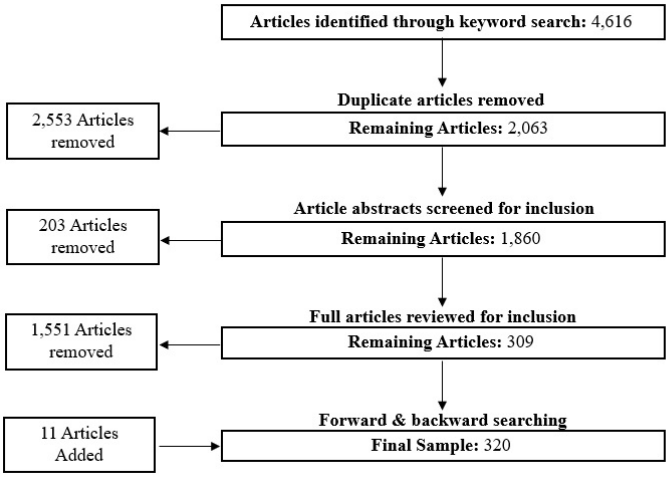
E-COMPANION

E-Companion A: Article Searching and Screening Process



Search Terms: "Healthcare" and "Health" in conjunction with "Analytics," "Business Analytics," "Machine Learning," "Big Data," "Artificial Intelligence," "Deep Learning," "Hospital Decision Support", "Healthcare Decision Support", "Medical Decision-Making", and "Clinical Decision Support Systems"

UT Dallas Journals Searched from 2007 to 2020:

OM Journals
Journal of Operations Management (*JOM*)
*Management Science** (*Mgt Sci*)
Manufacturing and Service Operations Management (*MSOM*)
Operations Research (OR)
Production and Operations Management (*POM*)

Management Journals
Academy of Management Journal (*AMJ*)
Academy of Management Review (*AMR*)
Administrative Science Quarterly (*ASQ*)
Organization Science (*Org Sci*)
Strategic Management Journal (*SMJ*)

IS Journals
Information Systems Research (*ISR*)
Journal of Management Information Systems (*JMIS*)
MIS Quarterly (*MISQ*)

Finance, Accounting, Economics Journals
Journal of Accounting and Economics (JAE)
Journal of Accounting Research (JAR)
Journal of Finance (JF)
Journal of Financial Economics (JFE)
The Accounting Review (AR)
The Review of Financial Studies (RFS)

Marketing Journals
Journal of Consumer Research (JCR)
Journal of Marketing (JM)
Journal of Marketing Research (JMR)
Marketing Science (Mark Sci)

Other Journals
Journal on Computing (JC)
Journal of International Business Studies (JIBS)

*We recognize that *Mgt Sci* is considered a multi-disciplinary journal and thus its classification as an OM vs. IS journal varies depending on the analysis.

Figure eC1. Article Searching and Screening Process

E-Companion B: Topic Modeling Setup

Here we provide a detailed description of the interactive article analysis web application (IWA) and all of its functionalities. The IWA can be accessed here: https://notredame-mendoza.shinyapps.io/msom-bah/. We organize this description by creating a subsection that corresponds to each tab in the IWA to help users navigate all of its features.

eC B.1 "About" Tab

The IWA loads to an "About" page that describes the features and functionality (see Figure eC2). More specifically, it describes topic modeling at a conceptual level and links to relevant articles and YouTube videos so users can learn more about the model fitting process and how to interpret the results. For more specific information on topic modeling, see E-Companion C. Figure eC2 also shows the sidebar where users select options for running their topic model analysis. The IWA is highly customizable; users can select from pre-curated datasets or upload a dataset of their choosing. Once a dataset is selected, users can modify start and end years, number of topics, and journal subsets before fitting the topic model.

Figure eC2. "About" tab from the IWA

eC B.2 "Instructions" Tab

The "Instructions" tab provides detailed instructions on using the IWA and interpreting results from each of the analyses. One of the best features of the IWA, which may significantly increase its impact across multiple research areas, is that scholars can upload their own custom datasets. For instructions that guide users through how to upload a custom dataset, click on "Show Upload Instructions" located at the bottom of the "Instructions" tab, which brings up step-by-step instructions about how to export abstracts from PubMed (see Figure eC3).

