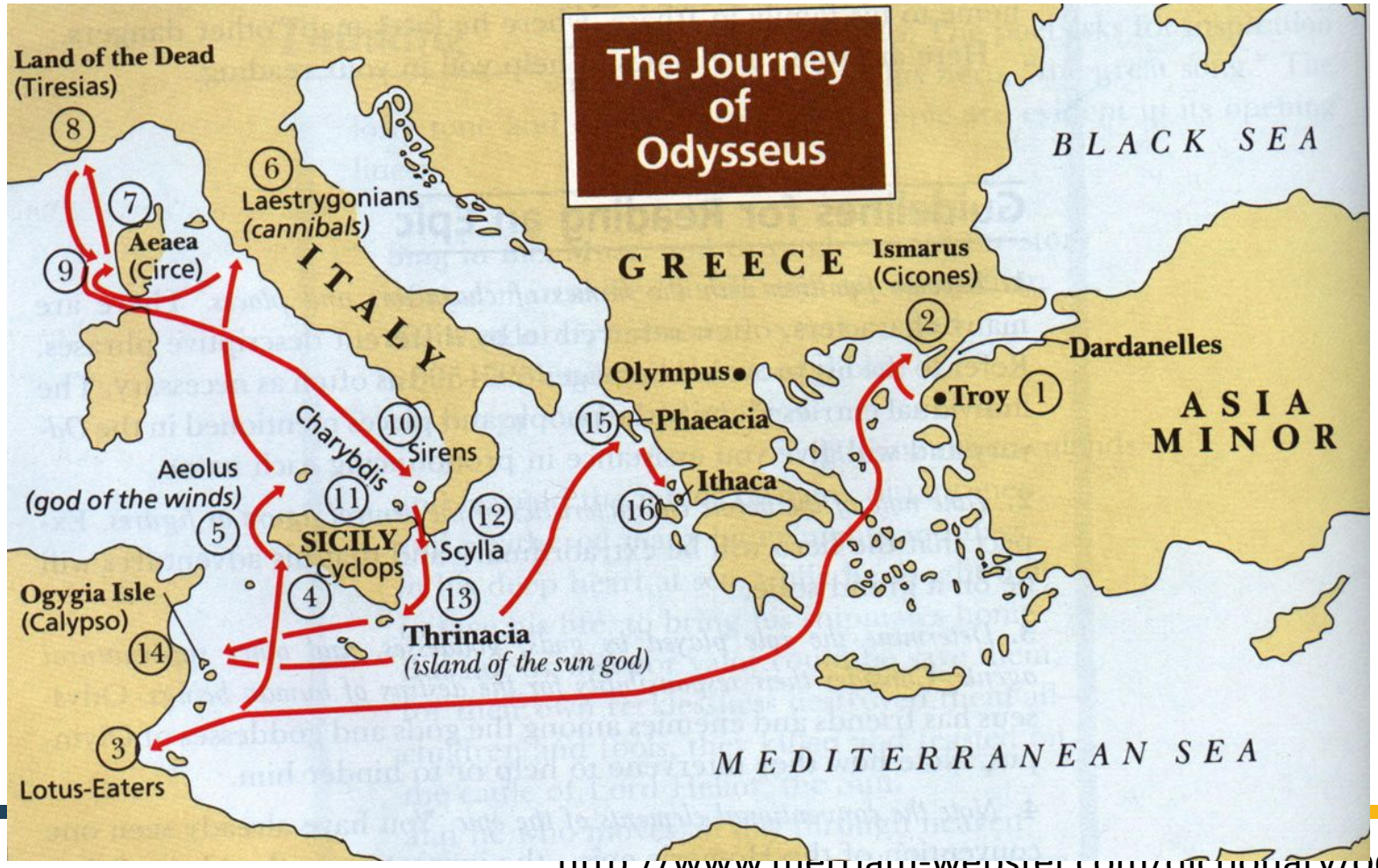
A Common Data Model- Why? Strengths and limitations of a common data approach

Patrick Ryan, PhD
Janssen Research and Development
Columbia University Medical Center

