Initial analysis of a large database: An investigator's perspective

Shengping Yang PhD, Gilbert Berdine MD

I am planning to investigate the longitudinal relationship between childhood obesity and cardiovascular diseases (CVD). Given the nuances of this relationship – involving life-course epidemiology, evolving diagnostic criteria, and complex confounders – I plan to analyze a large existing database. Any suggestions on what one needs to be careful about?

Using large existing databases can be a powerful approach, but it requires careful consideration of several conceptual, methodological, and practical issues.

1. Introduction

Before undertaking analyses, it is essential to understand how these databases were established and to be familiar with the major databases available within the relevant field. Such preliminary understanding helps investigators identify resources appropriate for their specific study objectives. In particular, it facilitates the assessment of measurement validity, the generalizability of findings, the suitability of analytic strategies, and the ethical and legal constraints governing data use. Critically, a database's origin and original purpose often have greater implications for research validity than its size alone.[1]

1.1 Types of biomedical databases

Large biomedical databases are systematically organized collections of health-related data derived from population-based studies, healthcare systems, or research cohorts. These databases are typically established to support:

- Epidemiologic research
- Clinical and health services research
- Public health surveillance
- Quality improvement
- Precision medicine

They often integrate biological, clinical, behavioral, and demographic data, collected either prospectively or retrospectively and, in many cases, longitudinally. Data originating from electronic health records (EHRs) represent a major and increasingly important component, capturing information from routine clinical care across large and diverse populations.

Corresponding author: **Shengping Yang**
Contact Information: Shengping.Yang@pbrc.edu
DOI: 10.12746/swjm.v14i58.1643

1.2 Database creation

These databases are generally established through one (or a combination) of the following pathways.

1.2.1 Population-based cohort studies

These databases are established by recruiting participants from defined populations and following them longitudinally with standardized data collection. Examples include the Bogalusa Heart Study and the Framingham Heart Study.

1.2.2 Healthcare system-derived databases (administrative & EHR)

These databases are derived from routine clinical care and billing systems. Examples include Kaiser Permanente EHR databases and the National Patient-Centered Clinical Research Network (PCORnet).

1.2.3 Disease- or condition-specific registries

These databases focus on patients with specific diseases or conditions. Examples include the Surveillance, Epidemiology, and End Results (SEER) Program, a population-based cancer registry, and the Pediatric Cardiac Quality Improvement Collaborative.

1.2.4 Biobanks and omics-integrated databases

These databases combine biological specimens with clinical and phenotypic data. Examples include the UK Biobank and the All of Us Research Program.

1.2.5 Surveillance and public health monitoring systems

These databases are designed for population-level monitoring and are often cross-sectional or repeated cross-sectional. Examples include the National Health and Nutrition Examination Survey (NHANES) and the Behavioral Risk Factor Surveillance System (BRFSS).

For a study of childhood obesity and CVD, an ideal database would be a longitudinal cohort or biobank with repeated anthropometric measures from childhood, long-term follow-up into adulthood, and validated CVD outcome assessments.

2. Data Quality, Validation, and the "Ground Truth"

Before advanced statistical models or causal inferences are attempted, investigators must ask a fundamental question: *What does the data truly represent?* When working with large existing databases, it is therefore necessary to move beyond a mindset of simple data cleaning toward rigorous data auditing and validation.

This stage is not about extracting answers but about determining whether the database can credibly support the research questions being asked. Establishing the most accurate possible representation of the underlying data-generating process – the closest feasible approximation to *ground truth* – is essential before conducting any hypothesis-driven analyses.[2]

2.1 Defining the investigative objective

Before examining the database itself, investigators must clearly articulate the objective of the analysis. This includes defining exposure, e.g., "childhood obesity," the outcome, e.g., "incident CVD," the key confounders, and what constitutes a clinically meaningful result.

The first interaction with a large database should focus on orientation rather than inference. Key questions include the scale of the database, its structure, the unit of analysis, and whether the data are relational, longitudinal, or transactional. Equally important is understanding data provenance: who collected the data, for what purpose, and under what operational or regulatory constraints. Data should be viewed as a form of evidence, and evidence cannot be separated from the process that produced it.