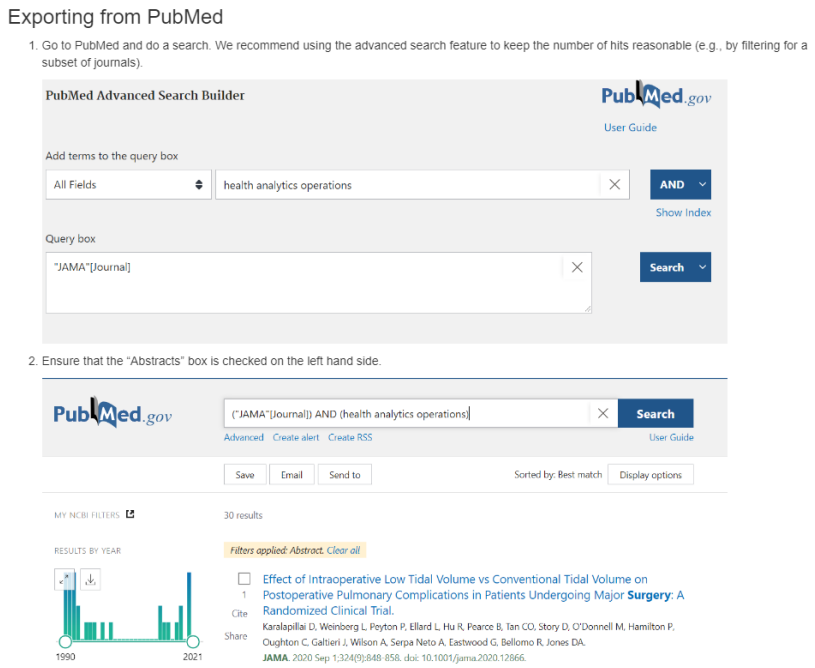


Figure eC3. Instructions for Exporting Abstracts from PubMed

eC B.3 "Examples" Tab

The "Examples" tab in the IWA provides pre-configured models for replicating the analyses in our paper and examples provided below. By clicking one of the buttons on this tab, the options on the sidebar are automatically adjusted to generate a specific topic model. The model and subsequent outputs can be used to follow along with the analyses specified in the corresponding manuscript section. This gives the reader an opportunity to use our manuscript as a starting point; by modifying the model from one of the examples, a researcher can extend our analyses to incorporate other themes, journals, etc.

eC B.4 Topic Model: A Brief Description

A topic model analyzes article abstracts, which reveals topics examined in the literature as well as trends of how topics change over time. Considering the collection of article abstracts as our data set, the IWA fits a Latent Dirichlet Allocation (LDA) topic model (for more information on LDA topic models see E-Companion C and Blei et al. 2003). LDA is a generative probabilistic model, which means that the model is designed based on a "story" of how the documents in the data set (in our case, the article abstracts) are written. For LDA, there exists some probabilistic distribution over topics, or themes, that describe the topics in a set of abstracts. Each of these topics has an associated probabilistic distribution of words. For example, a topic about diabetes will have certain words such as sugar, glucose, and insulin as high probability words for that topic. LDA assumes each individual abstract is written in a four-step process: 1) start with a blank document; 2) randomly select one of the available topics, according to their probabilities; 3) for that topic, randomly select a word based on the probabilities; and 4) write the word down, and move to the next position. Steps 3 and 4 repeat themselves until the entire abstract is written.

This is, of course, not how abstracts are actually written, but by making this assumption and starting with a large data set of abstracts, the LDA model can learn a coherent and interpretable set of topics that describe the abstract data set. A nice property of LDA and topic models in general is that they are *unsupervised* models that do not require any expensive human annotation. The only input required is the data set of abstracts itself, which also makes them less susceptible to human biases. The model learns by estimating the probabilities described above, given the words in the abstracts, specifically, the counts of how often each word occurs in each abstract. For our IWA implementation of LDA, we use a standard LDA model with typical model-fitting conventions. Consistent with the literature (Wallach, Mimno, and McCallum 2009), we remove stop words (e.g., "and," "the," "a," etc.) as well as overrepresented words in our data set. For example, "abstract" appears in most abstracts in our data set and therefore is not useful for distinguishing between topics.

There are three main outputs from the topic model analysis that correspond to a tab in our IWA: 1) the topic model itself, 2) trend analysis, and 3) abstract scores. As an illustration, we discuss each of these outputs below with a 4-topic model generated by our analysis. We offer a general discussion of the

insights that can be extracted from each output, but reserve our main discussion for overarching trends in the literature to Section 4.2 of the paper. We reproduce visuals from our IWA when discussing each output, but encourage readers to engage with the web application on their own while following along with the examples provided to identify additional interesting trends beyond those discussed below.

eC B.5 "Topic Model" Tab

The first output of the IWA is a topic-word probability chart (Figure eC4), which is displayed in the "Topic Model" tab. For each topic identified by the model, there are certain words that are highly probable for the topic. This means that the conditional probability of a word, given a topic, is high. For example, the plot in Figure eC4 shows a 4-topic model fit on the UTD24 data. To follow along using the IWA, choose Example 4 in the Examples tab, but note that the default condition for Example 4 excludes reviews and commentaries, whereas the ones shown in this E-Companion include reviews and commentaries. Simply check the "Include reviews and commentaries?" box to replicate the results presented here. It is also important to note that we are continually updating the IWA app to make it more useful by adding more merge-words and eliminating stop words and non-descriptive words. Therefore, results may vary, and topic numbers may not align with what is shown in below. For the first topic, the most probable words are "schedule," "appointment," and "wait." Given the first topic, these words are more likely to appear in the text than, for example, "expect" or "current" or other words that appear lower in the list. These probabilities effectively define each topic (recall topic modeling is an *unsupervised* process). Each abstract in our data set consists of a distribution over these four topics, and the percentage next to each topic represents the topic probabilities or how frequently they appear in our sample. For example, topic 1 represents 27.7% of the abstract data set (see Figure eC4).

The topic model provides objective insights on common themes in the literature. For example, topic 1 centers on the question of *scheduling appointments*[1] as a *resource* allocation or *capacities* problem, and deals with *wait* times. Topic 2 focuses more on *firm financial* performance or *business*, inter-organizational exchanges of *supplies* between *manufacturers* and other *market* participants. Topic 3 deals with *social, online, networks* and how they change *behaviors* or *relationships* within *communities.* Topic 4 centers on *risks* or change in *readmissions* according to various *chronic diseases*. Below the topic model is a chart on method analysis (see Figure eC5) that shows the average topic proportion associated with an analytical technique (see Figure 9). Note that this method analysis chart appears only for the UTD24 dataset and not for medical journals by subcategory because most journals do not require that the analysis methods be reported as keywords. We hope future enhancements of the IWA can provide this feature. Finally, the "Topic Model" tab also includes the top scoring articles for each of the topics. The articles that have the highest probability weight associated with a single topic are included in this table.

[1] Words in italics are taken directly from the topic model. Tense of words may be modified to enhance readability.

