Communications of the Association for Information Systems

Volume 47 Article 41

10-11-2020

Using Secondary Data to Tell a New Story: A Cautionary Tale in Health Information Technology Research

Sumantra Sarkar
SUNY, Binghamton, ssarkar@binghamton.edu

Kaushik Ghosh
Suffolk University, kghosh@suffolk.edu

Stacie Petter
Baylor University, stacie_petter@baylor.edu

Recommended Citation
Sarkar, S., Ghosh, K., & Petter, S. (2020). Using Secondary Data to Tell a New Story: A Cautionary Tale in Health Information Technology Research. Communications of the Association for Information Systems, 47, pp-pp. https://doi.org/10.17705/1CAIS.04705

Communications of the Association for Information Systems

Research Paper DOI: 10.17705/1CAIS.04705 ISSN: 1529-3181

Using Secondary Data to Tell a New Story: A Cautionary Tale in Health Information Technology Research

Sumantra Sarkar
State University of New York, Binghamton
ssarkar@binghamton.edu

Kaushik Ghosh
Sawyer Business School
Suffolk University

Stacie Petter
Information Systems
Baylor University

Abstract:

Through the growth of big data and open data, new opportunities have presented themselves for information systems (IS) researchers who want to investigate phenomena they cannot easily study using primary data. As a result, many scholars have "retooled" their skills to leverage the large amount of readily available secondary data for analysis. In this confessional account, we share the story about how the first and second authors faced challenges when using secondary data for a research project in the health information technology domain. Through additional analysis of studies on health information technology that have used secondary data, we identified several themes of potential pitfalls that can occur when collecting, appropriating, and analyzing secondary data for a research project. We share these themes and relevant exemplars to help IS researchers avoid mistakes when using secondary data.

Keywords: Secondary Data, Health Information Technology, Meaningful Use, Confessional Account.

This manuscript underwent peer review. It was received 08/04/2019 and was with the authors for four months for one revision. Fred Niederman served as Associate Editor.

1 Introduction

In late 2008, Google launched a new offering called Google Flu Trends (GFT) (Helft, 2008). Researchers from Google and United States' (US) Centers for Disease Control (CDC) shared their success in predicting influenza outbreaks based on search terms that individuals used in Google's search engine related to the flu (Ginsberg et al., 2009). Years later, GFT emerged as a case study among practitioners and academics regarding the potential for error when analyzing and interpreting secondary data because, while the data and resulting algorithm performed quite well in predicting past outbreaks, GFT underestimated the influenza level during the pH1N1 pandemic in 2009 (Cook, Conrad, Fowlkes, & Mohebbi, 2011). Furthermore, in 2013, GFT overestimated peak influenza levels by an elevated margin (Butler, 2013). The errors in the GFT analysis arose due to a statistical fallacy called the "semi-attached figure" (Huff, 1993). The semi-attached figure occurs when one considers data collected for one purpose (i.e., data collected via search engine queries about flu-like symptoms) a proxy for a variable that does not concur with the data's original intent (i.e., data assumed to represent influenza cases). After scholars and the public critiqued Google for "big data hubris" (Lazer, Kennedy, King, & Vespignani, 2014) and poor model prediction (Snyder, 2013), the company stopped publishing Google Flu Trends in 2014. If organizations and thought leaders that pride themselves on the collecting, analyzing, and interpreting data can make mistakes when repurposing data, then we, as scholars, could make similar mistakes in our work when using secondary data for research purposes.

If researchers identify a research question and then collect data to answer it (using quantitative or qualitative means), we refer to the dataset as primary data. As researchers, organizations, governments, or other sources make datasets available for public consumption, researchers with experience using primary data have begun to "retool" their skills to use secondary data for research projects. When scholars use, analyze, and interpret existing data collected for a different purpose to answer a new research question, we refer to the dataset as secondary data (Hox & Boeije, 2005). Researchers across disciplines (including the information systems (IS) discipline) rely on secondary data in their research studies, such as the analysis of secondary data in econometric modeling (Chi, Ravichandran, & Andrevski, 2010).

Although many IS researchers prefer to collect primary data for their studies, the collection of primary data is not always possible due to challenges associated with time, cost, and effort. Many researchers can access only a limited number of informants depending on availability and cost, which leads to small datasets with multiple limitations. Furthermore, researchers may find it difficult to examine other phenomena, particularly research that focuses on constructs that exist at a macro level (e.g., organization-level, industry-level, or society-level), using primary research methods. The "big data" revolution (McAfee & Brynjolfsson, 2012) has enabled researchers to access copious amounts of data that others have previously collected and stored. Moreover, improvements in technology have reduced the cost and effort required to collect, store, and share large amounts of data (Murdoch & Detsky, 2013; Martin, 2015; Clarke, 2016).

An increasing push towards "open data" in information systems and beyond has complimented the big data phenomenon (Link et al., 2017). Open data refers to "publicly available data structured in a way to be fully accessible and usable" (NSF, 2019), and researchers often use it "beyond the purpose of what the data was originally collected for" (Link et al., 2017, p. 592). Various attributes (e.g., volume, variety, veracity, velocity) characterize big data, while the data's availability defines open data. Initiatives that encourage open data promote it as a "public good" that enables researchers to extract new insights (Link et al., 2017). Over the last two decades, several developments have helped open data efforts, repositories, and policies emerge (Link et al., 2017), such as the international open access movement (Chan et al., 2002) and calls for more transparency in governments and organizations. The big data revolution and open data initiatives offer researchers an opportunity to analyze expanded datasets (Mayer-Schönberger & Cukier, 2013) rather than relying on limited data sets obtained via traditional primary data-collection methods (e.g., surveys).

While using and analyzing secondary data has promise for many research projects, when researchers consider using data that a third party collected for a different purpose, they must analyze, evaluate, and interpret secondary data differently as compared to primary data (Cowton, 1998). For example, the parties who collected the datasets used in secondary data research projects may not have collected data for scientific research purposes. The measures in many secondary datasets may not meet the validation and reliability requirements that scientific research requires, and scholars should carefully consider if the dataset has sufficient recency, relevance, and accuracy (Patzer, 1995; Shmueli,, 2010). In an editorial

note in *Information Systems Research,* Agarwal and Dhar (2014) state: "data that the researcher does not generate herself is often an imperfect observation for the real world concept that is being referred to" (p. 447). Yet, opportunities exist for researchers to advance science if they use secondary data appropriately (Lazer et al. 2014).

Keeping these issues about using secondary data in perspective, we address the following research question (RQ):

RQ: What cautions should researchers exhibit when using and analyzing secondary data?

We developed our research question when the first and second authors were unable to publish a research study using secondary data from the healthcare industry. Through discussions with the third author, we all realized the potential for researchers to make errors in using, analyzing, and interpreting secondary data. We had a desire to examine issues in secondary data that may weaken research projects. While many issues we identified also apply to projects that use primary data, we focus on issues that are particularly relevant in secondary data projects.

For some scholars, secondary data supports a bottom-up approach to data collection, exploration, and subsequent analysis (Constantiou & Kallinikos, 2015). Such an approach may be analogous to inductive research that "starts from data and then seeks to generate theoretical explanation" (Bholat, 2015, p. 4). Other scholars use secondary data for deductive research to support activities, such as theory testing. Despite why researchers use secondary data, challenges arise if they stretch the data's assumptions beyond what it reasonably permits (as in the case of Google Flu Trends). Reflecting on our experiences, we have concerns that the precipitous availability of secondary data may create a desire among researchers to mine as much secondary data as much as possible to generate publications, which can lead to inaccurate inferences.