2.2 Understanding the data-generating process (DGP)

In biomedical research, the DGP is typically a complex combination of biological mechanisms, clinical workflows, human decision-making, and technological systems. Understanding this process requires investigators to ask context-specific questions such as:

- Was a child's BMI percentile calculated from a measured height and weight during a well-child visit, or was it self-reported?
- If a sudden spike is observed in a laboratory value, does it reflect a true physiological event or a systematic artifact, such as a batch effect due to changes in reagents?
- Was the data captured automatically by a bedside monitor, or manually transcribed hours later by clinical staff, introducing potential recall or transcription errors?

Failure to understand the DGP can lead to confident but fundamentally flawed conclusions.

2.3 Proxies, triangulation, and reconstructing "Ground Truth"

In many large databases, clinical states are often not directly observed but represented through proxies. For example, an ICD-10 billing code for obesity reflects administrative documentation rather than a definitive clinical diagnosis. A patient may carry the

code for reimbursement purposes while not meeting formal diagnostic criteria.

Approximating ground truth therefore requires triangulation across multiple data sources. A robust definition of "childhood obesity exposure" might require:

- Diagnosis codes for obesity.
- Anthropometric data: BMI percentiles derived from recorded height/weight.
- Medication records: Prescriptions for weight management.
- Natural Language Processing (NLP): Keywords in clinical notes.

Similarly, a "CVD outcome" might combine diagnosis codes, procedure codes (e.g., stent placement), medication initiations (e.g., anticoagulants), and elevated cardiac biomarkers. This approach to outcome validation is a cornerstone of valid research using real-world data.[3]

2.4 Missing data as information

Standard data cleaning approaches often treat missing data as a technical nuisance to be corrected. Rigorous data auditing instead treats missing data as a potential source of information.

Missing values may reflect informative censoring – for example, missing follow-up BMI data because an obese adolescent disengaged from the healthcare system, a potential risk factor for later CVD. Alternatively, missing data may arise from structural issues (e.g., a clinic changed EHR systems). Data are often not missing at random, and naive handling (e.g., complete-case analysis) can bias results by systematically excluding vulnerable subpopulations.[4]

2.5 Data integrity, provenance, and structural validation

Beyond individual variables, investigators must assess overall data integrity. Discrepancies between data dictionaries and actual values are common and may signal quality issues. Manually entered data are susceptible to human error, while automatically captured data may suffer from sensor malfunctions or software glitches.

Structural validation should be performed to confirm that:

- The unit of analysis is correctly understood.
- Measurement units are consistent (e.g., inches vs. centimeters)
- Duplicate records are identified and removed.
- Temporal logic is preserved (e.g., admission dates preceded discharge dates)

Exploratory data analysis at this stage is used not to identify associations, but to detect anomalies, implausible values, and internal inconsistencies. Multivariate visualizations, such as scatter plots between biologically related measures, are often more informative than univariate summaries.

2.6 Documentation, ethics, and reproducibility

Throughout the initial analysis, careful documentation is essential. Assumptions, exclusions, data limitations, and analytical decisions should be recorded systematically to support transparency and reproducibility. Analyses should be conducted on a frozen snapshot of the database, especially for data with ongoing data collection, to prevent results from shifting as data are updated.

All data auditing and validation procedures must be conducted in compliance with relevant regulatory frameworks, such as the Health Insurance Portability and Accountability Act (HIPAA) in the United States or the General Data Protection Regulation (GDPR) in the European Union. Analyses should be performed within secure institutional environments, and raw data should never be altered directly. Any identified systemic data issues should be communicated to data providers to support future data quality improvements.

3. Data Standardization, Harmonization, and Scalability

Following rigorous data auditing and validation, it is often necessary to perform data standardization

and harmonization, particularly when working with multi-source data.

To enable large-scale analyses, databases commonly rely on:

- Standardized coding systems, such as ICD, SNOMED and LOINC.
- Common data models, such as OMOP and PCORnet.
- Harmonization protocols, which align variables across cohorts, institutions, or study waves.

While these frameworks are essential for integrative analyses, they may introduce interpretive uncertainty. Differences in coding practices, measurement protocols, or clinical workflows can persist even after harmonization, potentially affecting comparability across data sources. Initiatives like the OHDSI/OMOP network provide frameworks but still require careful application to specific research questions.[5]

In addition, large biomedical databases impose computational and scalability constraints that influence analytic choices. Investigators must consider whether data can be processed in memory or require database-backed or distributed computing approaches. Maintaining complete audit trails for all data transformations is essential for transparency and reproducibility.

4. Statistical Power Consideration

In the context of large biomedical databases, the traditional concern – "Is my sample size too small?" – often changes to "How do I interpret findings from this massive sample?" With data from hundreds of thousands, the risk shifts from Type II error (missing a real effect) to Type I error (inflation and spurious significance).

