Title

Secondary Data Analysis of Large Data Sets in Urology: Successes and Errors to Avoid

Permalink

https://escholarship.org/uc/item/0cr6f7cz

Journal

Investigative Urology, 191(3)

ISSN

0021-0005

Authors

Schlomer, Bruce J
Copp, Hillary L

Publication Date

2014-03-01

DOI

10.1016/j.juro.2013.09.091

Peer reviewed

Published in final edited form as:
J Urol. 2014 March ; 191(3): 587–596. doi:10.1016/j.juro.2013.09.091.

Secondary Data Analysis of Large Data Sets in Urology: Successes and Errors to Avoid

Bruce J. Schlomer[*] and **Hillary L. Copp**[†]
Baylor College of Medicine and Texas Children's Hospital, Houston, Texas (BJS), and University of California San Francisco, San Francisco, California (HLC)

Abstract

Purpose—Secondary data analysis is the use of data collected for research by someone other than the investigator. In the last several years there has been a dramatic increase in the number of these studies being published in urological journals and presented at urological meetings, especially involving secondary data analysis of large administrative data sets. Along with this expansion, skepticism for secondary data analysis studies has increased for many urologists.

Materials and Methods—In this narrative review we discuss the types of large data sets that are commonly used for secondary data analysis in urology, and discuss the advantages and disadvantages of secondary data analysis. A literature search was performed to identify urological secondary data analysis studies published since 2008 using commonly used large data sets, and examples of high quality studies published in high impact journals are given. We outline an approach for performing a successful hypothesis or goal driven secondary data analysis study and highlight common errors to avoid.

Results—More than 350 secondary data analysis studies using large data sets have been published on urological topics since 2008 with likely many more studies presented at meetings but never published. Nonhypothesis or goal driven studies have likely constituted some of these studies and have probably contributed to the increased skepticism of this type of research. However, many high quality, hypothesis driven studies addressing research questions that would have been difficult to conduct with other methods have been performed in the last few years.

Conclusions—Secondary data analysis is a powerful tool that can address questions which could not be adequately studied by another method. Knowledge of the limitations of secondary data analysis and of the data sets used is critical for a successful study. There are also important errors to avoid when planning and performing a secondary data analysis study. Investigators and the urological community need to strive to use secondary data analysis of large data sets appropriately to produce high quality studies that hopefully lead to improved patient outcomes.

Keywords

research design; outcome assessment

[*] Correspondence: Texas Children's Hospital, 6701 Fannin, CCC Suite 620, Houston, Texas 77030..

[†] Supported by the University of California, San Francisco KURe Career Development Program (National Institutes of Health Grant 1-K12-DK-083021).

Secondary data analysis involves the use of data collected by someone other than the investigator for research.[1] Primary data analysis is the use of data collected by the investigator. Secondary data analysis includes secondary analysis of data from randomized clinical trials and prospectively collected observational cohorts, as well as large administrative or survey data sets. The use of secondary data analysis has increased in clinical research and this increase has been seen in urology. There are urologists who have grown skeptical of the number of secondary data analysis studies, especially of secondary data analysis of large data sets that are often administrative. While the increase in secondary data analysis of large data sets has produced many interesting and well designed studies published in high quality journals, there has no doubt been an increase in studies that draw overreaching or inappropriate conclusions from poorly designed studies using inappropriate data sets or methods for the research questions.[2]

Secondary data analysis is appealing because of the generally large size and availability of many of the data sets, and the fact that primary data generation does not have to be performed.[1,3,4] In academic institutions where many of secondary data analysis studies are performed, secondary data analysis can be a way for fellows or junior faculty to create a foundation on which to build a research career.

In this review we will not discuss secondary data analysis of data from randomized clinical trials or cohort studies, but will focus on large data sets. The purpose of this review is to introduce the different types of large data sets commonly used for secondary data analysis in urology, and to discuss the advantages and limitations of research using these data sets. We also give examples of high quality urological secondary data analysis studies published within the last 5 years and discuss errors to avoid when performing secondary data analysis. Finally, we suggest an outline for performing a successful secondary data analysis study and ways for the urological community to ensure secondary data analysis studies are high quality.

ADVANTAGES OF SECONDARY DATA ANALYSIS

One advantage of secondary data analysis is that the data are already collected, which greatly increases the efficiency with which a researcher can perform a study. The use of secondary data analysis as an initial approach to a research question is also appealing for junior investigators without signifi-cant research funding because several large data sets are available free of cost from institutions or for a limited expense. Another advantage is the large size of many data sets, which allows for more precise estimates of trends or effects, especially for rare diagnoses. In addition, secondary data analysis can describe trends or findings on a much larger scale than a single center perspective, and the results may be more generalizable.[5] Secondary data analysis also allows investigators to search for answers to questions that could not be addressed by a randomized trial because that trial may be unethical or prohibitively costly.[6]

LIMITATIONS OF SECONDARY DATA ANALYSIS

There are several general limitations of secondary data analysis. The primary limitation is that the data set was not designed to answer the question an investigator is studying and was often created for administrative or billing purposes.[7] The investigator has no control over the types of patients included, which variables are captured, how the variables are collected and recorded, and the integrity of the data. Therefore, an investigator must be familiar with the data set and the types of hypotheses that can be tested.

