



National Institute of
Diabetes and Digestive
and Kidney Diseases

Central Repository

NIDDK-CR Resources for Research

Data Science and Meet the Expert Webinar Series



May 29, 2025



NIDDK Central Repository Overview

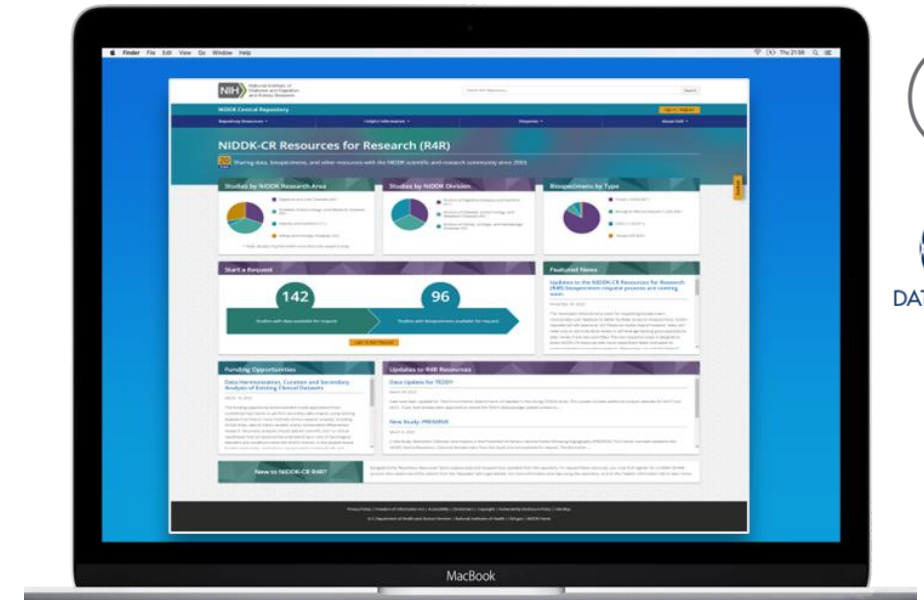
Mission


Established in 2003 to **facilitate sharing of data, biospecimens, and other resources** generated from studies supported by NIDDK and within NIDDK's mission by making these **resources available for request to the broader scientific and research community**.


- Supports receipt and distribution of data and biospecimens in a manner that is ethical, equitable, and efficient
- Enables investigators not involved with the original work to test new hypotheses without the need to collect new data or biospecimens
- Promotes FAIR (Findable, Accessible, Interoperable, and Reusable) and TRUST (Transparency, Responsibility, User focus, Sustainability, and Technology) principles





Recorded past tutorials, webinars, and other educational resources can be found on the NIDDK-CR website



Imaging Data Files

15.8 M

Clinical Datasets

>8,500
from 190 clinical studies

Biospecimens

>16 M

Registered Users

>6,900

Weekly Users

>5,000

Public Releases

>875

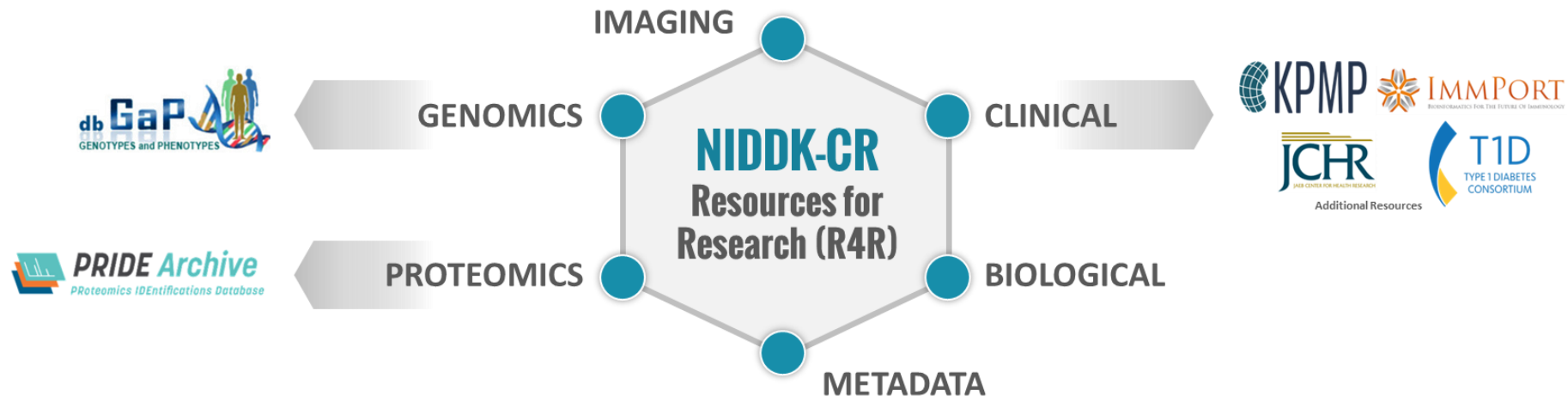


National Institute of
Diabetes and Digestive
and Kidney Diseases

Central Repository

NIDDK Data Sharing Ecosystem

The NIDDK-CR is a part of the broader NIH-funded biomedical data ecosystem and plays a key role in NIH's FAIRness and TRUSTworthiness goals. The NIDDK-CR houses a broad range of data types for secondary research, provides access to biospecimens, and direct links to other repositories with additional resources such as genomics data.



FAIRsharing.org
standards, databases, policies

DataCite
FIND, ACCESS, AND REUSE DATA

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES



Google Dataset Search

Schema.org

NIH U.S. National Library of Medicine
ClinicalTrials.gov

Vivli
CENTER FOR GLOBAL CLINICAL RESEARCH DATA

PLANNING
PHASE

figshare

NIH
HEAL
INITIATIVE



National Institute of
Diabetes and Digestive
and Kidney Diseases

Central Repository

Future Functionality: Analytics Workbench

Streamlining end-to-end data science lifecycle
and discovery of data-driven biomedical insights.