In this paper, we focus on the healthcare industry as a source of secondary data for research projects. The healthcare industry produces big and open datasets that vary in content from financial information to information about patient care (Raghupathi & Raghupathi, 2014). For example, hospitals create datasets based on patient data to support 1) patient care, 2) compliance and regulatory requirements, and/or 3) policy decisions. Historically, hospitals have recorded patient data on paper charts and archived data using microfilms. As part of a global call to digitize patient records, the World Health Organization (WHO) encourages "the development, evaluation, implementation, scale-up and greater utilization of digital technologies, as a means of promoting equitable, affordable and universal access to health for all" as one of its priorities (WHO, 2019). Many national governments have passed laws to encourage or mandate healthcare organizations to use electronic medical records in the past decade, such as the HITECH Act in the US (Murdoch & Detsky, 2013) and the National Health Service's Long Term Plan in the United Kingdom (UK) (NHS, 2019). These mandates have dramatically increased how much healthcare data organizations produce, store, and, analyze (Raghupathi & Raghupathi, 2014). As healthcare and other organizations release and share data with the public, scholars repurpose these secondary datasets to further their research.

We write this paper as a confessional account (e.g., Kaplan & Duchon, 1988; Barley, 1990; Schultze, 2000; Mathiassen & Sandberg, 2013) to demonstrate the potential pitfalls that may occur when researchers new to secondary data use these datasets in a research project. We base the account on our experience of analyzing secondary data from the healthcare industry. By reviewing our experience, we provide insights for researchers using secondary data. In this paper, we focus on research using secondary data related to the healthcare industry primarily because 1) we originally failed to publish a research study using secondary data from the healthcare industry and 2) researchers can increasingly access many datasets in the healthcare domain due to the HITECH Act (Blumenthal, 2010). Although healthcare represents a specific subdiscipline of the IS discipline with its own nuances and contexts, the insights we gained about using and analyzing secondary data apply to non-healthcare contexts as well.

We organize this paper as follows: in Section 2, we offer a confessional account that explains why we chose to write this paper after the first two authors failed to publish a study for which they used secondary data from the healthcare industry. In Section 3, we review the IS literature related to healthcare that uses secondary data to identify common themes. In Section 4, we use our findings to share lessons learned to help researchers better leverage secondary data in their research. Finally, in Section 5, we conclude the paper.

2 A Cautionary Tale

We follow a confessional (van Maanen, 2011) or vulnerable (Behar, 1996) writing style in this paper as we describe a "self-revealing and self-reflexive account of the research process" (Schultze, 2000, p. 4). One typically writes a confessional to "challenge readers to examine their own informing practices" (Schultze, 2000, p. 4). By sharing our own vulnerabilities, we seek to inform others through this process.

We began this research project when the first two authors were exploring how to revise a manuscript that a leading healthcare journal rejected. They used secondary data to analyze a research question in the research project. The editors and reviewers expressed concerns that the manuscript did not offer substantial findings. Frustrated, the first two authors approached a colleague (the third author) for guidance. As the third author read the paper, she expressed concerns regarding the potential mismatch between the measures that the first two authors used for their analysis and the goals they outlined in their paper. When the third author asked the first two authors questions about the dataset, the discussion focused on the limitations of the measures in the secondary dataset. The measures in this dataset affected how the first two authors analyzed and interpreted the results. Subsequent meetings and discussions among all three authors led to this manuscript. We examined other IS research related to healthcare that used similar datasets and considered how researchers used, analyzed, and interpreted secondary data to arrive at plausible conclusions. For the confessional writing process, we iterated between two phases: a reflective phase and an exploratory phase. In the reflective phase, we discussed 1) how to improve on the rejected paper and 2) how to avoid some mistakes the first two authors faced in the initial project. In the exploratory phase, we reviewed how other published IS healthcare papers that use secondary data to identify practices that may weaken secondary data projects. Based on the insights we gathered from iterating between these phases, we developed this paper.

2.1 Reflective Phase

In their original research project, the first two authors examined the relationship between healthcare information technology (HIT) implementation and patient health outcomes. They sourced data from Hospital Compare, a publicly available data service that the United States Center for Medicare and Medicaid Services (CMS) created and maintains. This dataset included observations for 4,807 hospitals for 2016 with information about 1) HIT implementation state, 2) patient outcomes in terms of 30-day mortality rates and unplanned re-admission rates for five medical conditions (i.e., acute myocardial infarction, congestive heart failure, pneumonia, chronic obstructive pulmonary disease, and stroke), and 3) hospital quality ratings on a five-point scale that CMS assessed based on 57 measures along seven dimensions (safety of care, timeliness of care, effectiveness of care, mortality, readmission, patient experience, and efficient use of medical imaging). The first two authors used HIT implementation state as the primary independent variable, which they operationalized as meaningful use (MU). The dataset they used for the study contained a measure for MU—a binary (yes/no) variable that identified whether a given hospital met the criteria for MU as CMS guidelines outline in the calendar year. The first two authors chose this dataset in large part because it provided an objective measure for meaningful use. This dataset allowed the first two authors to further investigate the previously established and sometimes inconsistent relationship between MU and patient outcomes.

MU has a specific meaning that the CMS has established. The CMS defines meaningful use (of technology) as when a healthcare provider uses "certified" electronic health record (EHR) technology to 1) exchange clinical data between other healthcare providers, patients, and insurers; 2) improve related clinical processes such as care coordination; and 3) realize improved clinical, quality, safety, and efficiency outcomes over time (HealthITGov, 2016). Furthermore, the CMS designates MU as an outcome that comprises three stages. Each stage builds on the previous one to improve the efficacy and outcomes of EHR usage (Cohen & Adler-Milstein, 2016). The CMS specifies the core objectives for each stage and measures for achieving MU, which healthcare providers self-report. If a healthcare provider succeeds in achieving a minimum number of objectives in a specific stage, it can obtain MU attestation from the CMS and receive financial incentives associated with the outcome. The first stage (beginning in 2011 and now completed) encouraged healthcare providers to meet fourteen objectives related to using certified EHR technology for capturing and sharing data. The second stage (starting in 2014 and now completed) required healthcare providers to use certified EHR technology to advance clinical processes and measured progress based on sixteen objectives. The final stage (beginning in 2016 and ongoing) advocates for using certified EHR technologies for improved health outcomes and uses six objectives and measures to measure meaningful use.

The healthcare literature has widely analyzed meaningful EHR technology use (Lammers & McLaughlin, 2017), and studies have extensively used secondary data to link MU with outcomes, such as hospital performance and patient health outcomes (Blumenthal & Tavenner, 2010; Furukawa, Spector, Rhona Limcangco, & Encinosa, 2017; Kwon & Johnson, 2018). By embracing MU, a provider using certified EHR technology to electronically prescribe and exchange health information can improve its efficiency and care (Brice et al. 2018). Patient data, such as vital signs, demographics, current and active diagnoses, active medications, allergies, and smoking status, recorded and stored in an EHR helps physicians make better decisions, avoid preventable errors, and, thus, improve patient care (Blumenthal & Tavenner, 2010; Lin, Lin, & Chen, 2019). Prior research has found conflicting results about the relationship between MU and outcomes. While some studies have found a positive relationship between the adoption of MU capabilities and improved patient care outcomes (e.g., Furukawa et al., 2017), other studies have failed to observe a relationship between MU and quality outcomes such as patient readmissions (e.g., Nakamura, Harper, Castro, Yu, & Jha, 2015). Researchers have issued calls for further studies to examine if MU does yield the expected benefits (Brice, Joynt, Tompkins, & Ritter, 2018).

