2026 Data Challenge: Vital Statistics

DAIR^3

Jackson State U · U Michigan · UT San Antonio

DAIR^3 Team

Faculty:
Clifton Addison, Associate Professor of Biostatistics, Jackson State University
Yalanda Barner, Assistant Professor of Health Policy and Management, Jackson State University
Johann Gagnon-Bartsch, Associate Professor of Statistics, University of Michigan
Juan B. Gutiérrez, Principal Investigator; Professor of Mathematics, University of Texas at San Antonio
Gregory Hunt, Assistant Professor of Mathematics, College of William and Mary
Brenda Jenkins, Director of Training and Education, Jackson State University
Erin Kaleba, Director, Data Office for Clinical and Translational Research, University of Michigan
Jing Liu, Principal Investigator; MIDAS Executive Director, University of Michigan
Jodyn Platt, Associate Professor of Learning Health Sciences; Associate Professor of Health Management and Policy, University of Michigan Medical School
Suraj Rampure, Lecturer III in Electrical Engineering and Computer Science, University of Michigan
Kerby Shedden, Professor of Statistics; Professor of Biostatistics, University of Michigan
Teaching Assistants:
Taofeq... complete
Sim... complete
Admisnitrative Support:
Kelly Psilidis, Faculty Training Program Manager, University of Michigan
Michele Randolph, Evaluation Specialist, Marsal School of Education, University of Michigan

Document prepared by: Juan B. Gutiérrez, Ph.D. Professor of Mathematics, juan.gutierrez3@utsa.edu

The University of Texas at San Antonio

February 17, 2026

Contents

Chapter 1

2026 Data Challenge - Vital Statistics

1.1 Problem Description

Your team is part of a consultancy supporting a planning commission for the State of Texas. The commission is planning budgetary requirements for various State services in 2030. The commission requests the following:

- Projections of underweight newborns by county in Texas.
- Projections of newborn mortality by county in Texas. According to the CDC, a stillbirth is classified as either early, late, or term. An early stillbirth is a fetal death occurring between 20 and 27 completed weeks of pregnancy. A late stillbirth occurs between 28 and 36 completed pregnancy weeks. A term stillbirth occurs between 37 or more completed pregnancy weeks.
- Identification of the socioeconomic factors associated with these two outcomes.
- Comparison to other states.

Your team will produce an executive report accompanied by an appendix of technical material.

1.2 Data Description

The commission has approved specific data sources for this analysis. You can only use approved data available at this hyperlink.

National Center for Health Statistics (NCHS): Since 1969, all births recorded in the US are available as digital records from the NCHS, from the Centers for Disease Control and Prevention (CDC). The date and time of birth is publicly available only between 1969 and 1988; starting in 1989, only the week of birth is recorded. The NCHS keeps records of place of birth, assistance during delivery (at home, with doctors, with midwives), level of education of the parents, place of residence, weight at birth, number of weeks of gestation, number of siblings, birth order, etc. In total, over 100 variables are recorded. You have access to records form 1969 to 1988.

Surveillance, Epidemiology, and End Results Program (SEER): The U.S. Census Bureau annually releases unabridged population estimates for five-year age groups and race at the county level. The Census Bureau does not release bridged race estimates by single year of age at the county level due to concerns about the reliability of these estimates. However, these estimates are provided to the National Cancer Institute through SEER to meet programmatic needs such as the creation of age groupings that differ from the standard groupings used by the Census Bureau. Users of the single-year-of-age county-level interpolated race population estimates should carefully consider the limited reliability of these estimates. County-level population files with 19 age groups (<1, 1-4, ..., 80-84, 85+) and with 86 single-year age groups (<1, 1, 2, ..., 84, 85+) are provided.

Socioeconomic Data and Applications Center (SEDAC): SEDAC, the Socioeconomic Data and Applications Center, is one of the Distributed Active Archive Centers (DAACs) in the Earth Observing System Data and Information System (EOSDIS) of the U.S. National Aeronautics and Space Administration (NASA). SEDAC contains georeferenced U.S. county-level population projections, total and by sex, race and age, based on shared socioeconomic pathways (SSPs). This data set, produced by Mathew E. Hauer, consists of county-level population projection scenarios of population in five-year intervals for all U.S. counties for the period 2020 - 2100. Obtain the data description from https://doi.org/10.1038/sdata.2019.5.

The NCHS data set covers from 1969 through 1986; it provides individual birth data. The SEER data set provides age-bracketed population estimates from 1969 to 2020. Since for this exercise we do not have birth data beyond 1986, you will have to use the SEER data to infer births in the period 1987-2020; alternatively, you could go to the NCHS source and obtain more recent data, however the NCHS data is complex and downloading additional years is not advised in the short time available to complete the Rowdy Datathon (but we will not stop you). The SEDAC data has total population estimates per county from 2020 through 2100 categorized in four ethnicities: Hispanic, white, black, and other. A viable sequence of analysis is NCHS → SEER → SEDAC.

1.3 Identifying Underweight Cutoff Points

The CDC offers a data table of infant weight for age, available at:
https://www.cdc.gov/growthcharts/data/zscore/wtageinf.xls
You can extract from this table the information related to weight at birth.

The Excel file `wtageinf.xls` contains the LMS parameters and selected percentile values needed to compute exact percentiles and z-scores for anthropometric measurements. The parameters are provided by sex (1 = male, 2 = female) and by single month of age. Age is listed at the half-month point representing the entire month (e.g., 1.5 months corresponds to 1.0–1.99 months). The only exception is birth, which represents the point at birth.

The LMS method summarizes the distribution of a measurement at a given age using three parameters:

- L: the Box–Cox power transformation,
- M: the median,
- S: the generalized coefficient of variation.

