HEALTH SYSTEMS
2020, VOL. 9, NO. 4, 317–325
https://doi.org/10.1080/20476965.2019.1581433

THE OPERATIONAL RESEARCH SOCIETY
Taylor & Francis
Taylor & Francis Group

ORIGINAL ARTICLE

Key considerations when using health insurance claims data in advanced data analyses: an experience report

Renata Konrad[a], Wenchang Zhang[b], Margrét Bjarndóttir[b] and Ruben Proaño[c]

[a]Fosie School of Business, Worcester Polytechnic Institute, Worcester, MA, USA; [b]Robert H. Smith School of Business, University of Maryland College Park, College Park, MD, USA; [c]Industrial Systems Engineering, Rochester Institute of Technology, USA

ABSTRACT

Health claims have become a popular source of data for healthcare analytics, with numerous applications ranging from disease burden estimation and policy evaluation to drug event detection and advanced predictive analytics. Independent of the application, a researcher utilising claims information will likely encounter challenges in using the data, which include dealing with several coding systems and coding irregularities. We highlight some of these challenges and approaches for successful analysis that may reduce implementation time and help in avoiding common pitfalls. We describe the experiences of a group of academic researchers in using an extensive seven-year repository of US medical and pharmaceutical claims data in a research study, and provide an overview of the challenges encountered with handling claims records for data analysis while sharing suggestions on how to address these challenges. To illustrate our experiences, we use the example of defining episodes of care for a bundled payment reimbursement system in the US context.

ARTICLE HISTORY
Received 18 October 2017
Revised 6 February 2019
Accepted 6 February 2019

KEYWORDS
Claims data; bundled payment system; analytics

1. Introduction

Claims data are extensively used for healthcare studies. Claims data comprise information entered on bills (claims) submitted by healthcare providers to third-party payers. Such claims commonly include information about medical diagnosis, procedures and treatments performed, as well as prescription information (Birman-Deych et al., 2005; Yan, Birman–Deych, Radford, Nilasena, & Gage, 2005). The data also provide details on several financial metrics, such as costs, charges, and reimbursement amounts.

Claims data appeal to researchers, as they are structured, plentiful, and inexpensive; moreover, they are widely available in electronic format and can be anonymised (Hicks, 2003). Furthermore, these data are free from nonresponse and dropout (Wolf, Harvell, & Jha, 2012). Researchers have reported that claims data exhibit high congruence with medical records data (Fowles, Fowler, & Craft, 1998).

Over the past four decades the analytical approaches applied to claims data have evolved, from simple counting to sophisticated machine learning algorithms (Bjarndóttir, Czerwinski, & Guan, 2016). In the late 1970s and early 1980s claims-based research began to emerge (see e.g. Roos, Nicol, Johnson, and Roos 1979), models using diagnosis from claims data to predict future healthcare costs were introduced in the late 1980s (Ash, Porell, Gruenberg, Sawitz, & Beiser, 1989; Newhouse, Manning, Keeler, & Sloss, 1989) and the health status of a population (Mossey & Roos, 1987). In the late 1980s and 1990s, researchers used claims data as a data source to examine provider services and resource utilisation (De Coster et al., 2006; Lewis, Patwell, & Briesacher, 1993; Mossey & Roos, 1987; Wennberg, Roos, Sola, Schori, & Jaffe, 1987). Since the early 2000s, claims data have demonstrated to be a valuable data source for numerous health systems, and medical studies (Ferver, Burton, & Jesilow, 2009). Bjarndóttir et al. (2016) provide an overview of the history of claims data in healthcare research.

Claims-based studies can be much more inclusive than medical records, and they are significantly less expensive. As claims data contain rich cost information concerning medical services, researchers have used such data widely for studies of healthcare utilisation (Ypinga et al., 2018) and medical expenditures (Tyree, Lind, & Lafferty, 2006), and more recently, in applications from quality measurement to drug surveillance and forecasting (Bjarndóttir et al., 2016). Nation-wide collections of claims data have enabled researchers to explore the health state of large populations with unprecedented precision and extent, such as the outlook of age-related macular degeneration in Japan (Kume et al., 2016), the measurement of age- and gender-dependent relative risks for diabetes in Austria (Klimek, Kautzky-Willer, Chmiel, Schiller-Frühwirth, & Thurner, 2015) and a comprehensive study addressing multi-morbidity of the elderly population in Germany (van den Bussche et al., 2011).

Reliance on claims data for research, policy-making and decision-making continues to grow. For example, in the United States (US), the Affordable Care Act

CONTACT Renata Konrad rkonrad@wpi.edu Fosie School of Business, Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA 01609

(Public Law 111–148) mandated that after 2015 healthcare claims data be used extensively to assess resource use and quality of care. Confirming the growing relevance of such research globally, the German government increasingly funds claim data research (Kreis, Neubauer, Klora, Lange, & Zeidler, 2016).