Another limitation of several commonly used data sets is that longitudinal followup is not available and, therefore, long-term outcomes cannot be studied. Even in the data sets that have longitudinal followup, the followup may be limited, which raises the potential for bias when reporting outcomes. As with all observational studies, establishing causality is not possible and only associations can be reported.[8] Residual confounding is always a potential cause of any association seen in secondary data analysis studies. Residual confounding occurs when there is an unmeasured variable that is associated with the predictor and outcome of interest. The inability to include that variable in analysis can lead to a spurious association between the predictor and outcome. Multivariate analysis, propensity scores and instrumental variable analysis have been used to help control for confounding in secondary data analysis. However, the potential for residual confounding cannot be eliminated and is only minimized.

Many data sets used for secondary data analysis rely on ICD-9 and/or CPT codes, and several studies have questioned the reliability of these codes.[9–11] It is safe to assume that there is some measurement error for any study using a data set that contains ICD or CPT codes. Many secondary data analysis studies report the incidence or prevalence of a certain diagnosis, procedure, complication or demographic factor. These results are directly affected by the measurement error in the data set and the amount of error likely varies by research question. For example, if an investigator were to look at trends in robotic surgery by using billing codes, and many hospital billing systems were not using codes to indicate the use of robotic surgery, the study would underestimate the number of robotic surgeries. Studies have suggested that combining ICD-9 or CPT procedure codes with appropriate diagnosis codes increases the accuracy of identifying the target population.[12,13] One method to support the validity of findings from large data sets is to perform a review of patients at the investigator's own institution and compare results from the review to results from the large data set. This process increases the work required as some primary data collection is performed. Another method to support the validity of the findings is to address the same question with different large data sets. However, this method is not feasible with many research questions.

Finally, the results of secondary data analysis studies may fail to pass the “so what?” test. Because the data sets are large but the types of variables are not under control of the investigator, there can be statistically significant findings that do not seem clinically relevant. This specific limitation may be a major cause of skepticism for urologists reading secondary data analysis studies.

TYPES OF LARGE DATA SETS USED FOR SECONDARY DATA ANALYSIS

Large data sets used for secondary data analysis that we will discuss include national administrative data sets, national survey data sets, administrative data sets populated by private or public insurance claims and condition specific registries (Appendix 1). In this review we will only discuss United States based data sets. Countries such as Sweden and the United Kingdom have additional opportunities for secondary data analysis of their national health system data.

National Administrative Data Sets

National administrative discharge data sets are some of the most commonly used for secondary data analysis (Appendix 1). The HCUP (Healthcare Cost and Utilization Project) hospital discharge data sets, which include the NIS (National Inpatient Sample) and the KID (Kids' Inpatient Database), are nationally representative discharge data sets derived from samples of the SID (State Inpatient Databases).[14] Each observation in the KID and NIS corresponds to 1 discharge, and contains demographic information, hospital characteristics, diagnosis and procedure codes, illness severity measures, charge and cost information, and discharge weights. Discharge weights must be used to calculate national estimates. Discharge weights relate the number of observations in the NIS or KID to an estimated number of discharges in the entire United States (American Hospital Association universe, which includes all nongovernment hospitals in the United States).[15] A discharge weight is derived by dividing the total number of national discharges for a particular hospital stratum (defined by hospital characteristics such as location, size etc) by the number of discharges in the data set in that same hospital stratum. There is also a nationally representative administrative data set for emergency department visits called the HCUP National Emergency Department Sample (NEDS), which is derived from the HCUP State Emergency Department Databases and the SID.

Typical hypotheses include whether outcomes are associated with hospital or surgeon volume, whether outcomes are associated with various hospital or patient factors, and whether there is a significant trend with time of some measure of interest. These data sets include limited clinical information for inpatient admissions and there is no longitudinal followup. Therefore, a potential study testing the hypothesis that use of a urological procedure has changed with time could only use the NIS or KID if patients are typically admitted to the hospital after that procedure. Because individual patients are not identified in the NIS or KID, a single patient can contribute multiple discharges which may affect the results. Any multivariate analysis should account for clustering by hospital by using hierarchical modeling or regression models that account for the complex survey design of the data sets.[16] Clustering occurs when data can be organized into groups or clusters (eg hospitals) and each group contains multiple observations (eg patients). The observations within a group would be expected to be correlated due to measured and unmeasured factors. Accounting for clustering by hospital is important because there are unmeasured hospital practices or factors that are associated with outcomes, and not accounting for clustering may lead to biased associations.