These parameters allow computation of both percentiles and z-scores.

To obtain the measurement value X corresponding to a given z-score Z (or percentile), use:

$$X = M(1 + LSZ)^{1/L}, \quad L \neq 0$$

$$X = M\exp(SZ), \quad L = 0$$

where L, M, and S are taken from the Excel row corresponding to the child's age (in months) and sex. The z-score corresponding to common percentiles is:

$$\begin{aligned} -1.881 &\leftrightarrow 3^{\text{rd}} \\ -1.645 &\leftrightarrow 5^{\text{th}} \\ -1.282 &\leftrightarrow 10^{\text{th}} \\ -0.674 &\leftrightarrow 25^{\text{th}} \\ 0 &\leftrightarrow 50^{\text{th}} \end{aligned}$$

Example: For a 9-month-old male, the `WTAGEINF` table gives:

$$L = -0.1600954, \quad M = 9.476500305, \quad S = 0.11218624.$$

Using $Z = -1.645$ (5th percentile), the cutoff is:

$$X = 7.90 \text{ kg}.$$

This value represents the 5th percentile weight-for-age and can be used as an underweight threshold if the 5th percentile definition is adopted.

Computing a Z-Score from a Measurement: To obtain the z-score corresponding to a given measurement X:

$$Z = \frac{(X/M)^L - 1}{LS}, \quad L \neq 0$$

$$Z = \frac{\ln(X/M)}{S}, \quad L = 0.$$

The corresponding percentile is then obtained from the standard normal distribution.

Example: For a 9-month-old male weighing 9.7 kg:

$$Z = 0.207,$$

which corresponds approximately to the 58th percentile.

Identifying Underweight Cutoff Points in Excel: To identify an underweight threshold in the Excel file:

1. Filter the data by sex (1 = male, 2 = female).
2. Locate the row corresponding to the child's age in months.
3. Identify the desired cutoff percentile (e.g., 5th or 3rd percentile column).

4. Alternatively, compute the exact cutoff using the LMS formula above.

A child is classified as underweight if their measured value falls below the selected percentile cutoff (e.g., below the 5th percentile or below $Z = -2$, depending on the chosen clinical definition).

If finer age resolution is required, linear interpolation between adjacent months may be applied to the L, M, and S parameters prior to computation.

1.4 Data Files

Table 1.1: File sizes. If you have limited storage, plan your analysis in stages.

DATA SET	FILE	ZIPPED	UNZIPPED
NCHS	US1969-1986.zip	2.7 GB	22.8 GB
	natalityConfBackup_PostgreSQL.sql	2.42 GB	25 GB
	US1969.zip	32.6 MB	381 MB
SEDAC	hauer_county_NH_pop_SSPs.xlsx	N/A	15.1 MB
SEER	SEER Data Dictionary.pdf		73 KB
	tx.1969_2020.19ages.adjusted.txt.gz	5.3 MB	35 MB
	tx.1969_2020.singleages.adjusted.txt.gz	18.8 BM	19 MB
	tx.1990_2020.19ages.adjusted.txt.gz	6.5 MB	6 MB
	tx.1990_2020.singleages.adjusted.txt.gz	20.9 MB	21 MB
	us.1969_2020.19ages.adjusted.txt.gz	66.7 MB	430 MB
	us.1969_2020.singleages.adjusted.txt.gz	238.8 MB	1.6 GB
	us.1990_2020.19ages.adjusted.txt.gz	76.7 MB	520 MB
	us.1990_2020.singleages.adjusted.txt.gz	246.3 MB	1.8 GB
TOTAL		5.7 GB	52 GB

1.5 Criteria

Results will be evaluated according to requirements set by the commission:

- **Compelling presentation**: You must enable the commission to share your numerical and graphical results directly with legislators and citizens through executive summaries. This lay audience should find your summaries and implications to be understandable and convincing. Express your results in ways that can be acted on to plan e.g. funding of schools, care for the elderly, etc. The supporting documentation can be technical.

- **Analysis comprehension**: Before a single line of code is written, before a single byte of raw data is processed, you must be able to tell the story of what is the progression of steps that will be undertaken in analysis.
- **Sound technical methods**: You may cite the analyses of others, but the commission wants to see the methods that you have invented or adopted to calculate these projections (which should be accompanied by error bars, if possible). The commision must have confidence in your results in order to present those results to others.
- **Awareness of the data context**: All data have bias. Before, during and after analysis, it is essential to identify biases in the data and articulate clearly how these biases influence all steps of analysis and interpretation.
- **Reproducible results**: You must enable the commission to have your results confirmed by an independent team. That is, enable the independent team to replicate your results by describing your data and methods in detail.

Table 1.2: Evaluation Criteria

CRITERION	**% WEIGHT**
1. Compelling presentation - Informative	10%
2. Compelling presentation - Understandable	10%
3. Analysis comprehension	20%
4. Sound technical methods	20%
5. Awareness of the data context	20%
6. Reproducible results	20%

Chapter 2

Foundational Analysis Activities

2.1 A single year of vital statistics

A problem most people can relate to is demography. Millions of people are born in the US every year. Recording birth events is necessary for legal matters such as obtaining a driver's license or other forms of government-issued identification. However, recording, keeping, and using this information has challenges that exemplify many aspects of data analysis, as this exercise will demonstrate.

Since 1969, all births recorded in the US are available as digital records from the National Center for Health Statistics (NCHS) from the Centers for Disease Control and Prevention (CDC). Only between 1969 and 1988 the date and time of birth is publicly available; starting on 1989, only the week of birth is recorded. Why is the date and time of birth no longer recorded? Record your explanation as answer #1.