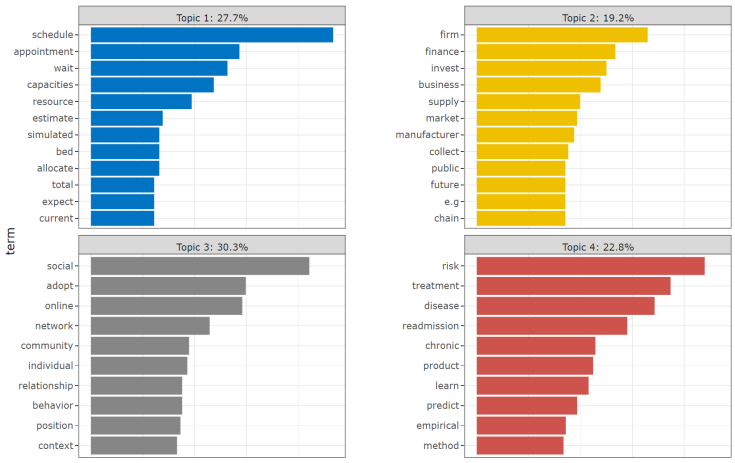


Figure eC4. Topic-word probabilities from IWA

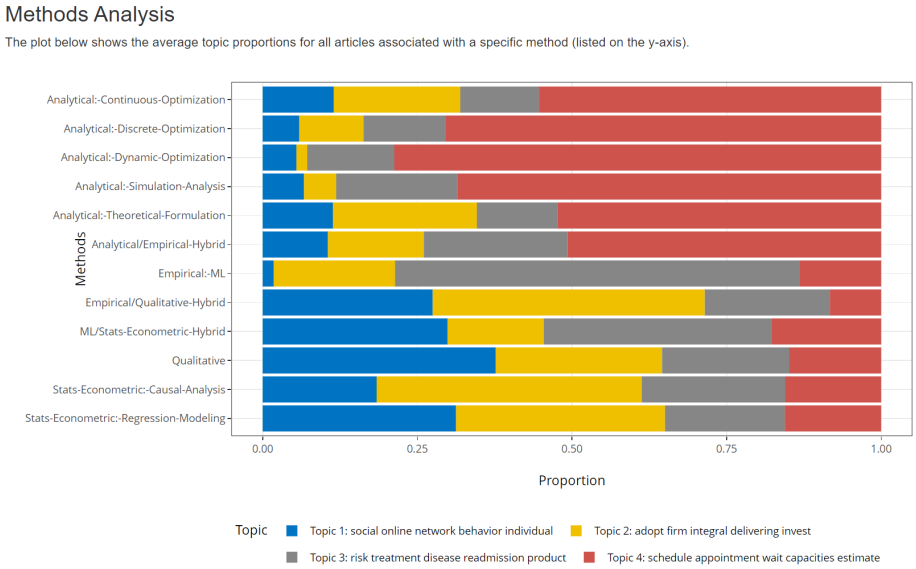


Figure eC5. Topic proportions associated with each method – see Example 4 topic model from IWA

eC B.6 "Trend Analysis" Tab

The second output of the IWA is a trend analysis (see "Trend Analysis" tab), which provides a more objective analysis of how research trends change over time. Figure eC6 provides an illustration of the same 4-topic trend analysis discussed above (see the "Trend Analysis" tab). The trend analysis considers all articles published in a single year, and calculates the distribution of topics across the articles (abstracts). To quantify how far the distribution of topics is from a uniform distribution (i.e., each topic equally represented that year), we calculate the Kullback-Leibler (KL) Divergence between the topic distribution and a uniform distribution for each year. This value is indicated by the number at the top of each column year. Smaller numbers indicate a distribution over topics that is closer to a uniform distribution. For example, the distribution of topics in 2020 is more balanced and much closer to a

uniform distribution than the distribution of topics in 2008. As shown in Figure eC6, at the beginning of our sample, topic 2 is by far the least examined – it typically represents less than 10% of the abstracts. However, over time, interest in topic 2 gains momentum and, as of 2012, represented 35% of the content of abstracts from papers published, when examining a 4-topic model. It eventually wains over time, but still is more than in the early years. On the other hand, popularity in topic 1 moves in the exact opposite direction. It is very strong in the early part of our sample and has since decreased in popularity, representing just over 24% of the content in the articles published in 2020. This plot is interactive; if any topic is double clicked, it will isolate that topic in the trend analysis. Below the topic trend analysis on the IWA is a plot of how many articles are published in each journal each year. If any specific journal is double clicked, the plot will isolate how many articles have been published in that journal over time.