For the pediatric population there is a large administrative data set that includes information from hospital admissions and observations, emergency department visits and outpatient surgeries from 43 children's hospitals called the PHIS (Pediatric Health Information System). The PHIS is different from the NIS or KID because a patient can be followed longitudinally and it contains additional information such as medication use. However, longitudinal followup is limited because outpatient clinic encounters are not included and not all hospital systems contribute full data sets. Because of the longitudinal nature of the PHIS, it is a popular database for use in pediatric urology. While it is a large data set, the PHIS is not designed to be nationally representative. Multivariate analysis should account for clustering by hospital and potentially by provider as well.

National Survey Data Sets

There are numerous national survey data sets and examples are given in Appendix 1. The NHANES (National Health and Nutrition Examination Survey), NAMCS (National Ambulatory Medical Care Survey) and NHAMCS (National Hospital Ambulatory Medical Care Survey) are some of the most popular. The NHANES is a group of studies designed to assess the health and nutritional status of children and adults in the United States. Data in NHANES are generated by health interviews conducted in participant homes along with physical examinations and blood tests performed in mobile centers.[17] NAMCS is a national survey of non-federally employed, office based physicians designed to obtain information about the provision of ambulatory medical care services. Interviewers visit physicians, and collect data regarding patient demographics, services provided, diagnoses made, medications prescribed and planned future treatment. NHAMCS is similar to NAMCS, but is designed to collect data regarding the provision of ambulatory care in the hospital emergency department, outpatient departments and ambulatory surgery centers. These surveys are designed to be nationally representative and investigators must be careful to use the appropriate statistical methods to account for the survey design. The types of hypotheses that can be tested vary based on the survey population and the data elements included in the survey.

Data Sets Populated by Private or Public Insurance Claims

Data sets populated by private or public insurance claims contain billing or claims information for a variety of encounters. An example of a private insurance claims database is the i3 Innovus database (primarily UnitedHealth), and examples of public insurance databases are Medicare claims data and Medicaid claims data through the Medicaid Analytic eXtract data set. Medicare claims data are powerful because information on inpatient, outpatient, nursing home and home care is available for more than 95% of the population older than 65 years in the United States. Medicaid claims data through Medicaid Analytic eXtract include similar information but for the adult and pediatric population with Medicaid claims.

A potential advantage of insurance claims data over administrative data sets such as the NIS is that a person can often be followed longitudinally with time. Other advantages of insurance claims data are that they allow analysis on a large scale (but not necessarily nationally representative). In addition, the tests performed or drugs prescribed are potentially more accurately captured than in administrative data sets since they are captured by a claim to or payment by an insurance entity. Some potential disadvantages are that the population of the insurance claims database may not be representative of other populations, there may be coding errors, patients may pay out of pocket for certain medications and not file a claim, and patients may switch insurance companies frequently, making longitudinal followup limited. Important factors to consider regarding insurance claims data sets is that the time it takes to obtain the information from the insurance entity (public and private) can be relatively long compared with other large data sets, and the cost can be prohibitively expensive.

Condition Specific Registries

Data sets that are generated from registries of patients with a certain type of condition are also used for secondary data analysis. While these data sets are different from administrative or claims based data sets because they were collected to study a particular condition, they are included in this discussion because they are large data sets that urological investigators often use for secondary data analysis. An example of this type of data set is the SEER (Surveillance, Epidemiology, and End Results) Program data set. The SEER data set is a product of the National Cancer Institute which began in 1973 and has expanded over time to cover almost 30% of the United States population. Typical information in the SEER data set includes patient demographics, primary tumor site and morphology, tumor stage at diagnosis, first course of treatment and vital status at followup. Linkage of the SEER data set with Medicare claims data combines cancer specific information in the SEER data set with longitudinal claims information in the Medicare data, and has been used by many investigators.

Another large condition specific registry used for secondary data analysis of patients with prostate cancer is the CaPSURE™ (Cancer of the Prostate Strategic Urologic Research Endeavor) data set.[18] The CaPSURE data set was designed to study outcomes in men with prostate cancer and includes more than 14,000 men with biopsy proven prostate cancer. While CaPSURE studies performed by the primary investigators could be considered primary data analysis, studies performed by other investigators who are granted access to the data would be considered secondary data analysis. The National Trauma Data Bank is another example of a condition specific registry maintained by the American College of Surgeons that has been used in secondary data analysis studies investigating urological trauma.

EXAMPLES OF SECONDARY DATA ANALYSIS SUCCESSES

While a successful study has a subjective definition, we wanted to give examples of secondary data analysis studies on urological topics that were published in journals with high impact factors. We performed a search for secondary data analysis studies on urological topics published since 2008 using search terms for the large data sets used for secondary data analysis discussed in this review and limiting the results to urological topics. The studies were ranked by the 5-year impact factor of the journal in which they were published.