Claims data are an attractive source to researchers, but since the data are intended for administrative purposes, their use for healthcare research requires substantial effort. Despite the increasing use of claims-based research, several studies have identified key challenges associated with using claims data: poor documentation and inaccurate coding (De Coster et al., 2006; Utter et al., 2010), nonidenpendence of the physician query process (Crews, Pronovost, Helft, & Austin, 2017), and absences of external processes to audit compliance or validate data accuracy (Crews et al., 2017). Rumsfeld, Joynt, and Maddox (2016) outline methodological issues such as validation, data inconsistency, data quality, and limitations in the observed data as it pertains to employing big data applications in cardiovascular care; however the discussion does not focus a specific type of data (e.g. claims) or offer solutions to the challenges identified. Gavrielov-Yusim and Friger (2014) discuss the biases and limitations of using administrative medical databases for epidemiological and biostatistical research.

The widespread use of claims data in research has led to a call for guidelines and best practices. Benchimol et al. (2011) developed guidelines for studies validating administrative data identification algorithms and went on to assess the quality of validation studies in the literature. Others have developed systematic reviews of validation methods for identifying patients for a specific diagnoses (Chung, Rohan, Krishnaswami, & McPheeters, 2013; Moores & Sathe, 2013) or health outcomes (McPheeters, Sathe, Jerome, & Carnahan, 2013) using claims data. Stein, Lum, Lee, Rich, and Coleman (2014) provide a comprehensive checklist for authors to use in reporting analysis involving claims. While there is a focus on developing best practices and guideline for validation of methods using claims data, little discussion exists on the process of generating analytical results from claims data. To this end, this paper aims to support practitioners, scholars, and decision makers in understanding the benefits, risks, and gaps concerning how to develop successful analytical projects using claims data. The goals of this article include the following:

- highlighting challenges in conducting healthcare-based big data research using claims data;
- sharing some lessons and guidelines on how to address these challenges; and
- suggesting opportunities for increasing the amount and impact of data-oriented research.

This experience report also focuses on several practical challenges for generating analytics results from data that are already in compliance with the Health Insurance Portability and Accountability Act (HIPAA), which is the legislation in the US that enforces data privacy and security protections for safeguarding medical information and patient identity. These challenges include grouping procedure, drug and diagnosis codes, and understanding and addressing coding variation. The report is based on the experiences of academic researchers in an ongoing research project using big-data analytics methods to analyse an extensive seven-year repository of medical claims data from two large insurance companies. It is set in the context of creating a mechanism that systematically defines and prices episodes of care for a community, using data-driven techniques. In what follows, we first summarise our research context and the data used. We then discuss the challenges a researcher utilising health claims data may encounter and offer practical solutions.

2. Research context: bundled payment

2.1. Identifying episodes-of-care

To provide context to our discussion, we use an example of how claims data research can provide recommendations for reimbursement reform in the US. Bundled payment systems represent a promising change in reimbursement systems to address the spiralling cost of healthcare in the US (Hussey, Eibner, Ridgely, & McGlynn, 2009). Bundled payments offer a single payment for an episode-of-care (the expected the set of services needed to treat a condition) rather than fee-for-service reimbursement, where every service is individually claimed and reimbursed and hold a great deal of promise (Hussey et al., 2009). Evidence from bundled-payment demonstration projects report reductions in hospital costs and improved quality (Campbell, Reeves, Kontopantelis, Sibbald, & Roland, 2009), as well as a reduction in unnecessary medical procedures (United States, 2010). Determining a single payment for an episode-of-care would be a trivial task if the set of treatments required for a given diagnosis were always the same. However, patient heterogeneity and variations in treatment practice lead to different treatment protocols for the same condition. As such, choosing which services should be included in an episode-of-care poses a significant challenge. Conventional processes for defining an episode-of-care rely on the consensus of expert panels, resulting in a collaborative, labour-intensive approach that fails to account for the abovementioned variation. Identifying episode-of-care patterns from insurance claims data would help to enhance and accelerate the process of defining an episode-of-care.

Our research team conducted a study on how to use claims data to carry out episode-of-care characterisation and cost inference. We investigated the potential of insurance claims data using data-mining approaches while considering their heterogeneous nature (Zhang, Bjarnadóttir, Proaño, Anderson, & Konrad, 2018). We proposed a data-driven clustering approach to automatically detect and explicitly represent homogeneous subgroups of services for a given condition. The automatically extracted clusters of services with different cost variations highlight the payer's expenditure and provider's financial risk under bundled payments. Using data analytics tools, we extrapolated meaningful insights about comorbidities, treatment qualities, and disease progressions within clusters.

2.2. Data source

The two largest insurance companies in the Rochester area of New York State, US, gave the research team access to a repository of fully de-identified HIPAA complaint data containing several years of historical claims records concerning insured customers from the Rochester area. This repository is administered by the Finger Lakes Health Systems Agency, and comprises data from commercial accounts, Medicare Advantage and Medicaid Managed Care accounts, and account data for which the two insurers serve as third-party administrators. The repository contained more than 300 million claims records related to outpatient, inpatient, and pharmacy services.

3. Challenges in using claims data

During our research project, we encountered several challenges in using the data. Below, we discuss some challenges that a researcher utilising claims data may face and describe how we addressed each issue. While the challenges we describe are general, the data was obtained from US health organisations. In Section Impact and Opportunities we discuss the structures of, and access to, other countries' claims data.

