up period, but has different lead surgeon allocations and estimated length of stay for each operation after the warm-up period. After the warm-up period, we assume that the surgeries and their outcomes remain identical to the original sample and that the team composition follows a rotation policy, but that the lead surgeon is allocated depending on the indicated policy (i.e., *Policy_Familiarity*). Once we create the team allocation for each operation, the length of stay is calculated using the model 3 in Table 5. This procedure is implemented in Python 2.7.10, using the Pandas library for data manipulations and the NumPy library for algebraic calculations. Having derived the estimated length of stay of each operation, we calculate the averages of *Recent Individual Failure, Leader Familiarity, Recent Individual Failure x Leader Familiarity,* and *Length of Stay* for observations after the warm-up period of our suggested policy. We then compare these averages with their actual counterparts (the current rotation-based policy in the hospital) and report the corresponding percentage changes in Table 14.

As expected, our policy performs better than the current assignment. Specifically, *Policy _Familiarity* has a high impact on *Leader Familiarity* and *Recent Individual Failure x Leader Familiarity*, and decreases Length of Stay by 28.80% (3.02 days). This simple simulation analysis illustrates that indeed practical policies based on our findings could have a major positive effect in improving individuals' task execution quality.

## 7. Conclusions

Failure is a common outcome in today's high-pressure and highly competitive work environments. Considering many sources of uncertainty (e.g., patient condition, clinical and no-clinical dynamics), failure is even more prevalent in healthcare settings. While a large body of research in operations management and healthcare focuses on the causes and antecedents of failure, our examines the consequences and effects of failure both in the long term and in the short term. In studying implications of failure, we concentrate on individuals' learning both individually and as part of a team. After all, with increasingly fragmented work and operations taking place in most organizational contexts, knowledge workers have taken a central role in organizational learning. As such, understanding the mechanisms through which they learn from failure is an important step towards achieving better organizational performance.

To provide a more nuanced understanding of the effects of failure on performance, we draw on the sensemaking perspective by considering the behavioral dynamics associated with failure. Through this novel theoretical perspective, we develop and test three hypotheses on the effects of failure over time. Our results indicate that failure promotes learning over time, which leads to greater task execution quality in the long term. Interestingly however, we find that failure can have the opposite impact in the short term, resulting in reduced performance on subsequent tasks.

We then consider relational related post-failure dynamics and its effects on performance. Our results suggest that shared work experience among individuals can be highly effective in reducing the detrimental effects of a recent failure. This finding provides potentially actionable prescriptions for how to better manage operations in highly critical healthcare settings. Our simulation analysis illustrates that significant performance improvements may be achieved through better allocation of existing healthcare personnel after failure incidences.

Our work comes with some limitations commonly found in empirical studies. First, while we employ a rare and highly granular surgical dataset in testing our hypotheses, this dataset comes from a single cardiac unit. As such, one should be careful in generalizing these findings. In addition, our information regarding the patients' condition prior the operations is limited with relatively coarse indicators for clinical condition (i.e., severe, medium, mild). Ideally, we would prefer to have more detailed preoperative risk information such as Higgins score or EuroScore, or additional patient-level information such as blood glucose levels, comorbidities, and cholesterol measures that could affect the quality of the surgical outcome, but such information was not available in our sample. Moreover, in examining the consequences of failure, we use a single measure for performance, task execution quality, as proxied by patients' length of stay in our study. Future work could consider the interplay of failure with other types of performance measures involving productivity or quality such as in-hospital mortality rates or number of readmissions after surgery. Our performance measure can also be affected by several unobserved factors that are not captured by our dataset. Nonetheless, our set of control variables and our extensive robustness checks (including two instrumental variable approaches) make us confident that our findings are not biased by such omitted variables problems. Furthermore, future work can examine how recent failure interacts with other team-level mechanisms such as team coordination or other surgery-related variables such as surgery preparation time. Additionally, while our theory draws on the sensemaking perspective with theoretical arguments based on emotions and learning, we cannot directly observe or measure, for example, negative emotions or learning processes in this study. Future work could examine and directly measure these after-failure dynamics in field or lab settings. Finally, although our policy simulations provide some actionable insights for hospital managers, ideally, we would like to see these and perhaps other competing policies to be implemented and monitored in a field experimental study.

Despite the aforementioned limitations, our paper addresses a highly interesting and important research question with a unique dataset and compelling theoretical arguments. Our study not only draws on the sensemaking perspective, but also contributes to the growing body of work in sensemaking by connecting the sensemaking process with emotions triggered by patient deaths and by introducing team familiarity as a key team-level dynamic during the sensemaking process in healthcare settings. Indeed, in their review article on sensemaking, Maitlis and Christianson (2014) make a strong case for the role of emotions during sensemaking and call for further research on emotions especially at team and organizational levels. In addition, as sensemaking is a social act that is grounded both at the individual level and at interactions with others (Weick, 1995; Maitlis 2005), Reuben (2007) calls for further investigations on the psychological processes and interrelations among team members that enable sensemaking. Our paper partly responds to these calls by highlighting after-failure emotional dynamics and team familiarity in a complex sensemaking process. Our findings also provide useful insights to both researchers and practitioners, and highlight the need for organizations to make sense of and deal with failures within fluid and uncertain settings such as surgical operations.

Our results in this paper are both statistically significant and economically noteworthy. For example, based on our sample of 4,306 operations (i.e., 70.59 operations per month) and hospital with current failure rates, it appears that through learning as a result of past failures, the total length of stay may be reduced by 70.59*2.38 = 168 days per month, which translates into around 168/10.48 = 16 additional new admissions on average every month. Additionally, our findings suggest that experiencing a recent failure (during the past week of the performed operation) can increase length of stay by 70.59*2.58 = 182 days per month, which is translated into around 182/10.48 = 17 additional patient admissions on average every month.

Also, it seems better to assign individuals that have experienced a recent failure to work in teams with which they have greater prior shared worked experience (a high level of familiarity). In this way, any possible negative effects of a recent failure stemmed from behavioral dynamics (e.g., negative emotions) could be mitigated. In addition, our suggested policy to allocate individuals who experienced a failure with other members familiar to them is beneficial for the individuals, as they have a smoother post-failure experience, since they interact more with members that are familiar to them.

