HOW SYNTHETIC DATA WILL TRANSFORM HEALTH RESEARCH AND INNOVATION

Making data more available to researchers using privacy enhancing technologies

July 7, 2021

Karen T.Cuenco, PhD, Senior Program Officer

DISCLAIMERS

The opinions and thoughts expressed in this presentation are

- my own views and should in no way be construed as representing the official position of the Bill & Melinda Gates Foundation

- not an endorsement or recommendation for any privacy enhancing technology or solution.

ALL LIVES HAVE EQUAL VALUE

We work to help all people lead healthy productive lives

Focus on addressing gaps in health, economic opportunity, and education.

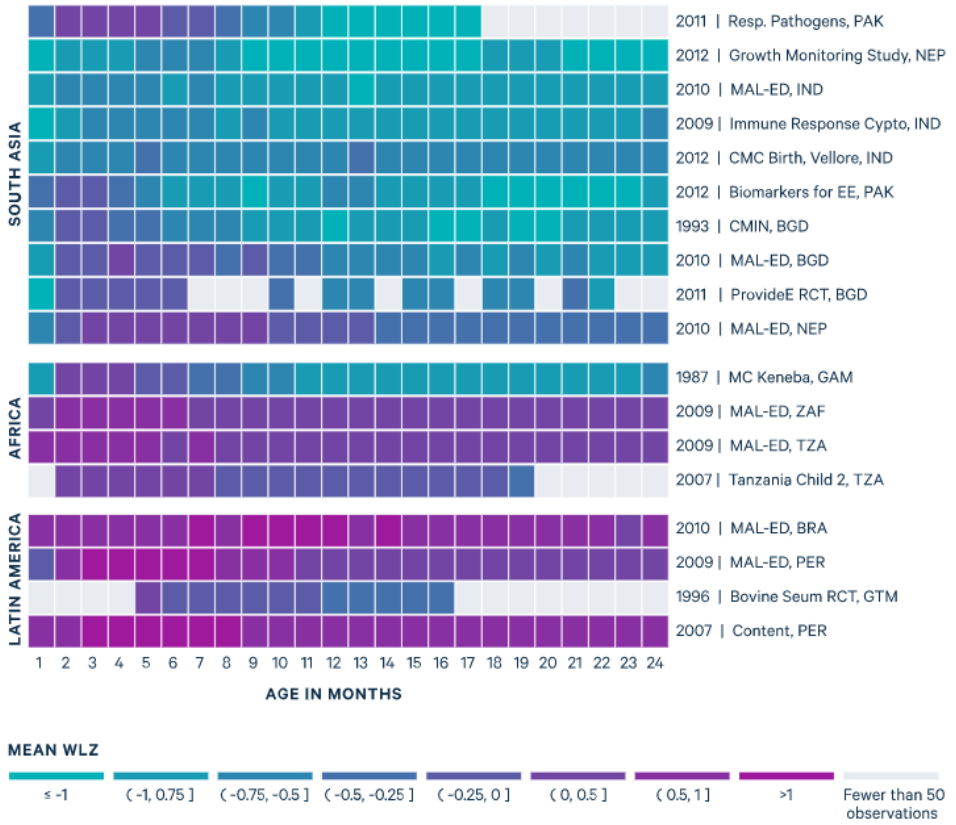
- Data and inference can transform how we engage in our work.

BMGF QUANTITATIVE SCIENCES/KNOWLEDGE INTEGRATION INITIATIVE (KI)

