Data quality assessments at Roche

Experiences and insights from the real-world data perspective

Spencer James, MD MPH
Principal Data Scientist
Genentech, a member of the Roche Group

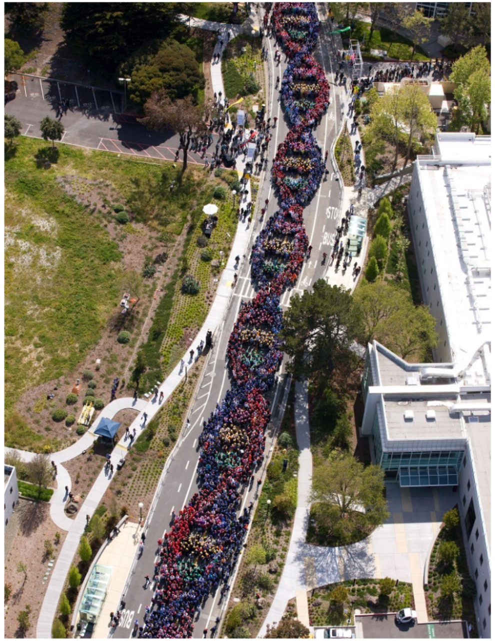
March 2023 | conference presentation at BES/BBS seminar – not for use or circulation elsewhere or by other parties

Data quality assessments at Roche

1. Background on RWD/E at Roche
2. Summary principles
3. Principles in a case study
4. Conclusion

Summarizing data quality assessments in RWD/E

General principles from Roche and Genentech experience



DNA replication involves data quality assessments, too!
Double helix created by Genentech staff in SSF

Summary principles

- Assessing data quality is integral to scientific research
- Assessments must themselves be systematic
- Data quality assessments in RWD must be comprehensive
  - Visualize every variable to inspect data distributions
  - Define cohort denominators to define sampling error
  - Rigorously evaluate risks of non-sampling error

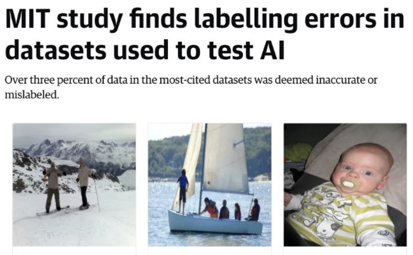
Background

Real-world data and evidence at Roche

- Roche and Genentech leverage large amounts of real-world data and evidence (RWD/E) to support various aspects of our R&D in Roche Pharma and Diagnostics divisions as well as across Roche affiliates
- We also work with historical clinical trials data, population health data, and genomics data
- We work with most major data assets in the RWD/E space to support various research initiatives
- Most major RWD assets are relational, complex, and require extensive domain knowledge and technical expertise to use responsibly and effectively
- We will use examples of data quality assessments with our Flatiron Health assets to convey key principles from this presentation

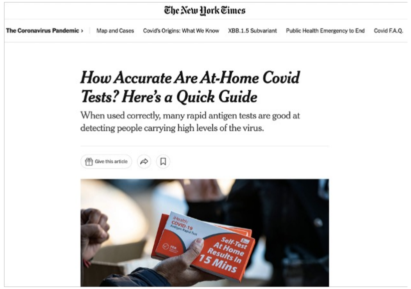
Assessing data quality is integral to scientific research

Define the need for data quality, which relate to everyday topics outside of R&D



MIT study finds labelling errors in datasets used to test AI

Over three percent of data in the most-cited datasets was deemed inaccurate or mislabeled.



The New York Times

The Coronavirus Pandemic › | Map and Cases | Covid's Origins: What We Know | XBB.1.5 Subvariant | Public Health Emergency to End | Covid F.A.Q.

How Accurate Are At-Home Covid Tests? Here's a Quick Guide

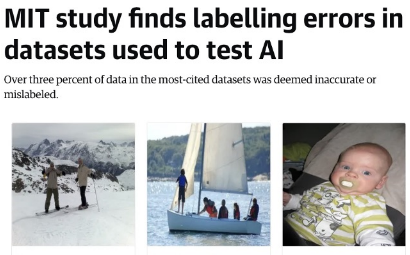
When used correctly, many rapid antigen tests are good at detecting people carrying high levels of the virus.