Finally, using real data, our simulation suggests that by considering the role of familiarity in making assignment decisions, average length of stay may be reduced by three days for each patient over a four-year time period. This can help hospitals to accommodate more patients with the same resources and can improve both clinical and cost performance (KC & Terwiesch, 2009).

Our study is a first attempt to examine the effects of failure over time. In addressing this question, we provide insights into learning from failures and offer suggestions into better failure management. Our work highlights that by embracing failure and understanding its implications, managers could make better use of team allocation strategies to improve their workers' short-term and long-term performance.

## References

Aiken, L. S., West, S.G., 1991. Multiple regression: Testing and interpreting interactions. Sage, Thousand Oaks, CA.

Akgun, A.E., Keskin, H., Lynn, G., Dogan, D. 2012. Antecedents and consequences of team sensemaking capability in product development projects. R&D Management 42(5) 473-493.

Antonakis, J., Bendahan, S., Jacquart, P., Lalive, R., 2010. On making causal claims: A review and recommendations. The Leadership Quarterly 21 (6), 1086-1120.

Argote, L., 1982. Input uncertainty and organizational coordination in hospital emergency units. Administrative Science Quarterly 27 (3), 420-434.

Ashmos, D., Nathan, M. 2002 Team sensemaking: a mental model for navigating uncharted territories. Journal of Managerial Issues 14(2), 198–217.

Avgerinos, E., Gokpinar, B. 2017. Team familiarity and productivity in cardiac surgery operations: The effect of dispersion, bottlenecks, and task complexity. Manufacturing & Service Operations Management 19 (1), 19-35.

Avgerinos, E., Gokpinar, B. 2018. Task variety in professional service work: When it helps and when it hurts. Production and Operations Management 27 (7), 1368–1389.

Bartunek J.M., Rousseau D.M., Rudolph J.W., DePalma J.A. 2006. On the receiving end: Sensemaking, emotion, and assessments of an organizational change initiated by others. Journal of Applied Behavioral Science 42(2), 182-206.

Baum, J.A.C., Dahlin, K.B., 2007. Aspiration performance and railroads' patterns of learning from train wrecks and crashes. Organization Science 18 (3), 368–385.

Breusch, T.S., Pagan, A.R., 1979. A simple test for heteroscedasticity and random coefficient variation. Econometrica: Journal of the Econometric Society 47 (5), 1287-1294.

Brown, A. D. (2000). Making sense of inquiry sensemaking. Journal of Management Studies 37(1), 45–75.

Byrne O., Shepherd D.A. Different strokes for different folks: Entrepreneurial narratives of emotion, cognition and making sense of failure. Entrepreneurship Theory and Practice 39(2), 375-405.

Cannon, M. D., Edmondson, A.C., 2005. Failing to learn and learning to fail (intelligently): How great organizations put failure to work to innovate and improve. Long Range Planning 38 (3), 299-319.

Carmeli A., Gittel J.H. 2009. High-quality relationships, psychological safety, and learning from failures in work organizations. Journal of Organizational Behavior 30, 709-729.

Catino M., Patriotta G. 2013. Learning from errors: Cognition, emotions and safety culture in the Italian Air Force. Organization Studies 34(4), 437–467.

Christianson M.K., Farkas M.T., Sutcliffe K.M., Weick K.E. 2009. Learning through rare events: Significant interruptions at the Baltimore & Ohio Railroad Museum. Organization Science 20(5), 846–860.

Christianson M.K., Sutcliffe K.M. 2009. Sensemaking, high-reliability organizing, and resilience. In P. Croskerry, K.S. Cosby, S.M. Schenkel, R.L. Wears (Eds.), Patient safety in emergency medicine (pp. 27–33). Philadelphia, PA: Lippincott Williams & Wilkins.

Chuang Y-T., Baum, J.A.C., 2003. It's all in the name: Failure-induced learning by multiunit chains. Administrative Science Quarterly 48 (1), 33–59.

Clark, J. R., Huckman, R.. 2012. Broadening focus: Spillovers, complementarities and specialization in the hospital industry. Management Science 58(4), 708–722.

Colville, I., Hennestad, B., Thoner, K. 2013. Organizing, changing and learning: A sensemaking perspective on an ongoing 'soap story'. Management Learning 45(2), 216-234.

Cornelissen J. 2012. Sensemaking under pressure: The influence of professional roles and social accountability on the creation of sense. Organization Science 23(1), 118–137.

Cornelissen, J., Mantere, S., Vaara, E. 2014. The contraction of meaning: The combined effect of communication, emotions and materiality on sensemaking in the Stockwell shooting. Journal of Management Studies 51(5), 699-736.

Cyert, R.M., March, J.G., 1963. A Behavioral Theory of the Firm. Prentice-Hall, Englewood Cliffs, NJ.

Daft, R.L., Weick, K.E. 1984. Toward a model of organizations as interpretation systems. Academy of Management Review 9(2), 284–295.

Dawson, J. F., Richter, A.W., 2006. Probing three-way interactions in moderated multiple regression: development and application of a slope difference test. Journal of Applied Psychology 91 (4), 917-926.

Dougherty D.S., Drumheller K. 2006. Sensemaking and emotions in organizations: Accounting for emotions in a rational(ized) context. Communication Studies 57(2), 215–238.

Easton, G.S., Rosenzweig, E.D., 2012. The role of experience in six sigma project success: An empirical analysis of improvement projects. Journal of Operations Management 20 (2012), 481-493.

Edmondson, A.C., 1999. Psychological safety and learning behavior in work teams. Administrative Science Quarterly 44 (2), 350-383.

Edmondson, A.C., 2011 Strategies for learning from failure. Harvard Business Review 89 (4), 48–55.

Edmondson, A.C., Winslow, A.B., Bohmer, R.M.J., Pisano, G.P., 2003. Learning how and learning what: Effects of tacit and codified knowledge on performance improvement following technology adoption. Decision Science 34 (2), 197-224.

Elkjaer B. 2003. Social learning theory: Learning as participation is social processes. In M. Easterby-Smith, & M.A. Lyles (Eds.), The Blackwell Handbook of Organizational Learning and Knowledge Management. Malden, MA; Oxford: Blackwell Publishing Ltd pp. 38-53.