Things we do: ki knowledge integration kiglobalhealth.org

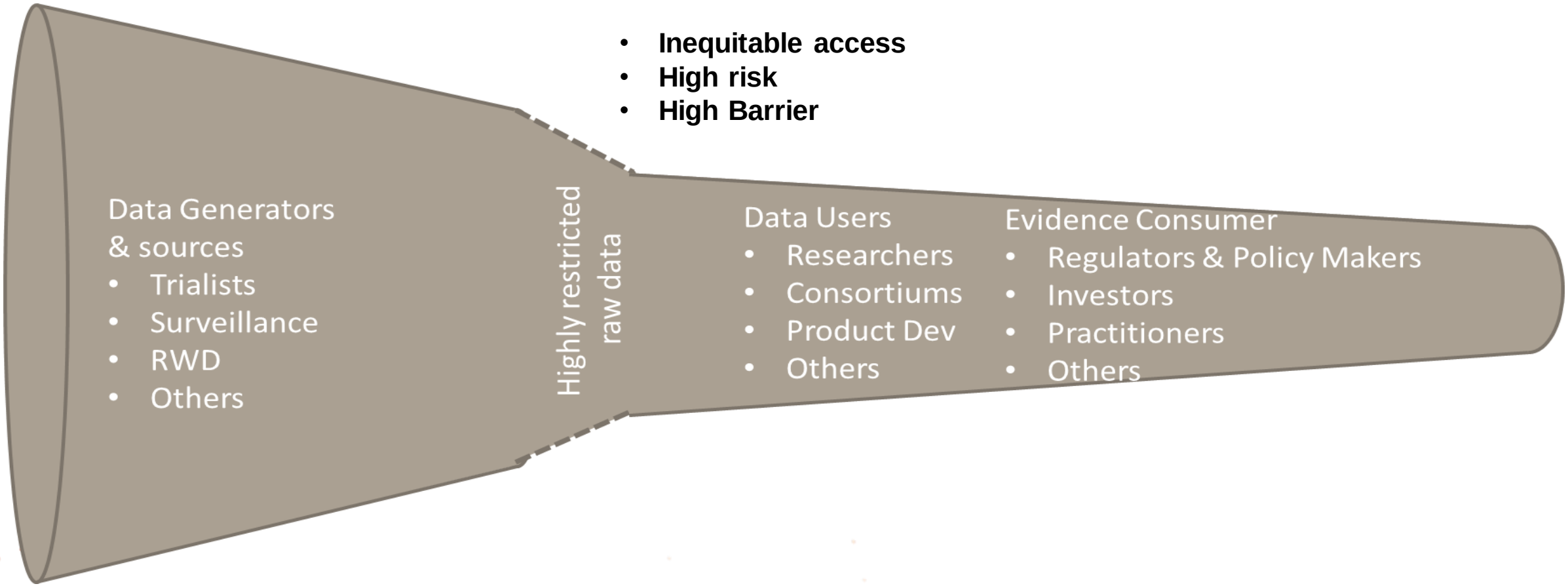
- Perform outcomes design & analyses primarily to inform investment strategy
- Advance and pilot data sharing capabilities to accelerate valuable data uses with the BMGF ecosystem
- Enable Foundation program teams to capture the unrealized potential of their funded data

