Databases for Clinical Research

Katrina Abuabara[1] and David J. Margolis[1,2]

Journal of Investigative Dermatology (2014) **135**, e35. doi:10.1038/jid.2015.213

DATABASES FOR CLINICAL RESEARCH

The growing availability of digital health data offers many opportunities for clinical research. Studies drawing on electronic data are often efficient, although the usefulness and validity of the data depend on the research question. We briefly review types of epidemiologic study designs commonly used with patient databases and then describe the types of electronic databases available, outline considerations for the ad hoc design of new databases, and discuss potential limitations to consider when performing database research.

WHAT IS THE RESEARCH QUESTION?

Epidemiologic questions are often framed around an exposure and an outcome used to answer a predefined question. The exposure can be an environmental exposure, medication, risk factor, or disease state. For example, does isotretinoin (exposure) cause irritable bowel disease (outcome)? Or, is severe psoriasis (exposure) associated with an increased risk of myocardial infarction (outcome)? Outcomes may refer to the onset of a disease (incidence), presence of a disease (prevalence), or severity or duration of a disease or symptom. The research question and nature of the exposure and outcome variables should guide the choice of epidemiologic study design.

STUDY DESIGNS

Epidemiologic study designs can be broadly categorized into descriptive and analytical studies (Figure 1). Descriptive studies, such as case reports and case series, tend to be hypothesis generating, and they ask questions about what, where, who, and when. Alternatively, analytical studies tend to test hypotheses and answer questions about why and how. They include both experimental studies (i.e., clinical trials) and observational studies (cross-sectional, cohort, and case–control designs) (Vandenbroucke *et al*., 2007). Cross-sectional studies assess all individuals in a sample at the same time point; the downside is that they can't ascertain the temporality of events and therefore can't be used to draw conclusions about causation. Cohort and case–control designs follow individuals over time to ascertain the relationship between an exposure and an outcome and differ in terms of whether the population is selected based on the exposure (cohort study) or the outcome (case–control study). Study design selection should be guided by the suitability of the design for the research question at hand and by feasibility constraints. For a more complete description of epidemiologic study designs, we recommend an introductory textbook (Gordis, 2013). Electronic databases are most commonly used for observational studies, but they can also be used for experimental studies (e.g., randomizing an intervention for patients within an electronic medical record (EMR)) or descriptive studies (searching an EMR for a case series).

WHAT DATABASES FOR CLINICAL RESEARCH DO

- Electronic health data are increasingly available and can offer an efficient means for clinical research if used appropriately given the data source's limitations.
- The usefulness and validity of the data depend on the research question.

LIMITATIONS

- Imprecision, potential biases (including information bias and selection bias), and generalizability of the results are all limitations of research conducted using databases.
- Currently, databases are generally more useful for studying the incidence or prevalence of a dermatologic disease than for studying disease resolution or changes in disease severity over time.

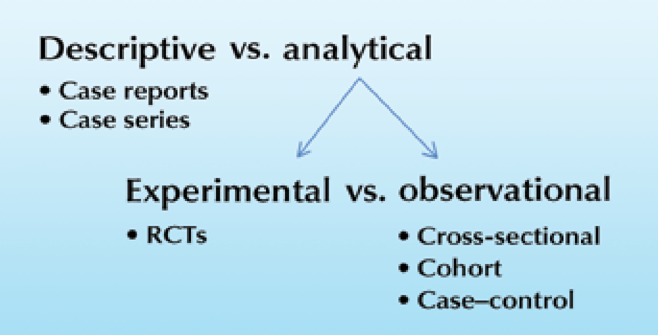

Figure 1. Common epidemiologic study designs. RCT, randomized controlled trial.

