DAiR3

2026 DAIR^3 Curriculum

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 24, 2026

Contents

Chapter 1

Unit 1: Responsible Conduct of Research

Total Time: 3 hours

1.1 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.

1.1.1 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.

1.1.2 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.

1.2 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.

1.2.1 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.

1.2.2 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.

1.3 Evaluation Rubric

Component	Excellent (90–100%)	Good (80–89%)	Satisfactory (70–79%)	Needs Improvement (<70%)
1. Peer Comparison & Issue Identification (10%)	Thoughtfully compares responses with peers and identifies multiple substantive issues; demonstrates critical reflection on discrepancies and their ethical or scientific implications; connects issues to broader reproducibility and data quality concerns.	Compares responses and identifies several relevant issues; discussion is mostly substantive with some reflection on implications; connection to reproducibility or data quality is partially developed.	Identifies some issues from peer comparison but analysis is surface level; limited reflection on why discrepancies matter for data science practice.	Minimal or no meaningful peer comparison; issues identified are vague or incorrect; little evidence of critical engagement with the data challenge.
2. Historical Data Context (10%)	Provides a well-reasoned, nuanced explanation for why date and time recording ceased after 1987; situates the change within broader sociotechnical, institutional, or policy contexts; demonstrates understanding of how historical decisions shape data availability and reproducibility.	Offers a reasonable explanation with some contextual grounding; demonstrates solid understanding of how institutional or policy factors affect data collection practices, with minor gaps.	Provides a basic explanation but lacks contextual depth; recognizes that external factors influence data collection without fully analyzing their implications.	Explanation is missing, or purely speculative; no grounding in sociotechnical or historical context; shows little understanding of how data collection decisions are made.

Continued on next page

Component	Excellent (90–100%)	Good (80–89%)	Satisfactory (70–79%)	Needs Improvement (<70%)
3. Data Governance, Bias, & Trends (15%)	Clearly articulates who decides what data to collect, store, and access; identifies multiple plausible sources of bias (e.g., rural infrastructure, collection gaps) with specific examples; provides a nuanced, accurate distinction between bias and trend supported by concrete examples; demonstrates sophisticated understanding of how governance decisions shape data quality and equity.	Identifies key decision-makers and discusses bias sources with reasonable examples; distinction between bias and trend is mostly clear and accurate; connection to equity or data quality mostly well developed.	Names some stakeholders but analysis is incomplete; identifies at least one source of bias with a limited example; distinction between bias and trend is attempted but imprecise or only partly illustrated.	Fails to address who governs data collection, storage, or access; bias sources missing or incorrectly described; no meaningful distinction between bias and trend; examples absent or irrelevant.
4. Stakeholder Engagement & Decision-Making (15%)	Framework identifies all relevant stakeholders with well-justified roles; describes specific, realistic engagement strategies; articulates a clear and equitable decision-making structure; anticipates conflicts of interest and addresses them thoughtfully.	Most stakeholders identified with reasonable engagement approaches; decision-making structure is mostly clear; minor gaps in equity considerations or conflict of interest analysis.	Key stakeholders named but engagement strategies are generic or underdeveloped; decision-making structure is present but lacks clarity or justification.	Stakeholders missing or poorly identified; no coherent engagement strategy or decision-making structure; shows limited understanding of governance design.

Continued on next page

Component	Excellent (90–100%)	Good (80–89%)	Satisfactory (70–79%)	Needs Improvement (<70%)
5. Ethical Considerations & Anticipatory Governance (20%)	Thoroughly evaluates ethical challenges specific to the chosen project; integrates anticipatory governance principles to address future uncertainties; differentiates among traditional bioethical, sociotechnical, and other frameworks with sophistication; analysis is proactive rather than reactive.	Ethical challenges well evaluated; anticipatory governance referenced and mostly applied; frameworks differentiated with minor inaccuracies or incomplete integration.	Some ethical challenges identified; anticipatory governance mentioned but superficially applied; framework differentiation is partial or inconsistent.	Ethical considerations minimal or inaccurate; no meaningful application of anticipatory governance; frameworks absent, conflated, or misapplied.
6. Monitoring, Unexpected Impacts, & Benefit-Sharing (15%)	Proposes specific, feasible monitoring mechanisms with clear indicators; thoughtfully anticipates unexpected impacts and describes adaptive responses; benefit-sharing plan is equitable, well-reasoned, and tied to the project-specific context.	Monitoring and benefit-sharing addressed with reasonable detail; unexpected impacts considered with some adaptive strategies; minor gaps in feasibility or equity analysis.	Monitoring and benefit-sharing mentioned but lack specificity or justification; unexpected impacts acknowledged without substantive adaptive planning.	Monitoring, benefit-sharing, or unexpected impacts missing or poorly addressed; framework is reactive or purely theoretical with no practical grounding.