Data quality in popular media are increasingly integral to life outside of science, too, as technology and statistical considerations enter the population conscience

- Data quality is conceptualized in most scientific domains as well as many everyday settings, e.g.:
  - Significant figures
  - Misclassification and mislabeling in AI/ML
  - Sensitivity, specificity, PPV, NPV throughout covid era
- Biopharma R&D necessitates strict measures for good reasons:
  - Patient safety
  - Regulatory considerations
  - Replicability and transparency

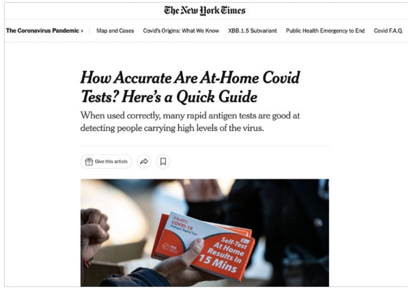
Assessing data quality is integral to scientific research

Define the need for data quality, which relate to everyday topics outside of R&D



MIT study finds labelling errors in datasets used to test AI

Over three percent of data in the most-cited datasets was deemed inaccurate or mislabeled.



The New York Times

The Coronavirus Pandemic › | Map and Cases | Covid's Origins: What We Know | XBB.1.5 Subvariant | Public Health Emergency to End | Covid F.A.Q.

How Accurate Are At-Home Covid Tests? Here's a Quick Guide

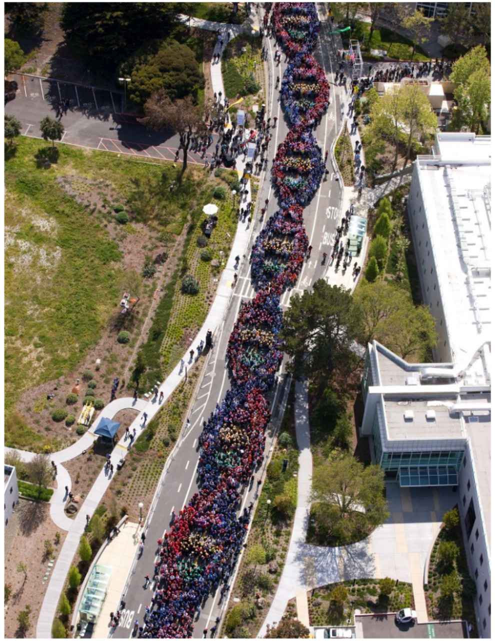
When used correctly, many rapid antigen tests are good at detecting people carrying high levels of the virus.

Data quality in popular media are increasingly integral to life outside of science, too, as technology and statistical considerations enter the population conscience

- Data quality should be a research paradigm as well as a mechanized process in conducting any study
- This paradigm spans the entire scientific R&D process from formulating data collection to sharing study results
- Many examples in biomedical research domain:
  - <u>Inclusive research</u>: how exactly are fields on race and ethnicity collected and reported at the point of data entry?
  - <u>External controls</u>: sensitivity analysis to assess loss in key variables (eg tumor grade coding "3" vs "3a" and "3b")
  - <u>Derived variables</u>: identifying lab values off by orders of magnitude suggesting incorrect unit data entry affecting calculation such as eGFR, FLIPI, etc

Assessments must themselves be systematic

Avoid creating further bias by approaching data quality assessments with an organized plan

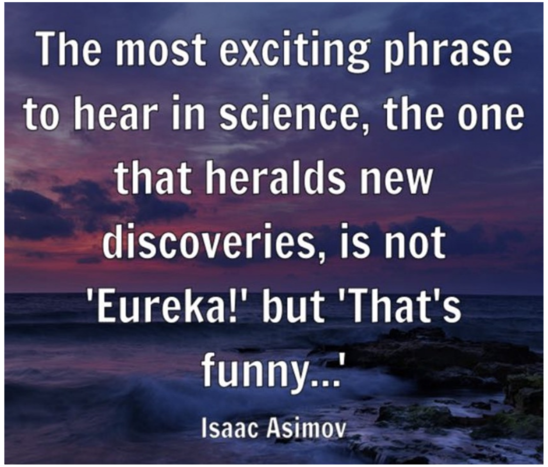


DNA replication involves data quality assessments, too

- Every row and column in a dataset is a potential source of error
- Error can be stochastic, biased, or both
- Data quality assessments should not be ad hoc
- Systematic approaches can be designed and peer reviewed
- Process ensures diverse expertise and input considered
- Critical in multimodal, interdisciplinary biomedical R&D
- Systematic, transparent approaches protect against bias

Assessments must themselves be systematic

Avoid creating further bias by approaching data quality assessments with an organized plan



The most exciting phrase to hear in science, the one that heralds new discoveries, is not 'Eureka!' but 'That's funny...'