Consistent with this call for additional research on MU and outcomes, the first two authors examined the relationship between meaningful HIT use with a wide set of clinical outcomes and quality metrics in their original research study. In analyzing the Health Compare dataset, they did not produce the discernible results they expected. For example, MU essentially had no significant direct effect ($p < 0.05$) on any patient outcome measures (i.e., the 30-day mortality rates and 30-day unplanned readmission rates). As an exception, they found a significant relationship between MU and pneumonia-related 30-day unplanned readmission rate[1]. They also examined the interaction between MU and hospital quality on patient outcomes. They found no significant interaction effects between MU and hospital quality with any patient outcome measures except pneumonia-related 30-day unplanned readmission rates. Interestingly, they noted that the hospital's overall quality rating had a consistent negative and significant effect on readmission rate for patient outcomes for all five medical conditions. This result differs to what they expected based on common understanding of hospital quality and patient care.

After the journal rejected the paper, the first two authors reflected on the manuscript's rejection with the third author's help by questioning its assumptions. First, we recognized that the common assumption that meaningful HIT use in a hospital may directly affect patient outcomes does not always apply. We realized that MU might not be the principal factor that affects patient outcomes in contrast to prior research on the topic (e.g., Appari, Johnson, & Anthony, 2013; Lammers & McLaughlin, 2017). Additional factors beyond HIT use in a hospital could influence clinical (mortality) and quality (readmission) outcomes. These additional factors vary and can include care providers' skill, the availability of diagnostic infrastructure, staff training on safety measures, patient acuity, healthcare provider staff levels, patient demographics, and the patient population's health conditions and socio-economic status. In sum, based on the first two authors' empirical findings, they could have concluded that MU does not relate to patient outcomes, but that assumption would be incomplete. MU in conjunction with factors that they did not include in their analysis (due to unavailable data) may serve an important role in determining patient outcomes (i.e., readmission and mortality rates). In reflecting on the project, we determined that limitations with existing datasets (such as the one the first two authors used in their original study) may limit one's ability to examine more complex models (e.g., polynomial or other types of relationships between variables) or develop predictive models. We also realized the limitations of using a binary independent variable (i.e., MU) to study the relationship between healthcare information technology (HIT) implementation and patient health outcomes. To learn from past research, we began exploring how other researchers have attended to similar issues in the context of research related to measuring meaningful HIT use.

2.2 Exploratory Phase

Realizing the limitations associated with the MU measure in many datasets, we chose to review other studies in a healthcare context that have used secondary data to measure MU. We identified several different approaches to measure MU across studies, which may explain the inconsistent findings regarding the relationship between MU and patient outcomes in prior research. Several studies measured

[1] Note that the first two authors examined the relationship between two independent variables: 1) meaningful use (MU) and 2) hospital quality measures on ten dependent variables (five clinical (30-day mortality rate) indicators and five quality (30-day unplanned readmission rate) indicators). Thus, they examined twenty relationships in their model. The only observed significant relationship between MU and pneumonia-related 30-day unplanned readmission rates may have arisen due to chance given they used using p-values of 0.05 to evaluate significance.

MU based on the attestation measures from the CMS website (e.g., Holmgren, Adler-Milstein, & McCullough, 2017; Brice et al., 2018; Kwon & Johnson, 2018). These studies recorded MU using CMS data related to whether a healthcare provider attained MU attestation for a certain stage (e.g., second stage meaningful use). After exploring how studies used MU attestation to measure MU (as secondary data) further, we developed some concerns. First, healthcare providers self-report to the CMS on their MU attestation (whether they reached a certain stage of meaningful use). Researchers may perceive the dataset is objective because a government agency (i.e., the CMS) records the data; yet, the data that healthcare providers share with CMS is self-reported and has limitations like any other self-reported data. Second, MU, a binary variable (yes/no), designates whether a healthcare provider has reached a certain MU state as designated by the CMS. The variable does not capture which objectives, among the list of many outlined objectives, a healthcare provider achieved. Furthermore, this binary measure does not identify how well a provider meets each objective and does not recognize more specific aspects of technology use. Third, because MU is a binary measure, researchers have limited options for data analysis.

To overcome these issues associated with measuring MU, some researchers have developed other means to measure it. For example, Appari et al. (2013) used multiple sources of secondary data to classify healthcare providers into one of five categories based on the types of clinical applications that they used. In their study on hospital performance, the authors tagged providers at level 0 if they had only primitive EHR capability. Level 1 hospitals used three ancillary information systems: laboratory, pharmacy, radiology. Level 2 hospitals used the applications in level 1 and clinical data repository and clinical decision support. Level 3 hospitals used the applications for level 2 and nursing documentation and electronic medication administration records. Finally, level 4 hospitals used the applications in level 3 and fully implemented computerized physician order entry (CPOE) and other optional applications. In this five-point scoring system, Appari et al. (2013) stated that hospitals that reached level 3 and above achieved MU since they had "the system capabilities required to meet 2011 MU objectives" (p. 358). Using a different approach, Furukawa et al. (2017) used an electronic medical record (EMR) proprietary scoring methodology that a third-party organization developed to measure MU. The Healthcare Information and Management Systems Society (HIMSS), an American not-for-profit organization promoting improvement in healthcare using HIT (HIMSS, 2020), has also attempted to measure MU in hospitals. The HIMSS approach measures the degree to which hospitals adopt and use EHR functions using an eight-stage electronic medical record adoption model (EMRAM) in which 7 signifies complete MU and 0 signifies a hospital has not implemented any HIT.

Reviewing these studies, we realized MU (as the CMS or any other organization defines it) represents a simplified proxy for an overly complex construct. While healthcare providers should meaningfully use HIT "to improve the quality of care" (CDC, 2017), they cannot easily capture data that they report about meaningfully using technology. Since MU, as the CMS defines it, only requires that a healthcare provider use a given technology to support a minimum number of objectives in each MU stage, the CMS does not require providers to, or even assume that they do, use the technology well (or meaningfully). Therefore, MU only measures that a healthcare provider uses technology to meet a minimum number of objectives among an extensive list. Thus, MU cannot serve as a proxy for more complex constructs, such as EHR/HIT adoption success or information quality.

In their initial research, the first two authors examined the relationship between MU and patient outcomes. In investigating this relationship, they placed more meaning and value on MU as a measure than what it actually provides. Due to the limitations with the secondary data they used, they could not know for certain if other antecedents to MU exist, which might be an opportunity for future research. Reflecting on the failed publication project and based on exploring studies that have examined MU, the three authors realized the potential for flaws and errors in analysis and interpretations when using secondary data. Hence, we expanded our literature review to identify other concerns that researchers may need to consider when using secondary data.

3 Secondary Data Analysis in HIT Research

After reflecting on the first two authors' experience with secondary data and exploring specific issues related to the original dataset, we reviewed HIT research with secondary datasets more extensively. We sought to identify potential issues that may arise that could weaken or limit how one uses secondary data in a research context, such as in drawing inferences based on potentially flawed assumptions (e.g., Huff, 1993; Cowton, 1998). We examined published IS research in the healthcare discipline using secondary

data. We searched for papers in the AIS Senior Scholar's "basket of eight" journals and the *Communications for the Association for Information Systems* that appeared between 2012 and 2019. We used the following keywords in our search: "healthcare", "health care IT", "health outcomes", "analytics", "EHR", "EMR", and "disease management". Based on the above criteria, we shortlisted 70 papers for further analysis. We eliminated conceptual papers or papers that used primary data, non-empirical modeling, or simulation methods. We only retained studies that used secondary data, which yielded 15 papers for further analysis. We analyzed and synthesized the 15 papers using qualitative research principles (Myers, 1997; Sarker, Xiao, & Beaulieu, 2013). Following methods used in qualitative data analysis (Miles, Huberman, & Saldna, 2014), the first two authors reviewed each paper to identify potential issues or concerns in the study and coded each paper independently. The researchers met regularly to discuss and standardize codes. The authors conducted this process in an iterative manner, which concurs with the norms for analyzing qualitative data. Based on commonalities in and interrelationships between the issues they identified, they grouped similarly coded papers to form categories. The third author provided feedback on the codes and final categories. Synthesizing patterns in and between categories led to our identifying parsimonious themes. We explain each theme and provide examples in Sections 3.1 to 3.3.