Continued on next page

Component	Excellent (90–100%)	Good (80–89%)	Satisfactory (70–79%)	Needs Improvement (<70%)
7. Consent & Secondary Data Use (10%)	Consent framework is clearly articulated and appropriately tailored to secondary data use; identifies specific privacy and confidentiality risks; proposes concrete strategies consistent with ethical standards and relevant policy or regulatory requirements.	Consent addressed with reasonable detail; privacy risks identified; strategies are mostly appropriate with minor gaps in policy alignment or specificity.	Consent mentioned but treatment is generic or incompletely applied to the secondary use context; privacy risks acknowledged without substantive mitigation strategies.	Consent absent, incorrect, or irrelevant to secondary use; privacy concerns not addressed; no evidence of understanding relevant ethical or regulatory standards.
8. Feasibility & Overall Framework Quality (5%)	Framework is coherent, well-structured, and realistic; all components integrate into a unified governance plan; writing is clear and precise; demonstrates thorough command of course concepts across the full framework.	Framework is mostly coherent and feasible; components are largely integrated; writing is clear with minor lapses; good command of course concepts throughout.	Framework is partially coherent; some components feel disconnected or underdeveloped; writing is adequate; course concepts applied unevenly.	Framework is incoherent, infeasible, or largely incomplete; components do not integrate; writing is unclear; limited evidence of course concept mastery.

Chapter 2

Unit 2: Data Management

Total Time: 7 hours

2.1 Data Collection

Pending...

2.2 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.

2.2.1 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.

2.2.2 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?

2.3 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.

2.3.1 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.

2.3.2 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?

2.4 Data Sharing

Time: 2.5 hours

2.4.0.1 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.

2.4.0.2 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.

2.5 Evaluation Rubric

Component	Excellent (90–100%)	Good (80–89%)	Satisfactory (70–79%)	Needs Improvement (<70%)
1. Metadata – Data About Data (20%)	Identifies 5 critical metadata categories with clear, relevant examples; provides a thorough analysis of metadata from Hauer's article, including how, on whom, and under what circumstances data were collected; differentiates high-quality vs. poor metadata for reproducibility.	Lists most metadata categories with relevant examples; analysis from the Hauer article is solid but lacks some detail; distinction between good/bad metadata is mostly clear.	Lists some metadata categories but with limited examples or partial relevance; Hauer's article analysis is basic; distinction is made at the surface level.	Fewer than 5 categories or inaccurate examples; superficial or missing analysis of Hauer article; shows little grasp of metadata quality for reproducibility.
2. Data Representation (20%)	Clearly describes alternative data representations and justifies which schema best supports analysis tasks; insightfully compares the flat file to the alternative, with thoughtful reasoning and examples; demonstrates a strong understanding of how representation affects research.	Describes an alternative schema and makes a reasonable comparison to the flat file, with some justification; demonstrates good understanding with minor gaps.	Identifies a basic alternative, but justification/comparison to the flat file is weak or incomplete; limited examples.	Incomplete, irrelevant, or unclear alternatives; little to no justification or analysis of schema versus flat file; shows limited understanding.

Continued on next page

Component	Excellent (90–100%)	Good (80–89%)	Satisfactory (70–79%)	Needs Improvement (<70%)
3. Data Sharing 101: Open Science, NIH Policy, FAIR (30%)	Thoroughly defines rationale for NIH Data Management & Sharing; lists 5 key components with well-explained relevance; accurately lists and explains all 4 FAIR principles with practical examples for each.	Defines rationale and lists the key NIH components with basic relevance; lists and explains all FAIR principles, with most examples being suitable.	Provides basic rationale and fewer than 5 NIH components; lists most FAIR principles with limited or superficial examples.	Incomplete rationale and components, missing several critical points; fails to list all FAIR principles or provides incorrect explanations/examples.
4. Data Sharing (The Reality): Privacy, Confidentiality, Types of Biomedical Research (30%)	Creates a complete Data Management & Sharing Plan tailored to the Texas birthweight challenge, following NIH requirements; thoroughly applies FAIR principles to the dataset; clearly articulates privacy concerns and differences among bench, clinical, and animal model research, with thoughtful analysis of sharing implications.	Plan meets most NIH requirements; applies FAIR principles with minor gaps; reasonably discusses privacy and research type differences.	Plan basic or incomplete; applies FAIR principles at the surface level; discusses privacy/research types with limited detail.	Missing/incomplete plan; fails to address FAIR principles, research types, or privacy concerns meaningfully.

