The use of big data for improving understanding of the natural history of neurosurgical disease

Katherine G. Holste, MD,1 Zoey Chopra, BA,1,2 Sara Saleh, MD,1 Yamaan S. Saadeh, MD,1 Paul Park, MD,1 and Cormac O. Maher, MD1

1Department of Neurosurgery and 2School of Medicine, University of Michigan, Ann Arbor, Michigan

Understanding the natural history of a condition is essential in determining appropriate patient management. Unfortunately, current understanding of the natural history for many conditions is quite poor. The impact of natural history studies and their importance is significant; since 1956, 15 of the 100 most-cited articles in neurosurgery journals are natural history studies.1 These studies have aided neurosurgeons in determining which treatment methodologies are best applied to certain patient populations with unruptured intracranial aneurysms,2 vestibular schwannomas,3 and dural arteriovenous fistulas.4

There has been an exponential generation of healthcare-related data in the past decade and, accordingly, increased usage of big data repositories in neurosurgical research.5,6 These databases store vast quantities of de-identified patient information that are readily available for querying. Examples of such repositories include, but are not limited to, the Healthcare Cost and Utilization Project Kids’ Inpatient Database (HCUP-KID), National Inpatient Sample, and American College of Surgeons National Surgical Quality Improvement Program.5,6 Big data studies are an increasingly important tool for identifying predictive factors associated with predetermined outcomes;6 however, the limitations of these databases are important to consider. Incomplete or missing data, the validity of diagnosis codes, and false-positive associations are concerns to be addressed in order to yield meaningful results in big data studies.

Large database reviews allow researchers the ability to examine both the treated and untreated natural histories of diseases. Each database is different. Some databases are created for the purpose of collecting clinical information, while others are administrative in nature or are insurance registries that have been adapted for use by clinicians. The aim of the database influences which patient populations are represented. Certain databases aim to be representative of state or national populations, whereas others aim to represent patients with a specific disease or treatment.6 As the data are de-identified and abstracted, extraction of the desired data and analysis is relatively fast and inexpensive compared with randomized controlled trials or observational studies. Simply put, if the data exist within the database, then they can be extracted. In addition to clinical outcomes, big data studies can look macroscopically and examine geographic and temporal trends. Another advantage of big data studies is the ease of assembling a large study cohort. Larger cohorts provide researchers with the ability to detect small differences within the study group or subgroups, which would be less feasible in a smaller cohort. This is especially helpful when investigating uncommon conditions. For example, a study that queried the HCUP-KID for pediatric blunt cerebrovascular injury found that 37.4% of patients experienced a posttrauma ischemic stroke with an in-hospital mortality rate of 20%.8 An analysis of risk factors for ischemic events in the setting of blunt cerebrovascular injury would be exceedingly difficult in a single-center review, as blunt cerebrovascular injury occurs in 0.3% to 0.9% of the pediatric trauma population.8,9

Understanding the limitations of large database studies aids both the clinician interpreting the results and the researcher designing the study. First, the accuracy of the data comes into question. Are the diagnostic or procedure codes assigned to the patient accurate? For example, if a practitioner is not familiar with the differences between Chiari malformation types I and II, they may mis-assign the diagnostic code. Second, data integrity comes into question: Are there incomplete or missing data? Searching a large database patient by patient to collect missing data is unrealistic. And finally, are the results valid? Effective presentation of validated findings is as important as the analysis itself. A variety of author-oriented advice has been published for reporting different domains of clinical research transparently for randomized controlled trials10 and meta-analyses of observational studies.11 These checklists may be helpful in framing discussions about internal and...
external validity, but they are not tailored toward the natural history of disease. For more direct guidance on reporting results from observational approaches, the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) statement provides a comprehensive checklist for delivering results effectively and unambiguously. Unfortunately, a guideline does not exist for responsible presentation of results from large database reviews.

Data accuracy and integrity are important to address when conducting a big data study. For a certain disease to be accurately studied within a database, first, the clinician must correctly diagnose and document that disease, and second, the abstractor must interpret and use the proper codes or terminology. It stands to reason that the more complex the patient case is, or the process for extracting data, the greater the chance of error. Woodworth and colleagues compared their single-institution prospectively collected Department of Neurosurgery database of treated intracranial aneurysms with the State of Maryland administrative database. Using ICD-9 codes, the authors found that the administrative database missed 17% of patients diagnosed with an intracranial aneurysm and 16% of patients treated surgically for an aneurysm. Misdiagnosis is not just an issue with primary diagnoses; surgical indications, comorbidities, and complications can be miscoded as well. In another study of patients who underwent spine surgery, the accuracy of ICD-9 comorbid obesity codes was examined at a single hospital. The authors found that the codes for obesity and morbid obesity, compared with the patient’s recorded BMI in the medical record, had sensitivities of 0.19 and 0.48, respectively. The authors concluded that diagnosis codes for morbidities introduced potential error due to code misuse. Codes for indications for spine surgery are also prone to error. When comparing the primary ICD-9 diagnosis, this matched the surgeon’s indication for surgery 48.4% of the time. With the addition of all secondary ICD-9 diagnoses, the indication for spine fusion matched the surgeon’s notes in 79.4% of cases.

Data accuracy and integrity are serious concerns in big data studies that need to be addressed within the study design to ensure that the results are meaningful. To address these concerns, we present a framework for designing a natural history study using big data.

