

RE09: You Just Can't Just Cut the Gordian Knot



Approaches to Real-World Data Quality

S. Robert Collins, SAS Institute



The Clinical Data
Science Conference

US 2026

22–26 March
Renaissance Austin Hotel



Our Voyage

- 1  Context
- 2  Capture Source Data
- 3  Extract and Transmit
- 4  Receive and Integrate
- 5  Create Analytic Data



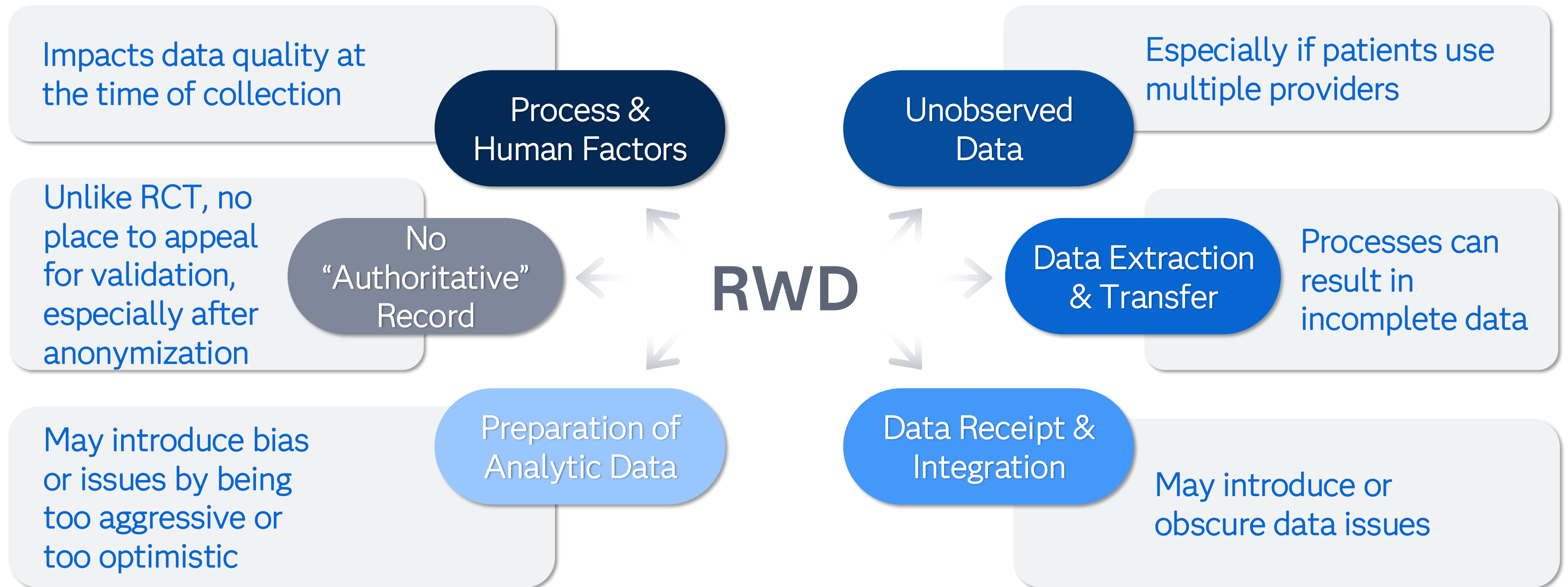
Context

RCTs Use Established Methods for Data Quality

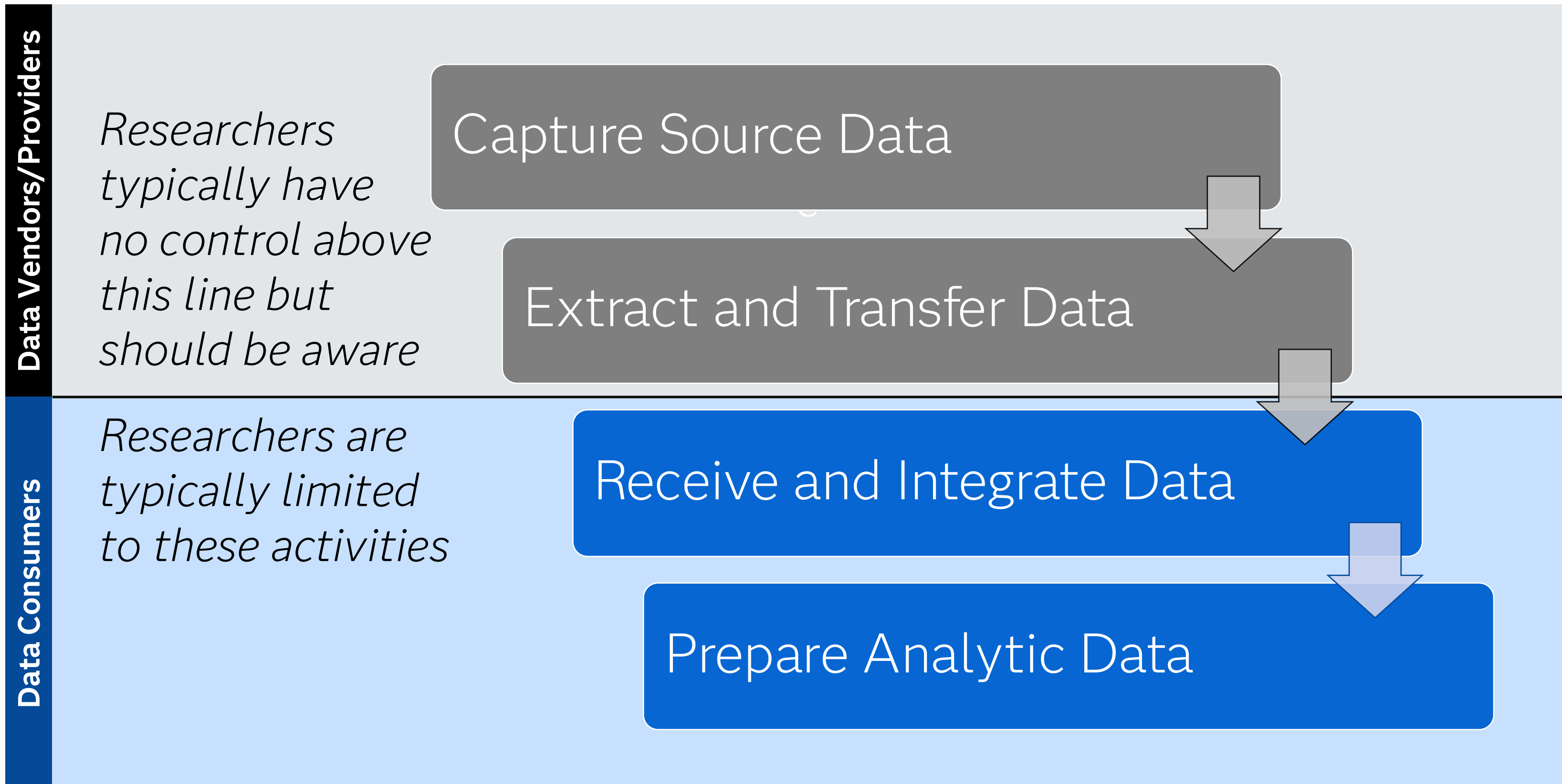
- Randomized controlled trials (RCTs) use well-established processes to ensure data are complete and correct.
 - Carefully designed case report forms (CRFs)
 - Standardized vocabularies and data dictionaries
 - Double-data entry and dedicated data management systems
 - Data monitoring and queries
 - Site training and monitoring
 - Predefined schedule for visits and other data collection
 - Enhanced collection, reporting and investigation of adverse events

Context

Real-World Data Does Not Benefit from Similar Set of Established Methods



What Can We Control?