3.1 Theme 1: Incomplete Formulations of Key Constructs

Researchers often analyze datasets that a third party collected for one intent and purpose for vastly different intents and purpose when they use such datasets as secondary data. We found that the measures available in a secondary dataset might not always appropriately represent the primary constructs relevant to the research question.

3.1.1 Subtheme A: Limited Measures for Key Independent Variable(s)

One or more variables may imperfectly measure a construct; however, with secondary data (as opposed to primary data), researchers may have to accept that they have to operationalize measures in a less-than-ideal way in their research studies. In our literature review, we found a strong reliance on binary variables when using secondary data (Bardhan, Oh, Zheng, & Kirksey, 2015; Angst, Block, D'Arcy, & Kelley, 2017a). However, while binary variables may be appropriate for the original data-collection effort (i.e., the original intent and purpose of the dataset), researchers may encounter limitations in the data-analysis process when they repurpose a dataset for a secondary data research project. For example, Angst et al. (2017a) created a binary measure for "entrepreneurial mindset", which they define as the "extent to which the hospital adopts innovative technologies" (p. 905) and measure with the Saidin index (Spetz & Maiuro, 2004), "a weighted metric of IT adoption" (Angst et al., 2017a, p. 903). In another study, Bardhan et al. (2015) examined the relationship between HIT use in hospitals and the risk of patient readmission. They measured the independent variable HIT use as a binary variable with 1 representing implemented and operational HIT and 0 representing not implemented or not operational HIT.

Using binary variables to represent HIT implementation or key constructs in a study has various limitations. For example, with HIT implementation, since HITs require a considerable amount of time (sometimes multiple years) to implement (Sykes, 2015), a binary variable does not capture how long the system has been operational. Second, a binary variable fails to identify which functional modules in a HIT a healthcare provider has implemented. In both primary data and secondary data, researchers seek to formulate independent variables appropriately for constructs (Clark, Zmud, & McCray, 1995). However, using a binary measure to represent HIT adoption, implementation, and/or operation limits the options available for data analysis. Researchers must then determine if they can address the independent variable's limitations through recoding the data or using different analysis techniques. Alternatively, researchers may determine that they cannot use a secondary dataset because key independent variables are measured or formulated inconsistently with their research goals.

3.1.2 Subtheme B: Need for Proxy Measures

Researchers often use proxy measures for key constructs in a research model when analyzing secondary data. However, in doing so, the proxy they use from the secondary dataset may not resemble how they conceptually define a construct. In their study, Angst et al. (2017a) examined how IT security adoption impacts the risk of data breaches in hospital settings. They counted the number of IT security systems in service for a given year as a proxy for measuring IT security investment. While the number of operational IT security systems may correlate with how much an organization spends on IT security, the two

phenomena differ. When using a proxy for a construct, researchers need to consider how the proxy they choose affects how they can interpret their results.

Another challenge that arises in using secondary data concerns finding measures among secondary datasets that apply to theory. In their study, Peng, Dey, and Lahiri (2014) identified factors that accelerate HIT adoption in hospital settings. They used absorptive capacity as a theoretical lens and argued that organizations with higher absorptive capacity will likely adopt technology more quickly. They defined absorptive capacity as an organization's ability to identify, recognize the value of, and assimilate external information and knowledge. They used current stock or inventory of HIT as a proxy for absorptive capacity. They explained that, as the inventory of HIT applications in a hospital increases, "so [too] do[es] the depth and breadth of expertise required to evaluate, acquire, deploy, operate, and maintain them" (p. 19). The authors acknowledge this need for proxy measures in their limitations section as the unfortunate "plight of a researcher working with secondary data" (p. 20).

Researchers often make difficult choices when choosing how to operationalize constructs with secondary data. When an organization collects primary data, often the intent is to achieve a practical goal as opposed to answering a theoretical research question. When researchers choose to repurpose a dataset, they must make difficult decisions, such as to not use a secondary dataset if the measures do not concur with theoretical definitions or to consider if the use of proxy variables will confound results or negatively influence how one interprets and uses the findings.

3.1.3 Subtheme C: Partial Measures of Key Constructs

Further, secondary data may only provide partial information about a construct. If one can only partially measure a construct, then one cannot completely measure key variables in a study. In a study on the value of hospitals sourcing software from a single vendor versus multiple vendors, Angst, Wowah, Handley, and Kelley (2017b) examined the number of vendors that each hospital used for five specific electronic medical record (EMR) modules. They selected five modules to include in the analysis based on "the most commonly adopted modules that makeup the EMR category as defined by HIMSS Analytics" (p. 1131). However, in practice, an EMR system can contain dozens of modules. Thus, in this case, the data source the authors used (i.e., HIMSS Analytics) likely only provided information for five specific EMR modules that the authors used.

When using secondary data, researchers should consider how partial data may affect the study. They may need to find additional support for using a partial measure based on research or practice or may need to find an alternative dataset that measures the construct more fully. Researchers must cautiously interpret their results to avoid overstating the lessons they learned the data if the dataset provides incomplete information.

3.2 Theme 2: Weak Relationships among Constructs

When analyzing secondary data, some relationships may initially appear plausible but, after further consideration, become uncompelling. Governments and organizations promote HIT use among healthcare providers to improve not only operational efficiency but also patient care (Blumenthal, 2010). In their original study, the first two authors expected to find a direct relationship between HIT use (operationalized as MU) and patient outcomes. They felt it represented a plausible hypothesis given the expectation that HIT use should produce positive patient outcomes. Upon further reflection, they later realized the problems with assuming there should be a direct relationship between HIT use and patient care.

In reviewing the literature, we came across similar instances in which researchers posited relationships among constructs that may not be directly related to one another. In their study, Bardhan et al. (2015) hypothesized a direct relationship between HIT use and patient readmissions similar to the first two authors' original research study. Bardhan et al. (2015) measured the impact of HIT adoption on readmission rates. They considered both clinical information systems and administrative applications that included accounts payable and patient billing systems. It remains unclear how back-office applications, such as accounts and billing systems, would influence patient readmission rates.

In their study, Kwon and Johnson (2018) found that hospitals that reach the first stage of MU attestation "observed decreases in external security breaches in the year after attestation" but noted that "they did not see any further reductions in the subsequent year" (p. 1063). In considering the role that the first stage of MU attestation has on security breaches, one needs to understand more about the stage's context. The first stage of MU primarily focused on data capture and sharing (Cohen & Adler-Milstein, 2016), and it

contained the (non-mandatory) objective to adequately protect patient data's privacy and security (CDC, 2017). If a hospital had little to no security infrastructure or processes, then one could expect this investment in HIT to reach the first stage of MU to sharply reduce the number of data breaches. Once the hospital implemented this initial security infrastructure, it would be unlikely to observe significant security improvements year over year.

When developing a research model using primary or secondary data, researchers need to consider the nature of the relationships and potential confounds that affect how they explain their findings. In the two examples above, the researchers found relationships among the constructs; however, other factors could justify their findings beyond what the research models suggest. Researchers should consider if alternative explanations can explain their findings, especially if such findings rely on secondary data. In considering the original study, in retrospect, the first two authors could have focused on theory building in their initial study (Agarwal & Dhar, 2014) rather than theory testing. Given that some relationships seemed plausible, not finding evidence for these relationships could offer a contribution by challenging past research claims (e.g., HIT use reduces the risk of patient readmission). Also, given that the measures did not reflect theoretical concepts well, the secondary data used by the first two authors could not adequately identify the theoretical relationships they posited. Finally, the original study could have acknowledged how to model relationships beyond simple linearity.