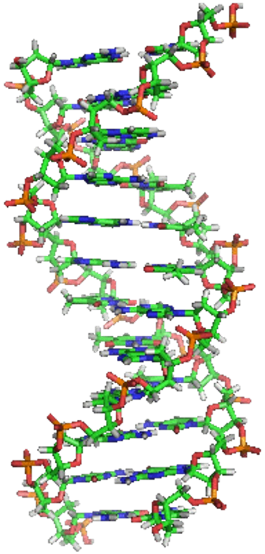
Isaac Asimov

Commonly heard in data quality assessments and should be treated as a discovery no matter how small

- Systematic approaches in our data quality assessments include developing an analysis plan in advance of the assessment
- This also includes adopting organizational measures such as Agile or Scrum to ensure systematic completion of all steps
- Systematic approaches can also be revised as more is learned about the nature of the data
- Data quality findings also prompt further questions and investigation and can unearth more systemic issues
- Unanticipated error discovery is an expected result and may lead to important insights

Data quality assessments in RWD must be comprehensive

Leave no stone unturned

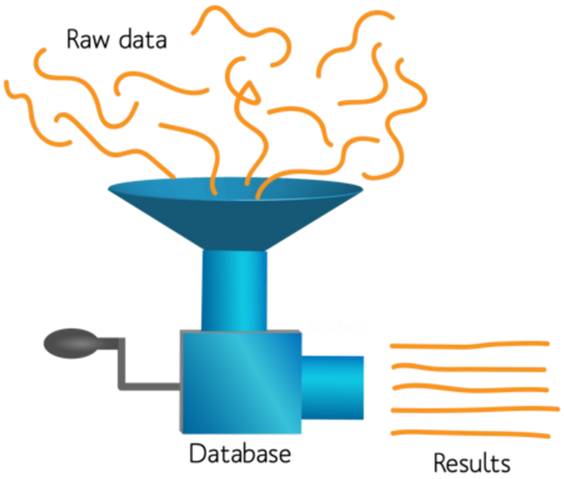


DNA replication involves data quality assessments, too

- Data quality assessments are systems to check for errors and biases
- It needs algorithmic checks and balances and methods to identify structural issues and problematic data
  - Preserve raw data
  - Inspect every variable
  - Sample and subsample among key strata
  - Consider other data error mechanisms and perspectives
  - Think broadly about data quality

Data quality assessments in RWD must be comprehensive

Leave no stone unturned



- Preserve all raw data
  - Ensure processes for converting raw data to usable format are transparent
  - Ensure data processing is entirely code-based
  - Avoid upstream processes involving data processing in Excel or other platforms where process cannot be replicated
  - Consider hidden processes in data storage process, eg floating point digit preservation in an Amazon Redshift enterprise database versus processing in R dataframe using `dplyr`

Data quality assessments in RWD must be comprehensive

Leave no stone unturned

Leave no stone unturned in assessing data quality, even if that means destabilizing a few others.

- Inspect every variable in the dataset
  - Every variable that has any relevance should be manually inspected using standard data analysis and visualization
  - Summary statistics are important for assessments and can provide a quick assay for problematic data through the strategic use of mean, median, IQR, SD, min, max, and statistics like median absolute deviation (MAD) estimator and Z-scores
  - Data visualization to discover new patterns and inspect distributions and outliers are also critical to this process
  - Automated systems are tempting and have their place, but visually inspecting the data remains foundational in data science
  - Domain knowledge is also a critical aspect of this process

Data quality assessments in RWD must be comprehensive

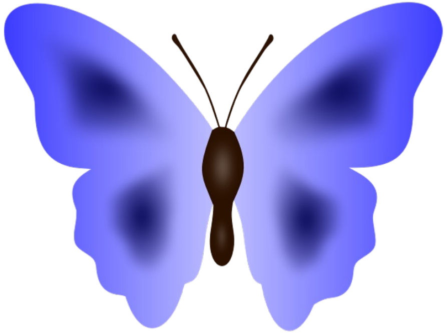
Leave no stone unturned

Inspecting, sampling, zooming in, zooming out, and then repeating are all parts of data quality assessments

- Continuously sample and resample to assess biases and errors
  - Data quality assessments are never complete
  - They represent an iterative process to support strategic use of a data asset to evaluate some research question or hypothesis
  - A different question may necessitate a different DQ process
  - Data scientists must understand the underlying research question being investigated to effectively conduct a data quality assessment