Evans, C.R., Dion, K.L., 1991. Group cohesion and performance a meta-analysis. Small Group Research 22 (2), 175-186.

Faraj, S., Sproull, L., 2000. Coordinating expertise in software development teams. Management Science 46 (12), 1554-1568.

Fugate, M., Knicki, A.J., Prussia, G.E., 2008. Employee coping with organizational change. Personnel Psychology 61, 1-36.

Gephart, R.P. 1993. The textual approach: Risk and blame in disaster sensemaking. Academy of Management Journal 36(6), 1465–1514.

George, J. M., Bettenhausen, K., 1990. Understanding prosocial behavior, sales performance, and turnover: a group-level analysis in a service context. Journal of Applied Psychology 75 (6), 698.

Gerber, D. E., Workman, D.P., 1995. Death in the operating room and postanesthesia care unit: Helping nurses to cope. Journal of Post Anesthesia Nursing 10 (2), 84-88.

Gibbons, C., Bruce, J., Carpenter, J., Wilson, A.P., Wilson, J., Pearson, A., Lamping, D.L., Krukowski, Z.H., Reeves, B.C., 2011. Identifications of risk factors by systematic review and development of risk-adjusted models for surgical site infection. Health Technological Assessment 15 (30), 1-156.

Gioia D.A., Chittipeddi K. 1991. Sensemaking and sensegiving in strategic change initiation. Strategic Management Journal 12(6), 433–448.

Gittel, J.H. 2002. Coordinating mechanisms in care provider groups: Relational coordination as a mediator and input uncertainty as a moderator of performance effects. Management Science 48(11), 1408-1426.

Gladstein, D. L., 1984. Groups in context: A model of task group effectiveness. Administrative Science Quarterly 29 (4), 499-517.

Goldstone, A.R., Mackay, J., Nashef, S.A.M. 2004. Should suurgeons take a break after an intraoperative death? Attitude survey and outcome evaluation. BMJ 328, 379.

Gully, S. M., Devine, D.J., Whitney, D.J., 1995. A meta-analysis of cohesion and performance Effects of level of analysis and task interdependence. Small Group Research 26 (4), 497-520.

Guru V., Anderson G.M., Fremes S.E., O'Connor G.T., Grover F.L., Tu J.V., 2005. The Canadian CABG Surgery Quality Indicator Consensus Panel. The identification and development of Canadian coronary artery bypass graft surgery quality indicators. Journal of Thoracic and Cardiovascular Surgery 130, 1257-1264.

Haas M.R. 2006. Knowledge gathering, team capabilities, and project performance in challenging work environments. Management Science 52(8), 1170–1184.

Hackman, J. R., 2002. Leading Teams: Setting the Stage for Great Performances. Harvard Business Press, Boston, MA.

Hannan E.L., Stone C.C., Biddle T.L., DeBuono B.A. 1997. Public release of cardiac surgery outcomes data in New York: what do New York state cardiologists think of it? American Heart Journal 134(1), 55-61.

Harrison, D.A., Klein, K.J. 2007 What's the difference? Diversity constructs as separation, variety, or disparity in organizations. *Academy of Management Rev*iew 32(4), 1199–1228.

Haunschild, P. R., Rhee, M., 2004. The role of volition in organizational learning: The case of automotive product recalls. Management Science 50 (11), 1545-1560.

Haunschild, P.R., Sullivan, B.N., 2002. Learning from complexity: Effects of prior accidents and incidents on airlines' learning. Administrative Science Quarterly 47 (4), 609–643.

Hipwell, A.E., P.A. Tyler, C.M. Wilson. 1989. Sources of stress and dissatisfaction among nurses in four hospital environments. Psychology and Psychotherapy: Theory Research and Practice 62(1), 71-79.

Holt R., Cornelissen J. 2014. Sensemaking revisited. Management Learning. 45(5), 525-539.

Huber, P.J., 1967. The behavior of maximum likelihood estimates under nonstandard conditions. In: Le Cam, L.M. (Ed.), Neyman, J. (Ed.), Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 2. University of California Press, Berkeley, CA, pp. 221-233.

Huckman, R.S., Staats, BR., D.M. Upton, D.M., 2009. Team familiarity, role experience, and performance: Evidence from Indian software services. Management Science 55 (1), 85-100.

Huy Q.N. 1999. Emotional capability, emotional intelligence, and radical change. The Academy of Management Review 24(2), 325-345.

Huy Q.N. 2002. Emotional balancing of organizational continuity and radical change: the contribution of middle managers. Administrative Science Quarterly 47, 31–69.

Jordan M.E., Lanham H.J., Crabtree B.F., Nutting P.A., Miller W.L., Stange K.C., McDaniel R.R. 2009. The role of conversation in health care interventions: Enabling sensemaking and learning. Implementation Science 4(15), 1-13.

Kayes, D.C. 2004. The 1996 Mount Everest climbing disaster: The breakdown of learning in teams. Human Relations 57(10), 1263–1284.

KC, D.S., Staats, BR., 2012. The effect of focal and related experience on surgeon performance. Manufacturing & Service Operations Management 14 (4), 618-633.

KC, D.S., Staats, BR., Gino, F., 2013. Learning from my success and from others' failure: Evidence from minimally invasive cardiac surgery. Management Science 59 (11), 2435-2449.

KC, D.S., Terwiesch, C., 2009. Impact of workload on service time and patient safety: An econometric analysis of hospital operations. Management Science 55 (9), 1486-1498.

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

Kiefer, T. 2005. Feeling bad: antecedents and consequences of negative emotions in ongoing change. Journal of Organizational Behavior 26, 875–97.

Kim J-Y., Kim J-Y.J., Miner, A.S., 2009. Organizational learning from extreme performance experience: The impact of success and recovery experience. Organization Science 20 (6), 958–978.

Kozlowski S.W.J., Ilgen D.R. 2006. Enhancing the effectiveness of work groups and teams. Psychological Science 7, 77–124.

Lenart S.B., Bauer, C.G., Brighton, D.D., Johnson J.J., Stringer T.M., 1998. Grief support for nursing staff in the ICU. Journal of Nurses in Staff Development 14 (6), 293-296.