3.2.1 Theme 3: Lack of External Validity

Scholars have discussed generalizability (Lee & Baskerville, 2003) as an issue related to external validity. Generalizing can occur 1) from a sample to a larger population (also called nomothetic generalization (Lincoln & Guba, 1985)), 2) to populations or/and settings other than those studied (also called transferability (Lincoln & Guba, 1985)), and 3) to places without reference to where one makes generalizations (Lucas, 2003). Researchers using secondary data may have a limited sample based on the availability of data, which raises questions about the results' generalizability (Calder, Phillips, & Tybout, 1982).

Menon and Kohli (2013) examined the impact that hospitals' IT spending had on their performance using secondary data from the state of Washington in the US. They found evidence that HIT spending had a direct effect on 1) malpractice insurance premium costs and 2) quality of patient care. By using data from a single U.S. state, the authors could "mitigate heterogeneity among hospitals that may result from narrow focus on patient care or differences in regulations, accounting practices, and other regional and environmental dissimilarities" (p. 923). However, since laws and regulations regarding healthcare vary across states, one might find it difficult to generalize these findings beyond Washington without knowing how its regulations vary from other states or locales. Since Menon and Kohli (2013) focused on a single location, their study may lack external rigor (Straub, 1989; Cowton, 1998). Depending on the available data, some authors have incorporated data from hospitals located across the US and mitigated heterogeneity bias using appropriate control variables and/or statistical techniques (e.g., Bardhan et al., 2015; Angst et al., 2017a).

While all research has generalizability limitations, we suggest that authors present their findings in a manner that acknowledges them. Furthermore, authors could generalize to theory by explaining how their findings support theoretical relationships that other studies have posited. By generalizing the results to theory, researchers can confirm theory's robustness and demonstrate its applicability in new context. Subsequent research can further test the theory's boundary conditions (Busse, Kach, & Wagner, 2016). Another means to improve external validity involves using industry-created measures when feasible. Lin et al. (2019) used prescribed measures that the Joint Commission[2] outlined to measure process quality and use guidelines that the Center for Disease Control and Prevention established to measure adoption. Using variables created with established industry guidelines provides a means to replicate results and more easily generalize findings across contexts.

[2] An independent and not-for-profit organization, the Joint Commission accredits and certifies nearly 21,000 healthcare organizations and programs in the United States. Joint Commission accreditation and certification is recognized nationwide as a symbol of quality that reflects an organization's commitment to meeting certain performance standards (see https://www.jointcommission.org/about_us/about_the_joint_commission_main.aspx).

4 Discussion

While one can find value in using large datasets derived from the healthcare industry or other sources, researchers new to using, analyzing, and interpreting secondary data should exhibit caution. Researchers need to understand what data a dataset contains, its limitations, and why its original collectors collected it. As Crawford (2013) states: "Data and datasets are not objective; they are creations of human design. We assign numbers their 'voice', draw inferences from them, and define their meaning through our interpretations". As researchers who use secondary data, we may use data that one collected for other purposes and conduct research based on the inferences we make based on that data. Given data's uncertain quality (Zhou, Chawla, Jin, & Williams, 2014), many researchers afford greater credibility to big data or open data than such data actually warrants (Clarke, 2016). Hidden biases in both collection and analysis present significant risks (Boyd & Crawford, 2012; Crawford, 2013).

Most IS research using quantitative methods focuses on theory testing and evaluating the explanatory power of underlying causal models (Shmueli & Koppius, 2011). In reviewing the HIT literature, we found two studies that used secondary data primarily for theory testing (Menon & Kohli, 2013; Peng et al., 2014), one study that used secondary data for prediction (Bardhan et al., 2015), and three studies used secondary data for theory testing and theorizing (Angst et al., 2017a, 2017b; Kwon & Johnson, 2018). In their original research project, the first two authors sought to use secondary data to test theory. They attempted to test a well-established presumption in the literature that MU positively influences clinical and/or quality outcomes (e.g., Furukawa et al., 2017). On reflection with the third author, the first two authors realized that they could not have used their secondary data to test this proposition and contribute to the literature. Further, they learned that one can use secondary data for not only theory testing but also theorizing or prediction (Rai, 2016). While several papers we identified in our literature review used secondary data for theory testing, researchers have many opportunities to use secondary data for theorizing by revealing "associations between variables that form the foundation for the development of theory that can be subsequently subject to rigorous testing" (Agarwal & Dhar, 2014, p. 446).

4.1 Lessons Learned from HIT Research

In reviewing the literature, we found studies that exemplify how one should use secondary data in a healthcare context (e.g., Du, 2015; Lin et al., 2019). Gaining insights from past literature and reflecting on what the first two authors could have done differently in their original project, we all learnt several lessons. First, researchers must identify whether they can best answer their research question(s) using theory building, theory testing, or prediction. Researchers can use secondary data in all these approaches, but they should recognize the potential limitations of secondary data for each approach. In the original research study, rather than assuming a theory testing approach, the first two authors would have examined the secondary data differently if they considered how to use data and theory in conjunction with one another to gain new and interesting insights (Maass, Parsons, Purao, Storey, & Woo, 2018). They could have used secondary data for theory building to discover new relationships between the different variables in the data (Agarwal & Dhar, 2014).

Second, based on the research question(s) for a study, researchers should carefully identify the purpose of the original data-collection effort and must examine the definitions and measures of the variables available in the secondary dataset. Researchers should spend significant time in exploring the nature of the secondary dataset(s) they use. Researchers should note that data exploration constitutes a continuous activity; one may discover secondary data's quality and nuances only after analyzing it in depth (Agarwal & Dhar, 2014). Researchers must take into account how well the variables in the dataset align with their research objective. Researchers commonly use proxy measures in secondary data projects because they use the data for a reason that differs from why it was collected in the first place. If researchers "force" a measure to fit a construct as a proxy, then most likely, the variable will not serve as a good fit for their research study. Therefore, when identifying proxy measures for a construct, researchers need to consider how well the measures 1) approximate closely with how they conceptually define the original variable, 2) correlate with the theoretical underpinnings of the variable of interest, and 3) appropriately indicate the variable of interest.

Third, researchers should understand the characteristics of the population or sample used to generate the dataset. The sample influences how researchers may generalize and present findings from their analyses (Huff, 1993). For example, the CMS website provides data for patients over 65 years old as "public use files" data. However, one could not plausibly use this data to answer research questions related to

whether HIT use improves readmission rates for *all* potential patients given the data that the CMS collected focuses on patients older than 65.

Fourth, researchers must be willing to acknowledge the constraints of their secondary dataset. The dataset may have limitations associated with the sample, the measures available, the type of data provided (e.g., self-report versus objective data), and so on. These limitations may affect findings' generalizability or a study's practical implications, and researchers need to recognize and state their results' limitations.

Finally, researchers need to identify whether the secondary data available constitutes the right data to explore their research question(s) when beginning their project. We realize that many researchers have heard about the promise of big and open data and may be tempted to use secondary data to answer research questions that they may best address using other methods. While some researchers may believe that secondary data represents an "easier" approach to gather data, this assumption is not necessarily valid. Given the general practice in the academic community to "publish or perish", especially for junior faculty (Larson, Nelson, & Carter, 2015), researchers can find it tempting to "mine" easily accessible secondary data to find relationships between constructs and then "retrofit" theory for hypothesizing and analysis, especially in studies that test theories[3]. We also acknowledge that editors sometimes ask authors to find theoretical explanations for data-driven research in the review process. Researchers face challenges in ensuring that the approach they use to analyze secondary data complements and enhances their research question(s).

