APPLIED HEALTH INFORMATICS

Research Informatics & Data Management

From Research Question to Trustworthy Evidence

Instructors/Presenters:
Ashley Landay, MBA/MS
Chamberlain Nwanne, PhD

Research Informatics & Data Management

One-Week Intensive | From Research Question to Trustworthy Evidence

Graduate-level applied course

Primary perspective: the Health Informaticist

Focus: information architecture behind trustworthy research

Five connected modules + one longitudinal LMIC case

A research finding is only as trustworthy as the information pathway that produced it.

Research Informatics & Data Management

Course Roadmap: Five integrated modules

This one-week course follows the research information pathway from question to trustworthy evidence.

- **From Research Question to Information Requirements:** Translating research concepts into computable variables.
- **Research Data Sources, Provenance & Lineage:** Understanding where data come from and what happens to them.
- **Research Data Structure, Meaning & Quality** Preserving meaning through data models, dictionaries, metadata and quality checks.
- **Governance, Ethics, Bias & FAIR Data** Ensuring responsible, authorized, and trustworthy use of research data.
- **Research Informatics in Resource-Constrained Settings** Designing reliable, practical research-data systems for LMIC realities

“Can We Trust the Dataset?” applied exercise
+
final Research Informatics Plan

What You Should Be Able to Do by the End of the Course

- ❑ Translate a research question into computable data requirements
- ❑ Distinguish clinical data from research-ready data
- ❑ Evaluate provenance, lineage, data quality and fitness for purpose
- ❑ Design a data dictionary, metadata and governance approach
- ❑ Recognize bias introduced before statistical analysis
- ❑ Design a practical LMIC research-data architecture
- ❑ Explain the Informaticist’s role in reproducible, trustworthy evidence

Question

↓

Data

↓

Meaning

↓

Evidence

APPLIED HEALTH INFORMATICS

Research Informatics & Data Management

MODULE 1 — FROM RESEARCH QUESTION TO INFORMATION REQUIREMENTS

Module 1 | Research Informatics: The Missing Layer

Research methods – Asks how to investigate; Research informatics – Asks if information can support the investigation.

Research question — What are we trying to learn?

Research methods — How will we investigate it?

Research informatics — How must information be represented, acquired, governed, transformed and preserved?

Analysis — What patterns or relationships are present?

Evidence — What conclusions are defensible from those data?

Teaching Pearl

A statistically sophisticated analysis cannot rescue poorly defined, poorly captured, or poorly governed information.

The Research Information Lifecycle

Every upstream decision constrains what can legitimately be concluded downstream.

Research Question → Data Requirements → Data Sources → Capture / Acquisition → Storage & Governance → Transformation & Quality → Analysis-Ready Dataset → Analysis → Evidence / Knowledge

At every transition ask: What changed? What meaning was added, removed, inferred, or lost?

From Clinical Question to Computable Variable

Conceptual variables must be operationalized before they can be queried.

Example question: Among adults receiving breast cancer treatment, is treatment interruption associated with poorer 12-month outcomes?

Population: What operational definition identifies an eligible adult with breast cancer?

Exposure: What event, duration or sequence constitutes a treatment interruption?

Outcome: What precisely counts as a poorer 12-month outcome?

Time: When does follow-up begin and end?

Covariates: Which factors could alter interpretation of the relationship?

Clinical concept ≠ computable variable

The definition must be represented in data.

Critical Thinking | What Can the Dataset Actually Answer?

Before extraction, ask whether the source contains the needed concept—not merely a related field.

Separate directly observed data from derived or inferred variables.

Identify required temporal relationships: before, during, after, within a defined interval.

Document assumptions used to transform clinical events into research variables.

If a concept cannot be represented reliably, revise the question or collect additional data.

Informaticist question:

“If we obtain an answer, can we trust what the data mean?”

APPLIED HEALTH INFORMATICS

Research Informatics & Data Management

MODULE 2 — RESEARCH DATA SOURCES, PROVENANCE & LINEAGE

ARGO

Module 2 | The Healthcare Data Ecosystem

Research data inherit the purpose and workflow of the systems that created them.

- ❑ EHR and clinical documentation
- ❑ Laboratory and pathology systems
- ❑ RIS/PACS and imaging repositories
- ❑ Pharmacy and medication systems
- ❑ Disease registries and public-health platforms
- ❑ Claims and administrative systems
- ❑ Surveys, mobile health and patient-generated data
- ❑ Genomic and dedicated research databases

Do not ask only:
“Where is the data?”

Ask:
Who created it? Why? When? Under what workflow and definition?