Innovation and ease of use

A cloud-based analytics environment
where researchers and data scientists
can access a suite of integrated analytics
tools and cloud computing resources to
participate in data challenges and AI
innovation.

Expected Benefits of Analytics Workbench:

Promote
Collaboration

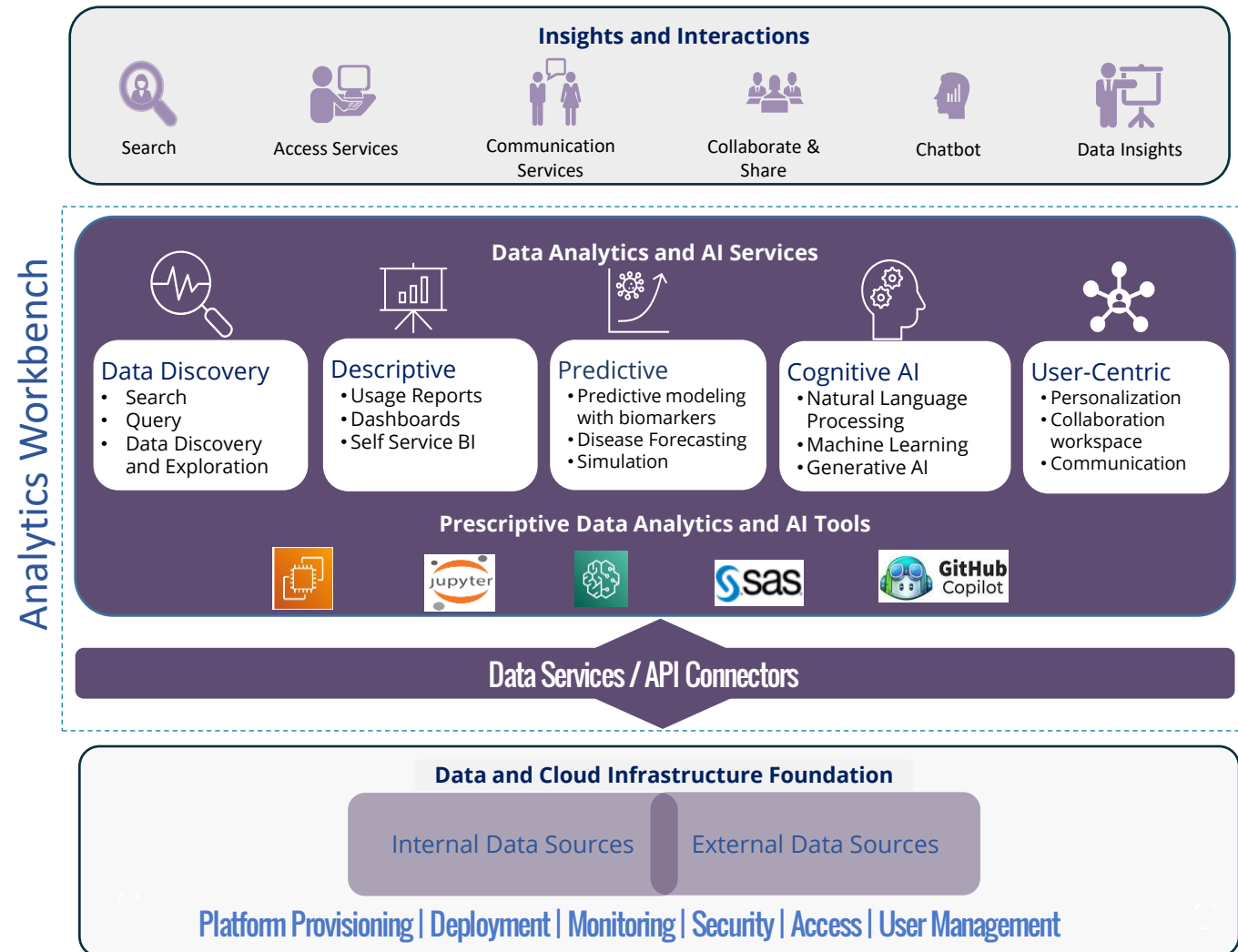
Support AI
Innovation

Minimize Data
Movement

Improve User
Experience

Discover
Data Insights

Advance NIDDK
Research Mission





National Institute of
Diabetes and Digestive
and Kidney Diseases

Central Repository

NIDDK-CR Data Science Centric Challenge Series

Goals of NIDDK-CR Data-science centric challenge series

- Develop tools, approaches, models and/or methods to increase data interoperability and usability for artificial intelligence (AI) and machine learning (ML) applications
- Augment and enhance existing data for future secondary research, including data-driven discovery by AI/ML researchers
- Discover innovative approaches to enhance the utility of datasets for AI/ML applications



Visit our website for more information on our data-centric movement and to learn more about our past data-challenges



National Institute of
Diabetes and Digestive
and Kidney Diseases

Central Repository

Secondary Data Science and Meet the Expert Webinar Series

About the Series

- Aims to accelerate data science and AI-driven biomedical research by fostering collaboration between biomedical researchers and experts in the field
- Monthly webinar held on the **last Thursday of each month**

Upcoming Webinars

- Today – FAIR and AI-ready data sharing
- June 26 – Different privacy preserving techniques and implications for researchers
- July 31 – Challenges, opportunities, and considerations for secondary researchers using electronic health records and real-world data sources
- August 28 – Impact and innovations realized



Learn more about the webinar series, register for future webinars, and access past webinars materials and recordings

Addressing Gaps, Challenges, and Opportunities Related to Data and Metadata Standards for NIDDK Research Priorities

Virtual • June 3–4, 2025



National Institute of
Diabetes and Digestive
and Kidney Diseases



Learn more and register for this workshop at:

