

5.2 Data Science

10x10 – Introduction to Biomedical Informatics & Artificial Intelligence
William Hersh, MD, FACMI, FAMIA, FIAHSI
Copyright 2026, All Rights Reserved



Data science

Data use

Data quality

Data bias

Best practices for data collection, sharing, and use

Data science workforce

Textbook on applications, uses, and challenges (Spector, 2023)

Reminder: data is plural; datum is singular

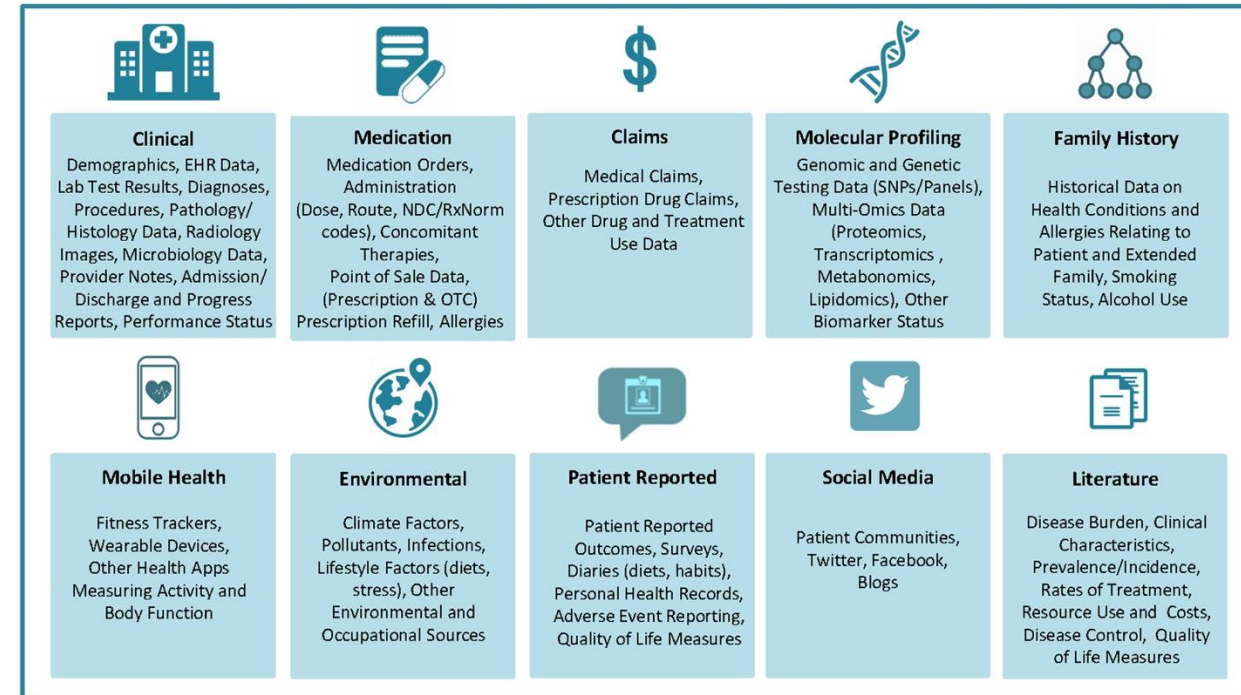
Data use

Even before widespread adoption of EHRs, recognized value for “secondary use” or re-use of data (Safran, 2007)

Data thought to be “new oil” but actually “new soil” (Smart, 2021)

Opportunities

- Beyond structured data of EHR in unstructured (narrative) data (Tayefi, 2021)
- Variety of uses, including beyond EHR (Liu, 2022)
- Including generation of “real-world evidence” from EHR and other clinical data (Bastarche, 2022)

