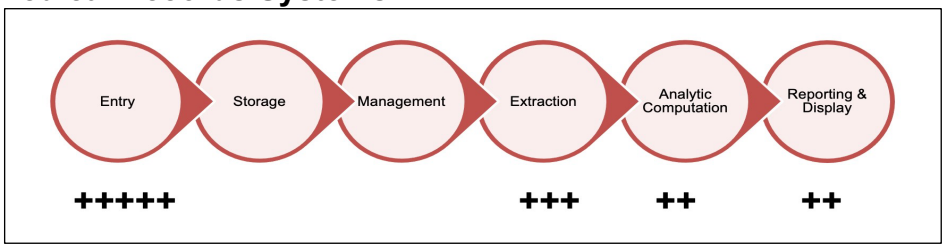
Data Quality Assessment Issues and Methods for Secondary Data Use

Michael G. Kahn MD, PhD (Michael.Kahn@cuanschutz.edu)

Handout, slides, example DQ reports, other links @
https://drive.google.com/drive/folders/11iJIG0AoS1KwgOjJM5RxPzqVaunZ6Kt9?usp=sharing

Sources of Quality Issues in Medical Records Systems

- The data lifecycle:
- Secondary data users rarely have access to source data systems.
- Operational systems focus on user efficiency, not data quality



Some key data quality "lingo" for framing your thinking & DQ activities

- Global data quality (DQ): A look at data quality across the entire data set irrespective of specific data use/analytics
- Fit for Use (F4U DQ) also called Fit for Purpose (F4P): A more-narrow view of data quality that is tailored to intended use/analytics.
    - F4U focuses on variables used to define cohort, exposure, outcomes, covariate.

- Intrinsic data quality: A look at DQ that doesn't depend on external data sources.
- Typically use local knowledge to determine data quality
- Extrinsic data quality: A look at DQ that compares DQ findings against some other data source (gold standard, relative gold standard, peer groups).
    - Peer group comparisons are common in multi-institutional data networks
      "How does my institution's data look compared to our peers"

- Data quality dimensions: An organizational model to break down of the wide range of data quality features that you could consider if relevant to your use case. The field uses terms inconsistently (sigh). I provide one attempt to try to harmonize DQ dimensions.
- Data quality measures: The actual computations used to quantify a specific data quality measure. The field has yet to develop a robust, reusable set of tools that is not dependent on the structure of a particular data set (sigh).
- Data quality rules: A set of "acceptance criteria" that if not met, will trigger a warning to investigate the data in more detail. These rules might be applied to global DQ measures or F4U DQ measures. The acceptance criteria can be different.
    - For a chronic renal disease study, it might be OK to have 100% missingness for psychiatric patients whereas this is (obviously) not so for a schizophrenia study.

- Non data quality features that impact F4U:
    - Are the data sufficiently timely for my needs (better data in 1 year vs poorer data today)?
    - Can I have access to the data elements I need, can I do the analytics I want, and can I present/publish as I wish (licensing and collaboration considerations)?
    - Can I afford access to the data and can I retain access as long as I need?

Getting Started

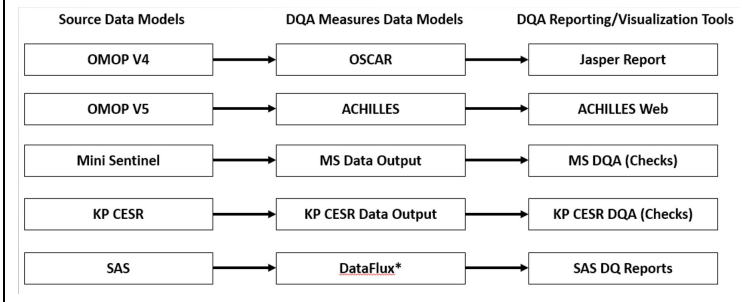
DQ Assessment is a big task. Align scope with resources. A rarely funded activity despite its importance in ensuring analytic validity. One (of many) data quality framework to use to scope your thinking/activities:

> Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, et al. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. eGEMs (Generating Evidence & Methods to improve patient outcomes) [Internet]. 2016 Sep 11 [cited 2016 Sep 12];4(1). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5051581/pdf/egems1244.pdf

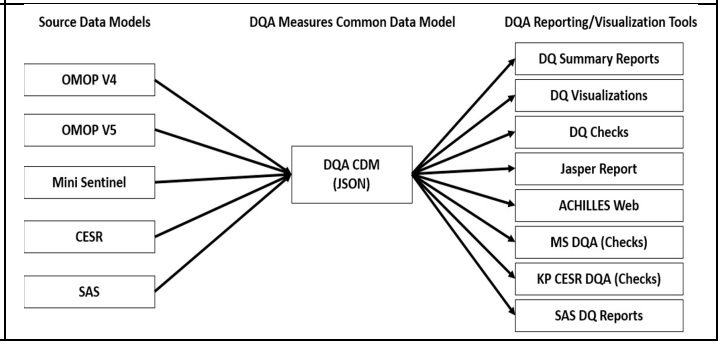
Focus on variables that matter: Work backwards from analytic plan. Consider interaction terms. List key variables on spreadsheet row. List data quality dimensions you feel are most impactful across columns. Start small, you can always grow. Convert DQ dimensions into DQ measures. Consider acceptability threshold (real world data never 100% clean so be realistic in your thresholds). Code or look for tools.