The NCHS keeps records of place of birth, assistance during delivery (at home, with doctors, with midwives), level of education of the parents, place of residence, weight at birth, number of weeks of gestation, number of siblings, birth order, etc. In total, over 100 variables are recorded.

I have made available two files for you: a compressed file with birth records from 1969, and a data dictionary. The uncompressed data file is about 380 MB in size. The data is contained in a "flat file". This means that every line of text in this file is a continuous chain of characters. We must extract information from this type of files with a "dictionary" that tells us the beginning and ending columns of a given variable. Why was this format used? Record your explanation as answer #2.

Please answer the following questions:

1. How many live births occurred in Texas in 1969 from mothers residing in Texas? Record your explanation as answer #3.

    - Bonus question: How would you visualize births from each state with respect to every other state?

2. Show graphically how the level of education of the mother is related to the birth order (1st born, second child, third, etc.) Record your explanation as answer #4.

    - Bonus question: How would you visualize each variable with respect to every other variable?

It is possible that you might not know how to answer some of these questions on first contact with this problem.

A simple program that can help you explore this file is

```
f = open("US1969.dat", "r", encoding="cp1252")
counter = 0;
for x in f:
    if x[25] == "7" and x[26] ==  "4": # other conditions?
        counter = counter + 1
print(counter)
```

The instruction *encoding="cp1252"* is necessary in non-Windows systems due the presence of single-byte character encoding of the Latin alphabet, used by default in the legacy components of Microsoft Windows for English and many European languages including Spanish, French, and German. If you remove this instruction, the following error might show up in non-Windows operating systems: "'utf-8' codec can't decode byte..."

A common approach to extract information from flat files is by importing it into Excel. The "*Text Import Wizard*" (typically shown as in Figure 2.1-1) would guide you through the process of identifying variables by column. However, this results in a problem. Describe it. Record your explanation as answer #5.

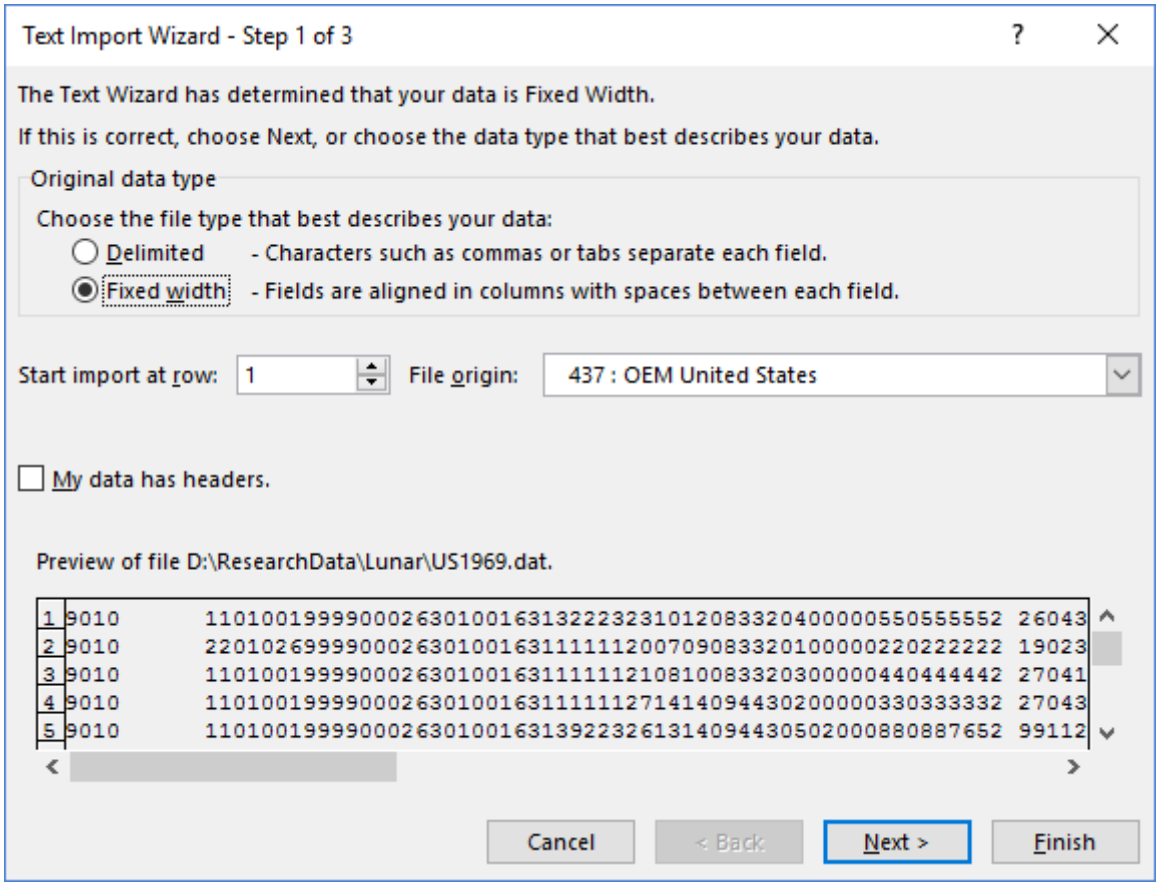

Figure 2.1-1: Excel's Import Wizard

Chapter 3

Unit 1: Responsible Conduct of Research

Total Time: 3 hours

3.1 RCR in the Context of Biomedical Data Science

This lesson examines the sociotechnical and ethical aspects of biomedical data science. We will consider ethical issues in the responsible conduct of research that are novel to or pose new challenges in the context of biomedical data science such as reproducibility and privacy. Students will also consider biomedical data science as a sociotechnical system and define roles for themselves and other key constituents.