Using our search criteria we identified 373 uro-logical secondary data analysis studies published since 2008. The most common journals were *The Journal of Urology*® (97), *Urology* (60), *Cancer* (41), *BJU International* (37) and *Urologic Oncology* (26). The highest impact factor journal in which a study was published was *JAMA*, with 3 articles since 2008.

Other high impact factor journals in which urological secondary data analysis studies were published include the *Journal of Clinical Oncology* (7), the *Journal of the National Cancer Institute* (5), *Archives of Internal Medicine* (3) and *Pediatrics®* (3).

In an article published in *JAMA*, Tan et al had the goal of comparing outcomes in patients treated with partial vs radical nephrectomy for early stage kidney cancer.[19] They used the SEER data set to identify patients with stage T1a kidney cancer and linked those patients to Medicare claims data to determine which patients were treated with partial or radical nephrectomy. On multivariate analysis they accounted for confounding by using an instrumental variable analysis that allowed pseudo-randomization and found that patients with early stage kidney cancer treated with partial nephrectomy had improved survival compared to those treated with radical nephrectomy. This article demonstrates how secondary data analysis can generate an answer to a clinically relevant question where a randomized clinical trial may be difficult or not feasible.

In another article published in *JAMA* Jacobs et al had the goal of assessing the use of advanced treatment technologies in men with a low risk of dying of prostate cancer.[20] This study also used the SEER data set and Medicare claims data to identify men treated for prostate cancer who were at low risk for death from prostate cancer. They found that the use of advanced technologies in treating men at low risk for death from prostate cancer had increased significantly with time. Overtreatment of prostate cancer is an important and timely research subject, and this study demonstrates how these data sets can be used to describe important trends in care that may or may not be appropriate. Appendix 2 lists several other examples of high quality secondary data analysis studies chosen to demonstrate the variety of urological topics, data sets and statistical methods that can be used.[19–29]

ERRORS TO AVOID WITH SECONDARY DATA ANALYSIS

The first error to avoid is not having a predetermined hypothesis or goal for the study. Data mining can be described as the process of running multiple hypothesis tests on a data set looking for a significant result that can be organized around a research question and presented as an abstract or publication. Unfortunately the large data sets we have described do lend themselves to data mining. While certain types of data mining may be appropriate (eg searching for candidate genes), data mining with large data sets should be avoided. We believe that investigators, mentors, program directors and the urological community as a whole must actively work to ensure high quality secondary data analysis studies are performed and improper data mining is not performed. Not all types of hypotheses can be tested with secondary data analysis and a working knowledge of the data sets is needed to know which types of hypotheses can be tested.

The second error an investigator can run into is choosing the wrong data set for the research hypothesis or goal. For example, an investigator has a hypothesis that orchiopexy is performed in children at an older age in rural areas, which suggests poorer quality of care for children with undescended testis in rural areas. The investigator chooses to use the KID. Unfortunately since the majority of orchiopexies are performed on an outpatient basis, those visits will not be captured in the KID. A much more appropriate choice would be the State Ambulatory Surgical Databases, which would include outpatient orchiopexies and zip code information.

Another type of database choice error would be if an investigator wanted to report national trends in admissions or surgeries and chose a data set that is not designed to be nationally representative. Data sets such as the PHIS and SEER are large, and the results may be highly generalizable, but those data sets are not specifically designed to be nationally representative.

The third error is inappropriate statistical analysis, which can be the result of not consulting a biostatistician. Many of these large data sets have complex survey designs, and are weighted to give correct estimates for nationally descriptive statistics and evaluating trends. It is preferable to use the survey structure and weights in multivariate regression analysis as well. This will account for data clustering and the weights, and is available on most statistical packages. Many other data sets do not have discharge weights but patients are usually still clustered by hospital and sometimes by physician. Accounting for clustering is important in obtaining accurate estimates of associations in multivariate analysis and biostatistician consultation should be used if investigators do not have training in biostatistics.

SUGGESTIONS FOR SUCCESSFUL SECONDARY DATA ANALYSIS

To perform an effective secondary data analysis study we have suggested a step-wise approach and have discussed errors to avoid at each step (Appendix 3).[8] For most investigators it is ideal to have a mentor who can help guide the investigator through all steps. In the first step, an idea for a hypothesis to be tested or the goal of the study is chosen, and this should be done before performing an analysis. Investigators should choose a topic in which they are interested and not choose an uninteresting question to fit a data set. Some flexibility in adapting the hypothesis or goal to an appropriate data set is acceptable, but the research question should come first. A literature review should be performed to ensure that a secondary data analysis study would add to the published literature, and that the results would be clinically relevant and pass the “so what?” test.

The second step is choosing an appropriate data set. A working knowledge of the types of data sets available for secondary analysis is needed to select which data set is appropriate to use to test the hypothesis. In Appendix 2 we have suggested some methods to become familiar with data sets.

Third, once a data set is chosen, the investigator should determine how the data set was created, what types of data are in the data set, how the data are assessed for reliability and what the limitations of the data set are. Reading studies that used the particular data set for a variety of topics can help with understanding and can reveal important analysis methods.

