The Academic Perspective: Opportunities to leverage Big Data and Analytic Methods to Promote the Health of Individuals and Communities

Beth Virnig, PhD, MPH
Professor
Why an “academic perspective”? 

- Identify modifiable risk factors/ability to implement change
- Generalizable knowledge
- Suggestive of a causal relationship*
- Strong level of concern about data quality, validity, representativeness

* yes, we can debate the use of the word “causality”
Big Data used in Oncology—Biggest...

- General healthcare data lacking oncology-specific measures:
  - Medicare
  - Medicaid
  - Optum
  - State All Payer Claims Databases/discharge databases...

- Opportunities for oncology-related inference
  - ICD diagnosis codes used to identify cancer
  - Measure treatment received

- But imperfect in important ways… are they too imperfect?
Smaller big data...

- Combine big data (previous slide) and cancer-specific detail
  - SEER-Medicare
  - NCDB
  - State cancer registries linked to state discharge data or APCDs
    - Restricted to a limited number of geographic areas and select populations
Single institution/EMR data

- Data from medical records
- Granular
- Include text notes, etc.

But...
- HIPAA!!!
- Cannot see care not received at that institution/system and limited interoperability across systems
- Generalizability to other organizations limited/unknown
- Likely only source direct of information on factors like language, immigration status, etc.
Putting it together

• Some individual level SDoH are potentially measurable using ICD codes

• ICD-10 codes exist for problems related to …
  - education and literacy (Z55)
  - Housing and economic circumstances (Z59)
  - social environment (Z60)
  - Primary support group (Z63)

BIGGEST issue: health care information is of higher quality for identification of the presence of a problem than an absence.
Putting it all together (individual level)

- Algorithms can be useful:
  - Hispanic surname
  - Hmong surname
    - But Muslim surname $\neq$ Somali

- Few “algorithms” are validated, when they are, typically small local sample
  - Is “face validity” really a good measure of validity?
Inference opportunities with geographic measures of SDoH

- Opportunities to combine data sources leads to increased number and range of potential measures
  - Are not limited to data where a particular individual is identifiable in both datasets
    - BRFSS
    - Census
    - % low income
    - % low education
    - % non-English speaking
    - Racial segregation
    - Minimum distance needed to travel to…
Challenges relying on geographic-level measures

- The larger the geographic area, the greater the chance that the mean value mis-represents the individuals in the area
  - ZIP vs. MSA level measures
  - Distances traveled measured ZIP to ZIP or county to county will have different levels of precision
Is there anything unique to oncology?

• Role of a particular SDoH may vary across the cancer continuum
  - Timeframe for measures—consider when can SDoH change over the life course?
  - Impact of SDoH may vary between hospital and community-based care, acute care vs. longer-term
  - Survival—not just due to the cancer
Opportunities for applying SDoH measures from big data to oncology

- Examining whether success of patient navigators relates to SDoH—how to optimally deploy them
- Application of models such as the Cumulative Complexity Model to oncology settings
- Assess whether it is possible to use SDoH data from administrative (billing) data as a screen for enrollment into programming
- Expand number and type of measures of SDoH available to researchers, particularly without requiring release of detailed geographic information
- Decision modeling—simulate experiments to guide real-time decision-making
Questions?