<https://www.niddk.nih.gov/news/meetings-workshops/2025/data-standards-2025>

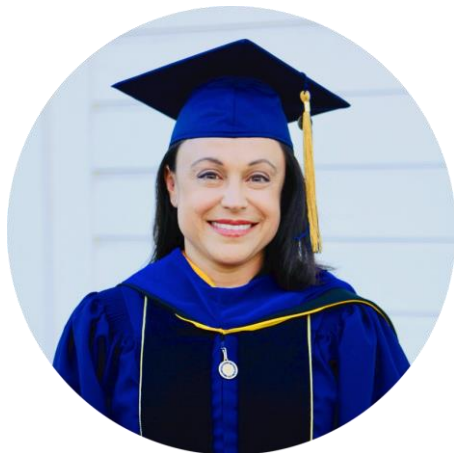




National Institute of
Diabetes and Digestive
and Kidney Diseases

Central Repository

Meet the Expert



Dr. Courtney D. Shelley, PhD, is a Health Data Scientist at Booz Allen Hamilton, where she focuses on data science education and AI-readiness of health-related data. She has supported the NIH Office of Data Science Strategy to develop online data science learning resources for pre-college and collegiate audiences, and to assess data science education across US universities to promote collaborative research between biomedical researchers and AI professionals. Prior to working at Booz Allen Hamilton, Dr. Shelley worked at Los Alamos National Laboratory, where she received the Postdoctoral Distinguished Performance Award for COVID-19 response efforts at local, state, and federal levels, as well as conducted research in suicide prevention with the support of the Department of Veterans Affairs and Million Veteran Program. She completed her PhD in Epidemiology with a focus on causal inference at University of California, Davis.



National Institute of
Diabetes and Digestive
and Kidney Diseases

Central Repository

FAIR and AI-Ready Data Sharing

NIDDK-CR Data Science Meet
the Experts Webinar Series

May 29th, 2025

Presented by: Booz Allen Hamilton





AGENDA

- A “Big Picture” and a “Common Foundation”
- Core competencies to support AI-assisted biomedical research
 - Competency 1: **Data Documentation**
 - Competency 2: **Ontology Usage & Data Encoding: Metadata**
 - Competency 3: **Data Cleaning and Formatting**
 - Competency 4: **Data Curation & Sharing : FAIR + CARE**
 - Competency 5: **Data Collaboration: Biomedical + AI Liaising**
- Training and educational resources available to researchers



Prepare and share AI-ready datasets that don't require you to be a Point of Contact for future users.

- Ideally, this can mean data you (or your institution) can maintain control of and host if necessary while still being FAIR (**F**indable, **A**ccessible, **I**nteroperable, and **R**eusable).
- Key concepts:
 - Data leakage
 - Handling missingness
 - Data generating processes
 - Data/Study support



AS A PRIMARY RESEARCHER: Infuse your data with your expert knowledge as the data generator who best understands the ***data generating process***.

- Provide as much detail as possible for future users, including study design and study population, machinery & methodology used to collect data. You are also best equipped to ensure the accuracy of your data values or to go back through records to correct errors.
- Key Concepts:
 - Documenting of data generating process
 - Metadata annotation and labeling
 - Data cleaning and formatting
 - Ethical considerations of data generating, handling, and sharing



AS A SECONDARY DATA USER: Beware of data leakage and introduced biases when combining datasets and study populations. Be sure you have permission and consent from data providers if possible/required.

- Key concepts:
 - Understanding study versus target populations
 - Understand biases associated with collected data
 - Techniques such as data harmonization & fusion to get larger datasets



AS A DATA COLLABORATOR: Understand your role of liaising between biomedical researchers and AI researchers.

- Supply duplicative terminology when possible.
- Help scientists understand that AI research is often iterative and help AI researchers understand that biomedical science is NOT iterative.
- Key concepts:
 - Biological data is precious
 - Respect populations from which data was derived and ensure they benefit from your research
 - Understand AI data requirements and what methodologies can and cannot



COMMON FOUNDATION

- Biomedical research aims toward the prevention and treatment of disease, as well as the genetic and environmental factors related to disease and health.
 - Researchers are predominantly **physicians and biomedical scientists**, and conduct **experimental, observational, and simulation studies**
 - Data may consist of:
 - **Time series** data from wearable devices and medical devices (e.g., ECGs, continuous blood glucose monitors)
 - **Longitudinal** data collected repeatedly over a time period
 - **Images**, such as radiologic or histopathologic
 - **Free-form text** (e.g., physicians' written notes)
 - **Quantitative** (e.g., measurements or laboratory values)
 - **Qualitative** (e.g., interpreted laboratory or radiologic findings, Likert-scale satisfaction scores, categorical data of race/ethnicity, family history of disease).



COMMON FOUNDATION

- Artificial intelligence (AI) uses computational problem-solving techniques to identify and learn from patterns across large and complex datasets.
 - AI methods include:
 - Machine learning for prediction and decision support
 - Natural language processing to interpret text or voice input
 - Automation and robotics to utilize and react to sensory input
 - Machine vision for image classification
 - Data mining to find correlations within large datasets
- **AI generally begins with the automated processing and analysis of data**



COMMON FOUNDATION

- Governmental and non-governmental organizations are increasingly moving toward **a state of “AI-readiness”** whereby AI use cases to be quickly developed and deployed.
- Secondary data analyses can maximize research dollars spent by **reusing data and associated metadata** collected in primary studies.
- Primary researchers generating data are in the better position to make data they generate AI-ready.
- Primary researchers can also benefit from knowledge of AI research and techniques to rigorously critique of methodology and findings.



AI-READINESS

AI-ready data are machine-readable, reliable, accurate, and explainable; reflective of the study population and predictive of target population; and are accessible for future AI applications when they arise to avoid the cost and time AI developers need to collect bespoke data.