The fourth step is statistical analysis. Knowledge of appropriate statistical methods and limitations of the particular data set is critical. Biostatistician consultation is often needed as the appropriate statistical methods are likely unfamiliar to most investigators without formal biostatistical training. Many if not most data sets will require a complex survey structure analysis or a method to account for clustering.

Finally, an accurate presentation of the findings in a clinically meaningful way so the findings pass the "so what?" test is critical. Results should be accurately put into context with prior literature and knowledge of the topic, and reasonable conclusions regarding the importance of the findings should be made. However, findings should not be over-interpreted and conclusions should not be overreaching (eg causality cannot be determined).

OTHER SUGGESTIONS FOR IMPROVING QUALITY OF SECONDARY DATA ANALYSIS

Reviewers and editors have a critical role in the quality of secondary data analysis studies. It is important for abstract and article reviewers to have a working knowledge of the data set that was used in the study as well as appropriate statistical methods that may be more complex than in other types of clinical studies. A study that has highly clinically significant results and passes the "so what?" test also needs reviewers to ensure that appropriate statistical analysis was performed. An attitude that writes off all secondary data analysis as data mining is inappropriate as many high quality studies have been published using appropriate secondary data analysis techniques. On the other hand, it is important for the reviewer community to identify and exclude poor quality studies.

Mentors and program directors at academic institutions also have a role in the quality of secondary data analysis as many of these studies are performed by trainees or junior faculty at academic institutions. The use of secondary data analysis requires a specific skill set and the development of those skill sets in interested individuals should be supported. Most academic institutions will have individuals familiar with secondary data analysis who can serve as additional mentors if the primary mentor is not familiar with secondary data analysis. A collaborative group of investigators across disciplines that meets to discuss projects will also potentially improve the quality of studies and research ideas. As previously mentioned, biostatistician consultation should be used and supported if needed.

CONCLUSIONS

Secondary data analysis of large data sets has increased in urology. As these studies have become more common it is important to understand their advantages and disadvantages. Secondary data analysis can be a powerful tool to answer important research questions and we have provided examples of high quality studies. There are important errors to avoid with secondary data analysis which have contributed to the skepticism of secondary data analysis studies. Investigators and the uro-logical community need to strive to use these important data sets appropriately to produce high quality studies that hopefully lead to improved patient outcomes.

Appendix

APPENDIX 1

Examples of large data sets used for secondary data analysis studies in urology

Data Set Examples	Key Points
HCUP National Administrative	
State Inpatient Databases (SID)	Administrative data for adult and pediatric inpatient discharges from 47 participating states comprising 97% of population. Contain all discharges from nonfederal hospitals. Data from SID are used to create the NIS and the KID but can be accessed themselves. No longitudinal followup.
State Ambulatory Surgery Databases (SASD)	Administrative data for adult and pediatric ambulatory surgery encounters. The completeness of data set varies from state to state, so a nationally representative sample of ambulatory surgery administrative data like the NIS, KID or NEDS is not possible to generate.
State Emergency Department Databases (SEDD)	Administrative data for adult and pediatric emergency department discharges from 31 participating states. All SEDD contain hospital affiliated emergency department discharges that resulted in discharge from the emergency department or transfer to another hospital. SEDD are used with SID to create the NEDS.
Nationwide Inpatient Sample (NIS)	Administrative data from adult and pediatric inpatient discharges. Contains a sample of discharges from a 20% stratified sample of all U.S. community hospitals in SID. Weighted to calculate national estimates.
Kids' Inpatient Data Set (KID)	Administrative data for pediatric inpatient discharges. Contains a sample of pediatric discharges from all U.S. community hospitals in the SID, 10% of normal births are sampled and 80% of all other pediatric discharges are sampled. Weighted to calculate national estimates.
Nationwide Emergency Department Sample (NEDS)	Administrative data for adult and pediatric emergency department visits. Contains a sample of discharges from a 20% stratified sample of hospital based emergency department visits in the SEDD and SID. Weighted to calculate national estimates.
Other Large Administrative	
Pediatric Health Information System (PHIS)	Administrative data from 43 freestanding pediatric hospitals in the U.S. Includes data for admissions, emergency department visits, outpatient surgeries and observation admissions. Outpatient clinic visits are not included. Not all hospitals contribute complete data sets. Medication and supply use available. Longitudinal followup available.
National Surgical Quality Improvement Program (NSQIP)	Preoperative through 30-day postoperative information collected on random sample of patients at participating hospitals. A pediatric version exists as well. More than 100 data points including preoperative risk factors, intraoperative factors, and 30-day postoperative morbidity and mortality outcomes are collected. Inpatient and outpatient surgeries are included.
National Surveys	
National Health and Nutrition Examination Survey (NHANES)	Nationally representative sample of adult and pediatric population in the U.S. Data include but are not limited to in-depth, in-person surveys, physical examinations, laboratory values, demographic data and dietary/ lifestyle data. Complex survey design with weights used for calculating national estimates.
National Ambulatory Medical Care Survey (NAMCS)	Nationally representative survey of outpatient, nonhospital affiliated physician visits. Physicians fill out forms regarding outpatient visits. Clustered and weighted design. No longitudinal followup. However, national trends can be evaluated.
National Hospital Ambulatory Medical Care Survey (NHAMCS)	Nationally representative survey of outpatient, hospital affiliated physician visits and emergency room physician visits. Physicians fill out forms regarding visits. Clustered and weighted design. No longitudinal followup. However, national trends can be evaluated.
Private Insurance Claims	
i3 Innovus database	Health care claims data from large U.S. commercial health plan (UnitedHealth). Longitudinal followup available if stay in plan.

