Comparison of Biopsy and Diagnosis Code Based Breast Cancer Phenotypes

Matthew Spotnitz, Thomas Falconer, Maura Beaton, Karthik Natarajan

Department of Biomedical Informatics, Columbia University Irving Medical Center

Background:
Gold standard phenotyping of cancer can improve the rigor and reproducibility of research about the disease. There are multiple candidate phenotypes for the same kind of cancer, which can have different strengths. Specifically, some phenotypes may have higher sensitivities and others may have higher antecedent data frequencies. We used ATLAS to characterize the antecedent data frequencies for biopsy and diagnosis code-based breast cancer phenotypes.

Methods:
Biopsy Based Phenotype
The index event was a procedure code for a breast biopsy, such as CPT 19102 ("Biopsy of breast; percutaneous, needle core, using imaging guidance"). Additionally, patients had a diagnosis code for breast cancer within 90 days following the biopsy, such as SNOMED-CT 93796005 ("Primary malignant neoplasm of female breast"), and no history of breast cancer at least 30 days prior, which was indicated by codes such as SNOMED-CT 429087003 ("History of malignant neoplasm of breast"). We restricted to patients who were women who were between 18 to 80 years old, had at least 90 days of prior observation in the databases and a biopsy on or after 01/01/2000.

Diagnosis Based Phenotype
The index event was a first diagnosis code for breast cancer such as SNOMED-CT 93796005 ("Primary malignant neoplasm of female breast"), and no history of breast cancer at least 30 days prior, which was indicated by codes such as SNOMED-CT 429087003 ("History of malignant neoplasm of breast"). We restricted to patients who were women who were between 18 to 80 years old, had at least 90 days of prior observation in the databases and had a breast cancer diagnosis code on or after 01/01/2000.

Data Frequency Plots
Using ATLAS characterization plots, we compared the one-year antecedent data frequencies of both phenotypes with data from multiple domains.

Data Sources
We implemented our analysis on the Columbia University Irving Medical Center (CUIMC) Electronic Health Record, IBM MarketScan Commercial Claims and Encounters (CCAE), and IBM MarketScan Medicare Supplemental Beneficiaries (MDCR) databases.

Results: The data frequency plots are shown in Figure 1.

Conclusion: On multiple databases, the biopsy-based breast cancer phenotype had higher one-year antecedent data frequencies and longer prior observation times, suggesting more

complete patient records. When choosing a cohort definition, reviewing such plots may help the investigators make data-driven decisions.

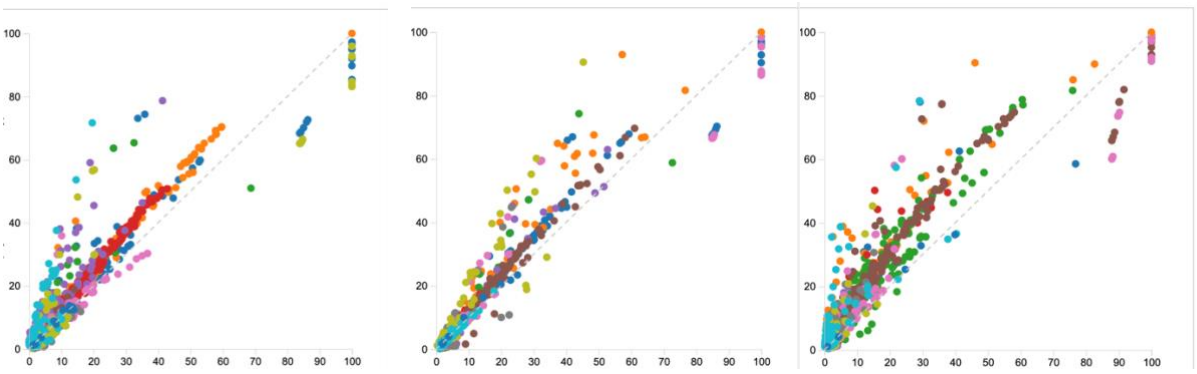

Figure 1: One-year antecedent data frequencies of the biopsy based (y-axis) and diagnosis based (x-axis) breast cancer phenotypes for the a) Columbia University Irving Medical Center (CUIMC) Electronic Health Record, b) IBM MarketScan Commercial Claims and Encounters (CCAE), and c) IBM MarketScan Medicare Supplemental Beneficiaries (MDCR) Databases. Each circle corresponds with a covariate and each color with a data domain.