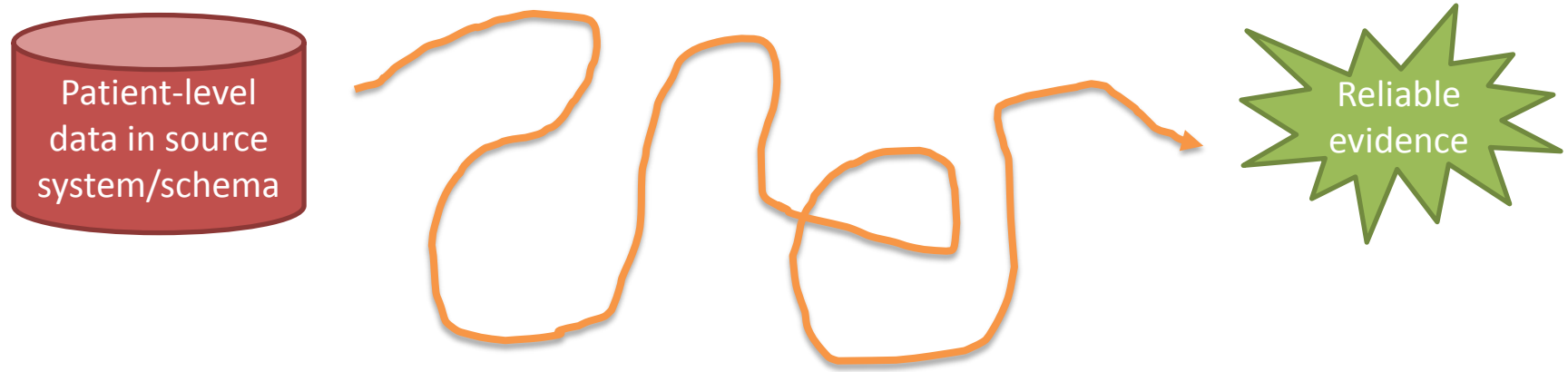
Odyssey (*noun*): \oh-d-si\

1. A long journey full of adventures



http://www.merriam-webster.com/dictionary/odyssey

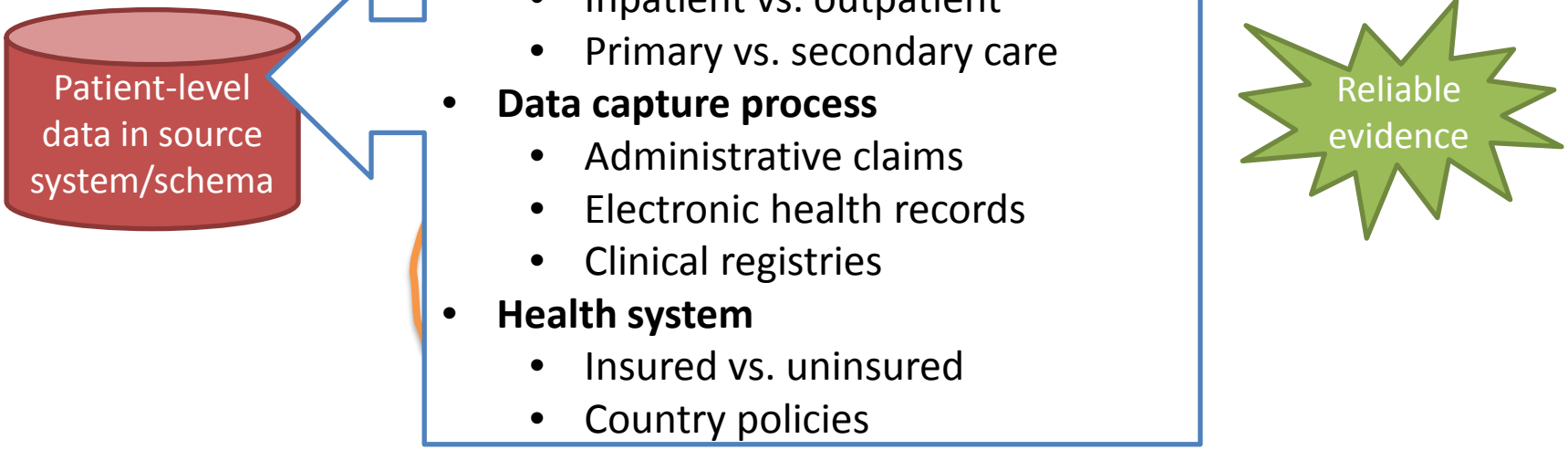
The journey to real-world evidence



One-time

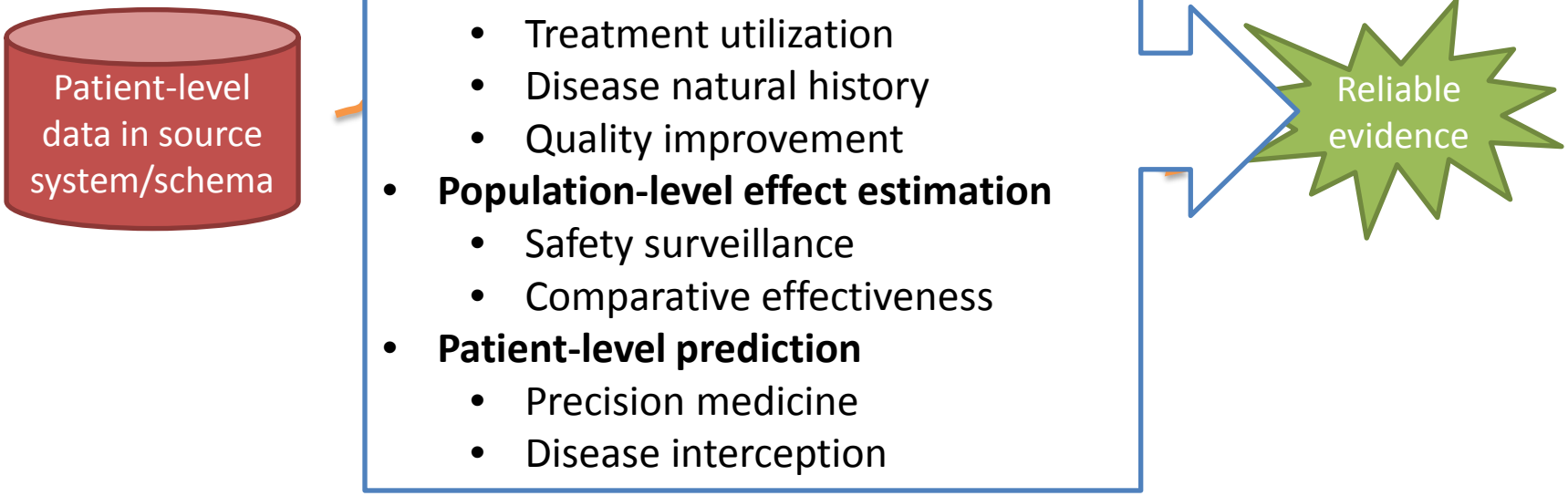
Repeated

The journey to real-world evidence



One-time | Repeated

The journey to real-world evidence



One-time Repeated

Opportunities for standardization in the evidence generation journey

Protocol

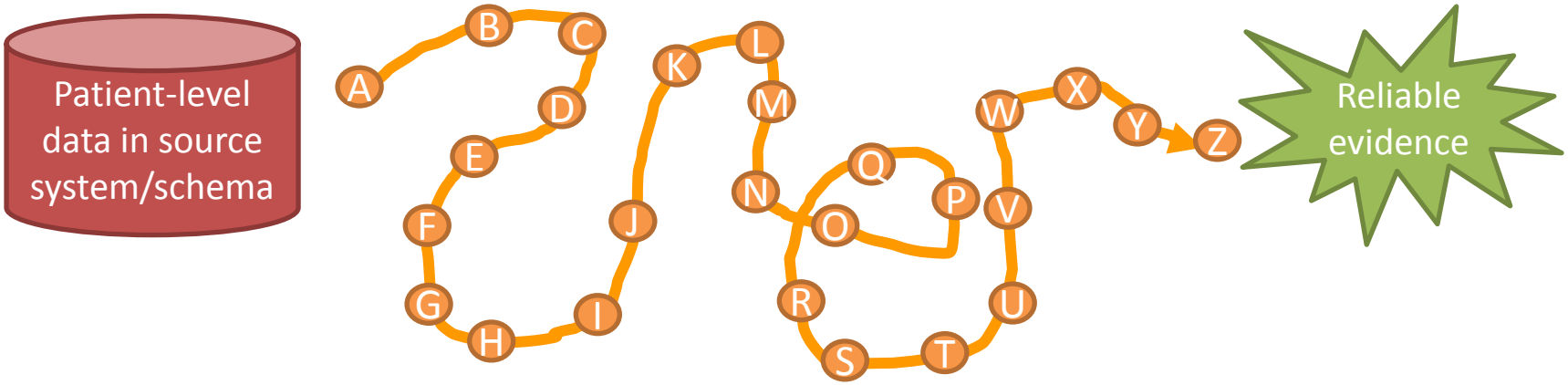
- **Data structure** : tables, fields, data types
- **Data conventions** : set of rules that govern how data are represented
- **Data vocabularies** : terminologies to codify clinical domains
- **Cohort definition** : algorithms for identifying the set of patients who meet a collection of criteria for a given interval of time
- **Covariate construction** : logic to define variables available for use in statistical analysis
- **Analysis** : collection of decisions and procedures required to produce aggregate summary statistics from patient-level data
- **Results reporting** : series of aggregate summary statistics presented in tabular and graphical form