Data Set Examples	Key Points
IMS LifeLink Health Plan Claims Database	Health care claims data on more than 60 million individuals from more than 80 different health plans. More than 75% are commercial insurance plans. Longitudinal followup available if stay in plan.
Public Insurance Claims	
Medicare claims data	ICD-9 and CPT claims information on 98% of population age 65 years or older. Longitudinal followup available. Obtaining data can be lengthy and costly process.
Medicaid claims data	Claims information for low income adults, disabled adults and children covered by Medicaid. Longitudinal followup available. Obtaining data can be lengthy and costly process. Obtained from Medicaid Statistical Information System or Medicaid Analytic eXtract.
Condition Specific Registries	
Surveillance, Epidemiology, and End Results Program (SEER)	Data on cancer incidence and survival rates. Covers almost 30% of population. Information about cancer diagnosis, demographics, initial treatment and survival. Often paired with Medicare claims data.
National Trauma Data Bank (NTDB)	Administrative data for adult and pediatric trauma admissions from participating hospitals. A nationally representative sample of adult patients in the NTDB is available to make national estimates. No longitudinal followup.
Cancer of the Prostate Strategic Urologic Research Endeavor (CaPSURE)	Longitudinal, observational study of prostate cancer outcomes in more than 14,000 patients that has been used for numerous secondary data analysis studies. Only certain investigators can request to use data set for secondary data analysis.

Appendix

APPENDIX 2

Examples of Secondary Data Analysis Successes

References	Hypothesis/Goal	Data Set Used	Statistical Issues	Important Findings
Tan et al[19]	Compare long-term survival after partial vs radical nephrectomy in early stage kidney cancer	SEER linked with Medicare claims	Instrumental variable analysis used to help account for residual confounding	Partial nephrectomy associated with decreased overall mortality compared to radical nephrectomy in early stage kidney cancer
Jacobs et al[20]	Assess use of advanced treatment technologies in men with low risk of dying of prostate cancer	SEER linked with Medicare claims	Evaluated for trends with time	The use of advanced treatment technologies has increased with time
Chang et al[21]	Assess the impact of common medications on prostate specific antigen	NHANES	Sampling weights used in multivariate analysis	Men using nonsteroidal anti-inflammatory drugs, statins and thiazides had decreased prostate specific antigen
Choe et al[22]	Anticoagulation use is associated with prostate cancer specific mortality	CaPSURE	Time varying covariates in Cox proportional hazards analysis	Aspirin use associated with decreased prostate cancer specific mortality
Elliott et al[23]	Assess effect of reduction in reimbursement on use of androgen suppression	SEER linked with Medicare claims	Evaluated for trends with time as well as multivariate analysis	Decrease in reimbursement was associated with decrease in use of androgen suppression therapy

References	Hypothesis/Goal	Data Set Used	Statistical Issues	Important Findings
	therapy for prostate cancer			
Yu et al[24]	Compare costs and outcomes between open and robotic assisted radical cystectomy	NIS	Survey weights used in multivariate analysis. Propensity score analysis	Fewer immediate postoperative complications with robotic assisted, but higher cost and similar length of stay
Sammon et al[25]	Examine trends in treatment for infected urolithiasis and compare outcomes of treatment modalities	NIS	Weighted estimates evaluated for trends. Propensity score analysis	Increasing use and more complications associated with nephrostomy tube
Copp et al[26]	Examine patterns of ambulatory antibiotic use for urinary tract infections in children and factors associated with broad spectrum antibiotic use	NAMCS and NHAMCS	Survey weights used in analysis	Use of 3rd generation cephalosporins has increased significantly for childhood urinary tract infections
Kokorowski et al[29]	Examine trends in timing of orchiopexies and factors associated with timing	PHIS	Accounted for patient clustering at the surgeon level	Only 43% of boys had surgery by age 2 years. Patient race, insurance status and hospital associated with timing
Gore et al[27]	Assess impact of delay of radical cystectomy on outcomes in national data set	SEER linked with Medicare claims	Cox proportional hazards analysis from time of bladder cancer diagnosis to avoid lead time bias	Delay of more than 12 weeks for radical cystectomy associated with decreased survival
Hollingsworth et al[28]	Assess association between being a self-employed urologist and use of imaging	NAMCS	Survey weights used in multivariate analysis	Being a self-employed urologist was highly associated with increased use of imaging