TYPES OF ELECTRONIC DATABASES

Electronic databases can be categorized by the source of the

[1]Department of Dermatology, University of Pennsylvania, Philadelphia, Pennsylvania, USA and [2]Department of Biostatistics and Epidemiology, University of Pennsylvania, Philadelphia, Pennsylvania, USA

Correspondence: Katrina Abuabara, Department of Dermatology, 1468 Penn Tower, One Convention Avenue, Philadelphia, Pennsylvania 19103, USA.
E-mail: katrina.abuabara@uphs.upenn.edu

Table 1. Categories of electronic databases

Category	Examples (with selected dermatology-specific references)
Repurposed data	
Claims data	
Government insurers	US Medicare, US Medicaid, national health insurers (Huang *et al*., 2012)
Commercial insurers	United HealthCare, Pharmetrics (Arellano *et al*., 2007), Humana, Aetna
Electronic medical record data	UK general practice research databases (Gelfand *et al*., 2009; Langan *et al*., 2012); institution-specific databases (e.g., Kaiser Permanente, Veterans Affairs Computerized Patient Record System)
Registry data	Surveillance, Epidemiology, and End Results (Linos *et al*., 2009); Swedish Family Cancer Database (Chen *et al*., 2014)
Ad hoc data	Pediatric Eczema Elective Registry (Mockenhaupt *et al*., 2008); EuroSCAR (Mockenhaupt *et al*., 2008)
Hybrid data	Nurses' Health Study, National Health Interview Survey, Veterans Affairs Million Veteran Program, HMO Research Network, PatientsLikeMe

data (Table 1). One major distinction is whether the data are repurposed (that is, originally generated for purposes other than clinical research) or the result of an ad hoc design specific to an individual study. Data that can be used for clinical studies but that were originally designed for a different research question fall somewhere in between and are referred to as "hybrid data."

Repurposed data

Repurposed data include both administrative claims data, which are generated for billing purposes, and EMR data, which are generated for the purposes of patient care. Additionally, repurposed data include public health registry data such as cancer registries like the Surveillance, Epidemiology, and End Results (SEER) program in the United States and death registries that may be linked to claims or EMRs.

Administrative claims data. Administrative claims data include inpatient and outpatient medical record codes, and these may be linked with pharmacy prescriptions and laboratory values. They often contain limited demographic and risk factor information, and they may have variable follow-up, especially in the United States, because patients frequently change insurers. Claims have been widely used in health services research and pharmacoepidemiology, and they are best suited to study outcomes that are easily captured by diagnostic codes such as procedures or acute events.

EMR Data. EMR data are essentially paperless, digital versions of patient charts generated for the purposes of clinical care. The number of office-based practices and hospitals using EMR systems is increasing, yet there is a lack of standardization and interoperability. Like claims data, EMR data are best suited to study outcomes that are easily captured by diagnostic codes, yet they may offer the possibility of more detailed data via manual review of physician notes or natural language processing systems. They are also likely to lack routinely collected social and behavioral variables, although efforts are underway to improve collection of these types of data (Adler and Stead, 2015). Some EMRs may be representative of the general population and capture all of a patient's health-care interactions, such as the Clinical Research Practice Datalink or the Health Improvement Network, both large UK general-practice research databases. Others may include only inpatient or specialty patient care.

Ad hoc data

Ad hoc data are generally designed for a particular study, and they often take the form of a prospective cohort study in which patients are selected for inclusion on the basis of a particular diagnosis or exposure. For this reason, they are often disease specific and may lack a control group.

Hybrid data

Large prospective cohort studies such as the Framingham Heart Study and the Nurses' Health Study may be considered hybrids between repurposed and ad hoc data because

Table 2. Considerations for the design of a patient database

Consistent data collection	Provide clear, operational definitions of data elements. Create and distribute standard instructions to data collections. Use standardized data element definitions and/or data dictionaries whenever possible—review the literature to identify existing, widely used definitions before drafting new definitions
Systematic patient enrollment and follow-up	Enroll patients systematically and follow them in as unbiased a manner as possible, using similar procedures at all participating sites. Describe how patients and providers were recruited into the study. Monitor and minimize loss to follow-up. Develop a patient retention plan that documents when a patient will be considered lost to follow-up and what actions will be taken to minimize such loss
Data quality assurance	Create structured training tools for data abstractors. Perform data quality checks for ranges and logical consistency for key exposure and outcome variables
Data safety and security	Provide transparency by describing data use agreements, informed consent, data security, and approaches to protecting security including risk of identification of patients

Adapted from PCORI (2013).

they represent large amounts of data that have been used to test many hypotheses beyond the one they were originally designed to study. Similarly, national survey data are available that typically offer cross-sectional snapshots of patient-reported risk factors and health outcomes such as the National Health Nutrition and Examination Survey (NHANES) or the National Health Interview Survey. Finally, a newer type of electronic health data is becoming available via crowdsourcing (Ranard *et al*., 2014; Wicks *et al*., 2010). These data are particularly useful for rare exposures or outcomes, but they are prone to selection bias and information bias because they rely on patient self-reporting.