Desired attributes for reliable evidence

Desired attribute	Question	Researcher	Data	Analysis		Result
Repeatable	Identical	Identical	Identical	Identical	=	Identical
Reproducible	Identical	Different	Identical	Identical	=	Identical
Replicable	Identical	Same or different	Similar	Identical	=	Similar
Generalizable	Identical	Same or different	Different	Identical	=	Similar
Robust	Identical	Same or different	Same or different	Different	=	Similar
Calibrated	Similar (controls)	Identical	Identical	Identical	=	Statistically consistent

Minimum requirements to achieve reproducibility

Desired attribute	Question	Researcher	Data	Analysis		Result
Reproducible	Identical	Different	Identical	Identical	=	Identical

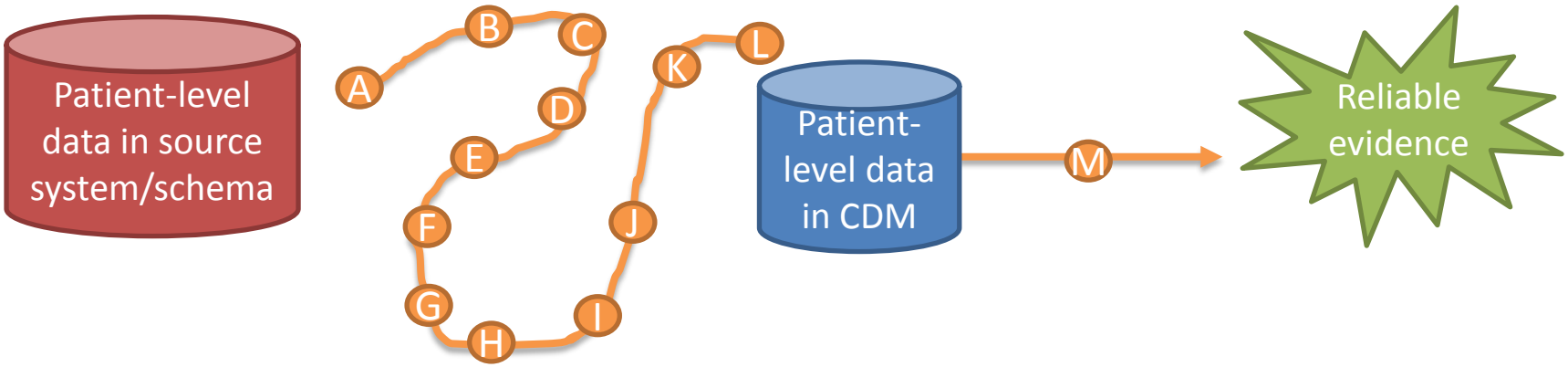


- Complete documented specification that fully describes all data manipulations and statistical procedures
- Original source data, no staged intermediaries
- Full analysis code that executes end-to-end (from source to results) without manual intervention

One-time Repeated

How a common data model + common analytics can support reproducibility

Desired attribute	Question	Researcher	Data	Analysis		Result
Reproducible	Identical	Different	Identical	Identical	=	Identical