Data quality assessments in RWD must be comprehensive

Leave no stone unturned



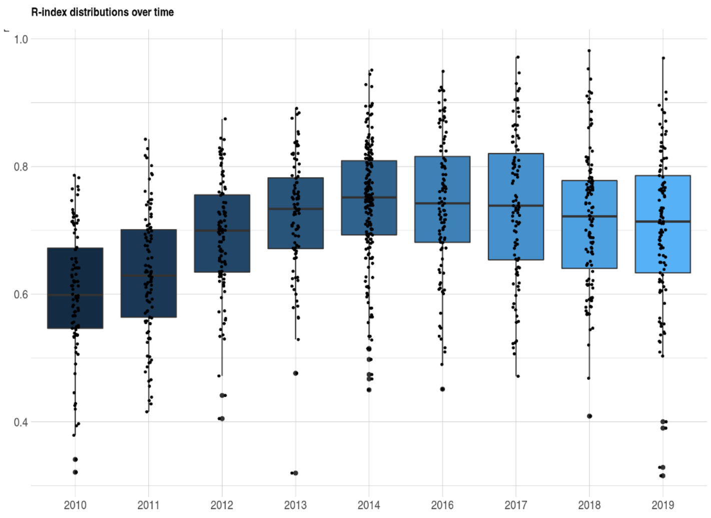
Seemingly minor data quality issues can have substantial downstream effects on studies and insights

- Include mechanisms to identify other data-related problems
    - Data entry errors or one class of error should not be the only consideration in how we conceptualize data quality
    - Most biomedical research draws on a large number of parameters and assumptions to deliver a result
    - Experienced data scientists are trained in looking for idiosyncratic patterns in data quality, which require technical expertise and domain knowledge, eg age heaping in survey data or using race, ethnicity, and ancestry variables in genomics data
    - This reiterates the need to include domain knowledge and historical perspective to enhance contextualization

Assessing representativeness for inclusive research

Developing new metrics for new challenges on the research horizon

- Representative sampling is foundational in statistics
- Increasing focus on inclusive research across health sector globally
- How representative are our data with respect to the populations that may benefit from research?
- We developed a new statistical measure R-index to systematically assess and compare distributions in categorical variables (eg race, ethnicity, gender) to measure the representativeness of our research

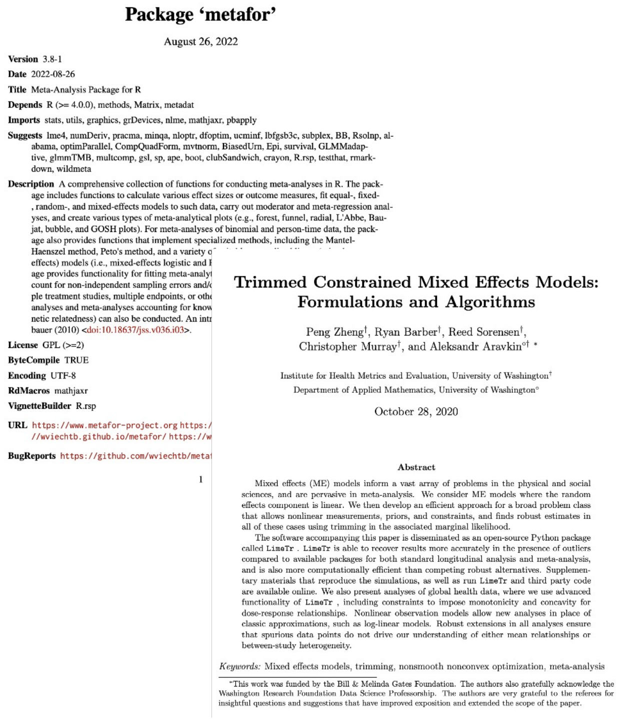


Simulated data demonstrating representativeness of clinical trial datasets over time compared with census-based population data

Harmonizing data to account for bias and maximize yield

Meta-regression to mitigate bias

- Undetected bias poses more risk than observable bias
- We can use meta-analysis and meta-regression as techniques for accounting for bias and harmonizing different sources of information
- In terms of data quality, these methods allow researchers to leverage data from various sources while accounting for bias
- These processes allow for more nuanced assessments of data quality rather than simply classifying data as bad or good



Package ‘metafor’

August 26, 2022

Version 3.8-1

Date 2022-08-26

Title Meta-Analysis Package for R

Depends R (>= 4.0.0), methods, Matrix, metadat

Imports stats, utils, graphics, grDevices, nlme, mathjaxr, pbapply

Suggests lme4, numDeriv, pracma, minqa, nloptr, dfoptim, ucminf, lbfgsb3c, subplex, BB, Rsolnp, alabama, optimParallel, CompQuadForm, mvtnorm, BiasedUrn, Epi, survival, GLMMadaptive, glmmTMB, multcomp, gsl, sp, ape, boot, clubSandwich, crayon, R.rsp, testthat, rmarkdown, wildmeta

