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.

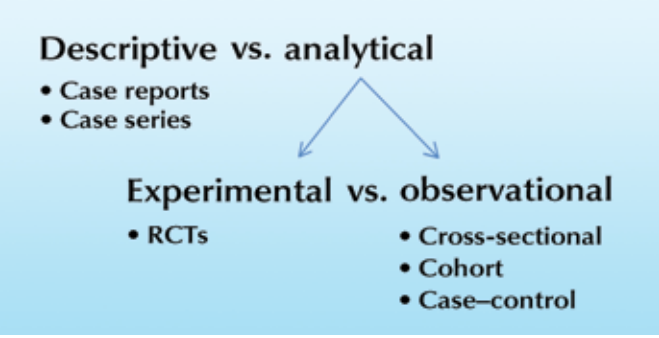

Figure 1. Common epidemiologic study designs. RCT, randomized controlled trial.

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).