Example Analysis: Analyzing Stunting and Wasting Incidence



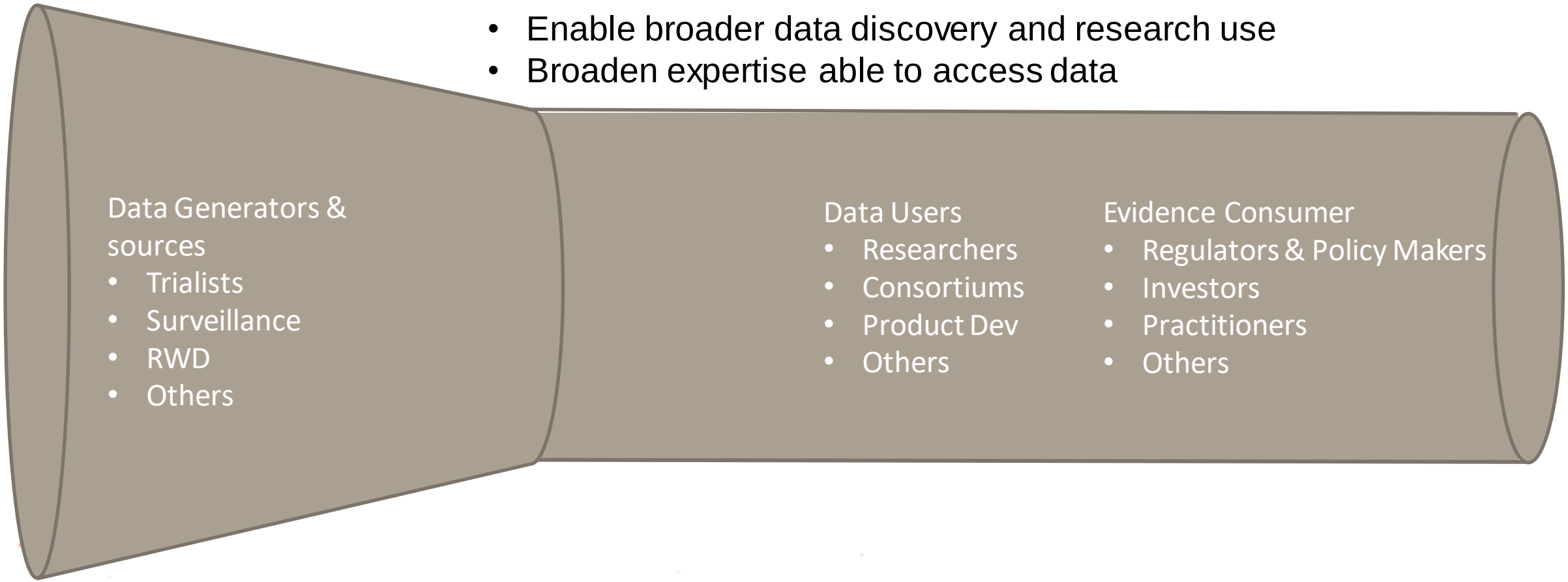
Graphic by UC Berkeley Data Science Team (A.Mertens, J.Benjamin-Chung, et al)
Data from BMGF grantees
Manuscripts under review

FRICTION REDUCES THE USE OF FUNDED DATA



SYNTHETHIC DATA BROADENS THE USES OF FUNDED DATA

- Enable broader data discovery and research use
- Broaden expertise able to access data



WHAT IS THE PERFORMANCE ENVELOPE FOR SYNTHETIC DATA?

Will synthetic data support analysis of low prevalence classes, novel research and other Foundation use cases?

How helpful are general utility measures in providing an assessment of the final inference had the real data been used?

Will privacy be sufficiently preserved to enable governments and other data custodians to increase access to researchers?

WHAT ARE THE PRIVACY PROTECTION LIMITS FOR SYNTHETIC DATA?

How to establish counterbalance of synthetic data from:

- **Learning some combination of characteristics (e.g., travel patterns, illness, likely residence locations) that identify a real individual or membership in the dataset**
- **Re-identification of a geographic site or institution (e.g., hospital, school) especially when desired**
- **Reconstruction of the original dataset from many synthetic dataset subsets or the synthetic data algorithm.**

SOME TYPES OF DATA WE THINK ABOUT

- Epidemiologic & biomedical
- Neuroimages and other image types
- Genetic sequence and omics
- Electronic medical records
- Mobile health
- Manufacturing quality control
- Survey, longitudinal, systems-based

EDUCATION-RELATED GOVERNMENT DATA FOR UNDERSTANDING WORKFORCE DEVELOPMENT

Connect K12, Higher Education, and Workforce data from longitudinal data systems with researchers

Approach:

- Create a privacy enhanced data-set using student, course and award data

Technical considerations:

- Must meet differential privacy expectations
- Apply Differential Privacy definition to each result individually
- Apply rounding and constraints
- Mask results under 10 students (FERPA)