4.2 Lessons Learned from other Disciplines

To complement the lessons we learned, we examined how other disciplines identify and explain limitations of research projects that use secondary data. For example, in social sciences, scholars urge other scholars to consider: 1) the fit between the research question and the dataset available, 2) the cost and time spent on getting familiar with the dataset, 3) the measurement precision for key constructs in the dataset, and 4) the need to resist the temptation to "fish" in the dataset to find relationships (Hofferth, 2005; Trzesniewski, Donnellan, & Lucas, 2011). In education, scholars advise other scholars to recognize limitations of secondary data by exhibiting a "healthy skepticism" about its origin, which they need to robustly analyze data and correctly interpret results (Smith, 2008). In the economics discipline, researchers must 1) have knowledge of secondary data's underlying sources, 2) have detailed documentation about its sources, and 3) explain any modifications they make while analyzing the secondary data (Atkinson & Brandolini, 2001). In medicine, researchers have highlighted concerns regarding the lack of control over data (e.g., types of variables in the dataset) that have already been collected to analyze a research question (Schlomer & Copp, 2014) and the inconsistency of coding schemes (e.g., schemes used to code medical diagnosis or surgical procedures) in secondary data (Khwaja, Syed, & Cranston, 2002).

Thus, the above discussion on how scholars in other disciplines use secondary data shows that both we and scholars in many disciplines have faced similar problems: not consistently operationalizing constructs, not appropriately representing variables of research interest, and mining data. To skillfully navigate between these pitfalls, Smith et al. (2011) suggest that researchers follow guidelines for using secondary data by 1) defining their research question(s) a priori, 2) selecting the dataset based on the research question(s) and resisting the temptation to mine data (Schlomer & Copp, 2014), 3) familiarizing themselves extensively with the dataset, and 4) interpreting the findings with caution (such as inferring statistical significance). More generally, researchers across disciplines should strike a balance between maintaining research rigor and ensuring prudent inferences based on the findings from secondary data.

Table 1 summarizes potential questions that both IS scholars and scholars from other disciplines should ask when using secondary data for research based on our insights from reviewing the HIT literature that relies on secondary data.

[3] We note that we did not find any evidence of the force fitting of theory to data in the manuscripts we reviewed for this paper.

Table 1. Questions for Researchers when using Secondary Data

IS research	Other disciplines
Why or for what research question do we use the secondary data?	Have we identified a research question a priori to collecting data?
Do the measures in the secondary dataset match our research question?	Did we choose a secondary dataset based on the research question as opposed to mining data for results?
Does the secondary dataset(s) sample concur with the population to which we seek to generalize?	Do we have confidence in the source and robustness of the secondary data?
Do the limitations or constraints of the secondary dataset(s) negatively influence the original research question or how we interpret the results?	Do we know the secondary dataset sufficiently enough to recognize its limitations?
Is secondary data the right approach to study our research question?	Have we interpreted the findings with caution?

5 Conclusion

Much empirical research relies on the fundamental assumption that the data researchers collect captures the theoretical constructs of interest. When designing a research project, researchers should identify and define the constructs, consider the relationships among them, and include an approach to collect data for the constructs of interest (van de Ven, 2007). For example, if researchers decide to collect primary data, their research plan should include survey items or scales that they pilot-test, pre-test, and validate to ensure the items measure the constructs appropriately (Straub, 1989). In a research plan for a secondary data project, researchers should also identify and define the constructs, consider the relationships among them, and ascertain the sources of data that they use to measure them. Without a carefully designed research plan for secondary data, researchers may encounter issues that weaken their results. For example, in the case of Google Flu Trends, the data collected for one purpose (i.e., search terms) failed to be consistent with the measurement construct (i.e., influenza cases).

Secondary data does not replace primary data; further, secondary data "is most useful in conjunction with other methodologies, such as experimentation, survey research, or clinical research" (Andersen, Prause, & Silver, 2011, p. 58). Secondary data opens new opportunities for researchers to provide answers to new and old research questions. The availability of secondary data has the potential to offer new insights to disciplines that researchers have not or not sufficiently explored due to difficulties in obtaining relevant primary data. In this confessional account, we narrate a story about two IS researchers who jumped into the latest trend of using secondary data in their research.

With this paper, we discuss how researchers can fall victim to problems that may weaken the impact of studies that employ secondary data. Being new to using secondary data, the first two authors learned several lessons along the way that would be helpful to others that also plan to embark on the quest to leverage secondary data for research. The first two authors began this journey after a journal rejected a manuscript that used secondary data. By exploring past literature that used similar data, methods, and goals, we identify issues that can weaken the results of secondary data research.

We acknowledge that the themes we identified about using secondary data among HIT studies do not necessarily only apply to secondary data. All research methods have limitations, and even primary data-collection efforts can struggle with operationalizing constructs appropriately, testing practical or useful relationships, and ensuring an appropriate level of external validity in a study. We do not present these themes or lessons learned to criticize existing research or as justification for rejecting research using secondary data. Rather, we share these themes and lessons learned to help researchers avoid potential pitfalls when using secondary data for their research projects.

We recognize that many subdomains in IS research and other disciplines use secondary data extensively. We acknowledge that the lessons we present in this paper may be trite or obvious to researchers that use these data sources extensively. However, we also recognize that many researchers see the promise of big and open data and are new to using secondary data for research. We share these insights with this audience to help them avoid the mistakes the first two authors made in their first attempt to use secondary data. We also acknowledge that many other issues that arise when one analyzes large datasets, such as the p-value problem (Lin, Lucas, & Shmueli, 2013), data management, and access control (Demchenko,

Grosso, De Laat, & Membrey, 2013). We purposefully scoped this paper for scholars in the preliminary stages of using secondary data. Researchers that collect and analyze primary data recognize and understand the various limitations with the choices they make in collecting and analyzing data. For researchers with training in collecting, analyzing, and using primary data, they need to recognize the limitations of secondary data as well. In this paper, we focus on enabling researchers new to using secondary data on issues to consider before analyzing data. Researchers must understand why a dataset's original collectors collected it, how the dataset's purpose affects the measures and data in the dataset, and how they can operationalize these measures to examine the relationships among constructs. In reviewing the secondary data, researchers should determine if the secondary dataset meets minimum requirements for veracity, credibility, and relevance (Patzer, 1995; Shmueli, 2010). Based on a review of the secondary dataset, researchers may need to make difficult choices, such as 1) reconsidering the research question or study objective, 2) including complementary secondary datasets, 3) not using the secondary dataset to begin with, and 4) recognizing the limitations of the inferences that they can make from their analyses.

Finally, we encourage more researchers in the IS discipline to be more reflective and confess their successes and failures. We write this paper because we have found other self-confessional papers useful (e.g., Kaplan & Duchon, 1988; Barley, 1990; Schultze, 2000; Mathiassen & Sandberg, 2013). As Barley (1990) acknowledges in his confessional account, "like all research methods, mine suffer from biases and limitations that should be made explicit" (p. 244). We hope that, by sharing our own failures and the lessons we learned "the hard way", we can provide guidance to others on their own journey of learning how to work with secondary data.

References

Agarwal, R., & Dhar, V. (2014). Big data, data science, and analytics: The opportunity and challenge for IS research. *Information Systems Research, 25*(3), 443-448.

Andersen, J. P., Prause J., & Silver, R. C. (2011). A step-by-step guide to using secondary data for psychological research. *Social & Personality Psychology Compass, 5*(1), 56-75.

Angst, C. M., Block, E. S., D'Arcy, J., & Kelley, K. (2017a). When do IT security investments matter? Accounting for the influence of instituitional factors in the context of healthcare data breaches. *MIS Quarterly, 41*(3), 893-A898.