Description A comprehensive collection of functions for conducting meta-analyses in R. The package includes functions to calculate various effect sizes or outcome measures, fit equal-, fixed-, random-, and mixed-effects models to such data, carry out moderator and meta-regression analyses, and create various types of meta-analytical plots (e.g., forest, funnel, radial, L'Abbe, Baujat, bubble, and GOSH plots). For meta-analyses of binomial and person-time data, the package also provides functions that implement specialized methods, including the Mantel-
Haenszel method, Peto's method, and a variety o
effects) models (i.e., mixed-effects logistic and P
age provides functionality for fitting meta-analyt
count for non-independent sampling errors and/o
ple treatment studies, multiple endpoints, or othe
analyses and meta-analyses accounting for know
netic relatedness) can also be conducted. An intr
bauer (2010) <doi:10.18637/jss.v036.i03>.

License GPL (>=2)

ByteCompile TRUE

Encoding UTF-8

RdMacros mathjaxr

VignetteBuilder R.rsp

URL https://www.metafor-project.org https:/
//wviechtb.github.io/metafor/ https://w

BugReports https://github.com/wviechtb/metaf

1

Trimmed Constrained Mixed Effects Models: Formulations and Algorithms

Peng Zheng[†], Ryan Barber[†], Reed Sorensen[†],
Christopher Murray[†], and Aleksandr Aravkin[°†] [*]

Institute for Health Metrics and Evaluation, University of Washington[†]
Department of Applied Mathematics, University of Washington[°]

October 28, 2020

Abstract

Mixed effects (ME) models inform a vast array of problems in the physical and social sciences, and are pervasive in meta-analysis. We consider ME models where the random effects component is linear. We then develop an efficient approach for a broad problem class that allows nonlinear measurements, priors, and constraints, and finds robust estimates in all of these cases using trimming in the associated marginal likelihood.

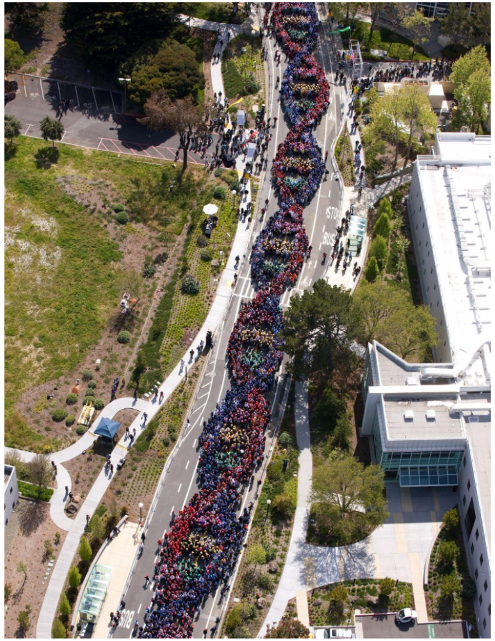
The software accompanying this paper is disseminated as an open-source Python package called `LimeTr` . `LimeTr` is able to recover results more accurately in the presence of outliers compared to available packages for both standard longitudinal analysis and meta-analysis, and is also more computationally efficient than competing robust alternatives. Supplementary materials that reproduce the simulations, as well as run `LimeTr` and third party code are available online. We also present analyses of global health data, where we use advanced functionality of `LimeTr` , including constraints to impose monotonicity and concavity for dose-response relationships. Nonlinear observation models allow new analyses in place of classic approximations, such as log-linear models. Robust extensions in all analyses ensure that spurious data points do not drive our understanding of either mean relationships or between-study heterogeneity.

Keywords: Mixed effects models, trimming, nonsmooth nonconvex optimization, meta-analysis

[*] This work was funded by the Bill & Melinda Gates Foundation. The authors also gratefully acknowledge the Washington Research Foundation Data Science Professorship. The authors are very grateful to the referees for insightful questions and suggestions that have improved exposition and extended the scope of the paper.

Summarizing data quality assessments in RWD/E

General principles from Roche and Genentech experience



DNA replication involves data quality assessments, too!
Double helix created by Genentech staff in SSF

Summary principles

- Assessing data quality is integral to scientific research
- Assessments must themselves be systematic
- Data quality assessments in RWD must be comprehensive
    - Visualize every variable to inspect data distributions
    - Define cohort denominators to define sampling error
    - Evaluate risks of non-sampling error

Thank you

Spencer James, MD MPH
Principal Data Scientist
Genentech, a member of the Roche Group

March 2023 | conference presentation at BES/BBS seminar