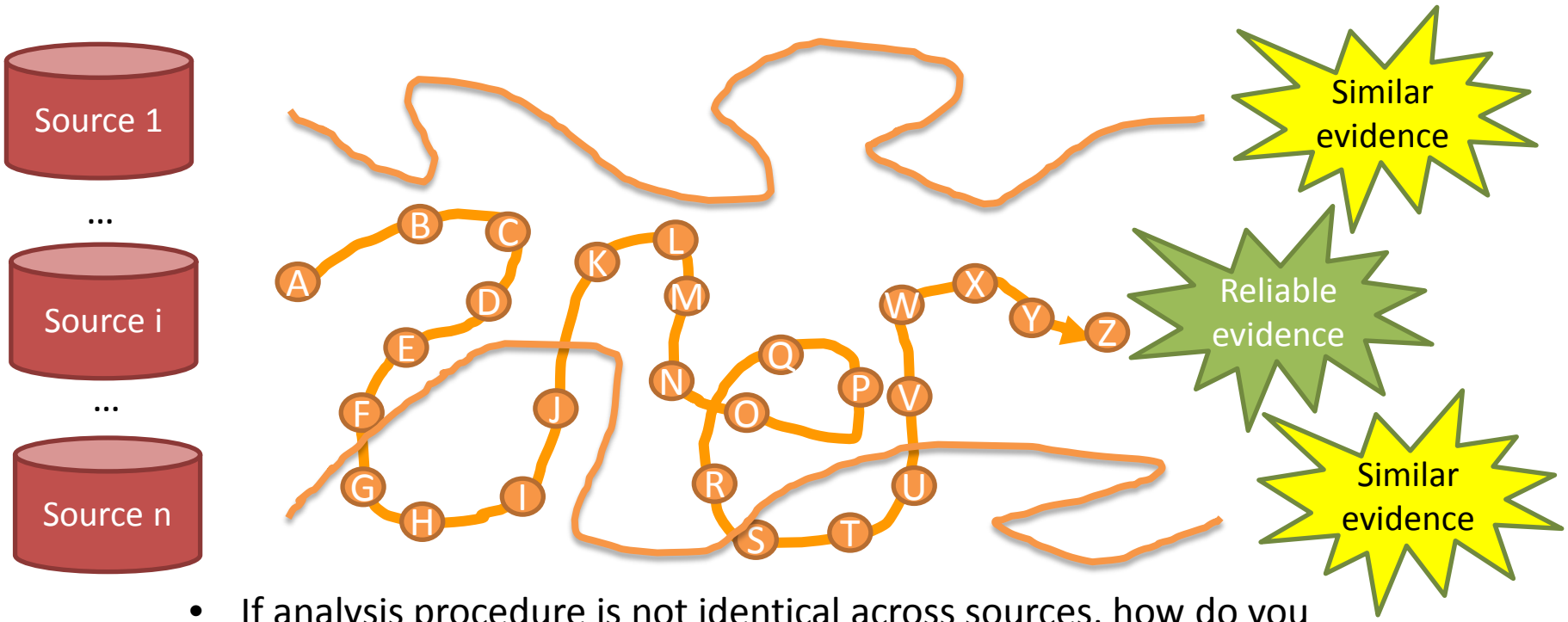


- Use of common data model splits the journey into two segments: 1) data standardization, 2) analysis execution
- ETL specification and source code can be developed and evaluated separately from analysis design
- CDM creates opportunity for re-use of data step and analysis step

One-time | Repeated

Challenges to achieve replication

Desired attribute	Question	Researcher	Data	Analysis		Result
Replicable	Identical	Same or different	Similar	Identical	=	Similar