3.1. Coding practices

Claims exist to facilitate an adequate transfer of funds between payers and providers; as such, there are differences between information contained in claims and information one may expect to find in medical records. Claims data depend on professional transcription services and ICD[1] coding for billing purposes. Therefore, the details of diagnostic information, for example, may vary from one provider to the next. Studies have shown that some diagnoses may be missed, different professional groups (e.g. physicians and nurse practitioners) may have different coding patterns, and not all coding may be accurate (Tyree et al., 2006). We focused on data from one provider to control for the heterogeneity in coding behaviours. Further, coders may enter incorrect information due to faulty decisions about what to code, misreading of the medical record, or typographical errors (Romano, 2000; Schneeweiss & Avorn, 2005; Tyree et al., 2006).

Claim submission practices change frequently, and this affects how claims are registered over time. The reason for such changes may be due to gradual modifications in medical knowledge, changes in medical practice, or in some cases, upcoding—the practice of misrepresenting conditions (possibly unintentionally) to maximise reimbursements or minimise penalties (Bastani, Goh, & Bayati, 2018; Rosenberg, Fryback, & Katz, 2000). Thus, coding changes responding to alterations in reimbursement policies or treatment practices represent an additional factor that adds to coding variability, complicating efforts to analyse patients whose history of services elapses over multiple periods. For example, in our data, we observed that the number of diagnosis codes related to congestive heart failure (CHF), specifically ICD-9 code 428.0, decreased significantly from 2007 to 2013; moreover, we noted that other codes indicating similar conditions increased in the same period. Assuming the patient base remained stationary, these variations may reflect changes in coding practices over time. In working with our data, we found that such changes were not retroactively adjusted in the databases when ICD-10 codes replaced ICD-9. To prevent these changes from affecting the resulting clusters, we limited our analysis to evaluating a single year of data. It is important for researchers to be aware of temporal patterns in their data and their effect on the definitions of variables.

In the US, how providers relate with its physicians can also result in claim variability. For example, we can consider specialty physicians, such as radiologists. A procedure offered by a radiologist will be coded and claimed differently, if the radiologist was directly employed by the hospital caring for a patient versus a radiologist who is not part of the hospital and offers the procedure as a contractor. Table 1 shows an example of two patients undergoing the same cardiac operation covered through Medicare; the difference in their claims only reflects the employment status of the radiologist.

There is no unique solution that will resolve these coding variabilities. Reducing the sample of claims to those occurring over a shorter period, and if possible, limiting the providers to similar types, will help in focusing the analysis on the effect of interest. Moreover, a commonly used strategy to reduce the effect of coding noise is to group individual diagnoses, as well as procedure and drug codes, into coding groups, as discussed in detail below.

Table 1. Selected columns from two electrocardiogram claims, differing only the radiologists' employment status.

Patient One		Patient Two	
Code Type	Description	Code Type	Description
Revenue	(Electrocardiogram)—General	Revenue	EKG/ECG, (Electrocardiogram)–General
		CPT/HCPCS	Cardiac Output Monitoring by Electrical Bioimpedance

3.2. Group coding and overlap

Grouping of similar codes for procedures, diagnoses, and drugs is necessary to reduce coding variability and sparsity in the data (Zhou, Wang, Hu, & Ye, 2014). For example, in our study, we group the Current Procedural Terminology (CPT) codes using Clinical Classification Software for Services and Procedures (CCSS). In the US, CPT codes are maintained by the American Medical Association and are a standard for documentation and reporting of medical, surgical, and diagnostic services. Table 2 provides an example of a coding group.

We also grouped hospital revenue codes[2] based on the first two digits of each code to specify their category.[3] Table 3 provides an example of all private room and board revenue codes that are included in a single group based on this criterion. In general, researchers utilise available groupers, for example, the CCSS for ICD-10 data, which is a part of the Healthcare Cost and Utilization Project (Bertsimas et al., 2008).

Another challenge is the overlap in the use of revenue and CPT codes. For example, electrocardiograms for example are commonly used in the treatment of CHF. There are both revenue codes and CPT codes to indicate that a patient received this service (i.e. revenue codes 730, 731, 732, and 739 and CPT codes 93000, 93005, and 93010). Depending on a provider's use of coding systems, for some patients, we observed both revenue and CPT codes for electrocardiograms. In other cases, only a CPT or revenue code was used for CHF patients' electrocardiograms. Therefore, unless the intricacies of why a provider may prefer to report a procedure using a revenue or CPT code are clear, the codes from the different coding systems that reflect the same situation should be combined. One may not necessarily combine the codes if, for example, the place of service (hospital vs. physician's office) is important. Notably, the charge for a procedure can be drastically different depending on where it is performed, even across different units of the same provider.

Table 3. Room and board revenue codes grouped into a single private room and board code ('11').