Appendix

APPENDIX 3

Step-wise approach to successful secondary data analysis study

Step	Key points	Errors to avoid
1. Determine goal or hypothesis of study	1.Select a question and topic in which you are interested. 2.Choose a mentor if you are unfamiliar with secondary data analysis to assist in all steps. Even if you are familiar with secondary data analysis, consider choosing a mentor or meeting with others interested in secondary data analysis to discuss projects. 3.Review published literature to ensure you are not duplicating work. 4.Decide if results to the study goal or hypothesis would pass the “so what?” test. 5.Reserve some minor flexibility in hypothesis or goal of study in order to adapt to best data set.	1.Choosing a question or topic you are not interested in just because it might be good for a secondary data set analysis. Most likely, others will not be interested in it either. 2.Not having a mentor if you are unfamiliar with secondary data analysis can lead to inappropriate methods.
2. Choose a data set	1.Read about available secondary data sets and the types of information in the data sets. Online resources	1.Poor or limited understanding of information in data sets can

Step	Key points	Errors to avoid
	(www.sgim.org/go/datasets) or reviews on secondary data analysis are good places to start. 2.Identify potential data sets based on whether your research question can be addressed with the variables in the data set. 3.Once potential data sets have been identified, read several studies that used those data sets on a variety of topics. This will help you understand how the data sets can be used. 4.Choose the data set most appropriate for your research question where there is minimal or no change to your original hypothesis or goal of study. It is also valid to use a new data set for a research question that has been addressed with other data sets to add additional perspective on a topic. 5.Consult with mentors or others familiar with data sets.	lead to inappropriate data set selection. 2.Attempts to force a research question to fit an inappropriate data set should be avoided. 3.Prior use of a data set for a particular type of research question does not necessarily mean that it was appropriate use.
3. Learn about the data set	1.Read information on the data set available from source. Learn why the data set was created, how the data were gathered, how the data were tested for reliability and who makes the data set. 2.Learn about the variables in the data set. 3.Learn about the limitations of the data set. 4.Learn about the structure of the data set. For example, some data sets have a complex survey design with strata and weights.	1. Not fully understanding the data set may lead to inappropriate analysis methods, results and conclusions.
4. Statistical analysis	1.Learn about the appropriate statistical methods for analyzing the data set. 2.Have low threshold to employ biostatistician consultation. 3.Read statistical methods for several studies that used the data set. 4.Many data sets will have clustered data and this should be taken into account in analysis. 5.Any complex survey design and/or observation weights should be used.	1.Not employing a biostatistician when not familiar with statistical methods. 2.Prior application of particular statistical methods in a publication does not guarantee appropriate methodology. 3.Not accounting for clustering of data or structure of data set.
5. Present results of study	1.State hypothesis or goal of study clearly in introduction. 2.Accurately describe statistical methods used and results. 3.Interpret the clinical meaning of the results in a balanced way, and discuss results in context of prior literature and knowledge. Do not make overreaching conclusions.	1.Editorializing when presenting findings in the results section. 2.Overreaching in interpretation of results and conclusions.

REFERENCES

1. Best AE. Secondary data bases and their use in outcomes research: a review of the area resource file and the Healthcare Cost and Utilization Project. J Med Syst. 1999; 23:175. [PubMed: 10554733]
2. Terris DD, Litaker DG, Koroukian SM. Health state information derived from secondary databases is affected by multiple sources of bias. J Clin Epidemiol. 2007; 60:734. [PubMed: 17573990]
3. Lewis NJ, Patwell JT, Briesacher BA. The role of insurance claims databases in drug therapy outcomes research. Pharmacoeconomics. 1993; 4:323. [PubMed: 10146871]
4. Wennberg JE, Roos N, Sola L, et al. Use of claims data systems to evaluate health care outcomes. Mortality and reoperation following prostatectomy. JAMA. 1987; 257:933. [PubMed: 3543419]
5. Guller U. Surgical outcomes research based on administrative data: inferior or complementary to prospective randomized clinical trials? World J Surg. 2006; 30:255. [PubMed: 16485067]
6. Porter GA, Skibber JM. Outcomes research in surgical oncology. Ann Surg Oncol. 2000; 7:367. [PubMed: 10864345]
7. Yiee JH, Copp HL. Database research in pediatric urology. Curr Opin Urol. 2011; 21:309. [PubMed: 21499106]
8. Smith AK, Ayanian JZ, Covinsky KE, et al. Conducting high-value secondary dataset analysis: an introductory guide and resources. J Gen Intern Med. 2011; 26:920. [PubMed: 21301985]