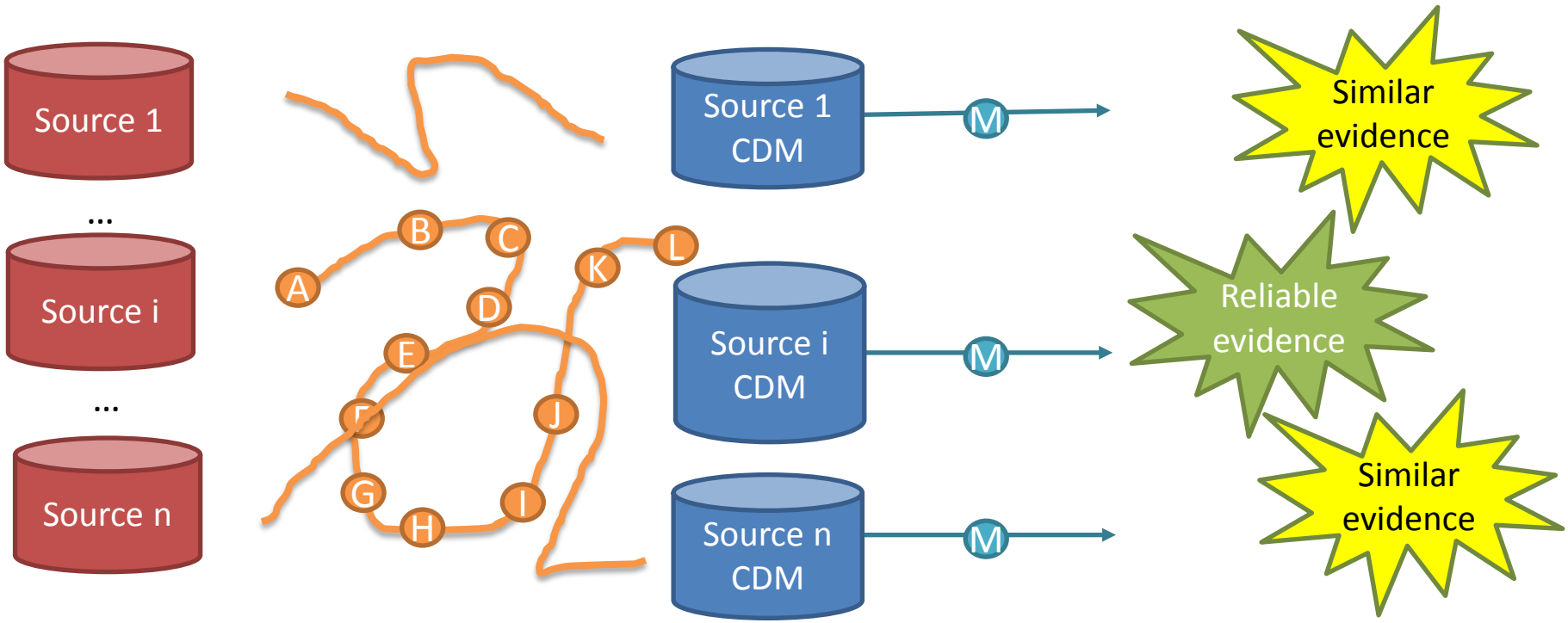


- If analysis procedure is not identical across sources, how do you determine if any differences observed are due to data vs. analysis?

One-time | Repeated

How a common data model + common analytics can support replication

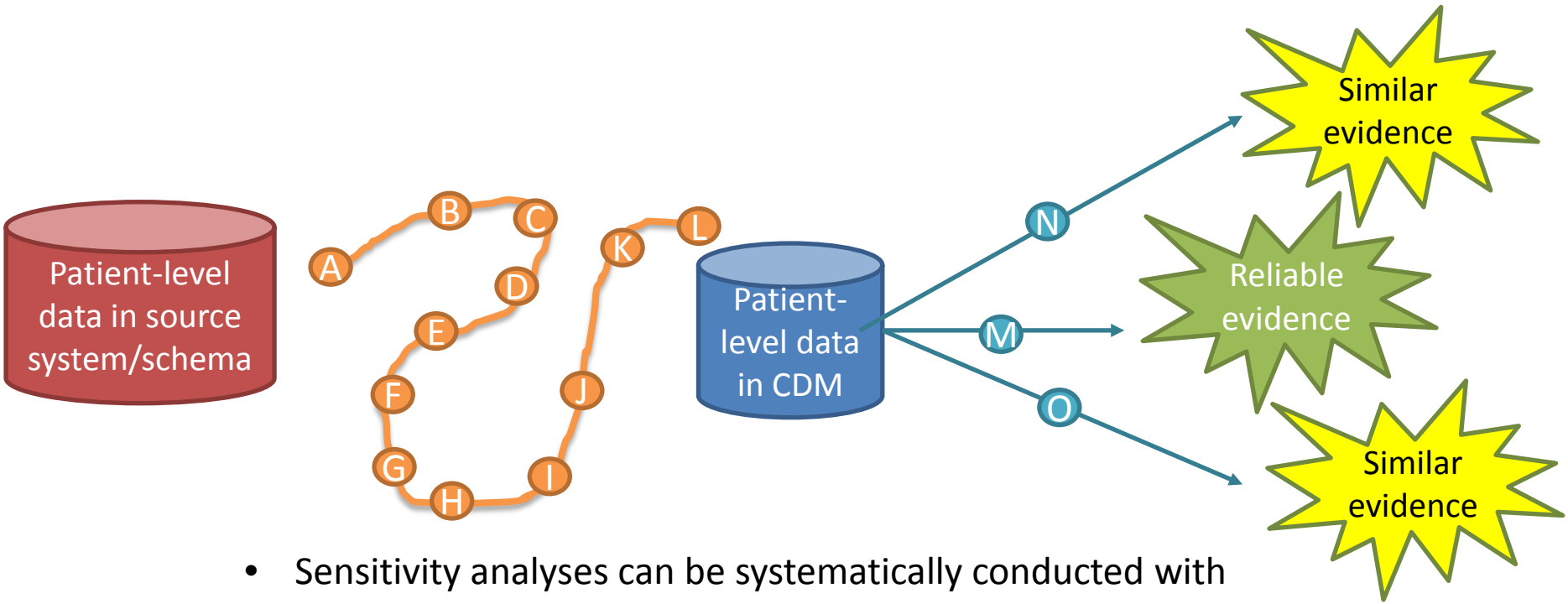
Desired attribute	Question	Researcher	Data	Analysis		Result
Replicable	Identical	Same or different	Similar	Identical	=	Similar



One-time | Repeated

How a common data model + common analytics can support robustness

Desired attribute	Question	Researcher	Data	Analysis		Result
Robust	Identical	Same or different	Same or different	Different	=	Similar