Lerner J.S. and Keltner D. 2001. Fear, anger and risk. Journal of Personality and Social Psychology 81, 146–59.

Liu F., Maitlis S. 2014. Emotional dynamics and strategizing processes: A study of strategic conversations in top team meetings. Journal of Management Studies 51(2), 202–234.

Locke, E.A., Shaw, K.N., Saari, L.M., Latham, G.P., 1981. Goal setting and task performance: 1969–1980. Psychology Bulletin 90 (1), 125–152.

Madsen, P.M., Desai, V., 2010. Failing to learn? The effects of failure and success on organizational learning in the global orbital launch vehicle industry. Academy of Management Journal 53 (3), 451–476.

Maitlis, S. 2005. The social processes of organizational sensemaking. Academy of Management Journal 48(1), 21–49.

Maitlis S., Christianson M. 2014. Sensemaking in Organizations: Taking Stock and Moving Forward. The Academy of Management Annals 8(1) 57-125.

Manian, F., Janssen, D. 1996. Curbside consultations: A closer look at a common practice. Journal of American Medical Association 275(2), 145–147.

March, J.G., Simon, H.A., 1993. Organizations. Blackwell, Cambridge, MA.

McGrath, J.E., 1984. Groups: Interaction and Performance. Prentice-Hall Englewood Cliffs, NJ.

Michalsky M., Teich S., Rana A., Teeple E., Cook S., Schuster D. 2013. Surgical risks and lessons learned: Mortality following gastric bypass in a severely obese adolescent. Journal of Pediatric Surgery CASE REPORTS 1, 321-324.

Mogg K., Mathews A., Bird, C., MacGregor-Morris R. 1990. Effects of stress and anxiety on the processing of threat stimuli. Journal of Personality and Social Psychology 59, 1230–1237.

Mullen, B., Copper, C., 1994. The relation between group cohesiveness and performance: An integration. Psychology Bulletin 115 (2), 210-227.

Nembhard, I.M., Edmondson, A.C., 2006. Making it safe: The effects of leader inclusiveness and professional status on psychological safety and improvement efforts in health care teams. Journal of Organizational Behavior 27 (7), 941-966.

Nolen-Hoeksema S., Morrow J. 1991. A prospective study of depression a distress after a natural disaster: the 1989 Loma Prieta Earthquake. Journal of Personality and Social Psychology 61, 105–21.

Oxford English Dictionary. 2018. Oxford University Press. Oxford, UK.

Parrino, T., White, A. 1990. Grand rounds revisited: Results of a survey of U.S. Departments of Medicine. American Journal of Medicine 89(4), 491–495.

Patterson J.M., Cary J. 2002. Organization justice, change anxiety, and acceptance of downsizing: preliminary tests of an AET-based model. Motivation and Emotion 26, 83–103.

Piaget J. 1963. The Psychology of Intelligence. Routledge, New York Reagans, R., Argote, L., Brooks, D., 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.

Rafaeli, A., Ravid, S., Cheshin, A. 2009. Sensemaking in virtual teams: The impact of emotions and support tools on team mental models and team performance. International Review of Industrial and Organizational Psychology 24, 151–182.

Raffensperger, J. 2019. A surgeon’s lessons, learned and lost. Strategic Book Publishing & Rights Agency, LLC.

Reagans, R., Argote, L., Brooks, D., 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.

Ron, N., R. Lipshitz, M. Popper. 2006. How organizations learn: Post-flight reviews in an F-16 fighter squadron. Organization Science 27(8), 1069–1089.

Schabram K., Maitlis S. 2017. Negotiating the Challenges of a Calling: Emotion and Enacted Sensemaking in Animal Shelter Work. Academy of Management 60(2).

Schneider E.C., Epstein A.M. 1996. Influence of cardiac-surgery performance reports on referral practices and access to care. A survey of cardiovascular specialists. New England Journal of Medicine 335(4), 251-256.

Schwandt D.R. 2005. When managers become philosophers: Integrating learning with sensemaking. Academy of Management Learning & Education 4(2), 176-192.

Shepherd D.A. 2009. Grief recovery from the loss of a family business: A multi- and meso-level theory. Journal of Business Venturing 24, 81–97.

Shepherd D.A., Cardon M.S. 2009. Negative Emotional Reactions to Project Failure and the Self-Compassion to Learn from the Experience. Journal of Management Studies 46(6) 923-949.

Sitkin S.B., 1992. Learning through failure: The strategy of small losses. In: Cummings, L.L (Ed.), Staw, B.M (Ed.), Research in Organizational Behavior, Vol. 14. JAI Press, Greenwich, CT, pp. 231–266.

Staats, BR., Gino, F. 2012. Specialization and variety in repetitive tasks: Evidence from a Japanese bank. Management Science 58 (6), 1141-1159.

Staiger, D., Stock, J.H. 1997. Instrumental variable regressions with weak instruments. Econometrica 65(3), 557-586.

Sutcliffe, K. M. (2013). Sensemaking. In M. Augier & D. Teece (Eds.), The Palgrave Encyclopedia of Strategic Management. Basingstoke: Palgrave Macmillan. Advance online publication.

Thomas, J.B., Clark, S.M., Gioia, D.A. 1993. Strategic sense making and organizational performance: linkages among scanning, interpretation, action, and outcomes. Academy of Management Journal 36, 239–270.

Thomas, J.B., S.W. Sussman, J.C. Henderson. 2001. Understanding "strategic learning": Linking organizational learning, knowledge management, and sensemaking. Organization Science, 12(3), 331–345.

Thompson, E.C. 2018. A trauma surgeon on trial. Bulletin of the American College of Surgeons.

Tucker, A.L., Edmondson, A.C., 2003. Why hospitals don't learn from failures: organizational and psychological dynamics that inhibit system change. California Management Review 45 (2), 55-72.

Urda, J., Loch, C.H., 2013. Social preferences and emotions as regulators of behavior in processes. Journal of Operations Management 31 (2013), 6-23.

Vinokur-Kaplan, D., 1995. Treatment teams that work (and those that don't): An application of Hackman's group effectiveness model to interdisciplinary teams in psychiatric hospitals. The Journal of Applied Behavioral Science 31 (3), 303-327.