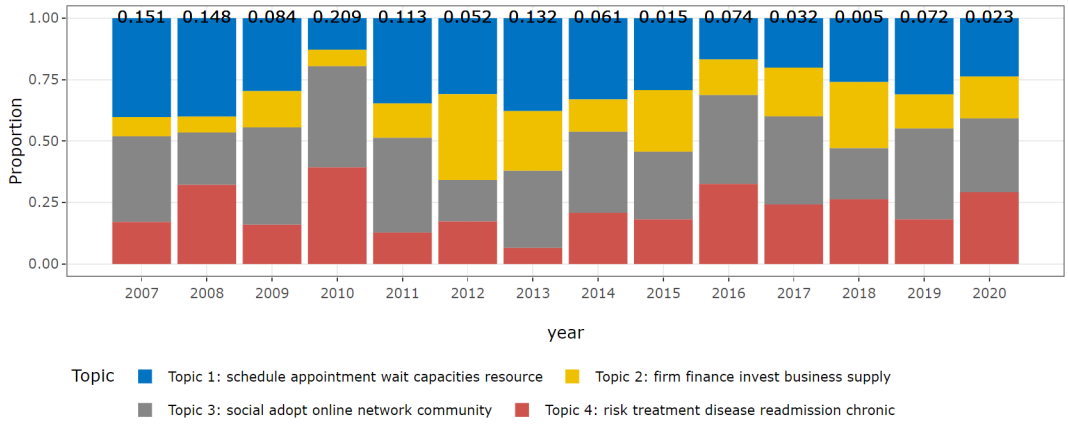


Figure eC6. Example of a 4-Topic Trend Analysis

eC B.7 "Abstract Scores" Tab

The third output of the IWA is abstract scores (see "Abstract Scores" tab), which offers insight about the specific abstracts (and, thus, papers) most associated with a particular topic. Each abstract has a score for each topic, indicating the weight of that topic for that abstract. The scores for all topics for an abstract sum to 1, so these scores can be thought of as a distribution over the topics for the abstract. Table eC1 provides an example from the same 4-topic analysis discussed above (see the "Abstract Scores" tab). In this table, there is a Score column for each topic searched. In the box on the upper left-hand side of the screen, users can select which collection of topics they would like displayed; in the specific example shown, it includes scores for all four topics. To identify the highest scoring article for Topic 1, select that column heading to sort. Article ID#145 (Carew et al., 2020) scores the highest on topic 1 based on the probabilities of the learned model, with low scores for topics 2, 3, and 4. This means that most of the text content in that paper's abstract fits within topic 1 (≈ 98.9%) and significantly less text fits in with the other topics. There is also a search box on the upper-right hand side where users can search for keywords in any of the columns such as Author, Title, Abstract, or Journal. This search feature can help users focus

on a variety of sub-categories from particular topics, authors, or certain journals. This table also provides direct access to each paper associated with a given abstract (assuming institutional or personal access).

Top Abstracts for Selected Topic

If you want to restrict the topics that are displayed, you can add/delete topics in the text box below. Click within the box and then hit your delete key to delete a topic. Once deleted, you can add a topic back in by selecting it from the dropdown.

Select a topic/topics:

1 2 3 4

Show 10 entries Search:

ID	Author	Title	Year	Abstract	Method	Journal	Topic 1 Score	Topic 2 Score	Topic 3 Score	Topic 4 Score	Paper Link
145	Carew, Stephanie and Nagarajan, Mahesh and Shechter, Steven and Arneja, Jugpal and Skarsgard, Erik	Dynamic Capacity Allocation For Elective Surgeries: Reducing Urgency-Weighted Wait Times	2020	Problem definition: given the variety of urgency levels in highly utilized operating rooms, capacity allocation decisions can have a major impact on how wait times are rationed. we examine a longer-t...	Qualitative	Manufacturing & Service Operations Management	0.9886	0.0038	0.0038	0.0038	click for online access

Table eC1. Example from Abstract Scores Analysis

eC B.8 "Network Analysis" Tab

The topic model provides insights about research trends over time (see Figure eC6), but falls short on revealing how frequently (or sporadically) research topics are examined together, which is why we provide a network analysis. The network analysis reveals research opportunities as it sheds light on under-examined topics in the literature and allows scholars to easily identify articles on a specific topic and/or articles that span multiple topics. Said differently, the topic model provides a *within* article analysis while the network analysis provides a *within* and *between* article analysis.

To conduct this network analysis, we leverage the article titles for keywords. We use this data in our network analysis to examine how frequently various keywords occur together across articles. To conduct this analysis, we built a keyword co-occurrence network, where each keyword is a node, and each edge indicates a connection between two keywords (i.e., both keywords are included in a single article). The size of the node reflects how many article titles include that specific keyword. Similarly, the edges in the network are weighted by the number of connections, so keywords that frequently co-occur in articles have a stronger connection and thicker edge than those that sporadically co-occur. Figure eC7 provides an illustration of a network analysis for the same collection of articles included in the 4-topic model discussed above (see the "Network Analysis" tab). To manage the size and readability of the network, the network shown in Figure eC7 only includes 14 keywords, but this number can be expanded or reduced in the interactive version.

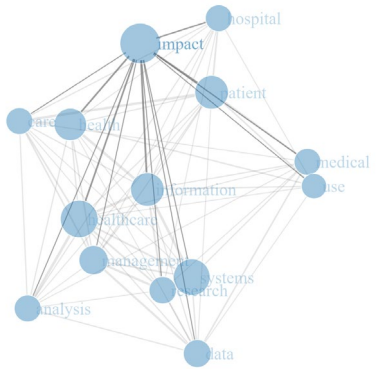


Figure eC7. Example of a Network Analysis from 4-Topic Model Analysis

The value of this network is multifaceted. First, it provides scholars with an understanding of commonly researched topics, topics less explored, and research that falls at the intersection of multiple topics. Second, it offers guidance on which relationships and/or topics are mature enough for synthesis via a meta-analysis or other statistical analysis. In order to extract meaningful insights from meta-analyses, there needs to be a certain number of studies on the relationships examined (Lipsey and Wilson 2001); the thickness of edges in the network analysis can help reveal if this threshold is met. Third, if users click on a node ("impact" in this example), a more fine-grained network analysis appears and a table is populated directly below the network analysis that shows all studies that include the specific keyword in the title (see Table eC2). This also occurs if scholars click an edge (i.e., link between two keywords). Specifically, if a researcher clicks on the edge between the hospital node and the patient node, all studies that include both of those specific keywords in their titles are identified and populated in a table below the network analysis. Both of these features allow scholars to focus their literature review efforts and enable them to easily identify all relevant research on a specific topic or at the intersection of two topics, which will help reveal under-examined research areas.