Chapter 3

Unit 3: Rigorous Statistical Design

Total Time: 5.5 hours

3.1 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.

3.1.1 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.

3.1.2 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.

3.2 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.

3.2.1 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.

3.2.2 Assessment Instrument

Carry out a comprehensive assessment of *a priori* power for the analysis conducted in Section 3.1. 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.

3.3 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 3.1. 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 3.2. 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.

3.3.1 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.

3.3.2 Assessment Instrument

Building on the assessments for Sections 3.1 and 3.2, 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 3.1.

3.4 Evaluation Rubric

Component	Excellent (90–100%)	Good (80–89%)	Satisfactory (70–79%)	Needs Improvement (<70%)
1. Research Aim & Study Design (30%)	Research aim is precise, impactful, and clearly motivated by literature; proposes a rigorous study design with well-justified choices; articulates data sources and methods clearly and non-technically; demonstrates deep understanding of design rigor and practical considerations.	Strong aim and generally appropriate design; minor gaps in justification or clarity; literature context adequate but not comprehensive.	Aim is understandable but narrow, vague, or missing justification; study design adequate but not clearly aligned to research aim or literature.	Aim poorly defined or unsupported; study design unclear, infeasible, or disconnected from research aim or available data.
2. Strengths, Weaknesses, and Practical Considerations of Design (10%)	Insightful discussion of methodological rigor, feasibility, and tradeoffs; identifies limitations and practical challenges with clear, evidence-based reasoning.	Good identification of strengths/weaknesses; minor omissions in feasibility or rigor reasoning.	Basic discussion present but shallow; misses key limitations or practical issues.	Minimal or incorrect discussion; fails to address rigor, feasibility, or weaknesses.

Continued on next page

Component	Excellent (90–100%)	Good (80–89%)	Satisfactory (70–79%)	Needs Improvement (<70%)
3. Analytic Plan (25%)	Analytic plan is technically sound, well-aligned to design, and matches study structure; demonstrates strong command of appropriate statistical methods; plan is reproducible, transparent, and scientifically rigorous.	Plan is coherent and largely appropriate; some methods need refinement or lack full justification.	Plan is partially appropriate but missing methodological details, clarity, or alignment to study design.	Analytic plan incomplete, inappropriate, or methodologically incorrect.
4. Power Analysis & Precision Justification (20%)	Power analysis well-constructed, clearly explained, and directly tied to study design; includes effect sizes, parameter precision, detectable differences, and assessment of reproducibility; demonstrates excellent reasoning about what can and cannot be detected.	Power analysis correct and relevant, though missing depth in effect sizes or reproducibility justification; interpretation generally sound.	Basic power analysis included but incomplete, overly simplistic, or lacking explicit connection to study design or precision.	Missing, incorrect, or poorly executed power analysis; no discussion of precision or reproducibility.
5. Bias, Uncertainty, and Non-Reproducibility (10%)	Clearly identifies key biases and uncertainty inherent to the observational design; provides thoughtful, feasible remedies; demonstrates strong understanding of causal and methodological threats.	Identifies major sources of bias and uncertainty; remedies adequate but may lack depth or completeness.	Identifies some biases but misses important ones; remedies are generic or under-developed.	Fails to identify important biases or misunderstands them; remedies missing or incorrect.

Continued on next page

Component	Excellent (90–100%)	Good (80–89%)	Satisfactory (70–79%)	Needs Improvement (<70%)
6. Interpretation of Findings (15%)	Interpretation directly and accurately addresses the research aim; integrates analytic results, design limitations, bias considerations, precision, and confidence; demonstrates sophisticated reasoning and nuance.	Interpretation mostly aligned with results and study aim; minor gaps in addressing design limitations or confidence.	Interpretation present but superficial or disconnected from study aim; limited reflection on confidence or limitations.	Interpretation incorrect, overly speculative, or not linked to results or study design.
7. Overall Professionalism, Structure, and Clarity	Writing is polished, well-organized, and concise; formatting adheres to instructions; figures/tables clear and correctly labeled; demonstrates strong professional communication.	Generally well-written with minor organizational or formatting issues; understandable and professional.	Writing understandable but includes stylistic or structural issues; formatting inconsistent.	Poor structure, unclear writing, numerous errors, or failure to follow instructions.