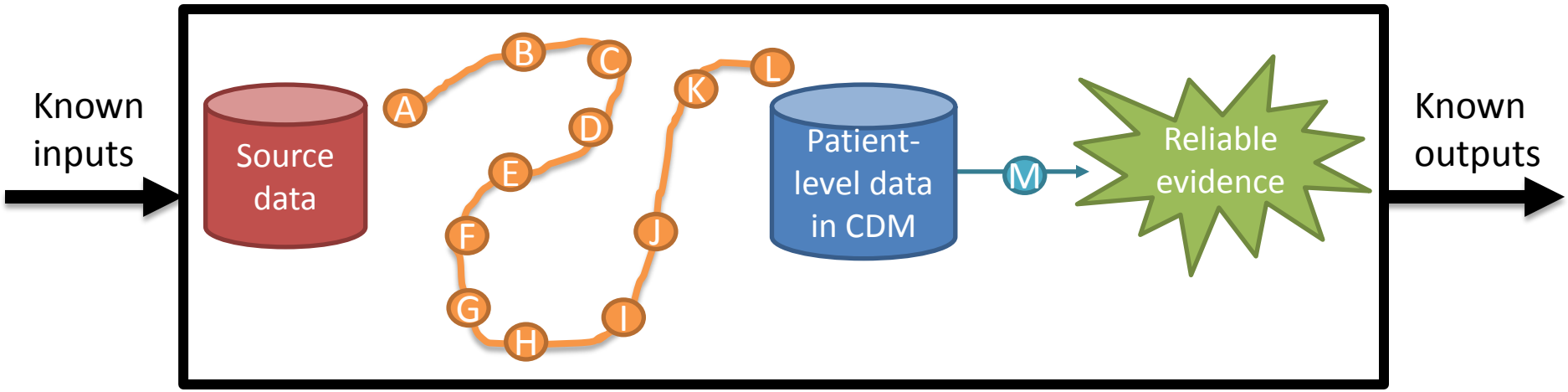


- Sensitivity analyses can be systematically conducted with parameterized analysis procedures using a common input

One-time | Repeated

How a common data model + common analytics can support calibration

Desired attribute	Question	Researcher	Data	Analysis		Result
Calibrated	Similar (controls)	Identical	Identical	Identical	=	Statistically consistent



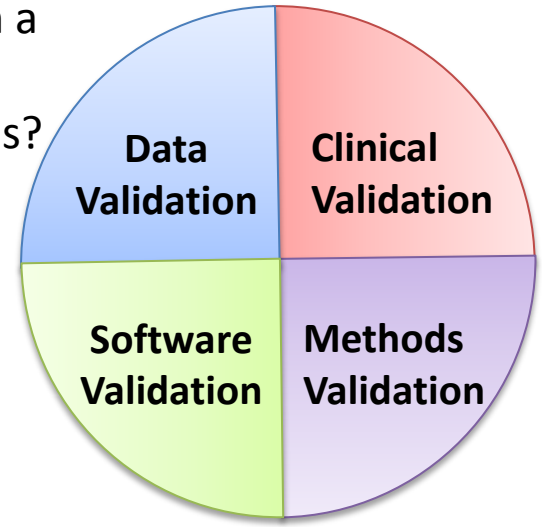
- With a defined reproducible process, you can measure a system's performance and learn how to properly interpret the system's outputs

One-time | Repeated

Flavors of validation throughout the evidence generation journey

Validation: "the action of checking or proving the accuracy of something"

Data : are the data completely captured with plausible values in a manner that is conformant to agreed structure and conventions?

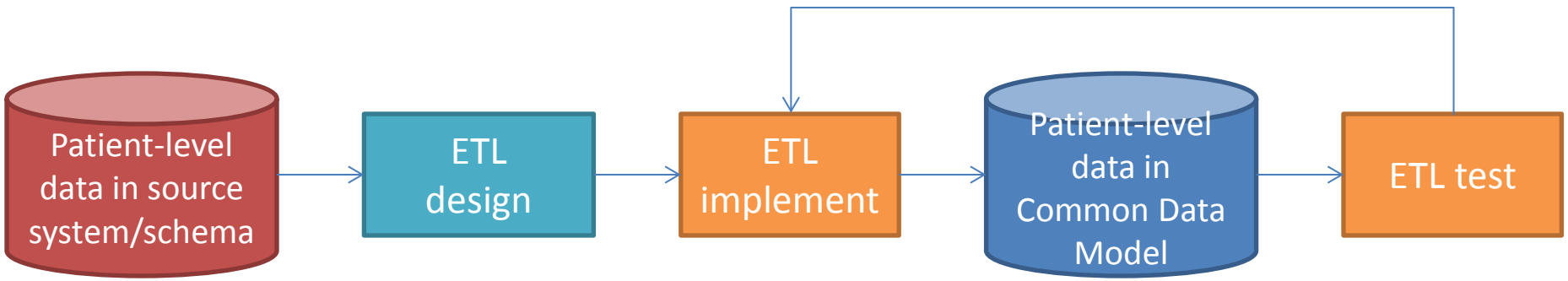


Clinical: to what extent does the analysis conducted match the clinical intention?

Software : does the software do what it is expected to do?

Statistical : do the estimates generated in an analysis measure what they purport to?