An AI-ready dataset consists of data that is:

- Reflective of the population from which it was drawn
- Well documented and FAIR (**F**indable, **A**ccessible, **I**nteroperable, and **R**eusable)
- Model-agnostic



AI-READINESS COMPETENCIES

Competency	Description
1. Dataset Documentation	<ul style="list-style-type: none">• <i>Recognize the ethical, legal, and social implications of poor or incomplete dataset documentation and its effects on current and potential secondary usage.</i>• Generate comprehensive documentation of dataset to include study design and selection biases of study population• Dataset documentation should be included within standard data element documentation in the form of datasheets, data dictionaries, or semantic models.
2. Ontology Usage & Data Encoding	<ul style="list-style-type: none">• <i>Ethics of self-identification in human-subject data collection (i.e., free response of race/ethnicity, sex/gender, and disability status) and compliance with Federal data collection standards</i>• Apply accurate metadata to datasets using standardized ontologies to enable secondary-use researchers to interpret and apply data without having to contact the publishing research team.• Researchers are proficient in using secure data/metadata entry software, and domain-specific ontologies
3. Data Cleaning & Formatting	<ul style="list-style-type: none">• <i>Ethics of handling self-identification in human-subject data collection including roll-up to Federal classification standards</i>• Import structured and unstructured forms of data that are typically used in their domains to an IDE, transform data into a structured, machine-readable data frame format, and recode columns and segments of data for consistency and missingness.
4. Data Curation & Sharing	<ul style="list-style-type: none">• <i>Ethics of human-subject data storage and sharing including need to de-identify or perturb data to remove PII and PHI prior to sharing or release</i>• Can curate, share, archive, and plan for long-term management, preservation, access, and reuse of data after publication, including NIH-supported open-source data sharing repositories
5. Data Collaboration (AI-Assisted Reuse of Existing Datasets)	<ul style="list-style-type: none">• <i>Ethics of <u>data collection and aggregation</u>, <u>data usage</u> and potential for incorporating historical bias and discrimination in algorithmic results; ethics of <u>algorithmic deployment</u>, the difference between statistically and clinically significant findings, model re-training after deployment</i>• Knowledge of fundamental AI concepts to assist biomedical researchers with data needs of algorithms and AI researchers to develop clinically meaningful research questions



COMPETENCY 1: Data Documentation

- Documentation of study design and data elements, including:
 - Methodology
 - Study population description
 - Inclusion/exclusion criteria
 - Sampling procedures
 - Expected relationships between variables through the use of **conceptual modeling** or **knowledge graphs**
- ***Data generators have an ethical responsibility to limit the ethical, legal, and social implications of poor or incomplete dataset documentation on the analysis of current studies and potential secondary usage.***



COMPETENCY 1: Data Documentation

Research Design

- Document study design:
 - Research methodology (e.g., animal/tissue experimentation, observational, clinical trial, computational/simulation)
 - Study population description including (where applicable) sampling schema, patient inclusion/exclusion criteria, mouse strain name/sex, cell culture line, diseases, organs/body parts, etc.
 - Study population representatives (i.e., relationship between study population and general population)
 - Intended research application
- Document data elements:
 - Methodology of data element collection, including machinery/laboratory equipment and laboratory testing cut-off values, name of collecting researcher, inclusion/exclusion criteria, sampling schema used, etc.
 - Motivation for collecting data elements including, if applicable, identification of predictors, outcomes, confounders, biomarkers, and proxy variables
 - Expected relationships between data elements, such as through a conceptual model, knowledge graph, or causal diagram
 - When possible, collection of additional data elements that are easily recorded during the study and that may be useful in addressing secondary research questions, such as patient demographic information not directly related to the study question
- Create FAIR data management and sharing plan:
 - Identify relevant domain-specific standardized ontology that research team will adhere to, an intended open-source data repository for final release, and available institutional resources (such as campus library services, translational science centers, and/or data science centers) available to aid in data release



COMPETENCY 1: Data Documentation

- **Datasheets:** a standardized way to convey a *conceptual model* of expert domain knowledge surrounding the data elements collected in a study.
 - **Motivations:** for creating the dataset, including funding, any specific tasks the authors had in mind, and who the authors are.
 - **Composition:** what kinds of data are in the dataset, how it was collected, whether labels are associated with the data, and whether the dataset contains sensitive information.
 - **Collection Process:** how the data was collected, where or who it was collected from, who was involved in the collection process, and, if people are involved, if consent was given for the data to be collected.
 - **Processing:** Whether the data was processed or labelled and how it was done.
 - **Uses:** The tasks the dataset is intended to be used for, how it has already been used, and limitations of use.
 - **Distribution:** How the dataset will be distributed and to who, and any restrictions on distribution.
 - **Maintenance:** Who and how the dataset will be maintained, and if and how others will be able to build on it.



COMPETENCY 1: Data Documentation

- The use of a ***conceptual model (or “framework”)*** to help secondary researchers understand how the study population generating the dataset relates to a wider population and to understand important factors that may not have been collected during the study but may be relevant in future applications.
 - In addition to aiding in data harmonization, a conceptual framework can help primary researchers identify additional data elements that can be easily collected at the time of study to avoid biases in secondary studies.
 - Failure to account for all possible confounders when collecting and annotating data can lead to unintended biases in datasets, which will be perpetuated in AI applications built with the dataset.



COMPETENCY 2: Ontology Usage & Data Encoding

- Domain-specific ontologies standardize terms and reduces ambiguity.
- Apply sufficiently granular and accurate metadata so future users can interpret and apply data elements to a question of interest without having to contact the publishing research team.
- The use of secure data/metadata entry software (such as REDCap)
- ***An important consideration here is self-identification in human-subject data collection (i.e., free response of race/ethnicity, sex/gender, and disability status) and compliance with federal data collection standards.***
- Data can then be made FAIR through efforts to standardize file formats and ensure machine readability of data and documentation. ***We emphasize an expectation that these tasks will be performed by the research team or their available resource centers, and not necessarily by a single individual.***
 - **Ontology:** Classification system to express the properties of a subject area and how they are related by defining a set of concepts and categories. Ontologies are more sophisticated versions of taxonomies because they track relations across entities. One of the world's most comprehensive repositories of biomedical ontologies is BioPortal, developed and is managed by the National Center for Biomedical Ontology (NCBO).
 - The **United States Core Data for Interoperability (USCDI)** is a standardized set of health data classes and interoperable data elements for nationwide health information exchange, developed by Office of the National Coordinator (ONC).