Walsh I.J., Bartunek J.M. 2011. Cheating the fates: Organizational foundings in the wake of demise. Academy of Management Journal 54(5), 1017–1044.

Waterman, A.D., Garbutt, J., Hazel, E., Dunagan, W.C., Levinson, W., Fraser, V.J., Gallagher, T.H. 2007. The emotional impact of medical errors on practicing physicians in the United States and Canada. The Joint Commission Journal and Quality and Patient Safety 33(8), 467-476.

Weick, K.E. 1988. Enacted sensemaking in crisis situations. Journal of Management Studies, 25(4), 305–317.

Weick, K.E. 1990. The vulnerable system: An analysis of the Tenerife air disaster. Journal of Management, 16(3), 571–593.

Weick, K.E. 1993. The collapse of sensemaking in organizations: The Mann Gulch disaster. Administrative Science Quarterly, 38(4), 628–652.

Weick, K.E. 1995. Sensemaking in organizations. Thousand Oaks, CA: Sage Publications.

Weick, K.E., Roberts, K.H., 1993. Collective mind in organizations: Heedful interrelating on flight decks. Administrative Science Quarterly 38 (3), 357-381.

Weick K.E., Sutcliffe K.M. 2007. Managing the unexpected: Resilient performance in an age of uncertainty (2nd eds.). San Francisco, CA: Jossey-Bass.

Weick K.E., Sutcliffe K.M., Obstfeld D. 2005. Organizing and the process of sensemaking. Organization Science, 16(4), 409–421.

White, H., 1980. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica: Journal of the Econometric Society 48 (4), 817-838.

Whitehead, P.R. 2012. The lived experience of physicians dealing with patient death. BMJ Supportive & Palliative Care 4, 271-276.

Wilson, J., M. Kirshbaum. 2011. Effects of patient death on nursing staff: a literature review. British Journal of Nursing, 20 (9), 559-563.

Wright, C.R., Manning, M.R., Farmer, B., Gilbreath, B. 2000. Resourceful sensemaking in product development teams. Organization Studies, 21(4), 807–825.

Zhao, B. 2011. Learning from errors: The role of context, emotion, and personality. Journal of Organizational Behavior 32, 435-463.

**Figure 1**

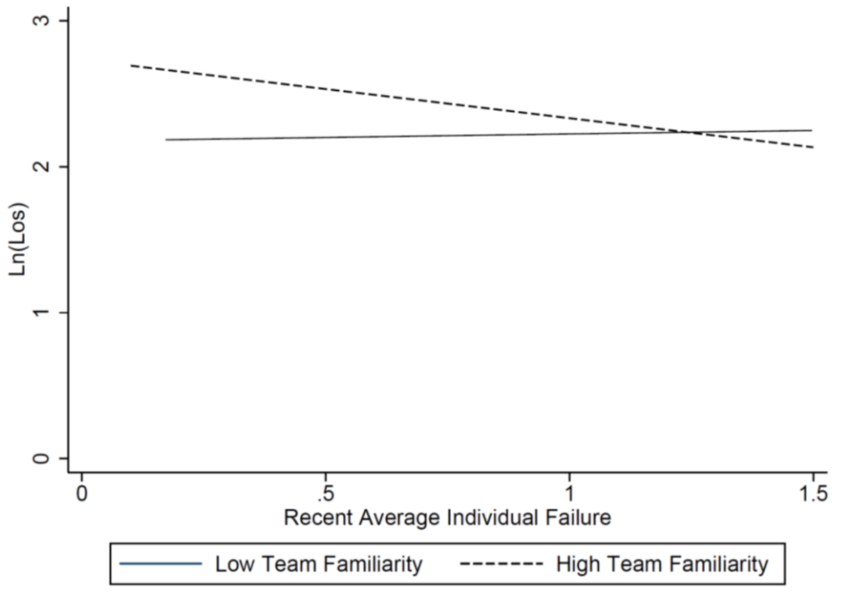

**Table 1. Team Size for Different Operations**

| Operation Type | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|
| Valve Operation | 5.547 | 0.864 | 3 | 8 |
| CABG | 4.963 | 0.697 | 3 | 8 |
| Congenital Surgery | 5.875 | 0.64 | 5 | 7 |
| Heart Failure | 5.245 | 1.011 | 3 | 7 |
| Tumor Removal | 5.714 | 0.726 | 4 | 7 |
| Routine Surgery | 4.356 | 1.005 | 2 | 6 |
| Other Normal Surgey | 5.044 | 0.999 | 3 | 7 |
| Double Surgery | 5.677 | 0.825 | 3 | 8 |
| Triple Surgery | 5.966 | 0.908 | 5 | 8 |

**Table 2. Summary Statistics in Raw Form**

| Variable | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|
| 1. Length of Stay | 10.478 | 11.469 | 0 | 194 |
| 2. Average Individual Failure | 11.779 | 12.753 | 0 | 20.451 |
| 3. Recent Average Individual Failure | 0.101 | 0.216 | 0 | 1.500 |
| 4. Team Familiarity | 296.297 | 423.503 | 0 | 828.811 |
| 5. Team Size | 5.200 | 0.849 | 2 | 8 |
| 6. Average Individual Direct Experience | 198.132 | 289.669 | 0 | 1019.424 |
| 7. Severe | 0.163 | 0.137 | 0 | 1 |
| 8. Medium | 0.606 | 0.239 | 0 | 1 |
| 9. Male | 0.726 | 0.445 | 0 | 1 |
| 10. Age | 65.265 | 11.311 | 10 | 96 |