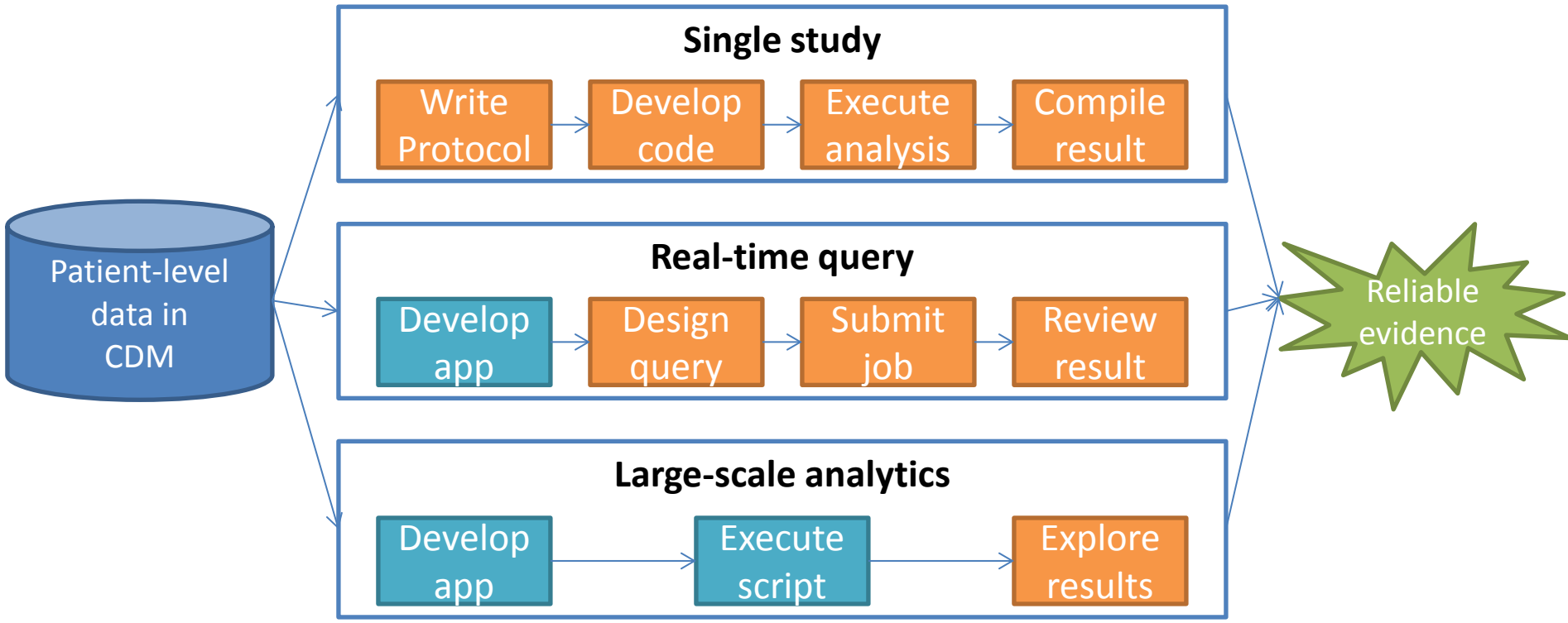
Structuring the journey from source to a common data model



Types of 'validation' required:
Data validation, software validation (ETL)

One-time | Repeated

Structuring the journey from a common data model to evidence



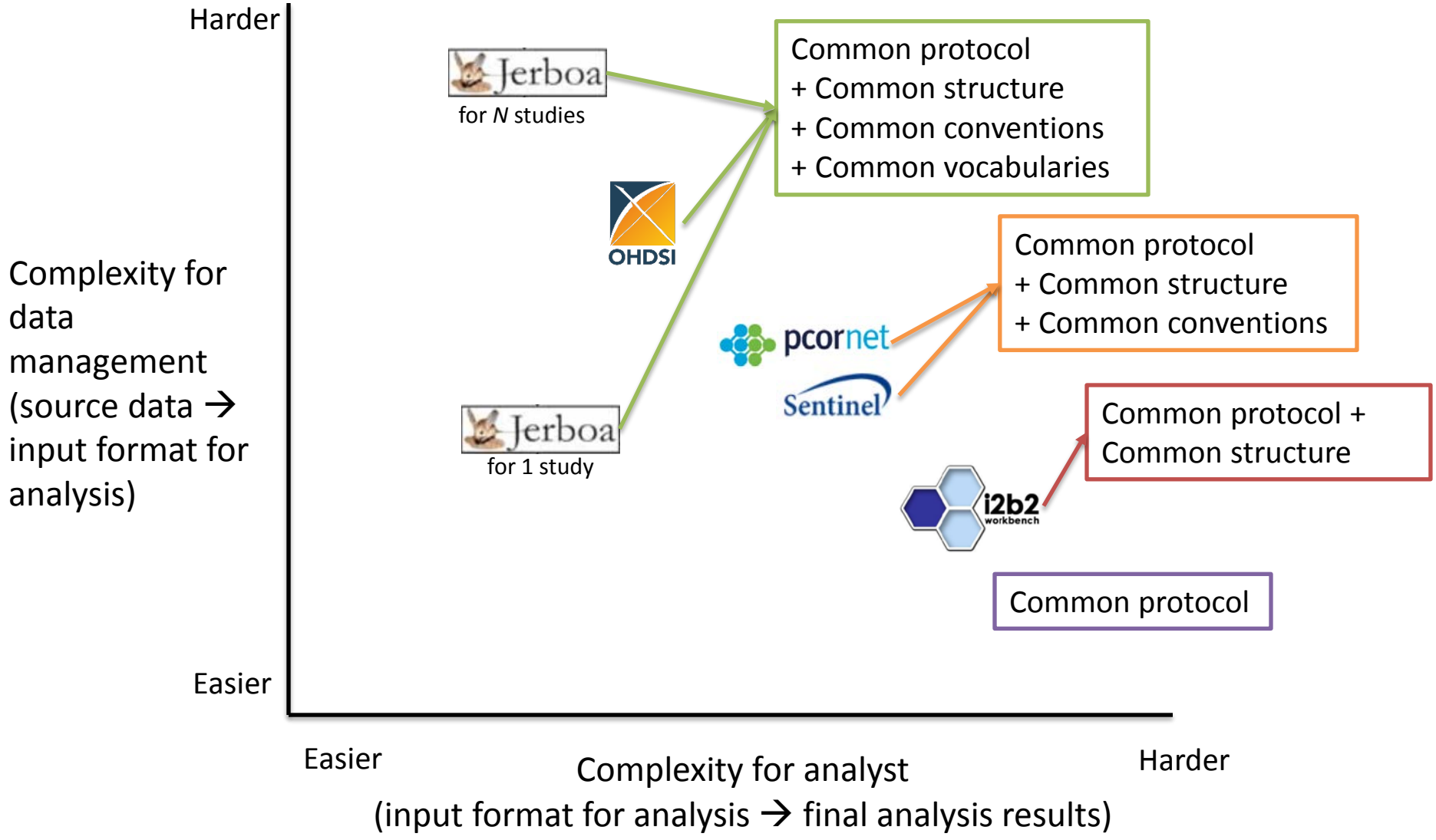
Types of 'validation' required:
Software validation (analytics), Clinical validation, Statistical validation