Code	Code Description
110	Room and board, private, general
111	Room and board, private, medical/surgical/gyn
112	Room and board, private, obstetrics
113	Room and board, private, paediatric
114	Room and board, private, psychiatric
115	Room and board, private, hospice
116	Room and board, private, detoxification
117	Room and board, private, oncology
118	Room and board, private, rehabilitation
119	Room and board, private, other

In the US healthcare system, pharmaceuticals are most commonly reported using the NDC system (U.S. Food & Drug Administration, n.d.). A typical NDC code contains 10 or 11 digits with three sections, including a labeller code (manufacturer, distributor), product code (strength, dosage form, formulation), and package code (package sizes, types). A user requires additional pharmacy expertise or access to commercial databases to pre-process the NDC codes and connect them to, for instance, pharmaceutical classes. In our dataset, less than 50% of patients had NDC details recorded during their inpatient stay, but limited drug information is available through the revenue codes. The NDC codes in our data contain 11 digits, requiring a conversion to map

Table 2. Current procedural terminology (CPT) codes included in a single group for knee arthroplasty.

CPT Code	Description
27420	Reconstruction of dislocating patella (e.g. Hauser-type procedure)
27422	Reconstruction of dislocating patella; with extensor realignment and/or muscle advancement or release
27424	Reconstruction of dislocating patella; with patellectomy
27427	Ligamentous reconstruction (augmentation), knee; extra-articular
27428	Ligamentous reconstruction (augmentation), knee; intra-articular (open)
27429	Ligamentous reconstruction (augmentation), knee; intra-articular (open) and extra-articular
27437	Arthroplasty, patella; without prosthesis
27438	Arthroplasty, patella; with prosthesis
27440	Arthroplasty, knee, tibial plateau
27441	Arthroplasty, knee, tibial plateau; with debridement and partial synovectomy
27442	Arthroplasty, femoral condyles or tibial plateau (S), knee
27443	Arthroplasty, femoral condyles or tibial plateau (S), knee; with debridement and partial synovectomy
27445	Arthroplasty, knee, hinge prosthesis (e.g. Walldius type)
27446	Arthroplasty, knee, condyle and plateau; medial or lateral compartment
27447	Arthroplasty, knee, condyle and plateau; medial and lateral compartments with or without patella resurfacing (Total knee arthroplasty)
27486	Revision of total knee arthroplasty, with or without Allograft; 1 component
27487	Revision of total knee arthroplasty, with or without Allograft; femoral and entire tibial component
G0428	Collagen meniscus implant procedure for filling meniscal defects (e.g. CMI, collagen scaffold, Menaflex)

them to a more common 10-digit ontology. Analyses regarding major treatment decisions related to specific drugs and their dosages (e.g. treatment of urinary tract infections) pose a challenge because of this encoding practice.

3.3. Missing information

As claims data are collected for reimbursement purposes, the medical details of a specific claim may be limited; for example, if a patient received an X-ray, but the results are not included in the data (partial), inference from subsequent data may be required. If outcomes are important to the analysis, the inferred results can be included. For instance, in this example, subsequent claims data may show that the patient received a cast following the X-ray, indicating a fracture. Some claims repositories have recently started to include laboratory results.

The design of the specific billing process from which our data originated resulted in missing information for our study. For example, the services provided during palliative care are grouped and coded as professional services. Further, in our data, we noted that among 14\% of surgical inpatient stays for knee replacement, only one revenue code ("room board") is evident; codes for other possible following procedures related to the visit (e.g. "anaesthesia," "surgical supplies devices") are missing. Depending on the provider, the missing information may be due to input errors, different codes bundled into others for convenience, or issues in the claims aggregation process.

Other researchers have reported similar problems. Paediatric immunisation and prenatal care visits are illustrative of such issues, as these services are often bundled with regular office visits for billing purposes, causing them to be vastly underreported in claims data (Dresser, Feingold, Rosenkranz, & Coltlin, 1997). Minor tests or routine hospital procedures associated with specific diagnoses are less likely to be recorded on a claims form because they are unlikely to qualify for additional reimbursement if billed separately (Dismuke, 2005). Similarly, non-operating room activities are often neglected in claims data because guidelines only require records of procedures that are surgical in nature, necessitate specialised training, or carry a procedural or anaesthetic risk (Quan, Parsons, & Ghali, 2004). To reiterate, understanding the specifics of the claims dataset being used, as well as its limitations, will help to guide any modelling decisions and set the appropriate scope for the study.

3.4. Claim costs

Providers have tailored procedures in place to interpret the appropriate way to process claim requirements and submit their charges to payers. In addition, providers negotiate different costs for similar services. Such costs are often confidential, which is the case for our data. Thus, costs for specific procedures are only available at an aggregated level (i.e. the average cost per service across providers). Researchers who may want to explore the effect of the cost structures that different providers employ would have difficulty using datasets with aggregated costs. In the case of our study, the inability to incorporate actual cost differences prevented us from investigating whether cost variation among providers can be an indicative factor for use in the characterisation of episodes of care.