Chapter 4

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.

4.1 Pre-reading Materials

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

4.1.1 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.

4.1.2 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.

4.2 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.

4.2.1 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.

4.2.2 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.

4.3 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.

4.3.1 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).

4.3.2 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.

4.4 Feature Selection and Model Explainability

Time: 1.5 hours (Instructional: 60 minutes)

Building upon Section 4.3, students will gain an understanding of the statistical approaches involved in selecting features.

4.4.1 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.

4.4.2 Assessment Instrument (30 minutes)

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

4.5 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.

4.5.1 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.

4.5.2 Assessment Instrument (30 minutes)

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

4.6 Evaluation Rubric

Component	Excellent (90–100%)	Good (80–89%)	Satisfactory (70–79%)	Needs Improvement (<70%)
1. Pre-Reading Foundations (10%)	Demonstrates clear understanding of core statistical concepts (central tendency, variability, visuals, best-fit lines); correctly loads data, fits sklearn linear regression, and accurately reports MSE; strong grasp of TRIPOD-relevant principles.	Minor errors in interpretation or MSE reporting; overall understanding solid; TRIPOD concepts mostly correct.	Basic understanding demonstrated; errors in implementation or conceptual explanation; shaky grasp of TRIPOD elements.	Incorrect or incomplete implementation; misunderstanding of statistical foundations; missing or incorrect MSE.
2. Supervised Learning Foundations (20%)	Clearly distinguishes supervised vs. unsupervised, regression vs. classification; correctly performs train-test split, trains model, evaluates it, and explains cross-validation; demonstrates strong understanding of overfitting and data leakage.	Good workflow with minor errors in reasoning or terminology; overfitting/data leakage described but not deeply.	Workflow partially correct; inconsistent understanding of splits, evaluation metrics, or model fitting.	Incorrect or incomplete workflow; poor understanding of model evaluation, overfitting, or data leakage.
3. Generalized Linear & Non-Parametric Models (15%)	Accurately explains and applies linear, logistic, and Poisson regression; correctly interprets coefficients; demonstrates understanding of assumptions; incorporates tree-based ideas if relevant.	Mostly correct application and interpretation; minor gaps in assumptions or coefficient explanations.	Partial understanding; misinterprets coefficients or uses GLMs incorrectly; assumptions' role unclear.	Major misunderstandings of regression types, assumptions, or interpretations; incomplete or incorrect analysis.

Continued on next page

Component	Excellent (90–100%)	Good (80–89%)	Satisfactory (70–79%)	Needs Improvement (<70%)
4. Feature Engineering (15%)	Insightfully uses visualizations to identify trends; appropriately applies encoding, scaling, and interaction terms; builds a clear sklearn Pipeline; thoroughly documents decisions and rationales.	Good feature engineering with minor justification gaps; Pipeline mostly correct; documentation adequate.	Basic feature creation with limited insight; Pipeline incomplete or justification superficial.	Poor or incorrect feature engineering; missing Pipeline; unclear or unjustified decisions.
5. Feature Selection & Explainability (20%)	Correctly applies correlation measures (Pearson/Spearman/Cramér's V), VIF, R^2; thoughtfully justifies feature selection; generates accurate SHAP explanations and interprets them clearly for model behavior.	Most statistical techniques applied correctly; SHAP computed with minor interpretation issues.	Partial or inconsistent application of selection methods; SHAP produced but poorly interpreted.	Incorrect statistical methods; missing or incorrect SHAP analysis; no justification of selected features.
6. Model Evaluation, Comparison & Reporting (20%)	Accurately computes and compares regression/classification metrics (MSE/RMSE/MAE, accuracy, precision, recall, ROC-AUC, F1); report is clear, well-structured, and communicates model design decisions and behavior with professional visuals.	Metrics computed correctly with minor errors; comparison and explanation clear; visuals mostly appropriate.	Metrics included but inconsistently computed or interpreted; limited explanation; visuals basic or unclear.	Metrics missing or incorrect; report unclear or incomplete; explanations not aligned with results.