Papers for keyword: impact

Show 10 entries Search:

Showing 1 to 10 of 21 entries Previous 1 2 3 Next

	Author	Title	Journal	Year	Paper Link
1	Dong, Jing and Yom-Tov, Elad and Yom-Tov, Galit B.	The Impact Of Delay Announcements On Hospital Network Coordination And Waiting Times	Management Science	2018	Click here
2	Chen, Hsinchun and Chiang, Roger H. L. and Storey, Veda C.	Business Intelligence And Analytics: From Big Data To Big Impact	MIS Quarterly	2012	Click here

Table eC2. Example of Article Table Populated by Clicking Node

eC B.9 "Advanced Options" Tab

We provide an "Advanced Options" tab for users who want to further customize their topic modeling. The options on this tab allow users to adjust model inputs, training parameters, and output display options. Specifically, users can modify:

- *Number of keywords per topic*: The IWA defaults to showing the top ten words for each topic. Users can adjust this to see as few as five or as many as fifteen words in the IWA outputs.[2]
- *Exclusion words*: For certain datasets there may be words that should not be included in the topic modeling, because they are extremely common or known to be not relevant for a particular research question. Users can add these words to an exclusion list, and they will be removed from the abstracts before fitting the topic model.
- *Merge words*: Alternatively, users may want to consider multi-word phrases as a single word (e.g., "business analytics"). By adding phrases to the merge word list, they will be considered as a single "token" in the topic model. Merge words that are already included are shown in the shaded box.
- *Alpha parameter*: When fitting a topic model, part of the model fitting process is making an assumption about how many topics are highly probable for a given document. Adjusting the alpha parameter lets the user determine if a document should be comprised of a small number of highly probable topics (low alpha value) or should have a more even distribution over the topic space (higher alpha).
- *Keywords for network*: If keywords are available for a network (e.g., from EBSCO), then by selecting this option those keywords are used for the network analysis instead of the words in the article titles.

eC B.10 "Resources" Tab

We added a "Resources" tab that lists numerous databases and provides information on data included in each database, sample size, data coverage (i.e., dates), cost categorization, and a link to most data sources. These tables are also included in the E-Companion D; see Table eC3. Including this information in the IWA makes it more accessible for users (academics and practitioners alike) and can be added to over time as new databases become available.

E-Companion C: Topic Modeling Details

An LDA model estimates a distribution of topics over documents in a corpus and a distribution of words over topics (Blei 2012). It is a probabilistic, unsupervised model that learns the distributions using a corpus of text documents as inputs. In contrast with a supervised model, there are no labels defining words that can be used as part of the model-learning process. Learned topics are latent variables and are learned according to predefined assumptions about the document-generating process. For example, the underlying intuition of LDA can be summarized by the generative assumptions made about how a single document is "written." A document consists of a distribution over topics and each topic consists of a distribution over words. To generate a document, the user samples a topic according to the document's topic probabilities and selects a word from the vocabulary according to the selected topic's word probabilities. This process repeats itself as many times as the number of words in the document. The first step in learning an LDA model is to preprocess the input data to put it in the necessary format. Specifically, each entry in the input data is represented as a "bag of words," where word order does not

[2] In our implementation, "ties" (words that are equally probable) are displayed together, so the specific number of keywords may be greater than the selected option for one or more topics.

matter, but word count is retained. The corpus is a collection of text from all abstracts reviewed. A document is a single article abstract in bag of words format and thus multiple documents make up a corpus. Documents are represented as a vector $d = \{0,1\}^{|v|}$, where $|v|$ is the length of the corpus vocabulary, each column represents a word in the vocabulary, and $d_i = 0$ if word i is not present in the document, and $d_i = n$ if word i is present in the document n times. Each document's bag of words can be considered a term-frequency vector, and the combined vectors for an entire corpus is a term-frequency matrix. As part of the process of converting abstracts to bag of words representations, a standard list of "stop words" are removed. We supplement a standard stop word list with corpus-specific words that also occurred in most documents (e.g., "data," "abstract"). This term-frequency matrix is the input to the LDA model, which estimates the distribution of topics for the abstracts, as well as the distribution of tokens for each topic. The output topics are interpretable in that the tokens with the highest probability for each topic "define" the topic. The top words can then be used to analyze documents, with the highest probability topic for each document indicating the major theme of the document.

E-Companion D: Data Availability

This section is devoted to discussing high-quality healthcare databases that, when combined with analytic techniques, can extract insights that enhance the quality of care provided and, thus, benefit all stakeholders. To help guide research in this area, we discuss both commonly used healthcare databases as well as underutilized databases; see Table eC3 for a list of available databases.

When beginning this study, we expected HIMSS Analytics to be the most widely used database for research in this area. While our intuition was confirmed, HIMSS Analytics was only used in 27 articles (or 8.4%), which was much lower than we expected. While HIMSS Analytics provides rich, longitudinal, hospital and health IT data, recent merger/acquisition activity resulted in a dramatic increase in price.[3] Prior to 2019, access to HIMSS Analytics required a contractual agreement, in which the partner institution agreed to participate in its annual survey, and a modest purchase price. This allowed their data to diffuse quite widely and rapidly. As a result of the price increase, it likely will not be feasible for most business school researchers to use this data because of resource limitations. Consequently, we anticipate that scholars will migrate away from using HIMSS Analytics data in their research as less expensive, novel big data sets become available. Data from the Centers for Medicare and Medicaid Services (CMS) was more commonly used than anticipated – it was used in 23 studies (or 7.2%).[4] The vast majority of studies used more specialized databases (e.g., Washington State Department of Health data, Breast Cancer Surveillance Consortium data), proprietary data (e.g., user activity logs), or qualitative data collected by authors. We were also surprised by the number of partnerships researchers

[3] While the authors do have access to pricing data, we are not revealing this information since the costs are likely to differ among institutions based on a number of factors such as: number of users, number of variables requested, number of years of data, data processing expenses, etc.
[4] A handful of studies used data from both HIMSS Analytics and CMS.

formed with one or two hospitals in order to gain access to data (for examples see Dai and Shi 2020; Liu et al. 2018; Ranjan et al. 2017; Shi et al. 2020; White, Torabi, and Froehle 2017; Zargoush et al. 2018).