Angst, C. M., Wowak, K. D., Handley, S. M, & Kelley, K. (2017b). Antecedents of information systems sourcing strategies in U.S. hospitals: A longitudinal study. *MIS Quarterly*, *41*(4), 1129-1152.

Appari, A., M., Johnson, E., & Anthony, D. L. (2013). Meaningful use of electronic health record systems and process quality of care: Evidence from a panel data analysis of U.S. acute-care hospitals. *Health Services Research, 48*(2), 354-375.

Atkinson, A. B., & Brandolini, A. (2001). Promise and pitfalls in the use of "secondary" data-sets: Income inequality in OECD countries as a case study. *Journal of Economic Literature*, *39*(3), 771-799.

Bardhan, I., Oh, J.-H., Zheng, Z, & Kirksey, K. (2015). Predictive analytics for readmission of patients with congestive heart failure. *Information Systems Research*, *26*(1), 19-39.

Barley, S. R. (1990). Images of Imaging: Notes on doing Longitudinal Work. *Organization Science*, *1*(3), 220-245.

Behar, R. (1996). *The vulnerable observer: Anthropology that breaks your heart.* Boston, MA: Beacon Press.

Bholat, D. (2015). Big data and central banks. *Big Data & Society*, *2*(1).

Blumenthal, D. (2010). Launching HITECH. *New England Journal of Medicine*, *362*(5), 382-385.

Blumenthal, D., & Tavenner, M. (2010). The "meaningful use" regulation for electronic health records. New *England Journal of Medicine, 363*(6), 501-504.

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, *15*(5), 662-679.

Brice, Y. N., Joynt, K. E., Tompkins, C. P., & Ritter, G. A. (2018). Meaningful use and hospital performance on post-acute utilization indicators. *Health Services Research*, *53*(2), 803-823.

Busse, C., Kach, A. P., & Wagner, S. M. (2016). Boundary conditions: What they are, how to explore them, why we need them, and when to consider them. *Organizational Research Methods*, *20*(4), 574-609.

Butler, D. (2013). When Google got flu wrong. *Nature News*, *494*(7436), 155-156.

Calder, B. J., Phillips, L. W., & Tybout, A. M. (1982). The concept of external validity. *Journal of Consumer Research*, *9*(3), 240-244.

CDC. (2017). *Meaningful use.* Retrieved from https://www.cdc.gov/ehrmeaningfuluse/introduction.html

Chan, L., Cuplinskas, D., Eisen, M., Friend, F., Genova, Y., Guédon, J.-C., Hagemann, M., Harnad, S., Johnson, R., Kupryte, R., La Manna, M., Rév, I., Segbert, M., de Souza, S., Suber, P., & Velterop, J. (2002). *Budapest open access initiative.* Retrieved from https://www.budapestopenaccessinitiative.org/read

Chi, L., Ravichandran, T., & Andrevski, G. (2010). information technology, network structure, and competitive action. *Information Systems Research, 21*(3), 543-570.

Clark, T. D., Zmud, R. W., & McCray, G. E. (1995). The outsourcing of information services: Transforming the nature of business in the information industry. *Journal of Information Technology*, *10*(4), 221-327.

Clarke, R. (2016). Big data, big risks. *Information Systems Journal*, *26*(1), 77-90.

Cohen, G. R., & Adler-Milstein, J. (2016). Meaningful use care coordination criteria: Perceived barriers and benefits among primary care providers. *Journal of the American Medical Informatics Association, 23*(e1), e146-e151.

Constantiou, I. D., & Kallinikos, J. (2015). New games, new rules: Big data and the changing context of strategy. *Journal of Information Technology, 30*(1), 44-57.

Cook, S., Conrad, C., Fowlkes, A. L., & Mohebbi, M. H. (2011). Assessing Google Flu Trends performance in the United States during the 2009 influenza virus A (H1N1) pandemic. *PLOS ONE*, *6*(8), e23610.

Cowton, C. J. (1998). The use of secondary data in business ethics research. *Journal of Business Ethics*, *17*(4), 423-434.

Crawford, K. (2013). The hidden biases in big data. *Harvard Business Review*. Retrieved from https://hbr.org/2013/04/the-hidden-biases-in-big-data

Demchenko, Y., Grosso, P., De Laat, C., & Membrey, P. (2013). Addressing big data issues in scientific data infrastructure. In *Proceedings of the International Conference on Collaboration Technologies and Systems.*

Du, K. (2015). Parenting new acquisitions: Acquirers' digital resource redeployment and targets' performance improvement in the U.S. hospital industry. *Information Systems Research*, *26*(4), 829-844.

Furukawa, M. F., Spector, W. D., Rhona Limcangco, M., & Encinosa, W. E. (2017). Meaningful use of health information technology and declines in in-hospital adverse drug events. *Journal of the American Medical Informatics Association*, *24*(4), 729-736.

Ginsberg, J., M., Mohebbi, H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature, 457*(7232), 1012-1014.

HealthITGov. (2016). *Certified health IT vendors and editions reported by hospitals participating in the medicare EHR incentive program, health it quick-stat #29.* Retrieved from http://dashboard.healthit.gov/quickstats/pages/FIG-Vendors-of-EHRs-to-Participating-Hospitals.php

Helft, M. (2008). Google uses searches to track flu's Spread. *The New York Times.* Retrieved from https://www.nytimes.com/2008/11/12/technology/internet/12flu.html

HIMSS. (2019). *Electronic medical record adoption model.* Retrieved from https://www.himssanalytics.org/emram

Hofferth, S. L. (2005). Secondary data analysis in family research. *Journal of Marriage & Family*, *67*(4), 891-907.

Holmgren, A. J., Adler-Milstein, J., & McCullough, J. (2017). Are all certified EHRs created equal? Assessing the relationship between EHR vendor and hospital meaningful use performance. *Journal of the American Medical Informatics Association*, *25*(6), 654-660.

Hox, J. J., & Boeije, H. R. (2005). Data collection, primary vs. secondary. In K. Kempf-Leonard (Ed.), *Encyclopedia of Social Measurement* (pp. 593-599). New York, NY: Elsevier Academic Press.

Huff, D. (1993). *How to lie with statistics.* New York, NY: W. W. Norton & Company.

Kaplan, B., & Duchon, D. (1988). Combining qualitative and quantitative methods in information systems research: A case study. *MIS Quarterly, 12*(4), 570-586.

Khwaja, H. A., Syed, H., & Cranston, D. W. (2002). Coding errors: A comparative analysis of hospital and prospectively collected departmental data. *BJU International*, *89*(3), 178-180.

Kwon, J., & Johnson, E. (2018). Meaningful healthcare security: Does meaningful-use attestation improve information security performance? *MIS Quarterly*, *42*(4), 1043-1067.

Lammers, E. J., & McLaughlin, C. G. (2017). Meaningful use of electronic health records and medicare expenditures: Evidence from a panel data analysis of U.S. health care markets, 2010-2013. *Health Services Research, 52*(4), 1364-1386.

Larson, E. C., Nelson, M. L., & Carter, M. (2015). The FIRST+ year information systems faculty experience. *Communications of the Association for Information Systems, 36*, 643-654.

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google flu: Traps in big data analysis. *Science, 343*(6176), 1203-1205.

Lee, A. S., & Baskerville, R. L. (2003). Generalizing generalizability in information systems research. *Information Systems Research, 14*(3), 221-243.

Lin, M., Lucas, H. C. J., & Shmueli, G. (2013). Too big to fail: Large samples and the p-value problem. *Information Systems Research*, *24*(4), 906-917.

Lin, Y.-K., Lin, M., & Chen, H. (2019). Do electronic health records affect quality of care? Evidence from the HITECH Act. *Information Systems Research, 30*(1), 306-318.