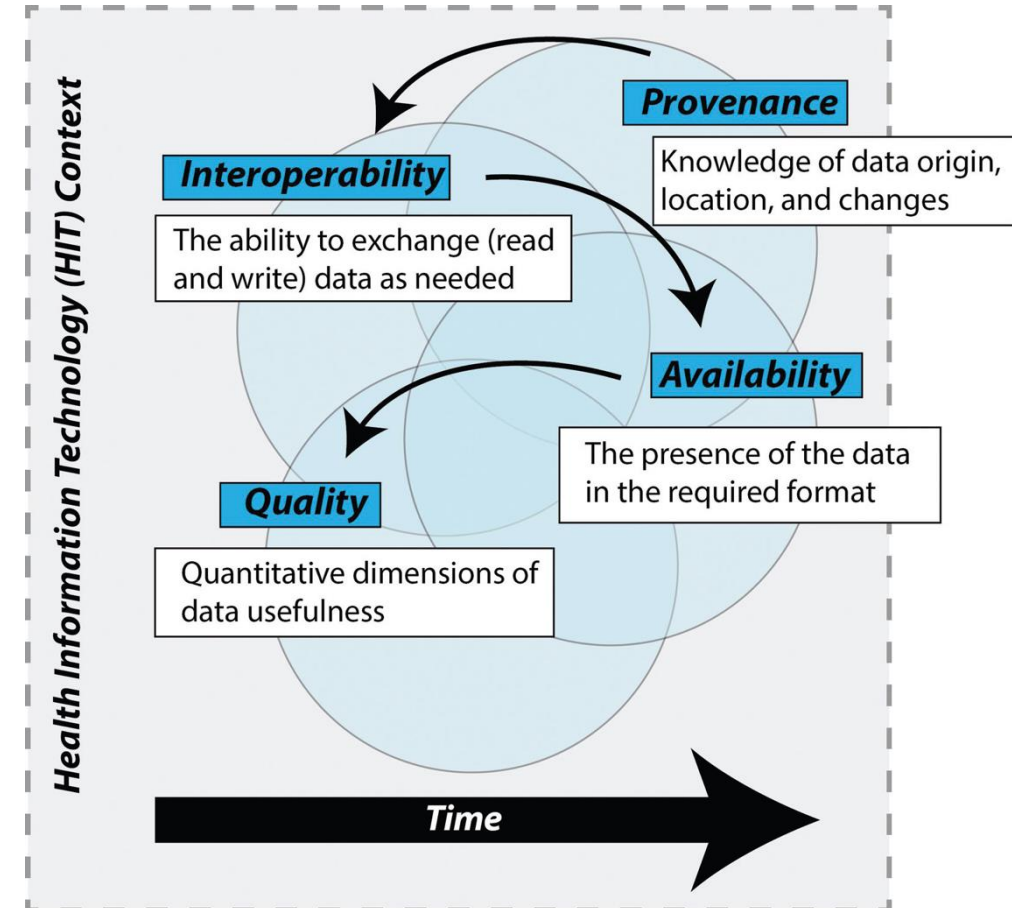


Data readiness and usability

Data readiness (Douhit, 2021)

Data usability (Sequoia Project, 2022)

- Data provenance and traceability of changes
- Effective use of codes in shared information
- Reduce impact of duplicates
- Data integrity/trust
- Data tagging/searchability
- Effective use of narrative for usability



Data quality

Long known that data quality and accuracy not a top priority for busy clinicians (de Lusignan, 2006)

Quality issues include

- Quantity
- Caveats
- Heterogeneity
- Missingness
- Inaccuracy
- Drift
- Bias

Data quantity

One challenge for data quality is its quantity

Large quantities of data, especially in EHR

- Average pediatric ICU patient generated 1348 information items per 24 hours (Manor-Shulman, 2008)
- Average hospital admission had 137,882 tokens (discrete pieces of data), which increased to 216,744 at discharge (Rajkomar, 2018)

Parts of EHR may be more difficult to process

- Much data “locked” in text (Hripcsak, 2013)
- Even structured data may not be usable in purely automated manner for clinical score calculation (Aakre, 2017)

Quantity of data in EHR varies across patients and conditions

Patients get care at different places (Bourgeois, 2010; Finnell, 2011)

Sicker patients have more data (Weiskopf, 2013; Rusanov, 2014)

Informed presence bias – people in EHR different from those not getting care (Goldstein, 2016)

Data completeness varies by patients; filtering by completeness may bias data (Weber, 2017)