4.1 Statistical power and effect size

With N > 100,000, even biologically trivial differences (e.g., a 0.2 kg/m² difference in childhood BMI) can yield $p < 0.001$. Therefore, the investigative focus must shift decisively from statistical significance to clinical or public health significance, as judged by the magnitude of the effect size (e.g., Hazard Ratio, Odds Ratio, Cohen's d). Pre-specifying minimal clinically important difference thresholds for key outcomes is crucial.

4.2 Sources of noise and heterogeneity

Large databases introduce unique challenges that can increase variance:

- Heterogenous population: Unlike randomized trials, large databases collect data from very heterogeneous groups, often including different age groups, health conditions, race/ethnicities, etc. Therefore, data variation can be substantially higher than that from a small trial with a focused population.
- Measurement Inconsistency: Data from different sources may use different devices, assays, or protocols (e.g., various scales and stadiometers for height/weight), adding non-differential misclassification and noise.

4.3 The "Effective" sample size

In large databases, the "analytic" sample size is rarely equal to the number of subjects initially enrolled. For example, for a datasets with a million subjects, after excluding records with missing laboratory values, insufficient follow-up time, or unconfirmed diagnoses, only 50,000 subjects may have valid data. In addition, when subgroup analyses are performed, some subgroups may contain unexpectedly small numbers of subjects, resulting in lower-than-anticipated statistical power. The missing data may not be random or uniformly distributed with respect to important variables introducing a selection bias and leading to invalid conclusions.

5. Common Limitations and Pitfalls

Several pitfalls could render the analysis of a large dataset invalid:

- Overreliance on data volume – neglecting data quality and context.
- Using simplistic proxies (e.g., a single obesity code) without validation or triangulation, leading to biased effect estimates.
- Confounding factors might not be accounted for – for example, obese children may have more clinical encounters, increasing the likelihood of detecting incidental findings or being tested for conditions like dyslipidemia, creating a false association with CVD.
- Time-Varying Confounding: Certain risk factors can change over time, which are not appropriately modeled.
- Selectively focusing on patterns that align with prior expectations or conducting unrestricted exploratory analyses without hypothesis testing, increasing false discovery rates.[6]
- Applying complex machine learning or causal inference models before understanding the basic structure and flaws of the data.

Avoiding these pitfalls requires disciplinary humility, a willingness to revise assumptions, and a methodology that prioritizes understanding over speed.

Initial analysis of a large database is both a technical and intellectual exercise. For investigators and data analysts, it represents the foundation upon which all subsequent findings rest. By prioritizing understanding over speed and context over complexity, analysts can ensure that their conclusions are grounded, defensible, and meaningful.

Article citation: Yang S, Berdine G. Initial analysis of a large database: An investigator's perspective. The Southwest Journal of Medicine. 2026;14(58):89–93
From: Department of Biostatistics, Pennington Biomedical Research Center, Baton Rouge, LA (SY) Department of Internal Medicine, Texas Tech University Health Sciences Center, Lubbock, Texas (GB)
Conflicts of interest: none

References

1. Wang W, Liu M, He Q, et al. Data source profile reporting by studies that use routinely collected health data to explore the effects of drug treatment. BMC Med Res Methodol. 2023 Apr 20;23(1):95. doi: 10.1186/s12874-023-01922-8. PMID: 37081410; PMCID: PMC10120171.
2. Ground-truth data cannot do it alone. Nat Methods. 2011 Nov;8(11):885. doi: 10.1038/nmeth.1767. PMID: 22148151.
3. Carter N, Bryant-Lukosius D, DiCenso A, et al. The use of triangulation in qualitative research. Oncol Nurs Forum. 2014 Sep; 41(5):545–7. doi: 10.1188/14.ONF.545-547. PMID: 25158659.
4. Cole SR, Zivich PN, Edwards JK, et al. Missing outcome data in epidemiologic studies. Am J Epidemiol. 2023 Jan 6; 192(1):6–10. doi: 10.1093/aje/kwac179. PMID: 36222655; PMCID: PMC10144620.
5. Reinecke I, Zoch M, Reich C, et al. The Usage of OHDSI OMOP – A Scoping Review. Stud Health Technol Inform. 2021 Sep 21;283:95–103. doi: 10.3233/SHTI210546. PMID: 34545824.
6. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society: Series B (Methodological) 1995;57(1):289–300.