3.1.1 Learning Objectives

1. Explain novel ethical issues in responsible conduct of research for data science such as reproducibility and privacy.
2. Describe the landscape of biomedical data science as a sociotechnical system and articulate roles.

3.1.2 Assessment Instrument

- Compare responses to the data challenge with your peers. What issues arise?
- Why was date and time no longer recorded after 1987?
- Who decides what data to collect, how to store it and how to access it? What biases could there be in the data (e.g., data collection in rural areas, existence of infrastructure)? What is the difference between bias and trend in these data? Provide examples.

3.2 What are Ethics? Ethical Issues in Biomedical Data Science

This lesson equips students to address ethical challenges in biomedical data science. Learners will identify strategies for ethical secondary data use, analyze engagement approaches, and develop frameworks for ethical project review, emphasizing anticipatory governance and responsible data science practices. Case studies will be used that draw on the group project selected for the 2026 cohort.

3.2.1 Learning Objectives

1. Differentiate between traditional bioethical, sociotechnical, and other ethical approaches to data science research and applications.
2. Evaluate key ethical challenges in biomedical data science.
3. Identify and formulate approaches to address ethical issues in secondary use, including anticipatory governance principles.
4. Develop a framework for ethical review of biomedical data science projects.

3.2.2 Assessment Instrument

The assessment involves designing a governance framework for a data science initiative, selecting one of three projects derived from the course-wide data activity. The framework must address stakeholder engagement, decision-making, monitoring, benefit-sharing, unexpected impacts, and consent, while evaluating ethical considerations, future challenges, and feasibility. The goal is to create an ethical, well-structured, and adaptable governance plan.

Chapter 4

Unit 2: Foundations of Data in Biomedical Research

Total Time: 7 hours

4.1 Data Management - Introduction to the Jackson Heart Study

Time: 3 hours

This lesson provides an overview of the Jackson Heart Study (JHS), focusing on its design, data collection, and variable interpretation. Students will describe the JHS's purpose, population, and exam structure, summarize clinical, survey, and genetic data collection methods across study phases, and learn to use JHS codebooks to identify variables for research questions.

4.1.1 Learning Objectives

1. Describe the JHS Study Design: Explain the purpose, population, and structure of the JHS, including its major exams.
2. Summarize Data Collection Methods: Identify the types of data collected (e.g., clinical, survey, genetic) and the methods used in different study phases.
3. Interpret Key Variables and Codebooks: Understand how to use JHS codebooks to find variables relevant to specific research questions.

4.1.2 Assessment Instrument

1. Describe in your own words the purpose of the JHS, cohort characteristics, and exam waves.
2. Describe how JHS collects specific data (e.g., CAC scores, lipid tests) and potential biases in data collection.

3. Answer the following question: "Association between hysterectomy and cardiovascular disease in the JHS, adjusting for covariates" by locating relevant variables in the JHS codebook. Describe the variables chosen and their rationale.

4.1.2.1 The Process of Manuscript Development in the Jackson Heart Study

This lesson covers the process for requesting and obtaining Jackson Heart Study (JHS) data. Students will learn to navigate data access procedures, including submitting manuscript or ancillary study proposals, completing data use agreements, and addressing ethical considerations to ensure responsible use of JHS data.

Learning Objective: Explain Data Access Procedures: Describe the process for requesting and obtaining JHS data, including data use agreements and ethical considerations.

Assessment:

1. In your own words, enumerate the steps involved in the process for developing a JHS manuscript.

2. Using the information acquired from the lecture, draft a mock JHS manuscript proposal, using the Manuscript Proposal Form provided and the sample manuscript proposal.

4.2 Metadata - Data About Data

Time: 0.75 hours

This lesson explores the essentials of metadata in biomedical research datasets, covering data collection methods, population, and context. Students will learn to identify high-quality metadata that supports reproducibility and distinguish it from inadequate metadata, ensuring robust and reliable research outcomes.

4.2.1 Learning Objectives

1. Understand standard components of metadata on biomedical science research datasets: how the data were collected, on what population, under what circumstances, etc.

2. Learn to distinguish between good and bad metadata for reproducibility.

4.2.2 Assessment Instrument

- List 5 critical metadata categories that researchers need to know when reproducing findings.

- From the Mathew E. Hauer article called, "Data Descriptors: Population projections for U.S. Counties by age, sex, and race," what metadata are provided about how the population data were collected?

4.3 Data Representation

Time: 0.75 hours

This lesson examines how data can be represented in multiple ways, highlighting that each representation impacts task efficiency. Students will learn to select optimal data representations tailored to specific research tasks, balancing ease and complexity.

4.3.1 Learning Objectives

1. Understand that the same data can be represented in many ways.
2. Appreciate that each representation choice makes some tasks easier, but others more difficult.
3. Learn how to choose a good representation for the task at hand.

4.3.2 Assessment Instrument

The NCHS data are provided as a flat file with more than 100 variables. What is an alternative representation of these same data? Is the original flat file or your alternative schema more conducive to analyses, and why?

4.4 Data Sharing

Time: 2.5 hours

4.4.0.1 Data Sharing 101

This lesson introduces the principles of Open Science, focusing on the NIH Data Management & Sharing Policy's rationale and key components. Students will explore the FAIR Guiding Principles (Findable, Accessible, Interoperable, Reusable), learning their definitions and practical examples to promote transparent and reproducible research.

Learning Objectives:

- Appreciate the foundations of Open Science
- NIH Data Management & Sharing Policy: Rationale and Key components
- FAIR Guiding Principles: Definition and examples

Assessment: Define rationale behind NIH Data Management & Sharing requirement; list 5 key components you should include in your 2-page Data Management Plan; list the 4 FAIR principles.