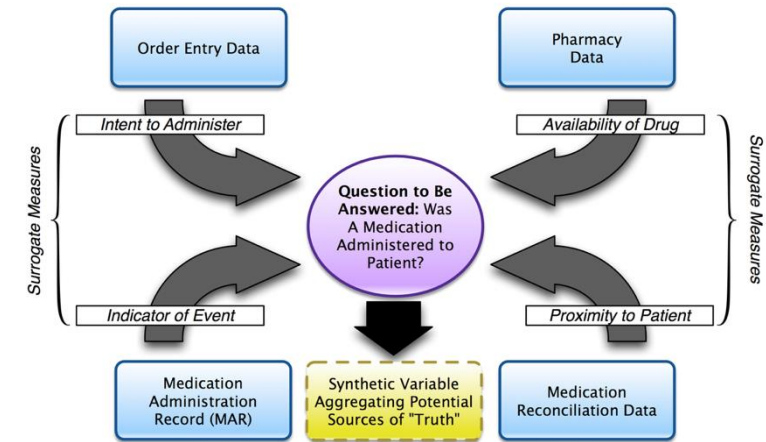
Absent and discordant EHR documentation of complementary and alternative medicine in cancer (Stan, 2018)

Some rare diseases have long lead-time for diagnosis, e.g., Peter Embi pheochromocytoma only in record of institutions after diagnosis (personal communication)

Caveats for use of operational EHR data

May be (Hersh, 2013)

- Inaccurate
- Incomplete
- Transformed in ways that undermine meaning
- Unrecoverable
- Of unknown provenance
- Of insufficient granularity
- Incompatible with research protocols



EHR data for secondary analysis – “best of the worst” (Agha-Mir-Salim, 2024)

- EHR is "forced structure"
- Limitations reduce reflections of real-world diversity of patients and their diseases
- EHR vendors do not design human-centered systems that encourage high quality data capture
- Data harmonization inadvertently adding a layer of bias
- Clinical translation of AI models trained on “high-quality” EHR data that may not reflect data captured in other settings

EHR data problems – heterogeneity

Challenges in interoperability with different and same EHRs (Glynn, 2019; Bernstam, 2022)

Difficulties in digital phenotyping from EHR data (Bastarche, 2021)

EHR-derived datasets like N3C difficult to identify and control for underlying cross-site heterogeneity, such as completeness and standards (Brown, 2021)