Continued on next page

Component	Excellent (90–100%)	Good (80–89%)	Satisfactory (70–79%)	Needs Improvement (<70%)
7. Overall Professionalism, Clarity, and Documentation	Writing polished and well-organized; notebooks/scripts clean, reproducible, and well-commented; outputs easy to interpret.	Mostly clear with minor structural or formatting issues; code readable.	Understandable but inconsistent structure or documentation.	Poorly organized, unclear writing, missing documentation, or sloppy code.

Chapter 5

Unit 5: Reproducible Workflows

Total Time: 5.5 hours

5.1 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.

5.1.1 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.

5.2 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.

5.2.1 Learning Objectives

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

5.3 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.

5.3.1 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.

5.4 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.

5.4.1 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.

5.5 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.

5.5.1 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.

5.6 Assembling a Full Analysis Pipeline

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

5.6.1 Learning Objectives

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

5.6.2 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.

5.7 Evaluation Rubric

Component	Excellent (90–100%)	Good (80–89%)	Satisfactory (70–79%)	Needs Improvement (<70%)
1. Goals of Reproducible Analyses (10%)	Provides a thorough summary, clearly articulating key goals/challenges of reproducible analyses; connects concepts to personal work and the CDC dataset; demonstrates careful reflection on transparency, usability, version control, and archiving.	Addresses most key challenges/goals clearly; provides some connection to personal work and dataset; demonstrates solid understanding with minor omissions.	Identifies basic challenges/goals, but with limited depth or reflection; may not connect well to examples or skip some elements.	Incomplete or superficial summary; lacks understanding of core concepts or fails to address relevance to own work.
2. Reproducibility via Code Notebooks (20%)	Creates a well-structured notebook with reproducible code, markdown, at least one plot and table; documentation is clear; script downloads CDC dataset; all work uploaded; demonstrates integration of tools.	Notebook includes most required elements (code, markdown, plot/table); documentation is mostly clear; dataset obtained; minor gaps in reproducibility or clarity.	Notebook has basic elements, but some required elements (plot/table, markdown, script) are missing or poorly integrated; minimal documentation.	Notebook missing key requirements; poorly documented or not reproducible; files not uploaded, or CDC dataset not included.

Continued on next page

Component	Excellent (90–100%)	Good (80–89%)	Satisfactory (70–79%)	Needs Improvement (<70%)
3. Best Practices for Reproducible Programming (15%)	Adds thorough documentation to notebook; creates a well-organized Makefile with clear, reusable commands; analysis file is uploaded; demonstrates strong understanding of reproducible scripting (functions, no magic numbers, caching, seeding randomness, refactoring).	Documentation and Makefile present and mostly clear; analysis file uploaded; most best practices followed, though minor omissions may be present.	Documentation or Makefile is incomplete or unclear; basic understanding of practices is evident, but with gaps in execution or clarity.	Documentation and/or Makefile missing or severely lacking; little evidence of understanding reproducible programming practices.
4. Version Control (15%)	Template project code on GitHub is complete and well-structured; the link is shared; meaningful commit history demonstrates staged progress and clarity; challenges are described in detail.	Project is on GitHub with most files complete; link is shared; some commit history and brief challenge discussion; mostly correct use of git.	Project present but incomplete or poorly organized; minimal commit history; challenges noted superficially.	Project not on GitHub or files missing; little/no evidence of version control understanding; challenges not described.
5. Containers & Dependency Management (20%)	The template project is successfully containerized (Docker/renv/venv); runs as expected; Dockerfile/other files uploaded; demonstrates a clear understanding of strengths/weaknesses; GitHub project is complete and well-documented; challenges discussed.	Container image runs with minor issues; most files uploaded; reasonable documentation and explanation of tool choices; challenges discussed briefly.	Container created but may not run as expected, or files/documentation incomplete; basic understanding of tools, but lacks depth.	Container not created or does not run; missing files/documentation; little/no understanding of dependency management.

Continued on next page

Component	Excellent (90–100%)	Good (80–89%)	Satisfactory (70–79%)	Needs Improvement (<70%)
6. Assembling a Full Analysis Pipeline (15%)	Creates a comprehensive template project with a clear directory structure and markdown documentation; the pipeline demonstrates thoughtful organization, reproducibility, and ease of use; meets all requirements outlined in class with innovative or robust elements.	Project structure and markdown are mostly clear; most requirements met, with minor organizational or documentation gaps.	Basic directory structure and markdown present; pipeline missing some components or documentation unclear; meets minimal requirements.	Template project missing core structure or documentation; does not demonstrate pipeline assembly skills; requirements not met.
7. Professionalism, Clarity, Structure (5%)	Work is polished, well-organized, clear, and follows all instructions; files are named and formatted consistently; writing is succinct and logical.	Work is mostly clear and organized; minor formatting or organizational issues.	Basic clarity but contains organizational weaknesses, unclear writing, or some failures to follow instructions.	Work is disorganized, unclear, or does not follow instructions; formatting and naming are inconsistent or missing.