**Designing a Natural History Study Using Big Data**

**Identification**

1) Identify the patient population of interest.
2) Identify the database that will best contain the data of interest.
3) Define specific diagnostic and procedural variables for the population of interest. ICD, Current Procedural Terminology (CPT), or other codes may be used.
4) Define specific variables for the outcomes of interest.
5) Define exclusion criteria, namely, what factors might confound results later. Typically, patients are excluded from studies when they carry important alternative risk factors for the outcome of interest.
6) Determine the standard of care for the disease in question, if applicable.
7) Perform a comprehensive literature review.

**Search Methodology**

1) Query the desired database to obtain a cohort of de-identified patients with the specified variables.
2) Exclude confounders.
3) Determine control groups, as necessary.

**Data Collection**

1) From the final cohort, query outcome codes and the time the service was provided. These data can be compared within subgroups.
2) Verify that the diagnosis and procedure codes that were used accurately reflect the desired population. This is accomplished by using these same codes to query a known population at the parent institution and by calculating the positive predictive value of these codes in identifying the desired cohort.

**Statistics**

1) Consider correcting the alpha size, the level of significance, for extremely large sample sizes in order to reduce the risk of false-positive findings.

There are several factors to consider when designing a natural history study within a large database to ensure that valid, generalizable, and accurate results are obtained. When done well, conclusions drawn from the natural history of disease can be significant in their clinical impact, affecting diagnostic choices, modifying therapeutic decisions, and inspiring the next generations of research relevant to the disease. As with any other study, the first step is to identify the patient population of interest and subsequently the database that best fits the population. Such databases can be within a single institution, multi-institutional, regional, national, or international. To choose the proper database, it is necessary to understand the potential utility as well as the limitations of each of the choices. For example, a big data study performed by Wilkinson and colleagues in 2020 examined the obstetric outcomes of patients with a diagnosed Chiari malformation before and after their Chiari diagnosis. The Clinformatics Data Mart (Optum Inc.) database, a large nationwide private insurer database with more than 58 million private insurance enrollees, was used for review. This database includes only patients who are privately insured; therefore, patients who are on Medicare or Medicaid or have lapses in insurance are excluded, which may exclude patients of a certain age or socioeconomic status from review.

Once a database is selected, the variables of interest must be defined. Diagnosis codes, such as ICD-9 and ICD-10 or CPT codes, are just a few options that may be used to search the database. Although ICD-9 and ICD-10 codes are appealing for their ease of search, their accuracy and validity may come into question for some diagnoses, as discussed previously. To address the accuracy of the diagnostic codes used, one can attempt to validate a large multi-institutional data set by looking more closely at a known subpopulation of the data set at a single institution. This validation step is underused in big data studies, with only 3 of 78 studies attempting to validate their findings from a known cohort at the authors’ institution in one review article. Another option is to query the population for...
other variables with a known association to the primary diagnosis in question. For example, patients with myelomeningocele will often not only have this diagnostic code, but also frequently carry codes associated with urological or orthopedic comorbidities. These techniques allow the investigator to be more certain that the cohort accurately examines the population in question.

After the database is queried for the specified variables of interest, confounders are excluded. For example, in a big data review of Korean adults taking low-dose aspirin to determine the risk of intracranial hemorrhage, confounders such as hypertension, cerebrovascular disease, other antplatelets or anticoagulants, and systemic corticosteroids were identified and classified as present or absent and were addressed within a subgroup analysis. Determining a control group is sometimes necessary, and the same database can be queried for patients without the diagnosis or procedure of interest. For example, in the study by Wilkinson and colleagues, a cohort of patients without a diagnosis of Chiari malformation was used to determine rates of cesarean sections and adverse obstetrical outcomes.

Large database reviews have been at times critiqued for being “fishing expeditions.” Exploratory analyses of large databases search hundreds of variables for any statistically significant risk factor. In a recent study examining characteristics of big data studies in pediatric neurosurgery, almost all 74 articles employed exploratory analysis to look for effect size, which increases the risk of false-positive findings. With a very large number of data points, it is more likely that a random variable will be flagged as “significant” just through chance. A p value of 0.05 is an arbitrary number and should be changed to fit the study design and statistics used. To adjust the p value, Bonferroni correction may be applied, which multiplies the raw p value by the number of tests to correct for the inflation of the false-positive rate.

When studying the natural history of disease, it is important to consider the many methodologies, ethical confounders, and statistical challenges that arise. For neurosurgery in particular, studying the natural history of disease is challenging, given the high proportion of rare diseases and difficulties in early identification. For this reason, big data and large database reviews have arisen as promising methodologies for neurosurgical natural history studies. Large database reviews have the benefit of including large cohorts relatively easily, although concerns for data validity, integrity, and false-positive statistical findings are important to address. When designed well, big data studies have significant advantages over other methodologies for studying the natural history of neurosurgical diseases.


Acknowledgments

Mr. Chopra receives support from the NIH T32 grant program (T32 GM-007863). Dr. Holste receives support from an NREF training grant.

References

Disclosures
Dr. Park: consultant for Globus, NuVasive, and DePuy Synthes; royalties from Globus; and support of non–study-related clinical or research effort from DePuy Synthes, Cerapedics, SI Bone, and ISSG.

Correspondence
Cormac O. Maher: cmaher@med.umich.edu.

INCLUDE WHEN CITING
DOI: 10.3171/2022.1.FOCUS21712.