DESIGN OF A NEW AD HOC PATIENT DATABASE

The Patient-Centered Outcomes Research Initiative has outlined a number of general considerations for the design of a patient database or registry (Table 2) (PCORI, 2013). Another important consideration concerns patient selection. If possible, researchers should enroll all individuals who meet the case definition (or a random selection of these individuals) to ensure the external validity or generalizability of the results. They may also consider using incident cases to help differentiate between exposures prior to and after the onset of disease. Finally, researchers should carefully consider whether a comparison group will be enrolled and strive to ensure that the comparison group is randomly selected from a comprehensive listing of the target population.

DISCUSSION

In using electronic data, several potential limitations must be considered, including imprecision, potential sources of bias, and the generalizability of the results.

Imprecision may arise from the study size or from the measurement of exposures, confounders, or outcomes. A variety of strategies, including detailed chart review and physician query, may be used to evaluate the validity of measurements in electronic databases. Dermatologic outcomes, in particular, tend to be less conducive to precise measurement in electronic databases because few diagnoses are based on routinely collected data. A researcher studying hypertension, for example, is likely to find more standardized data than a researcher hoping to study changes in acne lesion counts. Therefore, current databases are generally more useful for studying the incidence or prevalence of dermatologic disease than for studying disease resolution or changes in disease severity over time. Standardization of outcome scales and/or photographic assessments at regular intervals offer potential for improvement.

Bias is a systematic deviation of a study's result from a true value. Typically, it is introduced during the design of a study from flawed information or subject selection. There are many types of bias; two that may be particularly relevant to database studies include information bias and selection bias. Information bias occurs where there are systematic differences in the accuracy or completeness of data leading to differential misclassification of individuals regarding exposures or outcomes. For example, patients with a family history of mel-

QUESTIONS

This article has been approved for 1 hour of Category 1 CME credit. To take the quiz, with or without CME credit, follow the link under the "CME ACCREDITATION" heading.

For each question, more than one answer may be correct.

1. Which of the following is not an example of an analytical study?

A. Cross-sectional study.

B. Case series study.

C. Case–control study.

D. Cohort study.

2. Which of the following is the most important consideration when choosing a database for clinical research?

A. Whether the data come from an electronic medical record.

B. Loss to follow-up.

C. The research question.

D. The generalizability of the population.

3. Which of the following is *not* a recommendation regarding patient enrollment and follow-up in a new longitudinal database (sometimes called a patient registry)?

A. Use similar procedures at all sites.

B. Use validated outcome measures whenever possible.

C. Never obtain emergency contact information.

D. Develop a patient retention plan.

4. Which of the following is an example of information bias?

A. Patients with a family history of melanoma are more likely to be recruited into a study.

B. Patients with a family history of melanoma are more likely to receive skin checks and biopsies.

C. Patients with a family history of melanoma are more likely to know whether they have a history of atypical moles.

D. Patients with a family history of melanoma are more likely to use sunscreen.

5. Which of the following is an example of selection bias?

A. Eczema diagnostic codes are sometimes used for contact dermatitis.

B. Eczema patients with severe disease are more likely to have herpes cultures sent.

C. Eczema patients with mild disease are less likely to see their providers.

D. Mothers of children with bad eczema are more likely to recall infections during the perinatal period.

anoma may be more likely to receive skin checks and biopsies, making it appear that they have higher rates of atypical nevi. One potential way to assess the influence of some types of information bias is to measure the intensity of medical surveillance in the different study groups and to adjust for this in statistical analyses. Selection bias may be introduced if the probability of including subjects in the study (or probability of subjects being lost to follow-up) is associated with exposure or outcome. For example, a study of patients followed in clinics may overestimate the severity of a disease because patients with mild disease who seek medical advice less frequently are underrepresented. Selection bias affects the internal validity of a study, but it is often related to the external validity or generalizability of the results.