# Table 3. Descriptive Statistics for all Team Members

| Variable | Mean | Std. Dev. | Min | Max | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1. Length of Stay | 2.16 | 0.5 | 0 | 5.268 | 1 | | | | | | | | | | |
| 2. Average Individual Failure | 1.81 | 0.556 | 0 | 3.018 | -0.084** | 1 | | | | | | | | | |
| 3. Recent Average Individual Failure | 0.101 | 0.216 | 0 | 1.5 | -0.020 | 0.196** | 1 | | | | | | | | |
| 4. Team Familiarity | 4.858 | 1.117 | 0 | 6.72 | 0.013 | 0.545** | -0.021 | 1 | | | | | | | |
| 5. Team Size | 1.635 | 0.171 | 0.693 | 2.079 | 0.030+ | -0.087** | 0.026+ | 0.029+ | 1 | | | | | | |
| 6. Average Individual Direct Experience | 4.923 | 1.445 | 0 | 6.927 | -0.231** | 0.385** | 0.215** | 0.488** | 0.031* | 1 | | | | | |
| 7. Severe | 0.163 | 0.137 | 0 | 1 | 0.383** | -0.266** | -0.487** | -0.044** | 0.038* | -0.370** | 1 | | | | |
| 8. Medium | 0.606 | 0.239 | 0 | 1 | 0.047** | 0.076** | 0.221** | 0.053** | 0.034* | 0.153** | -0.462** | 1 | | | |
| 9. Male | 0.726 | 0.445 | 0 | 1 | 0.007 | 0.040** | 0.013 | 0.031* | -0.005 | -0.017 | -0.002 | 0.027+ | 1 | | |
| 10. Age | 65.265 | 11.311 | 10 | 96 | 0.025 | -0.009 | -0.023 | -0.003 | 0.002 | -0.021 | 0.034* | -0.015 | -0.074** | 1 | |
| 11. Recent Average Individual Failure x Team Familiarity | 6.801 | 1.878 | 0 | 10.080 | -0.045** | 0.550** | 0.537** | 0.722** | 0.032* | 0.628** | -0.300** | 0.166** | 0.033* | -0.016 | 1 |

+, * and ** denote significance at 10%, 5% and 1% levels respectively

Logged values of all variables except Recent Average Individual Failure, Severe, Mild, Male and Age

# Table 4. Regression on Length of Stay using all members

| Variable | Los | | |
|---|---|---|---|
| | Model: (1) | (2) | (3) |
| Average Individual Failure | | -0.1788** | -0.2007** |
| | | (0.0634) | (0.0660) |
| Recent Average Individual Failure | | 0.3661* | 0.3064+ |
| | | (0.1665) | (0.1660) |
| Recent Average Individual Failure x Team Familiarity | | | -0.0211** |
| | | | (0.0052) |
| Team Size | 0.1871** | 0.1763** | 0.1407* |
| | (0.0482) | (0.0557) | (0.0598) |
| Team Familiarity | -0.0808** | -0.0580** | -0.0535** |
| | (0.0173) | (0.0175) | (0.0180) |
| Individual Average Direct Experience | -0.0452** | -0.0454** | -0.0453** |
| | (0.0112) | (0.0114) | (0.0115) |
| Medium | 0.1144** | 0.1127** | 0.1143** |
| | (0.0102) | (0.0102) | (0.0101) |
| Severe | 0.3130** | 0.3934** | 0.4232** |
| | (0.0377) | (0.0373) | (0.0387) |
| Age | -0.0001 | -0.0000 | 0.0000 |
| | (0.0004) | (0.0004) | (0.0004) |
| Male | -0.0050 | -0.0080 | -0.0101 |
| | (0.0120) | (0.0117) | (0.0117) |
| Constant | 4.1366** | 4.7785** | 4.9327** |
| | (0.1783) | (0.2811) | (0.3021) |
| Observations (N) | 4,272 | 4,272 | 4,272 |
| Adjusted $R^2$ | 0.1645 | 0.1789 | 0.1823 |
| Month Fixed Effect | Yes | Yes | Yes |
| Lead Surgeon Fixed Effect | Yes | Yes | Yes |
| Procedure Fixed Effect | Yes | Yes | Yes |

+, * and ** denote significance at 10%, 5% and 1% levels respectively

## Table 5. Regression on Length of Stay using only the lead surgeons

| Variable | Los | | |
|---|---|---|---|
| | Model: (1) | (2) | (3) |
| Individual Failure | | -0.1554** | -0.1549** |
| | | (0.0514) | (0.0513) |
| Recent Individual Failure | | 0.3553* | 0.4004* |
| | | (0.1584) | (0.1556) |
| Recent Individual Failure x Leader Familiarity | | | -0.0084** |
| | | | (0.0018) |
| Team Size | 0.2446** | 0.1634** | 0.1591** |
| | (0.0586) | (0.0536) | (0.0536) |
| Leader Familiarity | -0.0292* | -0.0273* | -0.0298* |
| | (0.0126) | (0.0116) | (0.0116) |
| Individual Average Direct Experience | -0.0431** | -0.0426** | -0.0425** |
| | (0.0103) | (0.0104) | (0.0104) |
| Medium | 0.1076** | 0.1102** | 0.1108** |
| | (0.0100) | (0.0092) | (0.0092) |
| Severe | 0.4816** | 0.4456** | 0.4432** |
| | (0.0340) | (0.0326) | (0.0326) |
| Age | -0.0112 | -0.0147 | -0.0148 |
| | (0.0116) | (0.0108) | (0.0108) |
| Male | -0.0000 | 0.0001 | 0.0001 |
| | (0.0004) | (0.0004) | (0.0004) |
| Constant | 3.6797** | 3.0865** | 3.0156** |
| | (0.2386) | (0.2619) | (0.2592) |
| Observations (N) | 4,272 | 4,272 | 4,272 |
| Adjusted $R^2$ | 0.1810 | 0.1969 | 0.1997 |
| Month Fixed Effect | Yes | Yes | Yes |
| Lead Surgeon Fixed Effect | Yes | Yes | Yes |
| Procedure Fixed Effect | Yes | Yes | Yes |