Chapter 6

Unit 6: Meta-analysis

Total Time: 3.5 hours

6.1 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.

6.1.1 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.

6.1.2 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.

6.2 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.

6.2.1 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.

6.2.2 Assessment Instrument

Building on the work done for Section 6.1, 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.

6.3 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.

6.3.1 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 6.1, 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.

6.3.2 Assessment Instrument

Continuing the work from Sections 6.1 and 6.2, 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.

6.4 Evaluation Rubric

Component	Excellent (90–100%)	Good (80–89%)	Satisfactory (70–79%)	Needs Improvement (<70%)
1. Evidence Combination & Pooled Estimation (30%)	Accurately estimates event rates for birth outcomes (pre-term birth, underweight birth, infant mortality) at local spatial and temporal scales; provides thorough uncertainty assessments for all point estimates; correctly pools p-values, standard errors, and confidence intervals from independent sources; applies precision-weighted pooling to produce an optimal combined estimate; clearly justifies assumptions of independence and homogeneity.	Estimates event rates with mostly correct uncertainty assessments; pools evidence from independent sources with minor errors; weighting approach is reasonable with minor gaps in justification.	Estimates event rates but uncertainty assessments are incomplete or partially incorrect; pooling approach is basic or inconsistently applied; limited justification for assumptions made.	Event rate estimates missing or incorrect; uncertainty assessments absent or flawed; little to no evidence of understanding pooling methods or independence assumptions.

Continued on next page

Component	Excellent (90–100%)	Good (80–89%)	Satisfactory (70–79%)	Needs Improvement (<70%)
2. Heterogeneity Quantification & Partial Pooling (35%)	Clearly identifies and quantifies spatial and temporal heterogeneity in point estimates; correctly partitions effect size variability into statistical noise and true heterogeneity; identifies appropriate scales for partial pooling and calculates statistically efficient pooled estimates; demonstrates improved precision; identifies and quantifies explanatory factors for observed heterogeneity using stratification and/or regression.	Heterogeneity identified and mostly quantified correctly; partial pooling applied at reasonable scales with minor errors; some explanatory factors identified, though quantification of their contribution may be incomplete.	Heterogeneity acknowledged but quantification is superficial or incomplete; partial pooling attempted but scale selection or efficiency poorly justified; explanatory factors noted but not rigorously assessed.	Heterogeneity not meaningfully addressed; no attempt at partial pooling or efficiency assessment; explanatory factors absent or incorrectly handled; conflates statistical uncertainty with true heterogeneity.

Continued on next page

Component	Excellent (90–100%)	Good (80–89%)	Satisfactory (70–79%)	Needs Improvement (<70%)
3. Non-Independence & Network Effects (35%)	Clearly identifies possible sources of non-independence among evidence sources; accurately reassesses precision of partially pooled estimates in light of non-independence; correctly revisits heterogeneity findings under non-independence; applies marginal or multilevel regression techniques appropriately; demonstrates nuanced understanding of how non-independence affects bias, uncertainty, and statistical power.	Sources of non-independence identified with minor gaps; reassessment of pooled estimates mostly correct; regression techniques applied with minor errors; understands impact on uncertainty and power, though depth may be limited.	Non-independence acknowledged but sources poorly identified; reassessment of estimates or heterogeneity is superficial; regression techniques attempted but with notable errors or limited justification.	Non-independence not meaningfully addressed; no reassessment of pooled estimates or heterogeneity; regression techniques absent or incorrectly applied; little understanding of implications for bias or power.

Chapter 7

Unit 7: Transformer-based AI in Biomedical Research

Total Time: 3 hours

7.1 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.

7.1.1 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.

7.1.2 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.

7.2 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.

7.2.1 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.

7.2.2 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.

7.3 LLMs in Biomedical Research: Building Consensus Pipelines

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

Building on the foundations from Sections 7.1 and 7.2, 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 7.1.