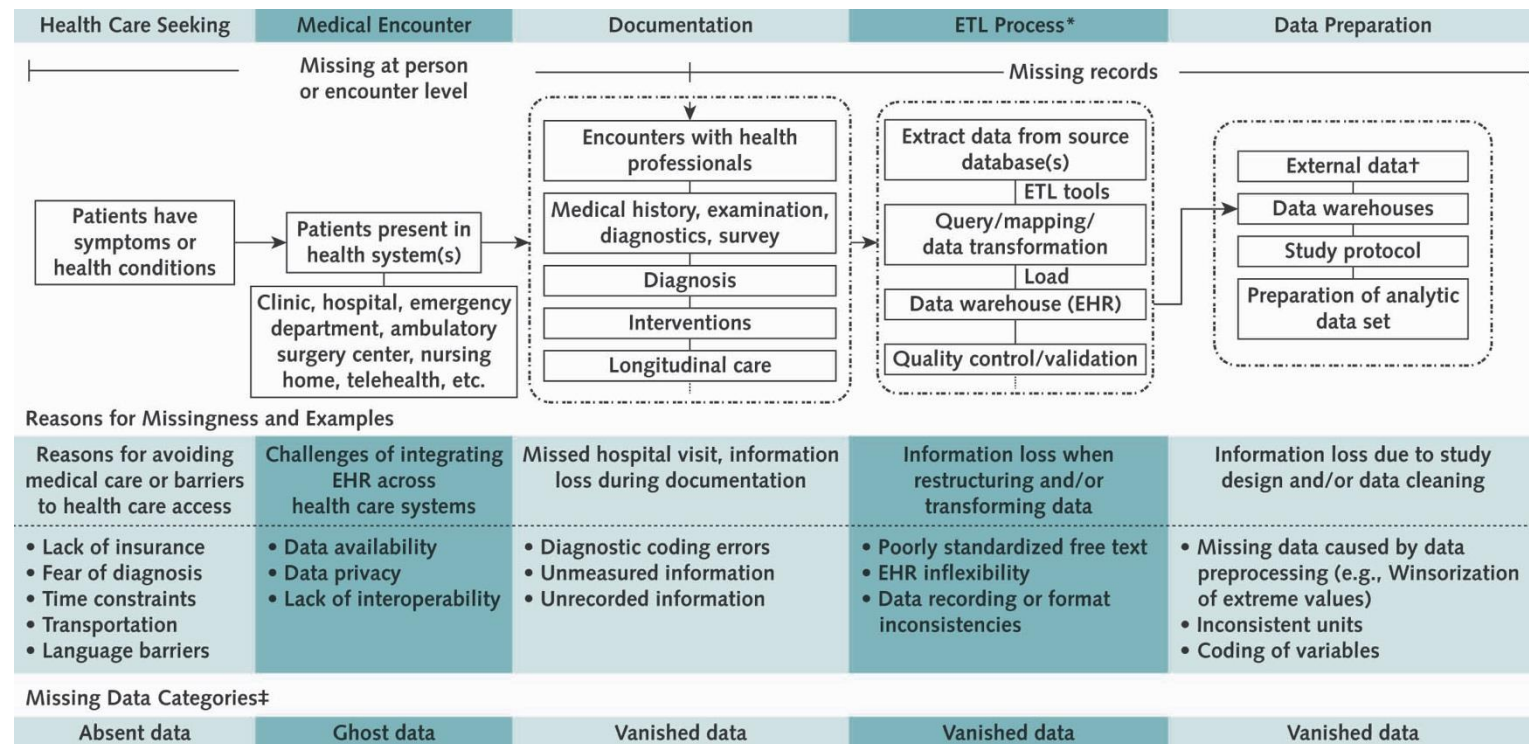
Missing and inaccurate data undermined response to COVID-19 pandemic (Arvisais-Anhalt, 2021)

Issues raised by patients not always captured in notes (Weiner, 2024)

Data missingness

Missing data in EHRs can lead to biased estimates of treatment effects and false-negative findings (Zhou, 2023)

Many reasons for missingness in EHR data – need methods to identify and control for predictive models (Lin, 2025)



EHR data problems – inaccuracy

Patient exam findings in notes – higher for EHR than paper (Yadav, 2017)

Mismatch between medications mentioned in notes vs. actually prescribed (Ashfaq, 2019)

Discordance between what physicians document vs. what they ask and examine (Berdahl, 2019; Weiner, 2020)

Patient-discovered errors

- One in five patients identify errors in their records, with 40% of errors “serious” (Bell, 2020)
- 6.5% of patients reading notes requested corrections (Nguyen, 2022)

Inaccuracy related to specific elements and situations

Inaccuracies in problem lists undermines risk adjustment (Daskivich, 2018)

Misclassification of diabetes type and episodes of ketoacidosis in clinical data warehouses (VanderWeele, 2018)

Pathologists make errors in assigning elements of LOINC codes (Stram, 2020)

Data quality for clinical decision support in hypertension inconsistent (Dorr, 2021)

Deceased patients not well-identified in EHR – one health system found 19% marked alive in EHR, a problem not only for data quality but also resulting in patients and families continuing to be contacted (Wenger, 2023)

Inaccuracy of specific elements (cont.)

Missing or inaccurate SDOH data (Cook, 2021)

In AllOfUs, inaccuracies of patient height and weight (Guide, 2024)

Errors in race or ethnicity in pediatric records (Freed, 2024)

Filtering for completeness of records may introduce race or ethnicity bias (Cortina, 2024)

Stigmatizing language may undermine accuracy (Barcelona, 2024)

Dataset shift/drift

Dataset shifts due to changes in technology, population and setting, and clinician or patient behavior (Finlayson, 2021; Knaus, 2022)

Impacts models using data (Nelson, 2025)