3.5. Patient comorbidity

Comorbidities are a patient's additional diseases beyond the condition under study (van den Akker, Buntinx, Metsemakers, Roos, & Knottnerus, 1998); these can be acute illnesses or chronic diseases, and they usually increase a patient's total illness burden (Shwartz, Iezzoni, Moskowitz, Ash, & Sawitz, 1996). Numerous studies have documented the significant effect of comorbid conditions on treatment selection and outcomes (Klabunde, Warren, & Legler, 2002). Hospitals usually report up to 15 secondary diagnoses in each claim to payers. However, in our data, we only had access to four diagnoses per claim record. The truncation of diagnoses in a claims repository is a crucial issue, as it may reduce the importance of secondary diagnoses in any analysis. Researchers interested in using primary and secondary diagnoses for potential analysis must understand the effect of a potential diagnosis truncation and whether the claims aggregator had a specific motivation or prioritisation policy to reduce the pool of diagnoses per encounter. In our study, using the first four diagnoses, we could determine whether a comorbidity existed, but as is common in claims data studies, we were not able to establish the severity of such conditions, which could potentially affect the use of services. Clearly, the truncation of the number of diagnoses dilutes the potentially rich source of information that comorbidities can provide.

3.6. What constitutes an episode-of-care?

A main challenge in the characterisation of episodes-of-care for a condition is defining its beginning and end. Mehta, Suzuki, Glick, and Schulman (1999) used differences in average daily charges to determine the duration of diabetic foot ulcer episodes. For certain conditions, such as pregnancy, the start and end dates of an episodes are well defined (Hornbrook et al., 2007). In other cases, anchor records that initiate a series of treatment are used to identify the start time of an episode (Dunn, Liebman, Rittmueller, & Shapiro, 2014). Even for relatively simple episodes, such as total knee replacement (TKR), correctly including all treatments is

not a trivial issue. Preliminary imaging and laboratory work are typically performed prior to a knee replacement procedure, but such services may not be included as part of the procedure. In addition, a primary diagnosis code typically defines a medical encounter, but it does not reflect the evolution of the condition. For example, in the case of TKR, the services prior to the actual knee replacement surgery may include trauma, emergency room services, outpatient visits, laboratory workups, and physical therapy, and they may not be associated with the actual encounter code related to the knee replacement. Furthermore, after the actual TKR takes place, the patient must undergo follow up and rehabilitation procedures. How to capture those services in an automatic way is still an open challenge.

4. Overcoming challenges in using claims data

We encountered the challenges described above as we tried to pursue our primary research goals. Here, we provide some key recommendations for anyone seeking to use large claims repositories. Although our focus was on studying bundles of services that a patient receives during an episode-of-care, the lessons learned have broad applications. Specifically, we recommend the following:

(1) Begin an analysis by generating the histories of a small sample of patients. This entails translating different codes into "plain language" to understand the level of detail the data provide and what a typical patient looks like in the data. Such a preliminary data scope will help to guide more detailed further analysis.
(2) Work with your data partner! Your data partner knows the process that generates the data and understands the intricacies of the claims. Working together with individuals and organisations that have a record of the changes regarding the processes of registering and submitting claims is extremely valuable for understanding the data.
(3) Take the time to analyse the temporal and provider effects of key events or codes in the analysis. This will minimise the risk of errors or missing results induced by changes in input practices. In some cases, it may be appropriate to group the data by year and provider.
(4) Aggregate claims data by "families" of service codes to mitigate the risk associated with input errors and high variation in the recorded services, as well as to overcome sparsity and fitting issues. Specifically, we recommend relying on CCSS and families of revenue codes for procedure codes.
(5) Seek full, untruncated data to prevent complications due to the truncation of secondary diagnoses; consider studying patient cohorts that may have fewer effects of comorbidities and whose claims tend to have lower number of secondary diagnoses; or minimally, be aware that the picture may not be complete.

A gold standard for using claims data in research does not exist. The methodical approach utilised will always depend on the indication and the research question at hand. Thus, it is important to for researchers working with claims data to detail methods used. To improve the usefulness, potential reproducibility, and overall study quality we encourage researchers to consider the items in the checklist for reporting analysis involving claims data by Stein et al. (2014). We also suggest readers turn to Motheral et al. (2003) who provide a guide for decision-makers for selecting study methodologies in claims data research.

5. Impact and opportunities

Good structure and access to research is of high importance to meet the World Health Organization's (WHO) goals of high-performing healthcare systems. Claims data research can provide important recommendations to improve healthcare systems. Claims data are increasingly and effectively used to conduct epidemiological studies, international comparisons and evaluate healthcare utilisation by providing relatively rapid access to collected information (Tuppin et al., 2017). Findings from such studies are considered by health policymakers, third-party payers, and other decision makers. In 2013, over 380 unique healthcare claims databases have been used by researchers varying from those focused on large nationwide populations, specific communities or patients with specific diseases (Stein et al., 2014).

While aggregated claims repositories have proven to be important sources of information for research in healthcare delivery and healthcare economics, multiple sources of variation influence these data significantly, and this can affect the consistency of the data and its quality for research. Unfortunately, each data source is unique; thus, each research team needs to make a significant effort to understand the intricacies of the data; moreover, it is time consuming for researchers to replicate the results of one study using different datasets, and consequently, they do not commonly do so. Unless there is an effort to standardise the fusion of aggregated claims data for research purposes, the specific claims data used limit the research findings, as do the assumptions made to deal with such limitations. Standardised processes would improve the quality of the research and the effectiveness of research teams. For the US case, we call to insurance companies and Medicare to work together in setting up a large repository of de-identified claims that is

safe for and designed to support research; currently in the US, third parties are responsible for such claim repositories, and thus, they are proprietary. Moreover, a uniquely designed repository should take into consideration how to ensure that the data reflect changes in practices and coding retrospectively.