+, * and ** denote significance at 10%, 5% and 1% levels respectively

## Table 6. Regression on Length of Stay after removing the first 9 months

| Variable | Los | |
|---|---|---|
| | Model (1) | (2) |
| Average Individual Failure | -0.1409* | -0.1620* |
| | (0.0650) | (0.0687) |
| Recent Average Individual Failure | 0.2305** | 0.2811* |
| | (0.0683) | (0.1111) |
| Recent Average Individual Failure x Team Familiarity | | -0.0164** |
| | | (0.0057) |
| Team Size | 0.1686** | 0.1696* |
| | (0.0581) | (0.0625) |
| Team Familiarity | -0.0598** | -0.0563** |
| | (0.0184) | (0.0188) |
| Individual Average Direct Experience | -0.0451** | -0.0450** |
| | (0.0123) | (0.0123) |
| Medium | 0.1169** | 0.1180** |
| | (0.0107) | (0.0106) |
| Severe | 0.3926** | 0.4158** |
| | (0.0397) | (0.0411) |
| Age | -0.0000 | -0.0000 |
| | (0.0004) | (0.0004) |
| Male | -0.0018 | -0.0103 |
| | (0.0130) | (0.0124) |
| Constant | 4.1541** | 4.0569** |
| | (0.2668) | (0.2726) |
| Observations (N) | 3,696 | 3,696 |
| Adjusted $R^2$ | 0.1806 | 0.1825 |
| Month Fixed Effect | Yes | Yes |
| Lead Surgeon Fixed Effect | Yes | Yes |
| Procedure Fixed Effect | Yes | Yes |

+, * and ** denote significance at 10%, 5% and 1% levels respectively

## Table 7. Regression on Length of Stay after removing the first 12 months

| Variable | Los | |
|---|---|---|
| | Model (1) | (2) |
| Average Individual Failure | -0.3163** | -0.2007** |
| | (0.0594) | (0.0660) |
| Recent Average Individual Failure | 0.3661* | 0.2227* |
| | (0.1665) | (0.1089) |
| Recent Average Individual Failure x Team Familiarity | | -0.0272** |
| | | (0.0052) |
| Team Size | 0.1629* | 0.1382* |
| | (0.0634) | (0.0648) |
| Team Familiarity | -0.0253 | -0.0109 |
| | (0.0194) | (0.0193) |
| Individual Average Direct Experience | -0.0470** | -0.0470** |
| | (0.0122) | (0.0122) |
| Medium | 0.1064** | 0.1073** |
| | (0.0113) | (0.0112) |
| Severe | 0.3527** | 0.3923** |
| | (0.0388) | (0.0404) |
| Age | 0.0002 | 0.0002 |
| | (0.0004) | (0.0004) |
| Male | -0.0018 | -0.0051 |
| | (0.0130) | (0.0130) |
| Constant | 5.2471** | 5.0976** |
| | (0.2420) | (0.2471) |
| Observations (N) | 3,432 | 3,432 |
| Adjusted $R^2$ | 0.1987 | 0.2049 |
| Month Fixed Effect | Yes | Yes |
| Lead Surgeon Fixed Effect | Yes | Yes |
| Procedure Fixed Effect | Yes | Yes |

+, * and ** denote significance at 10%, 5% and 1% levels respectively

**Table 8. Regression on Length of Stay for 11 Lead Surgeons**

| Variable | Los | |
|---|---|---|
| | Model: (1) | (2) |
| Individual Failure | 0.0997** | -0.0964** |
| | (0.0231) | (0.0230) |
| Recent Individual Failure | 0.3051* | 0.3963** |
| | (0.1305) | (0.1324) |
| Recent Individual Failure x Leader Familiarity | | -0.0070** |
| | | (0.0025) |
| Team Size | 0.1965** | 0.1982** |
| | (0.0428) | (0.0426) |
| Leader Familiarity | -0.0294+ | -0.0326* |
| | (0.0158) | (0.0138) |
| Individual Average Direct Experience | -0.0343* | -0.0349* |
| | (0.0168) | (0.0168) |
| Medium | 0.1024** | 0.1128** |
| | (0.0124) | (0.0124) |
| Severe | 0.4352** | 0.4370** |
| | (0.0265) | (0.0266) |
| Age | -0.0002 | -0.0001 |
| | (0.0005) | (0.0005) |
| Male | -0.0001 | -0.0015 |
| | (0.0115) | (0.0115) |
| Constant | 4.0720** | 3.9853** |
| | (0.1560) | (0.1567) |
| Observations (N) | 3,220 | 3,220 |
| Adjusted $R^2$ | 0.1569 | 0.1592 |
| Month Fixed Effect | Yes | Yes |
| Lead Surgeon Fixed Effect | Yes | Yes |
| Procedure Fixed Effect | Yes | Yes |

+, * and ** denote significance at 10%, 5% and 1% levels respectively

## Table 9. Regression on Length of Stay after dropping the 75th percentile for Team Familiarity

| Variable | Los | |
|---|---|---|
| | Model (1) | (2) |
| Average Individual Failure | -0.1343+ | -0.1578* |
| | (0.0701) | (0.0733) |
| Recent Average Individual Failure | 0.3058* | 0.2836+ |
| | (0.1303) | (0.1442) |
| Recent Average Individual Failure x Team Familiarity | | -0.0244** |
| | | (0.0067) |
| Team Size | 0.1677** | 0.1345* |
| | (0.0582) | (0.0625) |
| Team Familiarity | -0.0462* | -0.0434* |
| | (0.0194) | (0.0197) |
| Individual Average Direct Experience | -0.0418** | -0.0411** |
| | (0.0127) | (0.0128) |
| Medium | 0.1160** | 0.1175** |
| | (0.0120) | (0.0119) |
| Severe | 0.4159** | 0.4439** |
| | (0.0422) | (0.0438) |
| Age | 0.0002 | 0.0003 |
| | (0.0005) | (0.0004) |
| Male | -0.0036 | -0.0045 |
| | (0.0137) | (0.0137) |
| Constant | 4.0295** | 3.8851** |
| | (0.2741) | (0.2828) |
| Observations (N) | 3,204 | 3,204 |
| Adjusted $R^2$ | 0.1603 | 0.1640 |
| Month Fixed Effect | Yes | Yes |
| Lead Surgeon Fixed Effect | Yes | Yes |
| Procedure Fixed Effect | Yes | Yes |