Clinical Data Are Not Automatically Research-Ready

Data availability is not equivalent to data usability.

Multiple measurements may exist for the same concept.

Reference ranges, units and methods may vary by laboratory or time.

Results may be corrected after initial posting.

Free text may contain clinically important context absent from structured fields.

Encounter duplication can create false event counts.

Timestamps may represent order, collection, result, verification or documentation time.

Example

A hemoglobin value is available. But which value, unit, method, time point and clinical context should the study use?

Provenance and Lineage

Trust requires knowing where information came from and what happened to it.

- Patient
- Clinician
- EHR
- Warehouse
- Extraction
- Transformation
- Research Dataset
- Analysis

Provenance = origin and circumstances of creation.

Lineage = transformations and movement from source to analytic dataset.

Reproducibility requires documenting both.

Longitudinal Case | The Missing Cancer Patients

A simple research question becomes an information-integration problem.

Question: What proportion of women diagnosed with breast cancer complete treatment within 12 months?

EHR contains diagnoses and encounters.

Pathology uses a separate database.

Chemotherapy is partly electronic; radiotherapy uses another system.

Some treatment occurs outside the hospital.

Deaths outside the institution may not be recorded.

Patient identifiers are inconsistent across systems.

Leadership asks:

"Can you simply extract all breast cancer patients and calculate treatment completion?"

APPLIED HEALTH INFORMATICS

Research Informatics & Data Management

MODULE 3 — RESEARCH DATA STRUCTURE, MEANING & QUALITY

Module 3 | Research Data Models Preserve Meaning

A value without context is only a number.

Entity — What or who is being represented?

Variable — What characteristic or event is being measured?

Value — What was observed or recorded?

Context — Unit, date/time, source, method, status and clinical meaning.

Patient 027
Hemoglobin
8.7 g/dL
14 Jun 2026
Pre-treatment

The data model determines whether meaning survives aggregation, exchange and analysis.

The Data Dictionary Is Research Infrastructure

A shared semantic contract for how variables are represented.

Variable	Definition	Type / Values	Unit	Source	Missing Rule	Transformation

If two researchers can interpret the same variable differently, the data dictionary has failed.

Metadata Preserves Meaning Over Time

Data can remain technically available while becoming semantically unusable.

“Stage = 3” is incomplete without the staging system and version.

Clinical stage and pathological stage are not interchangeable.

A date may represent diagnosis, specimen collection, result verification or data entry.

A code may change meaning across software versions or reporting standards.

Metadata makes interpretation, reuse and reproducibility possible.

Metadata answers:

What does this value mean?
How was it produced?
Under which rules?

Research Data Quality Is Multidimensional

“Clean data” is an inadequate description.

Completeness
Expected data are present

Accuracy
Values reflect reality

Consistency
Representations agree

Validity
Values follow defined rules

Timeliness
Available at the required time

Uniqueness
No inappropriate duplication

Integrity
Relationships are preserved

Fitness for purpose:
A dataset can be high quality for one use and inadequate for another.

Missingness Is Information

A blank field may reflect workflow, access, infrastructure or clinical reality.

- Test not ordered
- Patient could not afford the test
- Equipment unavailable
- Result exists only on paper
- Patient transferred or was lost to follow-up
- Field was optional or workflow bypassed
- Information captured in narrative rather than structured data

Do not ask only:
“How much data are missing?”

Ask:
“Why are they missing, and is missingness systematic?”

Structured–Unstructured Tension

Computability and clinical meaning are both valuable.

- ❑ Structured data support validation, querying, aggregation and comparison.
- ❑ Narrative text preserves uncertainty, reasoning, nuance and context.
- ❑ Over-structuring can force complex clinical reality into misleading categories.
- ❑ Under-structuring can make important information difficult to retrieve or analyze.
- ❑ Design principle: structure what must be computable; preserve narrative where context matters.

The goal is not maximum structure.

The goal is usable information without loss of meaning.

APPLIED HEALTH INFORMATICS

Research Informatics & Data Management

MODULE 4 — GOVERNANCE, ETHICS, BIAS & FAIR

Module 4 | Research Data Governance

Technical access is not the same as authorized use.

- ❑ Who is the data steward or accountable owner?
- ❑ Who may access which data?
- ❑ For what approved purpose?
- ❑ Where may data be stored or transferred?
- ❑ What safeguards are required?
- ❑ How long should data be retained?
- ❑ What happens at project completion or reuse?

Having access to data does not mean you have the right to use it.

