Advances in Data Linkages

Michaela A. Dinan, PhD
Yale University

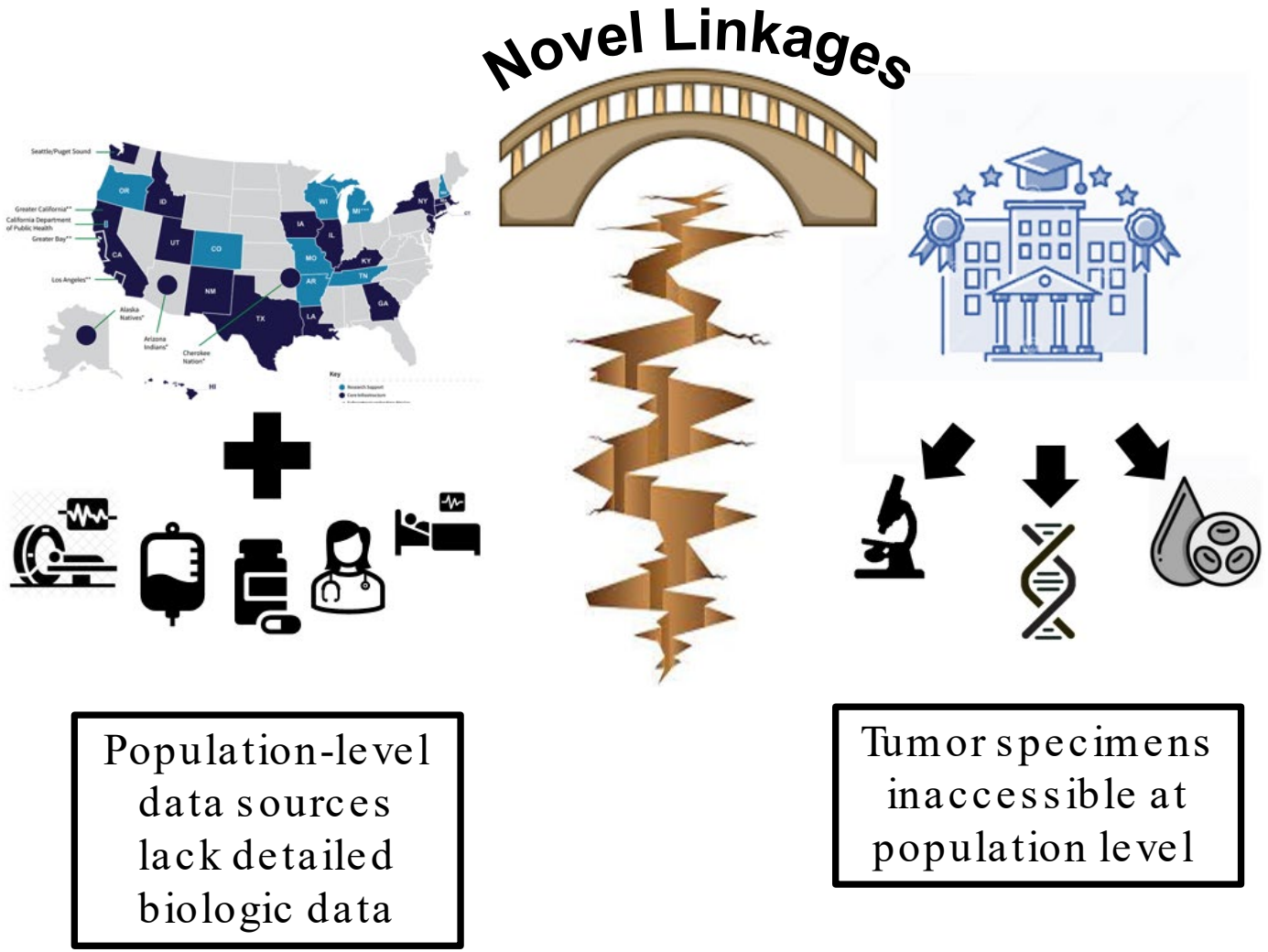
NCPF Workshop on Enabling Cancer Surveillance

July 29-30, 2024

Challenges in conducting population-level investigations that capture both social and biological factors

- Few (if any) population-level data sets include linkages to biological samples

- Why not?
  - Disconnect between population sciences and basic biology
  - Abundance of siloed genomic datasets
  - Privacy concerns



Opportunity for "Next-Generation" Population Science: Linking SEER-Medicare To Tumor Genomics

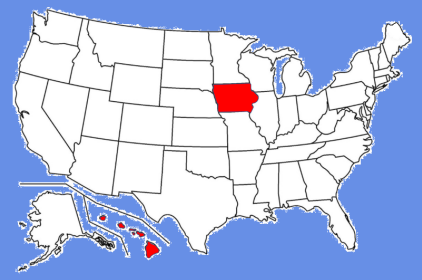
Newly dx ER+ breast cancer with linked SEER, Medicare, and genomic data (N = 130)

Medicare Claims

Symptomatic detection

Mortality HR 2.49

SEER Registry



Local (zip code) high-school graduation rate

Mortality HR 5.17

Tumor Samples (Genetic Analysis)

Androgen receptor, macrophage, cytotoxicity, and T-reg signaling.

Reduced Mortality

Factors related to socioeconomic status and screening access remain associated with mortality after adjusting for clinical and genomic factors

Robinson et al. Analysis of sociodemographic, clinical, and genomic factors associated with breast cancer mortality in the Surveillance, Epidemiology, and End Results (SEER)-Medicare linked database. JAMA-Network Open, August 23, 2021.

Barriers to Conducting Population-Level Genomic Analysis

- Restrictions on data use and sharing
  - SEER-Medicare (NCI)
  - Genomic Data Sharing Policy (NIH)
- Not all SEER registries participate
- Logistical workflows to maintain deidentification of data are complex
- Cost of obtaining FFPE samples is substantial

https://datascience.cancer.gov/data-sharing/genomic-data-sharing/about-the-genomic-data-sharing-policy