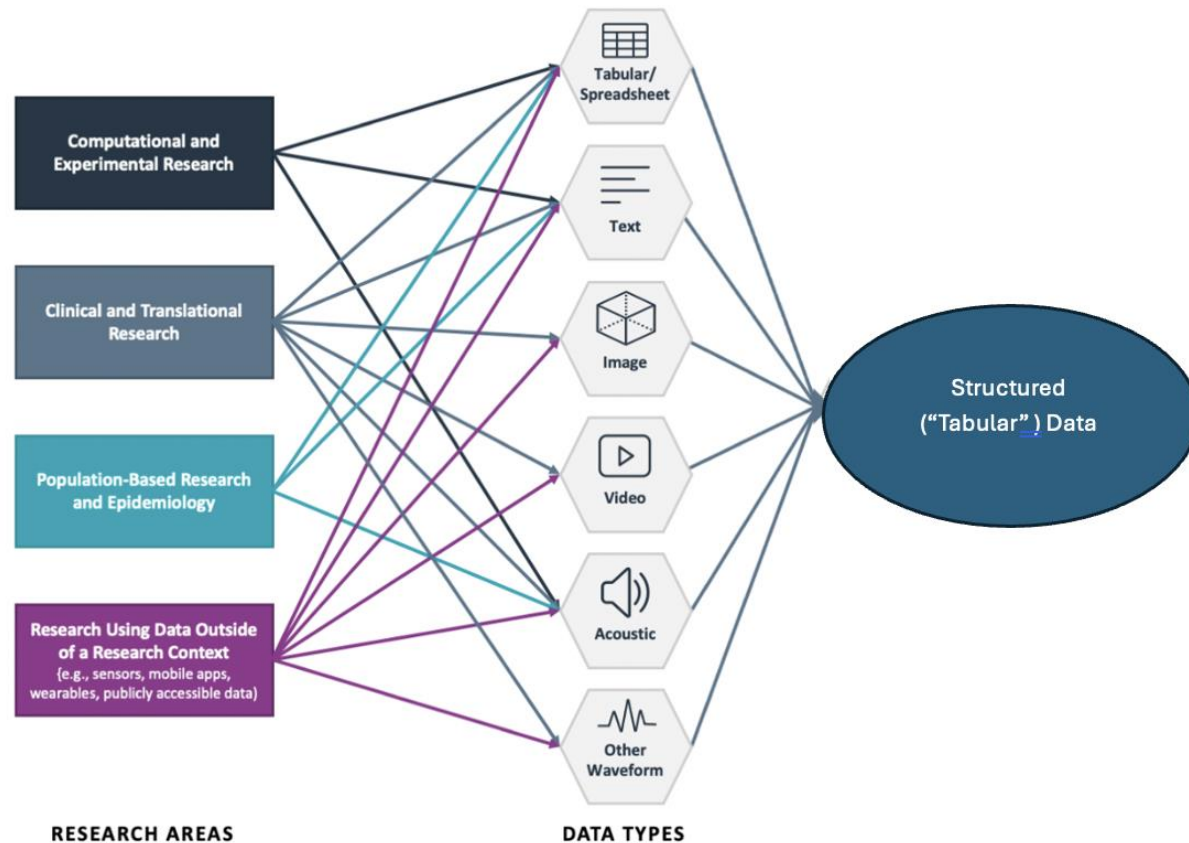
COMPETENCY 3: Data Cleaning & Formatting

“79% of a data scientist’s research time is spent finding and cleaning data. [Therefore, people] working intensively with data in their daily work ... apply their domain expertise [just one] day per week because in the other [four] days they are finding/collecting, cleaning and organizing data.” Dr. L.O. Bonino, Associate Professor
University of Twente

- ***Ethics of handling self-identification in human-subject data collection including ‘roll-up’ to federal classification standards and interpretation of classification categories in findings, as well as an understanding that these are evolving concepts that may not be accurately/consistently reflected in older datasets.***



COMPETENCY 3: Data Cleaning & Formatting



Data cleaning steps:

- 1) **Importing the data.** Ensuring it is read in correctly. Basic data exploration including variable types, number of observations per feature, missingness, potential coercion to incorrect type as a signal of errors.
- 2) **Visualizations, tables**
- 3) **“Pre-processing”:** typos, consistent labels, converting encoded values to labels, handling of missingness, feature engineering, dates, one-hot encoding
- 4) **Data harmonization and fusion.** MAY BE NECESSARY, ESPECIALLY IF YOUR DATA WAS COLLECTED AND STORED IN A RELATIONAL DATABASE.
- 5) **Data processing documentation**
- 6) **Data Models**

Step-by-step video tutorial and Jupyter notebook on “Performing Pre-Model Processing and Data Quality Checks” available on the [Informational/Instructional Information](#) page on the NIDDK-CR website



COMPETENCY 4: Data Curation & Sharing

- Data curation* is the process of managing and organizing data, and especially to make data **FAIR**:
 - Findable
 - Accessible
 - Interoperable
 - Reusable
- Data curators may need to curate existing data if standards change or anomalies are identified. They will need to archive data and plan for long-term management, preservation, and access, which may include NIH-supported open-source data sharing repositories or accessible storage through their home institution, which allows for long-term control of the data but also requires understanding of data sharing best-practices.
- ***Ethical expectations for human-subject data storage and sharing including need to de-identify or perturb data to remove PII and PHI prior to sharing or release and have sufficient understanding of study context and design to identify and document potential threats to inference.***

(*) Data governance is a similar concept focusing on quality, stewardship, protection/compliance, and management more in line with government or industry.