Data Source Issues

Many Factors Can Impact Data Accuracy Within an EHR

- Transcription and Human Error
 - Transpose or mis-enter
 - 157.3cm vs. 175.3cm
- Consistency and Standardization
 - Units, e.g., Fahrenheit vs. Celsius
 - Terminology
- Missing or Incomplete Data
 - Some elements may just not have been recorded
 - Intra- and inter-user scenarios
 - EHR data typically does not include some details such as nursing notes
- Historical Data in a New EHR
 - Was old data mapped in?
 - What process was used, how complete was it and how well was it done?

Data Source Issues

Systemic Concerns

- Consistency
 - Every instance of an EHR is “custom” to the institution
 - Many institutions use multiple different EHRs for different services
 - Coded values vs free text entries
 - Required vs. optional data
 - Other process controls
 - Some systems require manual physician approval for intake notes regarding changes to current meds
 - Older entries not updated despite newer information

Data Extraction and Transmission

- Correct and complete records selected for extraction?
- Unstructured data omitted for privacy reasons?
 - Free text such as physician notes, etc.
 - Notations from imaging, etc.
- If structured data extracted from unstructured data, understand the process, its approach and its limitations
- Work with the data provider to specify how data will be transferred and perform tests
 - Table format – proprietary, FHIR, OMOP, etc.
 - Actual transfer type, e.g., text (including tarballs, etc.), database dumps, cloud or APIs
- Understand update process, including subsequent corrections

Data Receipt and Integration

SQL is your friend



■ Check Documentation

- Data vendors do their best to harmonize their sources
- Solving EHR integration issues for provider networks is source of much RWD

■ Confirm Against Delivery Materials

- Simple SQL queries to summarize record and patient counts, etc.
- Check to ensure all fields are present with expected types

■ Perform Exploratory Data Analysis

- Check numeric and date ranges, categorical values, etc.
- Confirm primary keys – tables only supposed to have one row per patient or per claim, etc.

■ Avoid “Expensive” Operations

- Nested SQL queries to create grouped counts without resorting to full data sorts

Data Receipt and Integration

Risks Vary Based on Transfer Method

Text/ASCII Files

Greater care required

- Separators
- Missings representation; consecutive missings; etc.
- Quotation issues, including embedded quotes or apostrophes
- Field and record lengths
- Line endings and line wrappings
- Character sets

Binary Files

Typically “safer” but still can have issues

- Character sets and encodings
- Version issues and other “compatibility” issues
- Data not matching documentation

Preparing Analytic Data

Write and document clear code

Check to see if you need to add in missing categories for reporting

Check included & excluded records including joins

Consider hardening frequent actions for reuse and standardization

Check resulting record counts, patient counts, etc.

Macros, custom steps, functions, libraries, etc.

Preparing Analytic Data

Dealing With Questionable or Missing Data

- Types of Missingness, e.g., MCAR, MAR, MNAR
- Each case must be evaluated individually in the broader context
- Options include
 - Censoring
 - Value
 - Observation
 - Patient
 - Imputation or weighting
- Selected approach and rationale ***must*** be documented in results

Mitigation of Data Issues

For data source issues, researchers can only react

- With the heightened interest in RWD, additional attention and efforts are coming to bear
- AI may be able to help but any corrections need to be flagged and made in real time, otherwise they're just potential new issues

For transmission and receipt, employ transparent, standardized, reusable processes including data summarization

For analytic data, check and recheck work

- Where applicable, reuse of processes can reduce effort while improving results

Reuse tested, standardized code where possible at each step from receipt to analysis

Conclusions

- 1 RWD come with a variety of potential data quality issues
- 2 Unfortunately, we cannot duplicate methods from RCTs
- 3 Except in rare cases, RWD is anonymized - there is no way to request data value checks
- 4 There is no “golden record” to verify against as in RCTs
- 5 Participate in harmonization, guidance and standardization efforts across stakeholders
- 6 Carefully apply methods to addressable portion of data lifecycle to minimize issues

S. Robert Collins
robert.collins@sas.com



LinkedIn
Profile