Privacy, Consent and Secondary Use

Research use creates obligations beyond clinical access.

- ❑ Consent and/or appropriate ethics oversight
- ❑ Purpose limitation: use data for the approved question
- ❑ Minimum necessary data
- ❑ De-identification or pseudonymization where appropriate
- ❑ Role-based access and secure storage
- ❑ Data-use agreements for sharing
- ❑ Clear accountability for breaches or misuse

Scenario

A hospital has 10 years of oncology data.
Can a researcher use it simply because it is queryable?

Bias Can Enter Before Analysis

Information systems shape who becomes visible in the dataset.

Real Population

Reaches Care

Documented

Represented Digitally

Extracted

Eligible Cohort

Selection bias can begin with who reaches the health system.

Documentation practices determine who is visible and how.

Extraction and inclusion rules can further reshape the population.

FAIR Research Data

Findable • Accessible • Interoperable • Reusable

Findable — data and metadata can be located and identified.

Accessible — authorized users can retrieve them through defined processes.

Interoperable — meaning can be exchanged across systems and tools.

Reusable — documentation and context support appropriate future use.

FAIR ≠ open to everyone

Research data can be FAIR while remaining appropriately protected.

APPLIED HEALTH INFORMATICS

Research Informatics & Data Management

MODULE 5 — RESEARCH INFORMATICS IN RESOURCE-CONSTRAINED SETTINGS

Module 5 | Research Informatics in Resource-Constrained Settings

Adapt architecture to reality without lowering scientific standards.

- ❑ Paper–digital hybrid workflows
- ❑ Fragmented or inconsistent patient identifiers
- ❑ Limited interoperability
- ❑ Connectivity and power interruptions
- ❑ Variable documentation practices
- ❑ Multiple donor/program reporting requirements
- ❑ Limited longitudinal records
- ❑ Limited informatics and data-management workforce

Design question:

How do we preserve trustworthy information when infrastructure and workflows are imperfect?

Minimum Viable Research Data Architecture

Reliable simplicity is often more valuable than unnecessary complexity.

Source Systems / Paper

Standardized Capture

Validated Research Repository

Quality & Governance

Analysis-Ready Dataset

Research Output

Complexity should be justified by the research need—not imported from high-resource settings.

Practical Strategies for LMIC Research Data

Design for reliability, traceability and sustainable operations.

- ❑ Define a minimum dataset before collecting everything available.
- ❑ Use standardized data dictionaries and local coding conventions.
- ❑ Use offline-capable capture where connectivity is unreliable.
- ❑ Assign unique study identifiers and document linkage rules.
- ❑ Build validation rules into data entry.
- ❑ Use role-based access and routine backups.
- ❑ Schedule data-quality review rather than waiting until analysis.
- ❑ Document every transformation from source to analytic variable.

Goal:
Reliable simplicity — not technological sophistication for its own sake.

The Informaticist Role in Research

Align the scientific question with the information system that must answer it.

Research Question	**Clinical Meaning**	**Data Representation**	**Technology**	**Workflow**	**Governance**	**Analysis**

The Informaticist asks:

Can the question actually be answered with these data—and can we trust what those data mean?

Applied Exercise | Can We Trust the Dataset?

Move from identifying errors to evaluating fitness for purpose.

Review a 15–20 patient research extract containing deliberate problems.

Identify five data-quality or interpretation issues.

Decide which can be corrected algorithmically.

Identify which require source or clinical verification.

Identify which could introduce bias.

Conclude whether the dataset is currently fit to answer the research question—and justify your decision.

Examples embedded in the dataset:
Duplicates • inconsistent IDs • missing stage • conflicting dates • unit differences • outside treatment

Course Synthesis

Trustworthy evidence is engineered across the entire information lifecycle.

- ❑ Research questions must become explicit, computable information requirements.
- ❑ Clinical data require context, provenance and transformation before research use.
- ❑ Quality is multidimensional and always evaluated for a specific purpose.
- ❑ Governance determines not only what can be accessed, but what may be used responsibly.
- ❑ Bias can be introduced by healthcare access, documentation, systems and extraction—not only statistics.
- ❑ LMIC constraints require resilient, appropriately simple architectures—not weaker standards.

Final Teaching Pearl

The goal of Research Informatics is not merely to produce a dataset.
It is to preserve enough meaning, quality and accountability that the evidence can be trusted.

APPLIED HEALTH INFORMATICS

Research Informatics & Data Management

From Research Question to Trustworthy Evidence

END

Instructors/Presenters:
Ashley Landay, MBA/MS
Chamberlain Nwanne, PhD