7.3.1 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.

7.3.2 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.

7.4 Evaluation Rubric

Component	Excellent (90–100%)	Good (80–89%)	Satisfactory (70–79%)	Needs Improvement (<70%)
1. Ethics of AI Agents (30%)	Accurately distinguishes AI agents from chatbots and clearly explains why agentic workflows expand ethical risk; correctly identifies and explains all four invalidation types with biomedical examples; applies a thorough ethics checklist to an agentic pipeline, addressing all six risk areas (accuracy, authorship, privacy, bias, security, sustainability); produces a complete 5-bullet agent governance plan; drafts a precise, publication-ready disclosure statement aligned with COPE/ICMJE/WAME norms.	Mostly correct distinction between agents and chatbots; identifies most invalidation types with reasonable examples; ethics checklist and governance plan address most risk areas with minor omissions; disclosure statement mostly appropriate.	Basic understanding of agents vs. chatbots; identifies some invalidation types but with limited depth; governance plan incomplete or superficial; disclosure statement present but not well-aligned with publication norms.	Fails to distinguish agents from chatbots or misunderstands invalidation; ethics checklist or governance plan missing or incorrect; disclosure statement absent or unsuitable for publication.

Continued on next page

Component	Excellent (90–100%)	Good (80–89%)	Satisfactory (70–79%)	Needs Improvement (<70%)
2. LLM Consensus Framework (35%)	Demonstrates successful environment setup and authenticated API access to at least two LLM providers; clearly documents the prompt used, responses from each model, and areas of agreement/disagreement; produces a well-reasoned synthesis of the consensus recommendation; shows strong understanding of transformer architecture (attention, tokenization, context windows) and how it differs from simple ANNs.	Environment setup complete with minor issues; documents prompt and model responses with mostly clear comparison; consensus synthesis reasonable but lacking some depth; understands transformer architecture with minor gaps.	Environment setup attempted but incomplete; documentation of prompt or model responses is partial; consensus synthesis is superficial or poorly justified; basic understanding of transformer architecture with notable gaps.	Environment setup not completed or API access not demonstrated; prompt, responses, or synthesis missing; little to no understanding of transformer architecture or cross-model verification.

Continued on next page

Component	Excellent (90–100%)	Good (80–89%)	Satisfactory (70–79%)	Needs Improvement (<70%)
3. Consensus Pipeline Construction & Application (35%)	Constructs a complete, well-documented consensus pipeline incorporating prompt design, response collection, and synthesis stages; correctly applies the pipeline to either the NIH grant evaluation or JHS manuscript proposal; implements verification checkpoints that detect and flag invalidation; provides thorough provenance logging (prompts, model versions, timestamps); maintains clear human accountability for all outputs.	Pipeline mostly complete with minor structural gaps; applied to chosen assessment option with reasonable outputs; verification checkpoints present but may lack depth; provenance logging mostly complete.	Pipeline partially constructed; application to chosen option is incomplete or outputs are poorly synthesized; verification checkpoints minimal or inconsistently applied; provenance logging sparse.	Pipeline not constructed or non-functional; chosen assessment option not meaningfully addressed; no verification checkpoints or provenance logging; human accountability not demonstrated.

Chapter 8

Unit 8: Jackson Heart Study

Time: 3 hours

8.1 Introduction to the Jackson Heart Study

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.

8.1.1 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.

8.1.2 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, etc.) 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.

8.2 The Process of Manuscript Development in the Jackson Heart Study

This lesson covers the process for requesting and obtaining 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.

8.2.1 Learning Objectives

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

8.2.2 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.

8.3 Evaluation Rubric

Component	**Excellent (90–100%)**	**Good (80–89%)**	**Satisfactory (70–79%)**	**Needs Improvement (<70%)**
1. Variable Identification (50%)	Accurately identifies variables; clear rationale; demonstrates deep understanding of JHS datasets.	Minor omissions but overall strong alignment.	Adequate but limited justification.	Incomplete or incorrect variable selection.
2. Manuscript Proposal (50%)	Logically extends grant proposal; well-articulated and publication-ready.	Solid but needs refinement in framing or methods.	Basic understanding but lacks clarity or completeness.	Unclear, inconsistent, or incomplete proposal.
3. Overall Professionalism and Clarity	Polished writing, correct formatting, adherence to instructions.	Minor issues with organization or formatting.	Understandable but contains stylistic or format errors.	Poor structure, numerous errors.