4.4.0.2 Data Sharing - The Reality

This lesson examines privacy and confidentiality concerns in Open Science and data sharing. Students will learn to differentiate between biomedical research types: bench science, human clinical trials, and animal models, and understand the unique data sharing implications for each, including ethical considerations and strategies to protect sensitive data while promoting transparency.

Learning Objectives:

1. Learn about privacy/confidentiality concerns related to Open Science and data sharing.
2. Articulate difference in types of biomedical research (bench science, human clinical trials, animal models) and what implications data sharing has for each.

Assessment:

- Create a 2-page Data Management and Sharing Plan following the NIH requirements for your analyses of the birthweight data challenge.
- Apply the FAIR principles (Findable, Accessible, Interoperable, Reusable) to your datasets.

Chapter 5

Unit 3: Rigorous Statistical Design

Total Time: 5.5 hours

5.1 Principles of Study Design for Empirical Research

We will provide a practical introduction to conducting rigorous data-driven research in a health-science setting that is scientifically-grounded and analytically sophisticated. We will focus on methods for data analysis that can be utilized in the setting of population health research, leveraging official statistics or other systematically collected data with spatial and temporal structure.

5.1.1 Learning Objectives

1. The learner will be able to rapidly internalize the current state of scientific knowledge about a specific health-related topic. The goal here is not to master every detail that a specialist would know, but rather to develop fluency with the key known mechanisms, to recognize the quantitative strengths of established relationships, and to identify gaps in the current state of knowledge.

2. The learner will be able to rapidly internalize the structure and capacity of one or more datasets that can be used to conduct research on a specific health-related topic. This includes identifying the units of analysis (who is being measured) and the variables or attributes (what is being measured), understanding how the units of analysis were selected, what population they represent, how the variables were measured, and what types of measurement errors may be present.

3. The learner will be able to engage in a sophisticated discussion of quantitative relationships among measured quantities. This includes considering how such relationships can be assessed, how they contribute to achieving research

aims, what it means for an association to be causal, and the meanings of confounding and heterogeneity.

4. The learner will have a sophisticated understanding of the opportunities provided by temporal and longitudinal data. This includes understanding notions of spatial and temporal heterogeneity, the specific types of confounding that can be resolved with longitudinal data, and the implications of autocorrelation and other forms of dependence for estimation precision.

5. The learner will be able to develop a rigorous analytic plan to address a hypothesis that adds to the state of knowledge about a health science-relevant topic, using available data and employing sophisticated analytic methods.

6. The learner will have a sophisticated understanding of uncertainty, including (a) *a priori* notions of uncertainty as reflected in statistical power, and (b) uncertainty following data analysis as reflected in statistical measures of confidence, significance, and precision. The learner will understand how uncertainty is influenced by sample size, sampling and experimental design, collinearity, dependence of measurements, measurement error, and heterogeneity, among other factors.

7. The learner will be able to communicate in both speech and writing the high-level message as well as the technical details of a sophisticated, data-driven scientific inquiry in a health science setting. This will include strategies for effectively communicating to different audiences using precise but accessible language, providing complete documentation sufficient for reproducibility, while still having a narrative that does not lose sight of the forest for the trees.

5.1.2 Assessment Instrument

The assessment will be based on the vital statistics data challenge, centered on birth outcomes (underweight birth and infant mortality) in the state of Texas. Learners will be asked to develop a sophisticated but tractable research aim that can be addressed through careful analysis of public datasets from NCHS, SEER, SEDAC, and elsewhere. Learners will then (i) implement the analysis, (ii) interpret findings, and (iii) report their results. Methods and strategies for all three of these requirements will be covered in lectures and written materials provided to the learners.

5.2 Analytic Plans and Statistical Power

Further developing topic 6a above, this lesson equips learners with the skills needed to assess the likelihood of success of a rigorous analytical plan that is aligned with a specific scientific research aim. In most empirical research this is known informally as "power analysis," but here we take a much more expansive definition of this

concept in order to consider all forms of *a priori* vetting to which an analytic plan can be subjected.

5.2.1 Learning Objectives

1. Learners will understand the concepts of effect size, parameter estimation, and estimation precision, as the core theoretical ideas upon which *a priori* power analysis is based.

2. Learners will understand how estimation precision is related to the scope of available data, how these data were collected, and properties of the analytic approach. Learners will be able to articulate what types of data could be collected in the future to most efficiently improve the statistical power.

3. Learners will understand the conventional definition of power as the probability of research success contingent on the unknown true effect size. Learners will also be conversant in alternative notions of power that are directly related to estimation precision or prediction accuracy rather than hypothesis testing.

4. Learners will be able to articulate how assessments of statistical power do and do not reflect the real-world reproducibility of analytic findings.

5.2.2 Assessment Instrument

Carry out a comprehensive assessment of *a priori* power for the analysis conducted in Section 5.1. This analysis should not take into account the actual findings of this analysis, but rather should evaluate the proposed research design, situating yourself at the point before the analysis was implemented and executed.

5.3 Sources of Bias and Causal Interpretation

This section of the course will revisit and deepen learners' understanding of issues related to bias and causality introduced more briefly in Section 5.1. Students will understand how bias is directly related to fundamental notions of causality, while being complementary to the notion of estimation precision as developed in Section 5.2. Learners will understand why observational data is nearly always at risk for such bias, but why an ideal experiment is free of such concerns (but is subject to other limitations on validity). Learners will learn to identify potential sources of bias, uncertainty, and non-reproducibility in observational cohort studies, and suggest basic strategies to address these issues.