GRANULAR DATA FOR REPLICATING & EXTENDING PUBLISHED RESULTS

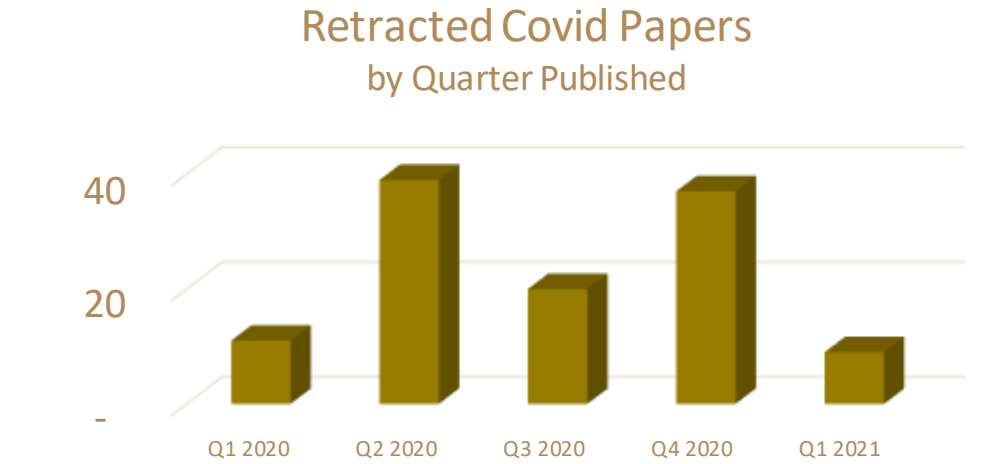
Published claims need to be evaluated to further evidence-based planning conversations

Approach:

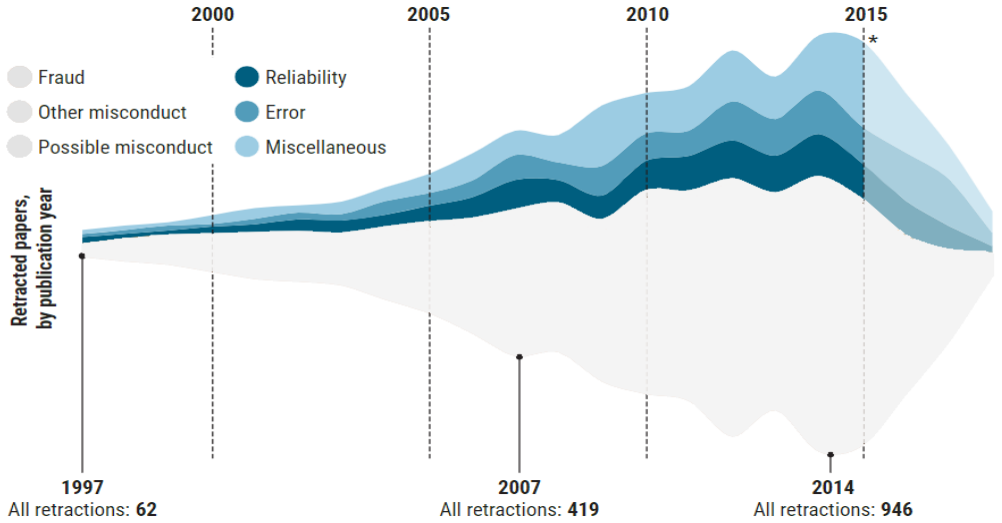
- Generate route to share underlying manuscript studies' data
- Replicate the results of mega-analysis focused on outlier data

Technical considerations:

- Evaluating partial and full data synthesis for privacy & analytics
- Parsimonious synthesis routes for heterogeneous studies with varied longitudinal visit schedules



Source: Retraction Watch Database July 1, 2021



Adapted from J Brainard, J. You; doi:10.1126/science.aav8384

NOVEL LONGITUDINAL ANALYSES FINDINGS WHILE ATTENUATING POTENTIAL RISKS

Sharing of rich subject level health data requires balance of increased privacy protection & data usability

Approach:

- Evaluate possibility of novel research using synthetic data

Technical considerations:

- Synthesize large cohort with rich patient history en masse
- Understand privacy risks associated with numerous synthesized covariates

DATA POWERS HOW WE DO WHAT WE DO

Grantees and partners are at the center of our work

Together, we take risks, push for new solutions and harness the power of science and technology

This work requires support from governments, the private sector, communities, nonprofits, and individuals

THANK YOU

Ki partners who contribute their study data and expertise

https://www.kiglobalhealth.org/data-contributors/

The ki Child Growth Consortium

PREVA GROUP

Tim Kinkead