9. Khwaja HA, Syed H, Cranston DW. Coding errors: a comparative analysis of hospital and prospectively collected departmental data. BJU Int. 2002; 89:178. [PubMed: 11856094]

10. Southern DA, Roberts B, Edwards A, et al. Validity of administrative data claim-based methods for identifying individuals with diabetes at a population level. Can J Public Health. 2010; 101:61. [PubMed: 20364541]

11. Wang MC, Laud PW, Macias M, et al. Strengths and limitations of International Classification of Disease Ninth Revision Clinical Modification codes in defining cervical spine surgery. Spine (Phila Pa 1976). 2011; 36:E38. [PubMed: 20975624]

12. Tanpowpong P, Broder-Fingert S, Obuch JC, et al. Multicenter study on the value of ICD-9-CM codes for case identification of celiac disease. Ann Epidemiol. 2013; 23:136. [PubMed: 23313264]

13. Tollefson MK, Gettman MT, Karnes RJ, et al. Administrative data sets are inaccurate for assessing functional outcomes after radical prostatectomy. J Urol. 2011; 185:1686. [PubMed: 21419458]

14. HCUP Databases. Healthcare Cost and Utilization Project (HCUP). Agency for Healthcare Research and Quality; Jul. 2013 Available at www.hcup-us.ahrq.gov/databases.jsp. [September 15, 2013]

15. Houchens, RL.; Elixhauser, A. HCUP Methods Series Report #2006-05 Online. U.S. Agency for Healthcare Research and Quality; Rockville, Maryland: 2006. Using the HCUP Nationwide Inpatient Sample to Estimate Trends (updated for 1988-2004)..

16. Houchens, R.; Chu, B.; Steiner, C. Hierarchical Modeling using HCUP Data. HCUP Methods Series Report #2007-01 Online. U.S. Agency for Healthcare Research and Quality; Rockville, Maryland: 2007.

17. National Health and Nutrition Examination Survey 1999-2012 Survey Content Brochure. National Center for Health Statistics; Atlanta, Georgia: 2012.

18. Lubeck DP, Litwin MS, Henning JM, et al. The CaPSURE database: a methodology for clinical practice and research in prostate cancer. CaPSURE Research Panel. Cancer of the Prostate Strategic Urologic Research Endeavor. Urology. 1996; 48:773. [PubMed: 8911524]

19. Tan HJ, Norton EC, Ye Z, et al. Long-term survival following partial vs radical nephrectomy among older patients with early-stage kidney cancer. JAMA. 2012; 307:1629. [PubMed: 22511691]

20. Jacobs BL, Zhang Y, Schroeck FR, et al. Use of advanced treatment technologies among men at low risk of dying from prostate cancer. JAMA. 2013; 309:2587. [PubMed: 23800935]

21. Chang SL, Harshman LC, Presti JC Jr. Impact of common medications on serum total prostate-specific antigen levels: analysis of the National Health and Nutrition Examination Survey. J Clin Oncol. 2010; 28:3951. [PubMed: 20679596]

22. Choe KS, Cowan JE, Chan JM, et al. Aspirin use and the risk of prostate cancer mortality in men treated with prostatectomy or radiotherapy. J Clin Oncol. 2012; 30:3540. [PubMed: 22927523]

23. Elliott SP, Jarosek SL, Wilt TJ, et al. Reduction in physician reimbursement and use of hormone therapy in prostate cancer. J Natl Cancer Inst. 2010; 102:1826. [PubMed: 21131577]

24. Yu HY, Hevelone ND, Lipsitz SR, et al. Comparative analysis of outcomes and costs following open radical cystectomy versus robot-assisted laparoscopic radical cystectomy: results from the US Nationwide Inpatient Sample. Eur Urol. 2012; 61:1239. [PubMed: 22482778]

25. Sammon JD, Ghani KR, Karakiewicz PI, et al. Temporal trends, practice patterns, and treatment outcomes for infected upper urinary tract stones in the United States. Eur Urol. 2013; 64:85. [PubMed: 23031677]

26. Copp HL, Shapiro DJ, Hersh AL. National ambulatory antibiotic prescribing patterns for pediatric urinary tract infection, 1998-2007. Pediatrics. 2011; 127:1027. [PubMed: 21555502]

27. Gore JL, Lai J, Setodji CM, et al. Mortality increases when radical cystectomy is delayed more than 12 weeks: results from a Surveillance, Epidemiology, and End Results-Medicare analysis. Cancer. 2009; 115:988. [PubMed: 19142878]

28. Hollingsworth JM, Birkmeyer JD, Zhang YS, et al. Imaging use among employed and self-employed urologists. J Urol. 2010; 184:2480. [PubMed: 20952030]

29. Kokorowski PJ, Routh JD, Graham DA, et al. Variations in timing of surgery among boys who underwent orchiopexy for cryptorchidism. Pediatrics. 2010; 126:e576. [PubMed: 20732947]