5.3.1 Learning Objectives

1. Learners will be able to identify exposures and outcomes in a research design, and identify plausible confounders (both observed and unobserved) of the exposure/outcome relationship.

2. Learners will be able to articulate how temporal and spatial structure enables deeper understanding of exposure/outcome relationships, and opens new opportunities to reduce or eliminate certain types of bias.

3. Learners will understand the different ways that an auxiliary variable can enter an analysis, such as being colliders and mediators, and will understand when an auxiliary variable is unlikely to introduce confounding bias.

4. Learners will understand how certain forms of bias can be reduced or eliminated analytically, while others cannot.

5. Learners will understand how various methods for sensitivity analysis can bound or provide conditional insights into the likely contributions of unobserved confounders.

5.3.2 Assessment Instrument

Building on the assessments for Sections 5.1 and 5.2, provide specific examples of potential measured confounders or other measured sources of bias. To the extent possible, provide analyses that do, and that do not attempt to compensate for such measured sources of bias, and compare these results. Then, provide some examples of unmeasured sources of bias, and conduct sensitivity analyses to assess the likely potential impact of these sources of bias. Finally, use negative control and/or negative exposure methods to provide further context into any causal associations identified through the analysis that you conducted in Section 5.1.

Chapter 6

Unit 4: Designing Interpretable Predictive Models

Total Time: 5 hours
Instructional Time: 3.5 hours
Unit Project Work Time: 1.5 hours

This unit will introduce the foundations of supervised machine learning models, with a focus on interpretability and communicating decision-making.

6.1 Pre-reading Materials

These pre-reading materials focus on foundational concepts in statistics and machine learning.

6.1.1 Learning Objectives

Learners will be able to:

1. Understand ideas from introductory statistics, including measures of central tendency and variability, visualization techniques, and lines of best fit.
2. Understand how to fit a basic machine learning model in Python using sklearn and real data.
3. Understand TRIPOD guidelines pertinent to this session.

6.1.2 Assessment Instrument

Learners will be asked to load a dataset into sklearn, fit a linear regression model, and report the model's mean squared error.

6.2 Foundations of Supervised Learning

Time: 1.5 hours (Instructional: 70 minutes)

We will start by providing a broad overview of the landscape of machine learning, and practice the general workflow for building a model, starting from raw data.

6.2.1 Learning Objectives

1. Understand the landscape of possible machine learning models (supervised vs. unsupervised, regression vs. classification).
2. Build a linear regression model, understanding how its optimal parameters were chosen and how they can be interpreted.
3. Practice the workflow for performing a train-test split, training a model on training data, evaluating a model on held-out test data, and the role of cross-validation.
4. Understand the risks of overfitting and data leakage.

6.2.2 Assessment Instrument (20 minutes)

Learners will build a basic linear regression model using the Vital Statistics dataset, which will serve as a baseline for future work.

6.3 Feature Engineering

Time: 1 hour (Instructional: 60 minutes)

Next, we will cover how linear regression can be extended to a variety of other tasks, and how to create new features that capture trends in the data.

6.3.1 Learning Objectives

1. Practice interpreting the coefficients of fit models.
2. Use visualizations to spot patterns in the data that inform feature engineering decisions, while keeping in mind the risks of overfitting.
3. Understand the differences between encoding strategies (one hot encoding vs. ordinal encoding).

6.3.2 Assessment Instrument (20 minutes)

Learners will be given a practical task and will need to build a small feature engineering Pipeline in sklearn and be asked to document their decision-making process.

6.4 Feature Selection and Model Explainability

Time: 1.5 hours (Instructional: 60 minutes)

Building upon Section 6.3, students will gain an understanding of the statistical approaches involved in selecting features.

6.4.1 Learning Objectives

1. Understand mutual information as a feature selection criterion.
2. Understand how to apply statistically appropriate techniques to select and justify features, e.g., Pearson, Spearman, and Cramér's V correlations, Variance Inflation Factor, multiple R^2.
3. Understand the role of interaction terms.

6.4.2 Assessment Instrument (30 minutes)

Learners will practice generating feature importance analyses and documenting their feature selection rationale.

6.5 Model Evaluation, Comparison, and Reporting

Time: 1 hour (Instructional: 30 minutes)

Finally, students will practice communicating the performance and design decisions behind a model to external stakeholders.

6.5.1 Learning Objectives

1. Understand how to report and compare different models on the same task (e.g., MSE/RMSE/MAE for regression models, accuracy vs. precision vs. recall vs. ROC-AUC vs. F1 for classification models).
2. Practice communicating modeling choices and model behavior.

6.5.2 Assessment Instrument (30 minutes)

Students will write a short report documenting and comparing the performances of two models, including relevant visualizations.

Chapter 7

Unit 5: Reproducible Workflows

Total Time: 5.5 hours

7.1 Goals of Reproducible Analyses

Learn the key goals and challenges of creating reproducible, transparent, and user-friendly analyses that are easy to share and reuse.

7.1.1 Learning Objectives

1. Awareness of key challenges and goals when creating reproducible workflows, including making analyses reproducible, user friendly, transparent, reusable, version controlled, and archived.

7.2 Reproducibility via Code Notebooks

Gain awareness of Markdown, Jupyter, and Quarto, and learn how these tools integrate to create clear, reproducible workflows for data analysis and reporting.

7.2.1 Learning Objectives

1. Awareness of Markdown, Jupyter, Quarto, and how these tools can be integrated into reproducible workflows.

7.3 Best Practices for Reproducible Programming

Learn essential best practices for reproducible programming, including writing clear scripts and functions, avoiding magic numbers, using caching and seeding for randomness, and refactoring code to enhance clarity, reliability, and repeatability.