In comparison to our experience in the US, some countries are much more progressive in making health insurance administrative databases available to researchers. A good example is the French *Système National d'Information Inter-Régimes* de *l'Assurance Maladie* (SNIIRAM) [National Health Insurance Information System] which fuses inpatient and outpatient medical reimbursements, sociodemographic and medical characteristics of 66 million people. Data can be analysed by accredited agencies and access to data is controlled by different permission levels depending on the type of data requested (Tuppin et al., 2017). Milea, Azmi, Reginald, Verpillat, and Francois (2015) provide a comprehensive study of 54 health administrative databases, including claims databases, in several Asia-Pacific studies. The authors characterise databases based on their accessibly to researchers to support work. Australia, Taiwan and Japan in particular were categorised as offering high levels of accessibility for researchers to their claims information.

6. Concluding thoughts

Despite the issues of working with claims data discussed in this paper, we anticipate that their use by researchers and policymakers will only increase and will put pressure on healthcare systems to reconsider the original billing purpose of the medical claims records. The development and adoption of more functional artificial intelligent systems and their use on healthcare data can help identify patients' and physicians' characteristics associated with high healthcare costs, redefining the notion of preventive care, as a data-driven system that makes sense of patterns in the patient's information to determine when patients should initiate care or if they would benefit from having a particular medical assessment. Due to their richness, claims data will also have a more active role in supporting data-driven efforts to identify inefficiencies and irregularities in the provision of care. Under this scenario, the urge to understand all the complications of working with claims data can provide the baseline to facilitate the redesign of this resource by considering how the data is used beyond billing purposes.

Notes

1. International Statistical Classification of Diseases and related Health Problems (ICD), a classification list maintained by the World Health Organization.
2. Three or four digit code used by providers in the USA for billing purposes, to describe to the payers the type of service provided to a patient, where it was provided, and an associated dollar amount.
3. See http://valuehealthcareservices.com/education/understanding-hospital-revenue-codes/.

Acknowledgments

The de-identified claims data were retrieved from the Finger Lakes Health System Agency (FLHSA); the authors would like to thank FLHSA for their support and feedback. Further, the authors acknowledge the financial support of the Center for Health Information and Decision Systems (CHIDS).

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This study was partially supported with University of Maryland's internal funds channeled through its Center for Health Information and Decision Systems

References

Ash, A., Porell, F., Gruenberg, L., Sawitz, E., & Beiser, A. (1989). Adjusting medicare capitation payments using prior hospitalization data. *Health Care Financing Review*, *10*(4), 17–29.

Bastani, H., Goh, J., & Bayati, M. (2018). Evidence of upcoding in pay-for-performance programs evidence of upcoding in pay-for-performance programs keywords: Medicare pay-for-performance upcoding asymmetric information quality control detection. *Management Science*, *Articles I*, 1–19.

Benchimol, E. I., Manuel, D. G., To, T., Griffiths, A. M., Rabeneck, L., & Guttmann, A. (2011). Development and use of reporting guidelines for assessing the quality of validation studies of health administrative data. *Journal of Clinical Epidemiology*, *64*(8), 821–829.

Bertsimas, D., Bjarnadóttir, M. V., Kane, M. A., Kryder, J. C., Pandey, R., Vempala, S., & Wang, G. (2008). Algorithmic prediction of health-care costs. *Operations Research*, *56*(6), 1382–1392.

Birman-Deych, E., Waterman, A., Yan, Y., Nilasena, D. S., Radford, M., & Gage, B. F. (2005). Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, *45*(5), 480–485.

Bjarndóttir, M., Czerwinski, D., & Guan, Y. (2016). The history and modern applications of insurance claims data in healthcare research. In H. Yang & E. K. Lee (Eds.), *Healthcare analytics: From data to knowledge to healthcare improvement* (pp. 541–560). Hoboken, New Jersey: John Wiley & Sons, Inc.

Campbell, S. M., Reeves, D., Kontopantelis, E., Sibbald, B., & Roland, M. (2009). Effects of pay for performance on the quality of primary care in England. *New England Journal of Medicine*, *361*(4), 368–378.

Chung, C. P., Rohan, P., Krishnaswami, S., & McPheeters, M. L. (2013). A systematic review of validated methods for identifying patients with rheumatoid arthritis using administrative or claims data. *Vaccine*, *31*, K41–K61.

Crews, H., Pronovost, P. J., Helft, P. R., & Austin, J. M. (2017). Improving the quality of data for inpatient claims-based measures used in public reporting and pay-for-performance programs. *Joint Commission Journal on Quality and Patient Safety*, *43*(12), 671–675.