One-time | Repeated

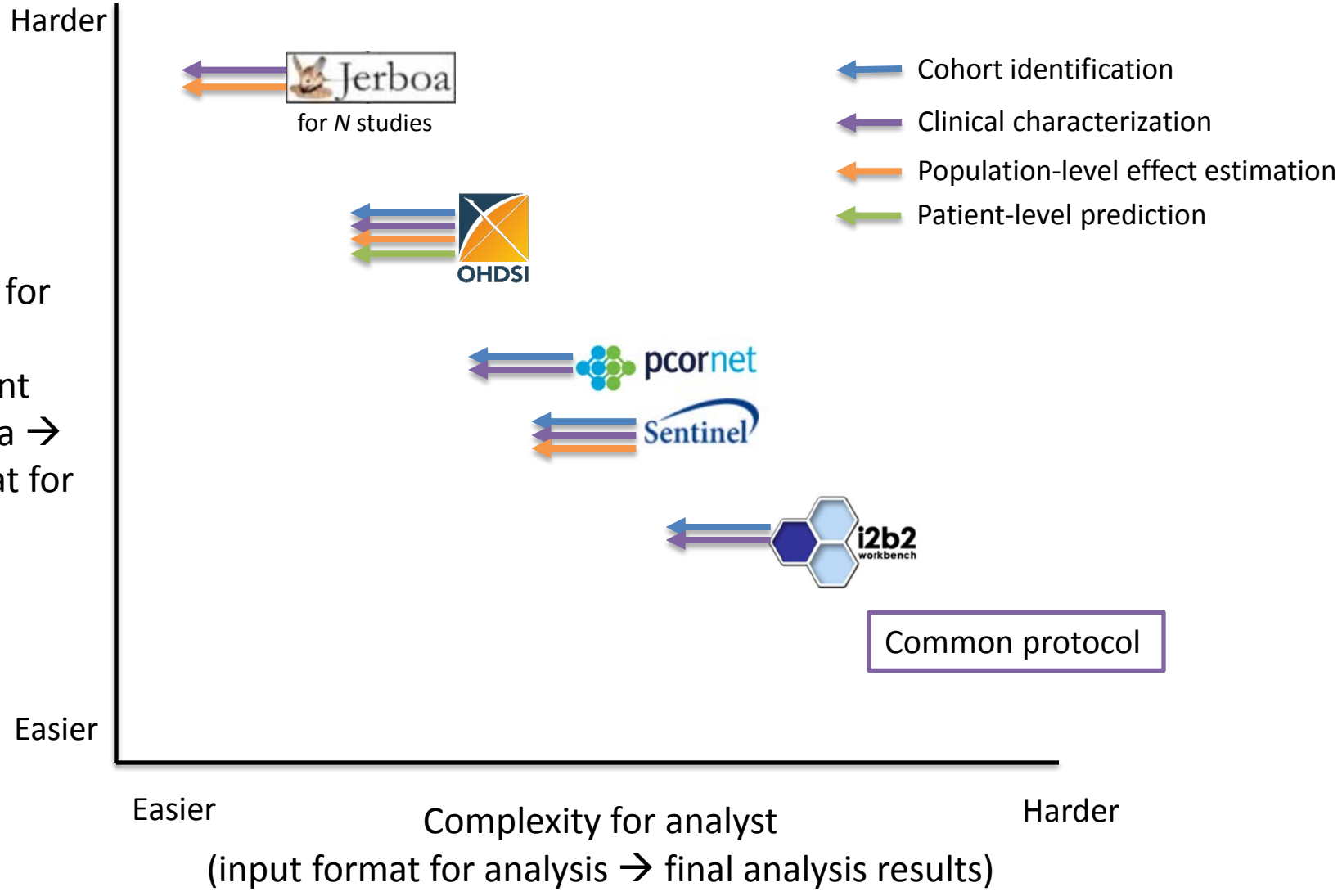
Motivations for developing different common data models

	Collaboration type	Data type(s)	Analytic use cases
I2b2	Grant -> Open-source project	EHR, 'omics cohorts	• Cohort identification • Translational research
Sentinel	Contract	US private-payer claims	• Clinical characterization • Safety surveillance
PCORNet	Grant	US EHR	• Cohort identification • Comparative effectiveness
EU-ADR (Jerboa)	Grant	European EHR, claims	• Clinical characterization • Safety surveillance
OHDSI (OMOP)	Open-science community	International claims, EHR, hospital, registries	• Cohort identification • Clinical characterization • Population-level estimation (safety + effectiveness) • Patient-level prediction

Balancing tradeoffs in data management vs analysis complexity



Common data model **+ common analytics** provides improved efficiency and reliability



Concluding thoughts

- On the journey from source data to reliable evidence, think about where you are starting and where you want to end up
- Common data model **+ common analytics** can help standardize parts of the journey
- The decision of whether (and which) CDM to apply to a EU network should be driven by the requirements around the reliability of the evidence and the efficiency of the evidence generation process

Questions?

Join the journey!

ryan@ohdsi.org