TYPE OF DRIFT/CHANGE	DESCRIPTION	EXAMPLES
Data Drift (Feature Drift)	Changes in the statistical properties or patterns of the input data.	▪ Introduction or withdrawal of medications. ▪ Changes in data collection or documentation (e.g. new EHR system). ▪ Shift in patient demographics (e.g. aging patient population) ▪ Adoption of new clinical practice protocols or guidelines.
Label Drift	Changes in the definition of the model output (labels).	▪ Redefined of clinical outcomes (e.g. how re-admissions are counted). ▪ Changes in diagnostic criteria. ▪ Updates in coding systems (e.g. ICD-9 to ICD-10 codes). ▪ Policy changes affecting quality metrics or treatment protocols.
Concept Drift	Changes in the relationship between input features and output labels.	▪ Evolving medical knowledge altering understanding of diseases. ▪ New treatment protocols becoming standard of care. ▪ Technological advancements in diagnostic tools. ▪ Emergence of new diseases or antibiotic-resistant infections.
Covariate Shift	Changes in input data distribution, when the input-output relationship remains the same.	▪ Applying a model trained on adults to pediatric patients ▪ Using the same model in different clinical settings. ▪ Expanding AI use from physicians to nurses and pharmacists altering input data distribution.
External Factors	Changes due to external conditions like policy changes or global events that affect the data.	▪ New healthcare regulations or drug formulary adjustments. ▪ Drug shortages leading to alternative treatments. ▪ Pandemics and epidemics.
Seasonal Variation	Predictable, time-based changes affecting data patterns.	▪ Increases in respiratory illnesses during certain seasons. ▪ Seasonal health campaigns impacting healthcare utilization patterns. ▪ Weather-related health issues (e.g. Heatwaves).

Data bias

Racial, ethnic, and other biases that clinicians have get reflected in medical records (Sun, 2022; Tahir, 2022)

De-identified data may obscure underlying health and economic disparities, undermining respect and justice (Gross, 2022)

Children with asthma of lower socioeconomic status had higher rate of error in algorithm predicting exacerbation, mostly due to missing information (Juhn, 2022)

Including social determinants of health data improved sepsis readmission prediction models (Amrollahi, 2022)

Data bias may undermine trust in studies and algorithms

Embedded pragmatic clinical trials using clinical data at risk from biases in data (Boyd, 2023)

- Incomplete and variable capture of data on social determinants of health
- Lack of representation of vulnerable populations that do not access or receive treatment
- Data loss due to variable use of technology

Many examples of data being “tortured” and transformed into disinformation (Smith, 2023)

May be competing interests between digital health and indigenous data sovereignty, e.g., “data colonization” (Cordes, 2024)

Best practices for data collection, sharing, and use

Principles for health information collection, sharing, and use (Spector-Bagdady, 2023)

- Make research generalizable, especially to historically underrepresented communities
- Collect and use data with highest respect and transparency
- Educate and involve patients in its value
- Make research and its data accessible, interoperable, and replicated
- Adherence to appropriate government laws and regulations

Technical

- Must understand clinical workflow to avoid common pitfalls (Sauer, 2022; An, 2025)
- Metadata standards, e.g., HHS (HealthData.gov, 2025)