De Coster, C., Quan, H., Finlayson, A., Gao, M., Halfon, P., Humphries, K. H., … Ghali, W. A. (2006). Identifying priorities in methodological research using ICD-9-CM and ICD-10 administrative data: Report from an international consortium. *BMC Health Services Research*, *6*(77), 1–6.

Dismuke, C. E. (2005). Underreporting of computed tomography and magnetic resonance imaging procedures in inpatient claims data. *Medical Care*, *43*(7), 713–717.

Dresser, M. V. B., Feingold, L., Rosenkranz, S., & Coltlin, K. (1997). Clinical quality measurement: Comparing chart review and automated methodologies. *Medical Care*, *35* (6), 539–552.

Dunn, A., Liebman, E., Rittmueller, L., & Shapiro, A. (2014). *Defining disease episodes and the effects on the components of expenditure growth*. Retrieved from http://www.bea.gov/system/files/papers/WP2014-4.pdf

Ferver, K., Burton, B., & Jesilow, P. (2009). The use of claims data in healthcare research. *The Open Public Health Journal*, *2*(1), 11–24.

Fowles, J. B., Fowler, E. J., & Craft, C. (1998). Validation of claims diagnoses and self-reported conditions compared with medical records for selected chronic diseases. *The Journal of Ambulatory Care Management*, *21*(1), 24–34.

Gavrielov-Yusim, N., & Friger, M. (2014). Use of administrative medical databases in population-based research. *Journal of Epidemiology and Community Health*, *68*(3), 283–287.

Hicks, J. (2003). *The potential of claims data to support the measurement of health care quality*. Santa Monica, CA: RAND Corporation. Retrieved from https://www.rand.org/pubs/rgs_dissertations/RGSD171.html

Hornbrook, M. C., Whitlock, E. P., Berg, C. J., Callaghan, W. M., Bachman, D. J., Gold, R., … Williams, S. B. (2007). Development of an algorithm to identify pregnancy episodes in an integrated health care delivery system. *Health Services Research*, *42*(2), 908–927.

Hussey, P. S., Eibner, C., Ridgely, M. S., & McGlynn, E. A. (2009). Controlling U.S. health care spending — separating promising from unpromising approaches. *New England Journal of Medicine*, *361*(22), 2109–2111.

Klabunde, C. N., Warren, J., & Legler, J. (2002). Assessing comorbidity using claims data: An overview on JSTOR. *Medical Care*, *40*(8), IV26–IV35.

Klimek, P., Kautzky-Willer, A., Chmiel, A., Schiller-Frühwirth, I., & Thurner, S. (2015). Quantification of diabetes comorbidity risks across life using nation-wide big claims data. *PLOS Computational Biology*, *11*(4), e1004125.

Kreis, K., Neubauer, S., Klora, M., Lange, A., & Zeidler, J. (2016). Status and perspectives of claims data analyses in Germany—A systematic review. *Health Policy*, *120*(2), 213–226.

Kume, A., Ohshiro, T., Sakurada, Y., Kikushima, W., Yoneyama, S., & Kashiwagi, K. (2016). Treatment patterns and health care costs for age-related macular degeneration in Japan: An analysis of national insurance claims data. *Ophthalmology*, *123*(6), 1263–1268.

Lewis, N. J. W., Patwell, J. T., & Briesacher, B. A. (1993). The role of insurance claims databases in drug therapy outcomes research. *PharmacoEconomics*, *4*(5), 323–330.

McPheeters, M. L., Sathe, N. A., Jerome, R. N., & Carnahan, R. M. (2013). Methods for systematic reviews of administrative database studies capturing health outcomes of interest. *Vaccine*, *31*, K2–K6.

Mehta, S. S., Suzuki, S., Glick, H. A., & Schulman, K. A. (1999). Determining an episode of care using claims data. Diabetic foot ulcer. *Diabetes Care*, *22*(7), 1110–1115.

Milea, D., Azmi, S., Reginald, P., Verpillat, P., & Francois, C. (2015). A review of accessibility of administrative healthcare databases in the Asia-Pacific region. *Journal of Market Access & Health Policy*, *3*(1), 1–11.

Moores, K. G., & Sathe, N. A. (2013). A systematic review of validated methods for identifying systemic lupus erythematosus (SLE) using administrative or claims data. *Vaccine*, *31*, K62–K73.

Mossey, J. M., & Roos, L. L. (1987). Using insurance claims to measure health status: The illness scale. *Journal of Chronic Diseases*, *40*, 41S–50S.

Motheral, B., Brooks, J., Clark, M. A., Crown, W. H., Davey, P., Hutchins, D., … Stang, P. (2003). A checklist for retrospective database studies—report of the ISPOR task force on retrospective databases. *Value in Health*, *6*(2), 90–97.

Newhouse, J. P., Manning, W. G., Keeler, E. B., & Sloss, E. M. (1989). Adjusting capitation rates using objective health measures and prior utilization. *Health Care Financing Review*, *10*(3), 41–54.

Quan, H., Parsons, G. A., & Ghali, W. A. (2004). Validity of procedure codes in international classification of diseases, 9th revision, clinical modification administrative data on JSTOR. *Medical Care*, *42*(8), 801–809.