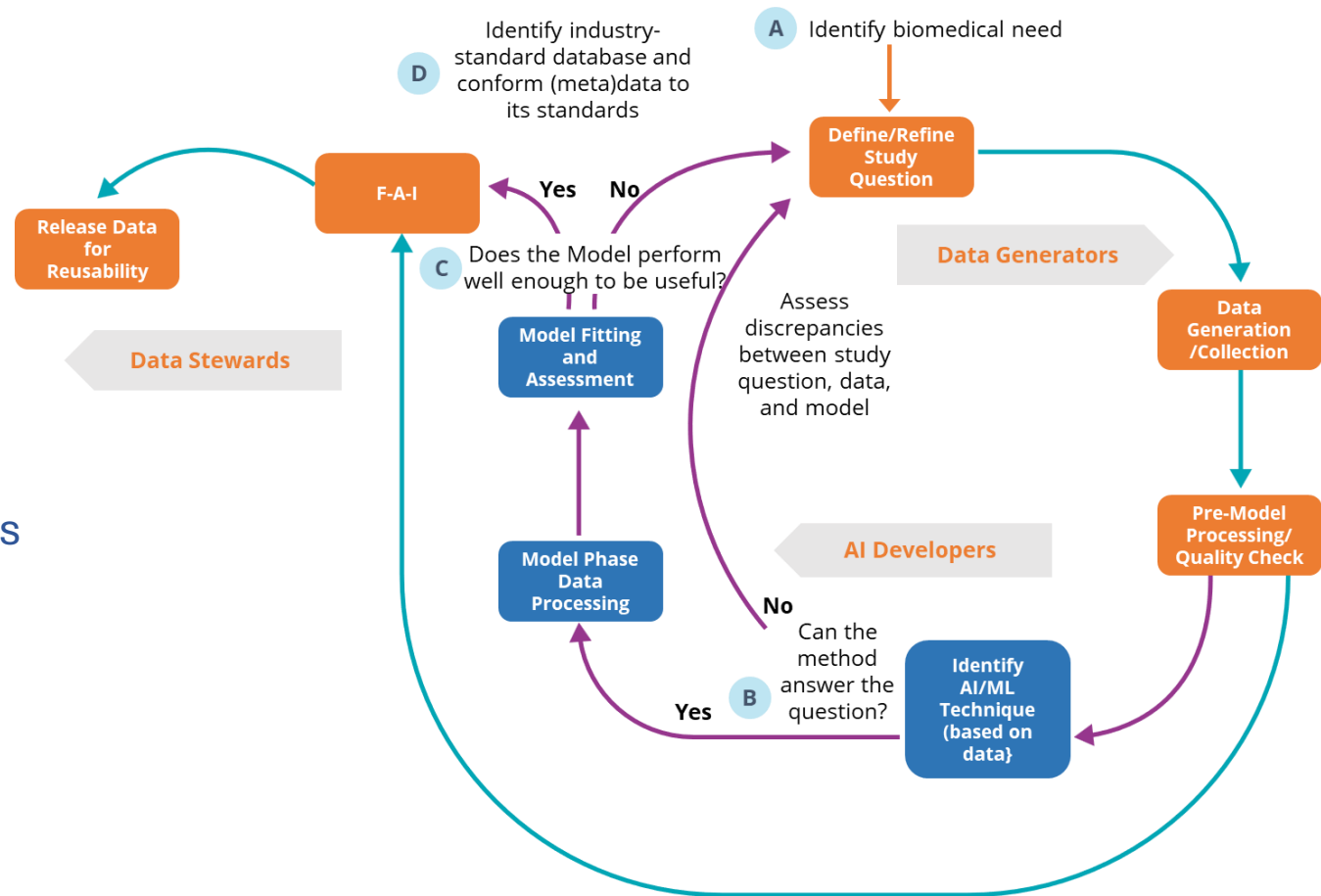
COMPETENCY 4: Data Curation & Sharing

- **An Additional Consideration for Population-Derived Data is CARE:**
- Data that is used for
 - the **Collective** benefit of those from which data was collected
 - whose populations maintain **Authority** to control the data
 - data collectors have a **Responsibility** to interact with minoritized populations respectfully
 - the **Ethics** of populations from which data was collected are respected



COMPETENCY 5: Data Collaboration/AI-Assisted Reuse

- **Facilitate interdisciplinary collaboration across biomedical and data science research teams** during iterations of study design, data collection, and model development/analysis
- Requires **sufficient knowledge of AI concepts** to assist biomedical researchers with algorithmic data needs and **sufficient knowledge of biomedical concepts** to convey to AI developers relevant and precise biomedical contextual information to generate and assess clinically meaningful results.





COMPETENCY 5: Data Collaboration/AI-Assisted Reuse

- Mitigation of the damaging ethical implications of:
 - Poor data collection and aggregation methods
 - Inadequately documented and/or labeled historical or discriminatory bias during model development
 - Insufficiently rigorous explorations of the effects that findings from algorithmic deployments could have on patients if incorporated into clinical workflows
 - Difference between statistically and clinically significant findings
 - How new treatment practices will require model re-training due to increasing noncomparability between pre- and post-treatment datasets.



COMPETENCY 5: Data Collaboration/AI-Assisted Reuse

Data Collaboration Considerations

- **Think very hard about who's "not there"**. Rural and impoverished areas and conflict zones will continue to use paper for the foreseeable future.
- **Think very hard about what's "not there"**. Consider data providers' cultural backgrounds, sensitivity to providing certain personal information, or biased responses depending on how questions are asked and in what order.
- **Think very hard about what's not "being asked"**. Similar to the idea of study populations versus target populations, consider who is providing the data versus who is benefiting from the data.
 - Engage junior researchers, students, and community members to join in data questions, bring in their own lived experiences and interests, and attempt to seek root causes or overlooked details.