7.3.1 Learning Objectives

1. Awareness of best practices for reproducible programming including writing scripts, functions, avoiding magic numbers, caching and seeding randomness, and how to refactor code to align with these practices.

7.4 Version Control

Gain a basic understanding of Git, its advantages, and learn to perform essential tasks such as cloning repositories, committing changes, and syncing with remote repositories using push and pull commands.

7.4.1 Learning Objectives

1. Familiarity with Git and its benefits, and the ability to begin using it for simple tasks, including cloning, committing changes, pushing and pulling.

7.5 Containers

Gain hands-on experience with key dependency management tools (Python virtual environments, renv, and containerization), understanding their pros and cons, and develop the skills to create and run basic Docker images.

7.5.1 Learning Objectives

1. Familiarity with various tools for dependency management, including Python virtual environments, renv, and containerization, and their respective strengths and weaknesses. Ability to create and run simple Docker images.

7.6 Assembling a Full Analysis Pipeline

Learn key factors in organizing an analysis pipeline and develop the skills to assemble a complete, reusable pipeline template.

7.6.1 Learning Objectives

1. Considerations when organizing an analysis pipeline, and the ability to assemble a full template pipeline.

7.6.2 Assessment Instrument

1. Describe your progress on the template workflow. What aspects did you find most confusing or challenging? Which tools (e.g., Git, Make, Docker) were hardest to implement, and why?

2. What is one thing you plan to change or do differently in your own projects after today's session? Give a specific example of an analysis or workflow improvement you intend to make.

Chapter 8

Unit 6: Meta-analysis

Total Time: 3.5 hours

8.1 Key Concepts in Research Synthesis

This lesson provides an introduction to meta-analysis as a tool for quantitative research synthesis, with the initial goal being to estimate robust consensus effects and statistical significance levels across methodologically diverse but independent research studies. We will focus on the setting of population health science research, using data sources with a longitudinal and temporal structure. We will consider how independent data reflecting similar or identical outcomes on different populations, potentially collected with disparate measurement methods and/or varying sampling designs can be integrated to improve precision and to uncover heterogeneity and effect modifiers.

8.1.1 Learning Objectives

1. Learners will understand how meta-analysis can be seen as a form of evidence combination.
2. Learners will be able to apply methods for evidence integration when complete data are available.
3. Learners will understand the basic mathematics behind pooling p-values, standard errors, and confidence intervals from independent sources.
4. Learners will be able to weight estimates with different precisions to produce an optimal pooled estimate.
5. Learners will be exposed to newer approaches for pooling evidence including E-values and empirical likelihood methods.

8.1.2 Assessment Instrument

The assessment will be centered on birth outcomes (pre-term birth, underweight birth, and infant mortality), in the United States and internationally, using official statistics or other sources of population-level data. Learners will be guided to appropriate sources of data, and will begin by estimating event rates at local spatial and temporal scales, and providing uncertainty assessments for all point estimates. For the purposes of this assessment, we will treat evidence sources as independent and homogeneous, even if this is unlikely to be the case.

8.2 Accounting for Heterogeneity

Students will learn to separate effect size variability into statistical noise (imprecision) and true heterogeneity. We will discuss several summary measures for unexplained effect heterogeneity, and discuss how stratification and regression methods can be used to understand heterogeneity that can be attributed to known factors.

8.2.1 Learning Objectives

1. Students will understand how variation in measured effect sizes can be partitioned into the component attributable to statistical variation, and the component attributable to heterogeneity in the true effects.

2. Students will understand how the level of heterogeneity in true effects can be partitioned into a component that is attributable to known factors, and a component that is unexplained by known factors.

3. Students will be able to articulate the difference between statistical uncertainty and effect heterogeneity.

4. Students will be able to utilize stratification and regression to estimate a consensus effect size and significance level from studies with heterogeneous designs and/or effect heterogeneity.

8.2.2 Assessment Instrument

Building on the work done for Section 8.1, learners will be asked to consider how and why spatially and temporally local point estimates may be heterogeneous, and to quantify this heterogeneity. They will then identify scales at which partial pooling may be appropriate, calculate statistically efficient pooled estimates of evidence, and assess the extent to which precision was improved. Finally, they will be asked to identify potential explanatory factors for any observed heterogeneity, and to quantify the extent to which the heterogeneity can be attributed to these factors.

8.3 Accounting for Non-independence and Network Effects

Grounded in the setting of population health outcomes assessed at different spatial and temporal scales, we will discuss how and why disparate sources of statistical evidence may be non-independent, and how this non-independence presents both obstacles and possibilities.

8.3.1 Learning Objectives

1. Learners will understand how non-independence of research results impacts research synthesis in terms of bias, uncertainty, and statistical power, and will be able to identify possible sources of non-independence.

2. Learners will revisit the task of combining p-values, Z-scores and other sources of summary evidence from Section 8.1, considering extensions that accommodate non-independent evidence measures.

3. Learners will be able to employ marginal and multilevel regression techniques to account for and leverage the presence of non-independence.

4. Learners will be exposed to modern methods of evidence summarization such as E-values that are robust to non-independence.

8.3.2 Assessment Instrument

Continuing the work from Sections 8.1 and 8.2, students will reconsider the precision of partially pooled estimates of event rates in light of possible non-independence. Then they will consider whether and how explained and unexplained heterogeneity should be re-evaluated if the evidence sources are non-independent.