The pervasive use of social media by society has resulted in firms accumulating massive amounts of data that can be used for research to benefit all healthcare stakeholders. Yet, very few scholars used data from social media platforms or crowd-sourcing websites such as Yelp. These platforms offer rich research opportunities as they allow scholars to answer novel questions with particularly large data sets. Abbasi et al. (2019), for example, analyzed more than 12 million tweets and 5 million forum postings to examine user-generated content signals for early adverse event warnings in the pharmaceutical industry. Similarly, Mejia et al. (2019) used more than 1.3 million Yelp reviews from 24,625 restaurants to examine hygiene inspection scores. Interesting data from social media platforms that has not been used for research in this area is the "check-in" feature, which shows when users are at a specific location (e.g., gym). Check-in data revealing when users work out at a gym or finish a run is particularly granular data that could be leveraged to examine fascinating research questions at the individual (patient) level.

The availability of social media data lies on a continuum with "no access" anchoring one side and "unlimited access" anchoring the other side. Typically, access to social media platform data is controlled via an application programming interface (API) or a curated collection of data sets that can be downloaded. APIs are made available by companies for programmatic access to their data. For example, while Facebook does not allow web scraping of their website,[5] they do offer curated data sets for researchers that can be downloaded via an application process.[6] Twitter is more open in terms of data access, offering full access to their data for research purposes via their API.[7] Google offers a wide variety of data sets for research, from company data around search, mobile, etc. to data sets collected by Google researchers and made available to the wider research community.[8] In particular, the Google Health Trends API allows researchers to analyze Google user search data to identify trends in health terms searched on the web over time (Zepecki et al. 2020). On the "full access" end of the continuum, researchers can scrape webpages such as online message boards to collect data themselves, in accordance with the page's terms of service. We are hopeful that researchers will form partnerships with social media companies that are willing to share data (even limited amounts of data), as such data can reveal rich insights into previously unexplored relationships in the BAH domain.

Another means of collecting large quantities of granular data is through the use of wearable devices (e.g., Fitbit or Apple Watch), which are devices equipped with sensors that collect data on heart rate, user activity, movement, etc. Wearables can be used to assess patient health conditions in real-time

[5] https://www.facebook.com/apps/site_scraping_tos_terms.php
[6] https://research.fb.com/data/
[7] https://developer.twitter.com/en/solutions/academic-research/products-for-researchers
[8] https://research.google/tools/datasets/

(Greiwe and Nyenhuis 2020), improve adherence to treatment protocols (Quinn et al. 2018), help smokers quit (Herbec et al. 2019), encourage people to exercise (Lambert et al. 2017), and generally positively impact health; yet the vast majority of the studies using such data have been published in medical, informatics, or computer science outlets rather than business journals. One challenge associated with conducting research in the healthcare domain is that the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule restricts researchers from acquiring and using Protected Health Information (PHI) unless the researcher's institution is a HIPAA Covered Entity (CE – often medical providers) or a Business Associate (often insurance companies). Even if a researcher was able to get a CE or Business Associate to share PHI, it does not make that researcher or his/her institution a CE or a Business Associate, but does transfer all or part of the data security requirements, depending on a data usage agreement (DUA), from the CE or Business Associate to the third party. DUAs vary quite extensively and various levels of anonymity can reduce security requirements, but these agreements require legal review. Finally, in some types of research, written authorization from all subjects involved in the study still may be required.[9] De-identified data is not protected by the HIPAA Privacy Rule, but acquiring such data from individual hospitals is extremely challenging due to risk aversion from hospital administrators and costs associated with de-identifying the data. In sum, HIPAA is clear for CEs and Business Associates, but the third party use of PHI is governed by a highly variable set of regulations enforced in DUAs.

[9] https://privacyruleandresearch.nih.gov/pr_08.asp

Table eC3. Healthcare Databases[10]