Key Concepts

Bias

- Advances in AI offer the potential to provide personalized care by taking into account individual differences. **At the same time, because machine learning algorithms aggregate and assess large volumes of real-world data, AI can reinforce bias in data, potentially reinforcing existing patterns of discrimination. Machine learning algorithms may work well for one patient group, but results may not be appropriate for others**
- **Sources of Bias**
 - Missing data
 - Sample size
 - Misclassification or measurement error
 - Mismatch between study population and target population. Statistical models have a concept of data support. Models are not accurate outside of the range of observations they have seen. This problem can extend to AI research when we attempt to apply models to populations we did not train the model on.



Key Concepts

Data Leakage

- ***Data leakage*** is the inadvertent process of training a model on data that will not be available when deployed.
- Since AI looks for patterns in data, it can find and utilize unintended patterns.
- Data leakage can also occur during model building.
- Preventing algorithms from making biased decisions is challenging and



Key Concepts

Data Harmonization and Fusion

- AI algorithms often need **A LOT** of data.
- One way to achieve this is by combining datasets, called ***data harmonization***.
 - Datasets to be combined must be ***very similar***. Ideally you are only reconciling column names and labeling schema.
- ***Data fusion*** combines dissimilar datasets to achieve insights that can't be gained from the individual datasets alone.
 - This is a new area of research and so best practices are evolving.
 - Considerations when creating a fusion dataset include spatiotemporal fallacies, incorrect use of proxy features, and fusing datasets representing differing populations



National Institute of
Diabetes and Digestive
and Kidney Diseases

Central Repository



Q&A and Poll



National Institute of
Diabetes and Digestive
and Kidney Diseases

Central Repository

Contacts:

- Dr. Courtney Shelley - shelley_courtney@bah.com

Upcoming Webinar: Different Privacy-Preserving Methodologies and Implications for Researchers

- **Date:** June 26th from 2-4pm ET
- **Experts:** Dr. Susan Tenney, Shruti Gautam, Datavant
- **Scan the QR code register**



Thank You!



National Institute of
Diabetes and Digestive
and Kidney Diseases

Central Repository

Training and Educational Resources





National Institute of
Diabetes and Digestive
and Kidney Diseases

Central Repository

Training and Educational Resources: Data Documentation

Course Title	Proficiency	Research Areas	Cost	Duration (Hours)	Institution
Creating API Documentation	Comprehension	Targets non-biomedical field; Research Using Data Outside of Research Context	29.99	0.5	Linkedin Learning
Data Management for Beginners - Main Principles	Comprehension	Generic	49.99	2	Udemy
Reproducible Templates for Analysis and Dissemination	Comprehension	Generic	0	20	Emory University
Writing in the Sciences	Comprehension	Clinical and Translational Research	0	30	Stanford
Big Data Integration and Processing	Basic	Generic	0	18	UC San Diego
Data Curation Foundations	Basic	Generic	29.99	5	Linkedin Learning
Data Dictionary (RC-201)	Basic	Clinical and Translational Research	N/A	1	University of Washington
Documentation and Usability for Cancer Informatics	Basic	Clinical and Translational Research	0	6	Johns Hopkins
Health Informatics Specialization	Basic	Computational and Experimental Research; Clinical and Translational Research	0	25	Johns Hopkins
Learn API Documentation with JSON and XML	Basic	Targets non-biomedical field; Research Using Data Outside of Research Context	34.99	1	Linkedin Learning
CERTaIN: Pragmatic Clinical Trials and Healthcare Delivery Evaluations	Foundational	Clinical and Translational Research	249	15	edX
Principles, Statistical and Computational Tools for Reproducible Data Science	Foundational	Computational and Experimental Research; Generic	0	40	Harvard University
Study Designs in Epidemiology	Full Performance	Population Based Research and Epidemiology	0	8	Imperial College London
Designing Big Data Healthcare Studies, Part One	Expert	Population Based Research and Epidemiology	39.99	1.5	Linkedin Learning
Designing Big Data Healthcare Studies, Part Two	Expert	Population Based Research and Epidemiology	44.99	2	Linkedin Learning



National Institute of
Diabetes and Digestive
and Kidney Diseases

Central Repository

Training and Educational Resources: Metadata Annotation and Labeling

Course Title	Proficiency	Research Areas	Cost	Duration (Hours)	Institution
Fundamentals of Data Warehousing	Comprehension	Generic	0	14	Learn Quest
Introduction to REDCap (RC-101)	Comprehension	Clinical and Translational Research	N/A	1.5	University of Washington
Metadata Repositories	Comprehension	Targets non-biomedical field	49.99	2	Udemy
REDCap Training Videos	Comprehension	Clinical and Translational Research	N/A	3.5	University of Chicago
Metadata Management Fundamentals	Basic	Generic	99.99	3	Udemy
Healthcare Data Literacy	Foundational	Clinical and Translational Research	0	13	UC Davis



National Institute of
Diabetes and Digestive
and Kidney Diseases

Central Repository

Training and Educational Resources: Data Cleaning and Formatting

Course Title	Proficiency	Research Areas	Cost	Duration (Hours)	Institution
AI for Scientific Research Specialization	Comprehension	Computational and Experimental Research	0	12	Learn Quest
Data Infrastructure and AI/ML	Comprehension	Generic	0	1	Linkedin Learning
AI Fundamentals for Non-Data Scientists	Basic	Targets non-biomedical field	0	7	University of Pennsylvania
Fundamentals of Scalable Data Science	Basic	Generic	0	22	IBM
Intro to R and RStudio for Genomics	Basic	Computational and Experimental Research	0	8	Data Carpentry
R Programming	Basic	Computational and Experimental Research	0	57	Johns Hopkins University
Descriptive Healthcare Analytics in R	Foundational	Computational and Experimental Research; Clinical and Translational Research; Population Based Research and Epidemiology	49.99	4	Linkedin Learning
Analyze Datasets and Train ML Models using AutoML	Full Performance	Generic	0	19	DeepLearning.AI AWS