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Beverly Hills, CA: Sage.

Link, G. J. P., Lumbard, K., Conboy, K., Feldman, M., Feller, J., George, J., Germonprez, M., Goggins, S., Jeske, D., Kiely, G., Schuster, K., & Willis, M. (2017). Contemporary issues of open data in information systems research: Considerations and recommendations. *Communications of the Association for Information Systems*, *41*, 587-610.

Lucas, J. W. (2003). Theory-testing, generalization, and the problem of external validity. *Sociological Theory*, *21*(3), 236-253.

Maass, W., Parsons, J., Purao, S., Storey, V. C., & Woo, C. (2018). Data-driven meets theory-driven research in the era of big data: Opportunities and challenges for information systems research. *Journal of the Association for Information Systems*, *19*(12), 1253-1273.

Martin, K. E. (2015). Ethical issues in the big data industry. *MIS Quarterly Executive, 14*(2), 67-85.

Mathiassen, L., & Sandberg, A. (2013). How a professionally qualified doctoral student bridged the practice-research gap: A confessional account of collaborative practice research. *European Journal of Information Systems, 22*(4), 475-492.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think.* New York, NY: Houghton Mifflin Harcourt.

McAfee, A., & Brynjolfsson, E. (2012). Big data: The management revolution. *Harvard Business Review, 90*(10). 60-68.

Menon, N. M., & Kohli, R. (2013). Blunting Damocles' sword: A longitudinal model of healthcare IT impact on malpractice insurance premium and quality of patient care. *Information Systems Research, 24*(4), 918-932.

Miles, M. B., Huberman, M. A., & Saldna, J. (2014). *Qualitative data analysis: A methods sourcebook.* Thousand Oaks, CA: Sage.

Murdoch, T. B., & Detsky, A. S. (2013). The inevitable application of big data to health care. *JAMA, 309*(13), 1351-1352.

Myers, M. D. (1997). Qualitative research in information systems. *MIS Quarterly*, *21*(2), 241-242.

Nakamura, M. M., Harper, M. B., Castro, A. V., Yu, J. F. B., & Jha, A. K. (2015). Impact of the meaningful use incentive program on electronic health record adoption by US children's hospitals. *Journal of the American Medical Informatics Association, 22*(2), 390-398.

NHS. (2019). *The NHS long term plan.* Retrieved from https://www.longtermplan.nhs.uk/wp-content/uploads/2019/01/nhs-long-term-plan.pdf

NSF. (2019). *Open data at NSF.* Retrieved from https://www.nsf.gov/data/

Patzer, G. L. (1995). *Using secondary data in marketing research: United States and worldwide.* Westport, CT: Greenwood Publishing Group.

Peng, G., Dey, D., & Lahiri, A. (2014). Healthcare IT adoption: An analysis of knowledge transfer in socioeconomic networks. *Journal of Management Information Systems*, *31*(3), 7-34.

Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: Promise and potential. *Health Information Science and Systems, 2*(1).

Rai, A. (2016). Editor's comments: Synergies between big data and theory. *MIS Quarterly*, *40*(2), iii-xi.

Sarker, S., Xiao, X., & Beaulieu, T. (2013). Qualitative studies in information systems: A critical review and some guiding principles. *MIS Quarterly*, *37*(4), iii-xviii.

Schlomer, B. J., & Copp, H. L. (2014). Secondary data analysis of large data sets in urology: Successes and errors to avoid. *The Journal of Urology, 191*(3), 587-596.

Schultze, U. (2000). A confessional account of an ethnography about knowledge work. *MIS Quarterly, 24*(1), 3-41.

Shmueli, G. (2010). To explain or to predict? *Statistical Science, 25*(3), 289-310.

Shmueli, G., & Koppius, O. R. (2011). Predictive analytics in information systems research. *MIS Quarterly*, *35*(3), 553-572.

Smith, A. K., Ayanian, J. Z., Covinsky, K. E., Landon, B. E., McCarthy, E. P., Wee, C. C., & Steinman, M. A. (2011). Conducting high-value secondary dataset analysis: An introductory guide and resources. *Journal of General Internal Medicine, 26*(8), 920-929.

Smith, E. (2008). Pitfalls and promises: The use of secondary data analysis in educational research. *British Journal of Educational Studies, 56*(3), 323-339.

Snyder, B. (2013). Google's panicky flu estimates were dead wrong. *CIO*. Retrieved from https://www.cio.com/article/2370689/consumer-technology/google-s-panicky-flu-estimates-were-dead-wrong.html

Spetz, J., & Maiuro, L. S. (2004). Measuring levels of technology in hospitals. *The Quarterly Review of Economics and Finance, 44*(3), 430-447.

Straub, D. W. (1989). Validating instruments in MIS research. *MIS Quarterly*, *13*(2), 147-166.

Sykes, T. A. (2015). Support structures and their impacts on employee outcomes: A longitudinal field study of an enterprise system implementation. *MIS Quarterly*, *39*(2), 473-495.

Trzesniewski, K. H., Donnellan, M. B., & Lucas, R. E. (2011). *Secondary data analysis: An introduction for psychologists.* Washington, DC: American Psychological Association.

van de Ven, A. H. (2007). *Engaged scholarship: A guide for organizational and social research.* Oxford, UK: Oxford University Press.

van Maanen, J. (2011). *Tales from the field: On writing ethnography.* Chicago, IL: University of Chicago Press.

WHO. (2019). *Global strategy on digital health 2020-2024.* Retrieved from https://extranet.who.int/dataform/upload/surveys/183439/files/Draft%20Global%20Strategy%20on%20Digital%20Health.pdf

Zhou, Z., Chawla, N. V., Jin, Y., & Williams, G. J. (2014). Big data opportunities and challenges: discussions from data analytics perspectives. *IEEE Computational Intelligence Magazine*, *9*(4), 62-74.

About the Authors

Sumantra Sarkar is currently an Associate Professor in Management Information Systems at the School of Management, State University of New York, Binghamton. He received his Ph.D. in Computer Information Systems from the J. Mack Robinson College of Business, Georgia State University. He has an MS in Computer Information Systems (Health Informatics), an MBA in Operations Research, and holds PMP and CISA certifications. His research interests include health information technology, organizational processes, agile development, IT governance and IT security. His work has appeared in premier IS journals and conferences like *Information Systems Research, Journal of Management Information Systems*, *European Journal of Information Systems*, *Information Systems Journal*, *Journal of Business Research*, *IEEE IT Professional*, *Academy of Management Best Papers Proceedings*, among others. He has over two decades of experience in the industry holding senior management positions in IT organizations of large multinational corporations like GEC, Novell, Hutchison Whampoa, and ABN AMRO Bank. Before he moved to research and academia permanently, he headed the IT delivery group for the Indian operations of Royal Bank of Scotland as Vice President, Head of Infrastructure and Shared Delivery.

Kaushik Ghosh is an Assistant Professor of Information Systems at the Sawyer School of Business, Suffolk University. He received his Ph.D. from The University of Mississippi. His research interests include topics relevant to the impact of technology on health care and quality, and the drivers of digital transformation. His work is published in journals, including *Communications of the Association for Information Systems*, *Journal of Computer Information Systems*, *International Journal of e-Collaboration*, among others.

Stacie Petter is the Ben H. Williams Professor of Information Systems & Business Analytics at the Hankamer School of Business at Baylor University. Her research focuses on positive and negative impacts of information systems, research methodology, and software project management. Her work appears in outlets including, *Journal of the Association for Information Systems, MIS Quarterly, Journal of Management Information Systems, European Journal of Information Systems,* among others. She has served as Editor-in-Chief of *The DATA BASE for Advances in Information Systems* and Vice President of Region 1 on the Association for Information Systems Council.