Name	Sample Size	Data Included and Dates	Categorization (Cost)[11]
i. AHRQ MEPS	Varies by state. Set of large-scale surveys of families and individuals, their medical providers, and employers across the U.S.	Provides data on specific health services that Americans use, how frequently they use them, the cost of the services, and how they are paid. Also includes insurance data covering cost, scope, and breadth of health insurance held and available to U.S. workers. Dates: 1996-Present	Open Source (Free)
ii. American Hospital Association (AHA)	Provides data on demographics, operations, service line, staffing, CEO/President, expenses, organization structures, beds, utilization, population health, and more from over 6,200 hospitals and 400 healthcare systems.	Annual survey of U.S. hospitals that targets those responsible for IT in hospitals (e.g., IO) and tracks the adoption of EHRs as well as hospital initiatives to demonstrate meaningful use. Dates: 1980-Present	Commercial or Open Source (Free) – cost depends on exact data requested
iii. Area Health Resources File (AHRF)	Varies	Includes data on health care professions, health facilities, population characteristics, economics, health professions training, hospital utilization, hospital expenditures, and environment at the county, state and national levels, from over 50 data sources. Dates: 2005-Present	Open Source (Free)
iv. Clarivate Analytics	Includes data on more than 260,000 global clinical trials from biomarkers, devices, biologics, and drugs. Recently acquired DRG, which includes data on 3,500+ Health Systems; 1.8+ million practitioners (including location, procedure volume and affiliations); and 7,100+ hospitals (including clinical and financial metrics).	Provides data on clinical trials including mechanism of action being pursued in the trial and adverse events by interventions as well as company collaborators and sponsors. DRG provides data on strength of affiliations within healthcare delivery systems. Includes data from individual physicians up to the facility and parent health system level. Dates: 2011-2019	Commercial ($-$$$$)
v. CMS HospitalCompare	Medicare/Medicaid patients for all hospitals serving this segment.	HospitalCompare allows consumers to select multiple hospitals and directly compare performance measure information related to heart attack, heart failure, pneumonia, surgery and other conditions. Dates: 2002-Present	Open Source (Free)
vi. County Health Rankings & Roadmaps (CHR&R)	Varies state by state	Provides data on the health of communities (e.g., length of life, quality of life) in nearly every county in the nation and provides insight on how to create healthier places to live. Dates: 2011-Present	Open Source (Free)
vi. Dartmouth Health Atlas	Varies – Primary care access and quality measures (2003-2015); Medicare mortality rates (1999-2017); Claims data (2003-2017)	Provides longitudinal data on healthcare spending Hospital Referral Regions (HRRs), quality measures, and mortality rates. Dates: 2003-2017	Open Source (Free)
vii. eHealth Initiative (eHI)	Includes data on over 100 health information exchanges (HIEs).	Annual survey of HIEs at the state and local levels. Dates: 2005-2010	Open Source (Free)
viii. Florida AHCA	Data from all Florida entities shown below: • Hospitals • Ambulatory Surgery Centers Emergency Departments	• Hospital Patient Data • Ambulatory Surgery Center Patient Data • Emergency Department Patient Data (starting in 2005) • Hospital Financial Data Dates: 1988-Present unless otherwise noted	Open Source (Free)

[10] To populate Table eC3, we attempted to contact data vendors in situations where we did not have information about variables, availability dates, and pricing. In some cases, we were unable to acquire this information prior to submission.

[11] $ = only pay for each year of data requested (ranges from $20/year for older years [e.g., 2006] to $1,000/year for more recent years)
$$ = less than ten thousand annually
$$$ = tens of thousands or more annually
$$$$ = hundreds of thousands

Table eC3. Healthcare Databases (continued)

Name	Sample Size	Data Included and Dates	Categorization (Cost)
ix. H-CUP	Varies depending on statewide inpatient data, ambulatory surgery and services data, or emergency department data.	Includes the largest collection of longitudinal hospital care data in the U.S. with all-payer, encounter-level information. Dates: 1988-2018	Commercial ($)
x. HIMSS Analytics (now owned by Definitive Healthcare)[12]	~ 85% of health providers, 6,000 hospitals, and 20,000 ambulatory care facilities. Definitive Healthcare includes all-payer claims information for 247 million U.S. patients, which can be combined with healthcare facility and provider data.	Health IT adoption data and hospital/ambulatory care facility meaningful use data. Dates: 2005-Present	Commercial ($$$)
xi. Leapfrog	2,300+ hospitals	Annual survey on U.S. hospital performance includes data on surgical volume, mortality rates, infection rates, C-section rates, and hygiene. Dates: 2001-Present	Commercial ($)
xii. Levin Associates Healthcare M&A	Includes data on 30,000+ mergers and acquisition deals across 13 seniors housing and healthcare sectors	Data includes quantitative metrics to measure transactions including EBITDA, price, units, pricing terms, revenue, and other metrics. Dates: 1993-Present	Commercial ($-$$)
xiii. MIMIC data	Includes data on 61,000+ ICU stays: 53,000+ stays for adult patients and 8,100+ stays for neonatal patients.	Includes detailed, but de-identified data on the clinical care of patients. Dates: 2001-2012	Open Source (Free)
xiv. National Center for Health Workforce Analysis (NCHWA)	13,000+ Nurse practitioners, number of registered nurses varies.	Provides data on the supply, use, access, need, and demand for health workers. Dates: periodic surveys 1977-2018	Open Source (Free)
xv. NEMSIS - National Emergency Medical Services Information System	43,488,767 EMS activations submitted by 12,319 EMS agencies serving 50 states and territories.	The dataset includes 640 data elements including but not limited to: dispatch and arrival times, patient condition, incident, etc. It does not contain information that identifies patients, EMS agencies, receiving hospitals, or reporting states.	Open Source (Free) but requires data request
xvi. OneKey by IQVIA	Includes data on 9.6 million healthcare providers across 242 specialties and 708,000 U.S. organizations	Provides email and physician mailing lists of healthcare providers at healthcare sites, specifically with reference to healthcare marketing. Also includes the affiliations that link together the professionals and organizations listed in database. Dates not available on the website and company did not respond when contacted multiple times.	Commercial ($)
xvii. Optum	200 Million de-identified patients	Insurance claims data, health risk assessments, and limited EHR data. Has data on administrative claims, patient and health plan costs, demographics, health behaviors, medical records, and self-reported health information. Dates: 2011-Present	Commercial ($$$$)
xviii. ProPublica Dollars4Docs	$12 billion in disclosed payments from 2,190+ companies to 1+ million doctors (medical doctors, dentists, osteopaths, optometrists, podiatrists, and chiropractors) and 1,249 teaching hospitals.	Provides data on "general" payments from pharmaceutical and medical device companies to doctors and teaching hospitals. Dates: 2009-2018	Commercial ($$) or Open Source (Free) depending on the exact data requested
xix. ProPublica Prescriber Checkup	Includes data on ~ 447,000 healthcare providers.	Includes data from Medicare's prescription drug benefit (Part D) for providers who wrote 50 or more prescriptions for at least one drug that year. Data organized by state, drug, claims, and cost. Dates: 2011-2017	Open Source (Free)
xx. State Health Information Exchanges (HIEs)	Varies by state	EHR data, depending on HIPAA agreements can be identified or de-identified. Dates: 2000s-Present	Commercial ($$) or Open Source (Free)
xx. Ready Signal	Demographic and COVID-19 data	Although not a healthcare database, researchers may find the data valuable as it provides demographic and COVID-19 data for a monthly fee	Exact fee varies depending on plan