Finding Re-usable Tools: Not so easy

What we are today…..



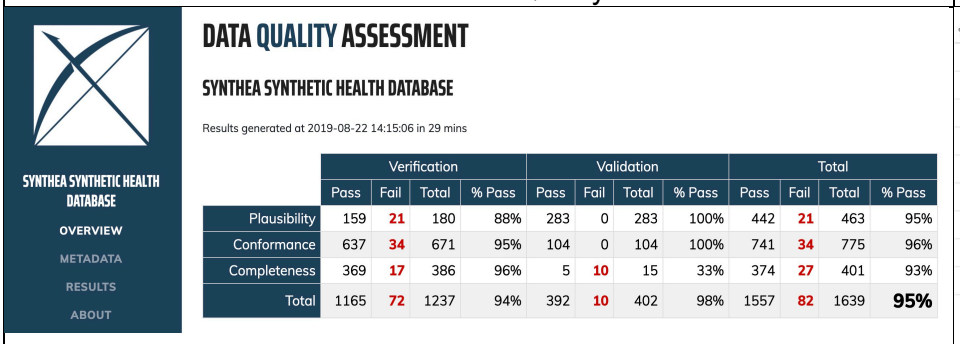
Where we want to be…..



- Many commercial tools (expensive): https://www.gartner.com/reviews/market/data-quality-solutions
  - New movement from "buy a tool" to "use a web service" ("DQ as a service"). Evolving but worth watching
- Open-source tools focused on health data.
  Large data networks have created E-X-T-E-N-S-I-V-E data quality tools. If you can use data in one of these formats, you can leverage their free (open access) DQ tools.
  - FDA Sentinel (claims oriented): https://www.sentinelinitiative.org/methods-data-tools/sentinel-common-data-model/data-quality-review-and-characterization-programs
  - PCORnet (medical records oriented): https://pcornet.org/data/
    SAS code @ https://github.com/PCORnet-DRN-OC/PCORnet-Data-Curation
  - OMOP (medical records oriented): https://github.com/OHDSI/DataQualityDashboard
  - A zillion "generic" (not health care focused) DQ tools on Github (https://github.com).
    Search "data quality" or "data profiling".
  - Other resources posted @
    https://drive.google.com/drive/folders/11iJIG0AoS1KwgOjJM5RxPzqVaunZ6Kt9?usp=sharing

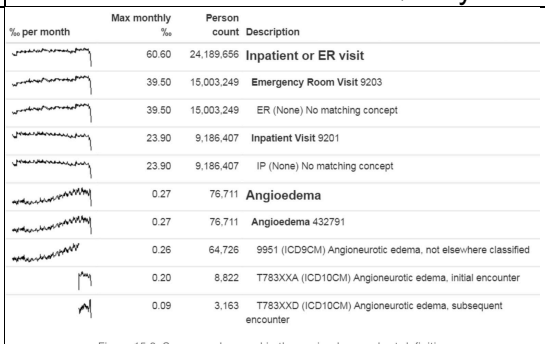
Data Quality Dashboard (OMOP) by Clair Blacketer: https://www.medrxiv.org/content/10.1101/2021.03.25.21254341v1.full.pdf

Global Data Quality



	Verification				Validation				Total			
	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass
Plausibility	159	21	180	88%	283	0	283	100%	442	21	463	95%
Conformance	637	34	671	95%	104	0	104	100%	741	34	775	96%
Completeness	369	17	386	96%	5	10	15	33%	374	27	401	93%
Total	1165	72	1237	94%	392	10	402	98%	1557	82	1639	**95%**

Fitness for Use Data Quality



‰ per month	Max monthly ‰	Person count	Description
	60.60	24,189,656	Inpatient or ER visit
	39.50	15,003,249	Emergency Room Visit 9203
	39.50	15,003,249	ER (None) No matching concept
	23.90	9,186,407	Inpatient Visit 9201
	23.90	9,186,407	IP (None) No matching concept
	0.27	76,711	Angioedema
	0.27	76,711	Angioedema 432791
	0.26	64,726	9951 (ICD9CM) Angioneurotic edema, not elsewhere classified
	0.20	8,822	T783XXA (ICD10CM) Angioneurotic edema, initial encounter
	0.09	3,163	T783XXD (ICD10CM) Angioneurotic edema, subsequent encounter

Figure 15.8: Source codes used in the angioedema cohort definition.

Data Quality Assessment Issues and Methods for Secondary Data Use

Michael Kahn, MD, PhD

[Notes]