Chapter 9

Unit 7: Transformer-based AI in Biomedical Research

Total Time: 3 hours

9.1 The Ethics of AI Agents

Time: 1 hour (Instructional: 50 minutes, Project Work: 10 minutes)

This lesson examines the ethical dimensions of using LLM-based AI agents in biomedical data science. Unlike single-turn chatbot interactions, agentic workflows involve planning, tool use, memory, and multi-agent coordination, expanding the ethical surface area considerably. Students will learn to identify invalidation risks (a broader concept than "hallucination"), understand accountability frameworks for agent-assisted research, and apply governance-by-design principles aligned with Responsible Conduct of Research (RCR) standards.

9.1.1 Learning Objectives

1. Define an AI agent in terms of planning, tool use, memory, and multi-agent interaction, and explain why agentic workflows expand ethical risk compared to single-turn model use.

2. Explain invalidation (factual, logical, normative, structural) as a framework broader than "hallucination," and identify at least three invalidation types relevant to biomedical data science.

3. Apply an ethics checklist to an agentic research workflow, identifying concrete risks in accuracy, authorship, privacy/confidentiality, bias, security, and sustainability.

4. Evaluate when multi-agent critique can improve reliability and when it may be inappropriate due to confidentiality constraints, compute burden, or epistemic homogenization.

5. Draft a compliant disclosure and accountability statement for agent-assisted work aligned with COPE/ICMJE/WAME publication-ethics norms.

9.1.2 Assessment Instrument

1. **Knowledge Check (6 minutes):** A short quiz covering: (a) distinguishing features of agents vs. chatbots, (b) types of invalidation, (c) authorship attribution for AI-assisted work, (d) confidentiality risks in peer review contexts, and (e) multi-agent compute tradeoffs.

2. **Case-Based Evaluation (12 minutes):** In small groups, analyze a scenario where an agentic pipeline generates a data dictionary, drafts methods text, proposes models, and summarizes results for the vital statistics data challenge. Identify risks and produce a 5-bullet "agent governance plan" addressing epistemic risk, confidentiality, authorship, bias/security, and sustainability.

3. **Disclosure Statement:** Write a one-sentence disclosure describing how AI tools were used in an analysis, suitable for inclusion in a manuscript methods section.

9.2 AI Agents for Technical Tasks: Consensus in LLMs

Time: 1 hour (Instructional: 45 minutes, Project Work: 15 minutes)

This lesson introduces the theoretical foundations of transformer models and demonstrates how multiple LLM systems can be orchestrated to achieve consensus on technical tasks. Participants will work hands-on with APIs from multiple providers (e.g., OpenAI GPT, Anthropic Claude) to compare model behaviors and understand how cross-model verification can reduce invalidation. The emphasis is on programmatic integration via APIs rather than web-based interfaces, preparing participants to build robust, verifiable analysis pipelines.

9.2.1 Learning Objectives

1. Describe the core architecture of transformer models (attention mechanisms, tokenization, context windows) and explain how transformers evolve into Large Language Models.

2. Compare and contrast simple Artificial Neural Networks (ANNs) with transformer architectures, articulating the advantages of attention-based models for sequential data.

3. Set up and authenticate programmatic access to multiple LLM APIs (OpenAI, Anthropic) using Python, demonstrating environment configuration and secure credential management.

4. Execute a consensus framework across multiple LLMs, either manually via parallel browser sessions or programmatically using provided source code templates.

5. Analyze convergent and divergent responses from multiple models to identify high-confidence outputs versus areas requiring human review.

9.2.2 Assessment Instrument

Pre-workshop requirement: Complete environment setup for local API access to at least two LLM providers.

In-session task (30 minutes): Using the consensus framework (manual or programmatic), query multiple LLMs about analysis routes for the Jackson Heart Study or vital statistics data. Document: (a) the prompt used, (b) responses from each model, (c) areas of agreement/disagreement, and (d) your synthesis of the consensus recommendation. Refer to step-by-step tutorial in the lesson PDF.

9.3 LLMs in Biomedical Research: Building Consensus Pipelines

Time: 1 hour (Instructional: 30 minutes, Project Work: 30 minutes)

Building on the foundations from Sections 9.1 and 9.2, participants will construct a complete consensus analysis pipeline using LLMs to address authentic biomedical research tasks. A step-by-step template is provided that participants can adapt and expand. The session emphasizes that LLMs serve as assistants; humans retain accountability for all outputs, consistent with the ethics framework introduced in Section 9.1.

9.3.1 Learning Objectives

1. Construct a multi-model consensus pipeline following a provided template, incorporating prompt design, response collection, and synthesis stages.

2. Apply the consensus pipeline to evaluate an NIH grant proposal, producing structured feedback aligned with review criteria.

3. Apply the consensus pipeline to draft a Jackson Heart Study manuscript proposal, using LLMs to generate structured content while maintaining human accountability for accuracy and originality.

4. Implement verification checkpoints within the pipeline to detect and flag potential invalidation (factual errors, logical inconsistencies, normative violations).

5. Document the pipeline with appropriate provenance logging (prompts, model versions, timestamps) to support reproducibility and accountability.

9.3.2 Assessment Instrument

Deliverable (choose one):

Option A: NIH Grant Evaluation Report. Using your consensus pipeline, evaluate a provided NIH grant proposal. Produce a technical report that includes: (a) consensus scores for each review criterion, (b) identified strengths and weaknesses with model agreement levels, and (c) a synthesis recommendation with confidence assessment.

Option B: JHS Manuscript Proposal Draft. Using your consensus pipeline and JHS guidelines, produce a draft manuscript proposal. Document: (a) the research question generated/refined by LLMs, (b) the proposed methods with model consensus assessment, and (c) a disclosure statement for AI assistance. Note: Participants do not write proposal text directly; all text is generated via the pipeline and reviewed for accuracy.

Refer to step-by-step tutorial in the lesson PDF. Estimated time: 1 hour.