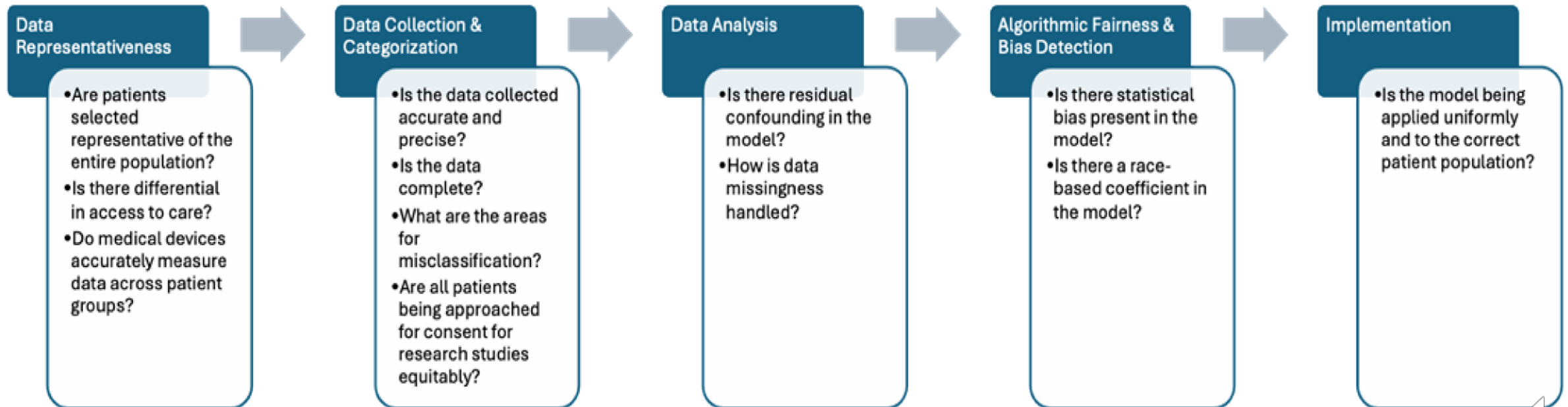
Governance

- Biomedical data repositories require governance, especially when used for AI/machine learning applications, at every step (Clayton, 2025)
- Ownership – governance more important than public vs. private ownership (Piasecki, 2022) but need trust in process (Dasgupta, 2026)

Best practices (cont.)

Equity

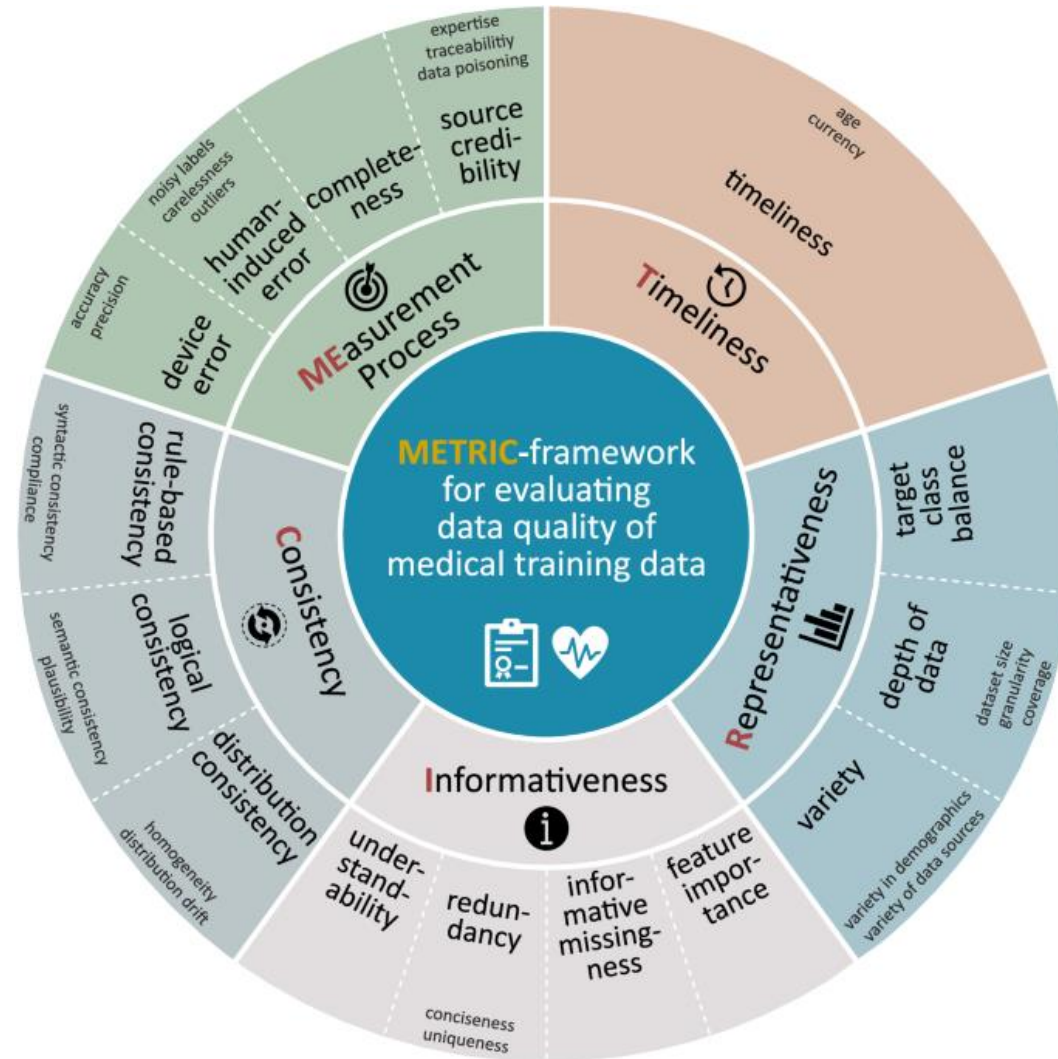
- Must address mechanisms that introduce disparities in data and modeling and steps to reduce them (Tjoeng, 2025)
- Must be accessible to all users, including researchers in low- and middle-income (LMIC) countries (Nakayama, 2025)