+, * and ** denote significance at 10%, 5% and 1% levels respectively

**Table 10. Regression on Length of Stay using all members after controlling for surgery duration**

| Variable | Los | |
|---|---|---|
| | Model: (1) | (2) |
| Average Individual Failure | -0.1887** | -0.2110** |
| | (0.0619) | (0.0645) |
| Recent Average Individual Failure | 0.3687* | 0.3010+ |
| | (0.1663) | (0.1643) |
| Recent Average Individual Failure x Team Familiarity | | -0.0202** |
| | | (0.0050) |
| Team Size | 0.1758** | 0.1400* |
| | (0.0556) | (0.0596) |
| Team Familiarity | -0.0576** | -0.0523** |
| | (0.0175) | (0.0179) |
| Individual Average Direct Experience | -0.0424** | -0.0422** |
| | (0.0116) | (0.0114) |
| Medium | 0.1124** | 0.1139** |
| | (0.0102) | (0.0101) |
| Severe | 0.3924** | 0.4222** |
| | (0.0374) | (0.0388) |
| Age | -0.0000 | 0.0000 |
| | (0.0004) | (0.0003) |
| Male | -0.0079 | -0.0100 |
| | (0.0117) | (0.0117) |
| Duration | 0.0002+ | 0.0002+ |
| | (0.0001) | (0.0001) |
| Constant | 4.1746** | 4.0409** |
| | (0.2651) | (0.2720) |
| Observations (N) | 4,272 | 4,272 |
| Adjusted $R^2$ | 0.1884 | 0.1919 |
| Month Fixed Effect | Yes | Yes |
| Lead Surgeon Fixed Effect | Yes | Yes |
| Procedure Fixed Effect | Yes | Yes |

+, * and ** denote significance at 10%, 5% and 1% levels respectively

## Table 11. Regression on Length of Stay with the IV approach for Lead Surgeons

| Variable | Recent Individual Failure | Los |
|---|---|---|
| | Model: (1) | (2) |
| Individual Failure | 0.0197 | -0.1646** |
| | (0.0233) | (0.0363) |
| Recent Individual Failure | | 0.7272** |
| | | (0.2867) |
| Recent Severe Dummy | 0.0574** | |
| | (0.0107) | |
| Team Size | 0.1453** | 0.1343** |
| | (0.0132) | (0.0464) |
| Leader Familiarity | -0.0065 | -0.0293** |
| | (0.0072) | (0.0111) |
| Individual Average Direct Experience | -0.0123** | -0.0422** |
| | (0.0046) | (0.0078) |
| Medium | -0.0046 | 0.1116** |
| | (0.0075) | (0.0115) |
| Severe | 0.0615** | 0.4298** |
| | (0.0151) | (0.0294) |
| Age | -0.0002 | 0.0001 |
| | (0.0003) | (0.0004) |
| Male | 0.0089 | -0.0141 |
| | (0.0068) | (0.0107) |
| Constant | 1.4894** | 2.8471** |
| | (0.0727) | (0.4409) |
| Observations (N) | 4,272 | 4,272 |
| Adjusted $R^2$ | 0.1068 | 0.1915 |
| Month Fixed Effect | Yes | Yes |
| Lead Surgeon Fixed Effect | Yes | Yes |
| Procedure Fixed Effect | Yes | Yes |

+, * and ** denote significance at 10%, 5% and 1% levels respectively

# Table 12. First Stage of the IV approach for Team Familiarity

| Variable | Team Familiarity | Recent Average Individual Failure x Team Familiarity |
|---|---|---|
| | Model (1) | (2) |
| Team Availability | 0.4533** | 0.5769** |
| | (0.0253) | (0.0857) |
| Team Familiarity Prediction x Recent Average Individual Failure | 0.4533** | 1.1480** |
| | (0.0253) | (0.0954) |
| Average Individual Failure | 0.0824+ | -0.3251+ |
| | (0.0486) | (0.1834) |
| Recent Average Individual Failure | 0.2991+ | 3.9399** |
| | (0.1567) | (0.4361) |
| Team Size | 0.4861** | -2.1117** |
| | (0.0263) | (0.0992) |
| Individual Average Direct Experience | -0.0070 | 0.0954** |
| | (0.0086) | (0.0326) |
| Medium | -0.0069 | 0.0766 |
| | (0.0141) | (0.0533) |
| Severe | 0.0507+ | 1.249** |
| | (0.0295) | (0.1114) |
| Age | -0.0002 | 0.0004 |
| | (0.005) | (0.0019) |
| Male | 0.0056 | -0.1525** |
| | (0.01289) | (0.0486) |
| Constant | 1.51094** | 2.5715** |
| | (0.2468) | (0.9304) |
| Observations (N) | 4,272 | 4,272 |
| Adjusted $R^2$ | 0.8880 | 0.6548 |
| Month Fixed Effect | Yes | Yes |
| Lead Surgeon Fixed Effect | Yes | Yes |
| Procedure Fixed Effect | Yes | Yes |

+, * and ** denote significance at 10%, 5% and 1% levels respectively

### Table 13. Regression on Length of Stay with the IV approach for Team Familiarity

| Variable | Los |
|---|---|
| | Model (1) |
| Average Individual Failure | -0.3346** |
| | (0.0526) |
| Recent Average Individual Failure | 0.6643* |
| | (0.3000) |
| Recent Average Individual Failure x Team Familiarity | -0.1449** |
| | (0.311) |
| Team Size | 0.1508+ |
| | (0.0865) |
| Team Familiarity | -0.0523 |
| | (0.0595) |
| Individual Average Direct Experience | -0.0447** |
| | (0.0091) |
| Medium | 0.1231** |
| | (0.0143) |
| Severe | 0.6006** |
| | (0.0498) |
| Age | 0.0000 |
| | (0.0005) |
| Male | -0.0217 |
| | (0.0135) |
| Constant | 3.3071** |
| | (0.2161) |
| Observations (N) | 4,272 |
| Adjusted $R^2$ | 0.1594 |
| Month Fixed Effect | Yes |
| Lead Surgeon Fixed Effect | Yes |
| Procedure Fixed Effect | Yes |

+, * and ** denote significance at 10%, 5% and 1% levels respectively

### Table 14. Policy Simulations

| | Policy_Familiarity |
|---|---|
| Percentage of Change in Recent Individual Failure | -90.68% |
| Percentage of Change in Leader Familiarity | 199.78% |
| Percentage of Change in Recent Individual Failure x Leader Familiarity | -75.86% |
| Percentage of Change in Length of Stay | **-28.80%** |