Romano, P. S. (2000). Using administrative data to identify associations between implanted medical devices and chronic diseases. *Annals of Epidemiology*, *10*(4), 197–199.

Roos, L. L., Nicol, J. P., Johnson, C. F., & Roos, N. P. (1979). Using administrative data banks for research and evaluation. *Evaluation Quarterly*, *3*(2), 236–255.

Rosenberg, M. A., Fryback, D. G., & Katz, D. A. (2000). A statistical model to detect DRG upcoding. *Health Services and Outcomes Research Methodology*, *1*(3/4), 233–252.

Rumsfeld, J. S., Joynt, K. E., & Maddox, T. M. (2016). Big data analytics to improve cardiovascular care: Promise and challenges. *Nature Reviews Cardiology*, *13*(6), 350–359.

Schneeweiss, S., & Avorn, J. (2005). A review of uses of health care utilization databases for epidemiologic research on therapeutics. *Journal of Clinical Epidemiology*, *58*(4), 323–337.

Shwartz, M., Iezzoni, L., Moskowitz, M., Ash, A., & Sawitz, E. (1996). The importance of comorbidities in explaining differences in patient costs. *Medical Care*, *34* (8), 767–782.

Stein, J. D., Lum, F., Lee, P. P., Rich, W. L., & Coleman, A. L. (2014). Use of health care claims data to study patients with ophthalmologic conditions. *Ophthalmology*, *121*(5), 1134–1141.

Tuppin, P., Rudant, J., Constantinou, P., Gastaldi-Ménager, C., Rachas, A., de Roquefeuil, L., … Fagot-Campagna, A. (2017). Value of a national administrative database to guide public decisions: From the système national d'information interrégimes de l'Assurance Maladie (SNIIRAM) to the système national des données de santé (SNDS) in France. *Revue d'Épidémiologie et de Santé Publique*, *65*, S149–S167.

Tyree, P. T., Lind, B. K., & Lafferty, W. E. (2006). Challenges of using medical insurance claims data for utilization analysis. *American Journal of Medical Quality*, *21*(4), 269–275.

U.S. Food & Drug Administration. (n.d.). *National drug code directory*. Retrieved from https://www.accessdata.fda.gov/scripts/cder/ndc/

United States, C. Compilation Of Patient Protection and Affordable Care Act: as Amended through November 1, 2010 Including Patient Protection and Affordable Care Act Health-Related Portions of the Health Care and Education Reconciliation Act of 2010, Pub. L. No. Public Law 111–148. (2010). Retrieved from https://www.congress.gov/111/plaws/publ148/PLAW-111publ148.pdf

Utter, G. H., Cuny, J., Sama, P., Silver, M. R., Zrelak, P. A., Baron, R., … Romano, P. S. (2010). Detection of postoperative respiratory failure: How predictive is the agency for healthcare research and quality's patient safety indicator? *Journal of the American College of Surgeons*, *211*(3), 347–354.e29.

van Den Akker, M., Buntinx, F., Metsemakers, J. F. M., Roos, S., & Knottnerus, J. A. (1998). Multimorbidity in general practice: Prevalence, incidence, and determinants of co-occurring chronic and recurrent diseases. *Journal of Clinical Epidemiology*, *51*(5), 367–375.

van Den Bussche, H., Koller, D., Kolonko, T., Hansen, H., Wegscheider, K., Glaeske, G., … Schön, G. (2011). Which chronic diseases and disease combinations are specific to multimorbidity in the elderly? Results of a claims data based cross-sectional study in Germany. *BMC Public Health*, *11*(1), 101–110.

Wennberg, J. E., Roos, N., Sola, L., Schori, A., & Jaffe, R. (1987). Use of claims data systems to evaluate health care outcomes. *JAMA*, *257*(7), 933–936.

Wolf, L., Harvell, J., & Jha, A. K. (2012). Hospitals ineligible for federal meaningful-use incentives have dismally low rates of adoption of electronic health records. *Health Affairs*, *31*(3), 505–513.

Yan, Y., Birman–Deych, E., Radford, M. J., Nilasena, D. S., & Gage, B. (2005). Comorbidity indices to predict mortality from medicare data: Results from the national registry of atrial fibrillation on JSTOR. *Medical Care*, *43*(11), 1073–1077.

Ypinga, J. H. L., de Vries, N. M., Boonen, L. H. H. M., Koolman, X., Munneke, M., Zwinderman, A. H., & Bloem, B. R. (2018). Effectiveness and costs of specialised physiotherapy given via ParkinsonNet: A retrospective analysis of medical claims data. *The Lancet Neurology*, *17*(2), 153–161.

Zhang, W., Bjarnadóttir, M. V., Proaño, R. A., Anderson, D., & Konrad, R. (2018). Accelerating the adoption of bundled payment reimbursement systems: A data-driven approach utilizing claims data. *IISE Transactions on Healthcare Systems Engineering*, *8*(1), 22–34.

Zhou, J., Wang, F., Hu, J., & Ye, J. (2014). From micro to macro: data driven phenotyping by densification of longitudinal electronic medical records. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '14 (pp. 135–144). New York, NY: ACM Press. https://www.kdd.org/kdd2014/