Standards for data provenance – Data & Trust Alliance

STANDARD	DESCRIPTION		
Lineage	Identifiers or pointers of metadata representing the data which comprise the current dataset	•	Provenance Metadata Unique ID Blockchain ledger ID Other:
Source	Identifies the origin (person, organization, system, device, etc.) of the current dataset	•	Username Organization name Software/system name Machine name Sensor/IoT device name File name IP Address Other:
Legal rights	Identifies the legal or regulatory framework applicable to the current dataset, along with the required data attributions, associated copyright or trademark, and localization and processing requirements	•	Attribution rights Copyrights or trademarks Required data storage or processing geolocation Applicable laws and regulations
Privacy and protection	Identifies any types of sensitive data associated with the current dataset and any privacy enhancing techniques applied	•	Protected data classification Applied privacy enhancing techniques
Generation date	Timestamp marking the creation of the current dataset	•	yyyy-MM-dd'T'HH:mm:ss
Data type	Identifies the data type contained in the current set, and provides insights into how the data is organized, its potential use cases, and the challenges associated with handling and using it	•	Structured Unstructured
Generation method	Identifies how the data was produced (data mining, machine-generated, IoT sensors, etc.)	•	Web scraping/Crawling Feeds Syndication Data mining Machine generated/MLOps Sensor and IoT output Social media User generated content Primary user source
Intended use and restrictions	Identifies the intended use of the data and which downstream audiences should not be allowed access to the current dataset	•	Intended use Restricted audience

METRIC model for data quality (Schwabe, 2024)



Concerns for US government data

Probably most prolific source of health (and other) data is US government

- Recent political developments may undermine

Many sources – [Data.gov](#), [HealthData.gov](#), [National Institutes of Health \(NIH\)](#), etc.

Recent actions

- Many sources disappearing (Freilich, 2025; Palmer, 2025; Carroll, 2025)
 - Disbanding of NHANES (Cooney, 2025)
- Alteration of data, e.g., changing names of fields or their data elements, including but not limited to gender/sex (Freilich, 2025)

Implications

- Undermines “data democracy” (Ponce, 2025)
- Potential loss of “unarchivable sites” (Cole, 2025)

Solutions

- Archive Data.gov (Goldstein, 2025)
- Need “more than snapshots,” e.g., data accessible via APIs (Rak, 2025)



Data science workforce

Data science called “sexiest profession of the 21st century” (Davenport, 2012)

Needs include “light quants” and “analytical translators” (Davenport, 2015)

- Early studies noted need not only for “deep analytical talent” but five-fold around them with skills in data management and interpretation (Manyika, 2011; IDC, 2014; Anaconda, 2022)

Occupational outlook is bright

- 245,900 jobs now, 82,500 anticipated in next 10 years (BLS, 2025)
- Including for clinical data managers (O*Net, 2025)
- Although may be impacted by AI (BLS, 2025)

Contrasting jobs in data science vs. health informatics (Hoffman, 2021)