National Institute of
Diabetes and Digestive
and Kidney Diseases

Central Repository

Training and Educational Resources: Data Curation and Sharing

Course Title	Proficiency	Research Areas	Cost	Duration (Hours)	Institution
Common Data Elements: Standardizing Data Collection	Comprehension	Generic	0	1	National Library of Medicine
CompTIA Data+ (DA0-001) Cert Prep: Domain 5.0 Data Governance, Quality, and Controls	Comprehension	Targets non-biomedical field	29.99	1	Linkedin Learning
Data Steward Foundations	Comprehension	Generic	29.99	2	Linkedin Learning
NIH Data Management and Sharing Requirements Series	Basic	Clinical and Translational Research	N/A	5	National Library of Medicine
Research Data Management and Sharing	Basic	Generic	0	14	The University of North Carolina at Chapel Hill and The University of Edinburgh
Open Science: Sharing Your Research with the World	Foundational	Generic	0	24	TU Delft
Reproducible Research	Foundational	Generic	0	8	Johns Hopkins University
Privacy by Design: Data Sharing	Full Performance	Targets non-biomedical field	34.99	1.25	Linkedin Learning
FAIR Data Management Training Course	Basic	Generic	N/A	N/A	University of Turin



National Institute of
Diabetes and Digestive
and Kidney Diseases

Central Repository

Training and Educational Resources: AI + Biomedical Liaising

Course Title	Proficiency	Research Areas	Cost	Duration (Hours)	Institution
AI for Everyone: Master the Basics	Comprehension	Generic	0	8	IBM Skills Network
Introduction To Artificial Intelligence	Comprehension	Generic	0	1	Udemy
The Essentials of Data Literacy Online Course	Comprehension	Generic	0	40	Davidson College
AI in Healthcare Specialization	Basic	Clinical and Translational Research	0	90	Stanford
Artificial Intelligence Foundations: Machine Learning	Basic	Generic	24.99	1.25	Linkedin Learning
Artificial Intelligence: Ethics & Societal Challenges	Basic	Generic	0	13	Lund University
Big Data in the Age of AI	Basic	Generic	29.99	2	Linkedin Learning
Biostatistics in Public Health Specialization	Basic	Generic	0	64	Johns Hopkins University
Data Management for Clinical Research	Basic	Clinical and Translational Research	0	17	Vanderbilt University
Introduction to Artificial Intelligence (AI)	Basic	Generic	0	11	IBM Skills Network
Introduction to Statistics & Data Analysis in Public Health	Basic	Population Based Research and Epidemiology	0	16	Imperial College London
Research Designs in Epidemiology	Basic	Population Based Research and Epidemiology	29.99	1.5	Udemy
The Data Science of Healthcare, Medicine, and Public Health	Basic	Clinical and Translational Research; Population Based Research and Epidemiology	34.99	1	Linkedin Learning
Demystifying Biomedical Big Data: A User's Guide	Foundational	Computational and Experimental Research	0	48	Georgetown University
Machine Learning	Foundational	Generic	0	140	Georgia Institute of Technology
Using Public Health Data Sources	Foundational	Population Based Research and Epidemiology	0	1	Madecraft
Data Science in Stratified Healthcare and Precision Medicine	Full Performance	Computational and Experimental Research; Clinical and Translational Research; Research Using Data Outside of Research Context	0	17	The University of Edinburgh



National Institute of
Diabetes and Digestive
and Kidney Diseases
Central Repository

Training and Educational Resources: AI + Biomedical Liaising (continued)

Course Title	Proficiency	Research Areas	Cost	Duration (Hours)	Institution
Essentials of Genomics and Biomedical Informatics	Full Performance	Generic	0	36	IsraelX
Machine Learning and AI Foundations: Prediction, Causation, and Statistical Inference	Full Performance	Generic	29.99	2	Linkedin Learning
Machine Learning and AI Foundations: Producing Explainable AI (XAI) and Interpretable Machine Learning Solutions	Full Performance	Generic	29.99	2	Linkedin Learning
Network Analysis in Systems Biology	Full Performance	Generic	0	30	Icahn School of Medicine Mount Sinai



National Institute of
Diabetes and Digestive
and Kidney Diseases

Central Repository

Training and Educational Resources: Ethical Considerations

Course Title	Proficiency	Research Areas	Cost	Duration (Hours)	Institution
Ethics in AI and Data Science	Comprehension	Generic	0	12	The Linux Foundation
AI Accountability Essential Training	Basic	Generic	49.99	2	Linkedin Learning
Artificial Intelligence Data Fairness and Bias	Basic	Generic	0	6	LearnQuest
Big Data, Artificial Intelligence, and Ethics	Basic	Generic	0	12	UC Davis
Data Science Ethics	Basic	Generic	0	15	University of Michigan
Debiasing AI Using Amazon SageMaker	Basic	Generic	39.99	2	Linkedin Learning
Ethical Issues in Data Science	Basic	Generic	0	24	University of Colorado Boulder
Machine Learning and AI Foundations: Predictive Modeling Strategy at Scale	Basic	Generic	39.99	1.3	Linkedin Learning
Tech On the Go: Ethics in AI	Basic	Generic	29.99	0.6	Linkedin Learning
The Total Data Quality Framework	Basic	Generic	0	12	University of Michigan
Data for Machine Learning	Foundational	Computational and Experimental Research	0	12	AMii
Validity and Bias in Epidemiology	Foundational	Population Based Research and Epidemiology	0	8	Imperial College London
Bias and Discrimination in AI	Full Performance	Generic	0	7.5	Université de Montréal
Foundations of Responsible AI	Full Performance	Generic	29.99	2.5	Linkedin Learning
Power and Sample Size for Multilevel and Longitudinal Study Designs	Full Performance	Generic	0	19	University of Florida