Generalizability refers to how representative the results from the study population are to the general population. Studies that only enroll patients from tertiary-care centers or that only include patients with particular demographic characteristics may have limited generalizability.

The ideal database depends on the research question. Generically speaking, an ideal database might include linked records from inpatient and outpatient care, emergency care, mental-health care, all laboratory and radiological tests, and all prescribed and over-the-counter treatments, as well as alternative therapies. The population would be large enough to permit discovery of rare events and interactions, would be stable over time, and would be representative of the general population from which it was drawn. It would include genetic, social, and physiologic information on all members, and there would be the ability to gather additional information, either from physicians or from patients themselves, to confirm outcomes.

CONFLICT OF INTEREST

The authors state no conflict of interest.

CME ACCREDITATION

This activity has been planned and implemented in accordance with the Essential Areas and Policies of the Accreditation Council for Continuing Medical Education through the joint sponsorship of the Duke University School of Medicine and Society for Investigative Dermatology. The Duke University School of Medicine is accredited by the ACCME to provide continuing medical education for physicians. To participate in the CME activity, follow the link provided. Physicians should only claim credit commensurate with the extent of their participation in the activity.

To take the online quiz, follow the link: http://continuingeducation.dcri.duke.edu/research-techniques-made-simple-journal-based-cme-rtms

SUPPLEMENTARY MATERIAL

A PowerPoint slide presentation appropriate for journal club or other teaching exercises is available at http://dx.doi.org/10.1038/jid.2015.213.

REFERENCES

Adler NE, Stead WW (2015) Patients in context—EHR capture of social and behavioral determinants of health. *N Engl J Med* 372:698–701

Arellano FM, Wentworth CE, Arana A *et al*. (2007) Risk of lymphoma following exposure to calcineurin inhibitors and topical steroids in patients with atopic dermatitis. *J Invest Dermatol* 127:808–16

Chen T, Fallah M, Kharazmi E *et al*. (2014) Effect of a detailed family history of melanoma on risk for other tumors: a cohort study based on the nationwide Swedish Family-Cancer Database. *J Invest Dermatol* 134:930–6

Gelfand JM, Dommasch ED, Shin DB *et al*. (2009) The risk of stroke in patients with psoriasis. *J Invest Dermatol* 129:2411–8

Gordis L (2013) Epidemiology, 5th edn. Elsevier/Saunders: Philadelphia, 398 pp

Huang YH, Kuo CF, Chen YH *et al*. (2012) Incidence, mortality, and causes of death of patients with pemphigus in Taiwan: a nationwide population-based study. *J Invest Dermatol* 132:92–7

Langan SM, Groves RW, Card TR *et al*. (2012) Incidence, mortality, and disease associations of pyoderma gangrenosum in the United Kingdom: a retrospective cohort study. *J Invest Dermatol* 132:2166–70

Linos E, Swetter SM, Cockburn MG *et al*. (2009) Increasing burden of melanoma in the United States. *J Invest Dermatol* 129:1666–74

Mockenhaupt M, Viboud C, Dunant A *et al*. (2008) Stevens–Johnson syndrome and toxic epidermal necrolysis: assessment of medication risks with emphasis on recently marketed drugs. The EuroSCAR-study. *J Invest Dermatol* 128:35–44

PCORI (2013) Patient-Centered Outcomes Research Institute Methodology Committee. Research methodology. http://www.pcori.org/content/research-methodology

Ranard BL, Ha YP, Meisel ZF *et al*. (2014) Crowdsourcing—harnessing the masses to advance health and medicine, a systematic review. *J Gen Intern Med* 29:187–203

Vandenbroucke JP, von Elm E, Altman DG *et al*. (2007) Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. *PLoS Med* 4:e297

Wicks P, Massagli M, Frost J *et al*. (2010) Sharing health data for better outcomes on PatientsLikeMe. *J Med Internet Res* 12:e19