[12] HIMSS Analytics does not include claims data

References

Abbasi, Ahmed, Jingjing Li, Donald Adjeroh, Marie Abate, and Wanhong Zheng. 2019. "Don't Mention It? Analyzing User-Generated Content Signals for Early Adverse Event Warnings." *Information Systems Research* 30 (3): 1007–28. https://doi.org/10.1287/isre.2019.0847.

Blei, David M. 2012. "Probabilistic Topic Models." *Communications of the ACM* 55 (4): 77–84.

Blei, David M, Andrew Y Ng, and Michael I Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3 (Jan): 993–1022.

Dai, Jim G., and Pengyi Shi. 2020. "Recent Modeling and Analytical Advances in Hospital Inpatient Flow Management." *Production & Operations Management* Forthcoming. https://doi.org/10.2139/ssrn.3310853.

Greiwe, Justin, and Sharmilee M Nyenhuis. 2020. "Wearable Technology and How This Can Be Implemented into Clinical Practice." *Current Allergy and Asthma Reports* 20: 1–10.

Herbec, Aleksandra, Jamie Brown, Lion Shahab, Robert West, and Tobias Raupach. 2019. "Pragmatic Randomised Trial of a Smartphone App (NRT2Quit) to Improve Effectiveness of Nicotine Replacement Therapy in a Quit Attempt by Improving Medication Adherence: Results of a Prematurely Terminated Study." *Trials* 20 (1): 1–12.

Lambert, Tara E, Lisa A Harvey, Christos Avdalis, Lydia W Chen, Sayanthinie Jeyalingam, Carin A Pratt, Holly J Tatum, Jocelyn L Bowden, and Barbara R Lucas. 2017. "An App with Remote Support Achieves Better Adherence to Home Exercise Programs than Paper Handouts in People with Musculoskeletal Conditions: A Randomised Trial." *Journal of Physiotherapy* 63 (3): 161–67.

Lipsey, Mark, and David Wilson. 2001. *Practical Meta-Analysis*. Thousand Oaks, CA: Sage Publications.

Liu, Xiang, Michael Hu, Jonathan E. Helm, Mariel S. Lavieri, and Ted A. Skolarus. 2018. "Missed Opportunities in Preventing Hospital Readmissions: Redesigning Post-Discharge Checkup Policies." *Production & Operations Management* 27 (12): 2226–50. https://doi.org/10.1111/poms.12858.

Mejia, Jorge, Shawn Mankad, and Anandasivam Gopal. 2019. "A for Effort? Using the Crowd to Identify Moral Hazard in New York City Restaurant Hygiene Inspections." *Information Systems Research* 30 (4): 1363–86. https://doi.org/10.1287/isre.2019.0866.

Quinn, Charlene Connolly, Erin C Butler, Krystal K Swasey, Michelle D Shardell, Michael D Terrin, Erik A Barr, and Ann L Gruber-Baldini. 2018. "Mobile Diabetes Intervention Study of Patient Engagement and Impact on Blood Glucose: Mixed Methods Analysis." *JMIR MHealth and UHealth* 6 (2): e31.

Ranjan, Chitta, Kamran Paynabar, Jonathan E. Helm, and Julian Pan. 2017. "The Impact of Estimation: A New Method for Clustering and Trajectory Estimation in Patient Flow Modeling." *Production & Operations Management* 26 (10): 1893–1914. https://doi.org/10.1111/poms.12722.

Shi, Pengyi, Jonathan E. Helm, H. Sebastian Heese, and Alice M. Mitchel. 2020. "An Operational Framework for the Adoption and Integration of New Diagnostic Tests." *Production & Operations Management* Forthcoming. https://doi.org/10.2139/ssrn.3430980.

Wallach, Hanna, David Mimno, and Andrew McCallum. 2009. "Rethinking LDA: Why Priors Matter." *Advances in Neural Information Processing Systems*, 1973–81.

White, Denise L., Elham Torabi, and Craig M. Froehle. 2017. "Ice-Breaker vs. Standalone: Comparing Alternative Workflow Modes of Mid-Level Care Providers." *Production & Operations Management* 26 (11): 2089–2106. https://doi.org/10.1111/poms.12743.

Zargoush, Manaf, Mehmet Gümüş, Vedat Verter, and Stella S. Daskalopoulou. 2018. "Designing Risk-Adjusted Therapy for Patients with Hypertension." *Production and Operations Management* 27 (12): 2291–2312. https://doi.org/10.1111/poms.12872.

Zepecki, Anne, Sylvia Guendelman, John DeNero, and Ndola Prata. 2020. "Using Application Programming Interfaces to Access Google Data for Health Research: Protocol for a Methodological Framework." *JMIR Research Protocols* 9 (7): e16543. https://doi.org/10.2196/16543.