TuftsMedicine
Tufts Medical Center
NEWDIGS

LEAPS Methods Innovation Team

Case Study Background

For the initial advanced NSCLC case study, develop a predictive model(s) that can improve decision-making for all stakeholders related to immune checkpoint inhibitor use in patients with advanced NSCLC.

Objectives

Assess feasibility of using federated (machine) learning methods, leveraging diverse data types (e.g., EHR, administrative claims, social determinants of health, biologic, clinical trials, patient-generated, etc.) to:

- Identify signals, generate hypotheses about clinically meaningful sub-populations
- Define next step in corroborating/validating promising hypotheses
- Reduce bias in algorithm development through the use of diverse data sets
- Establish federated learning environment (technology enablers, cross-functional expertise, governance) that is scalable

Purpose of Data & Technical Skills Framework

- Develop list of organizations with data & their corresponding technical skills needed for the successful completion of a prototype
- Define the data privacy and security framework to attract and address any concerns of the data providers
- Develop & apply a framework for characterization & evaluation of each data source in order to identify the initial set of datasets.

Framework Application and Approach

The Data & Technical Skills Characterization Framework (Framework) approach begins with the application of the NEWDIGS 5-Layer Cake to facilitate prioritization in the search of the data sources under consideration for inclusion into the Predictive Outcomes Platform (POP). A primary goal of this approach is to ensure strategic inclusivity of data sources.

The Framework seeks to identify and determine the strengths & limitations of the data sources and needed technical skills to the Advanced NSCLC Use Case specifically. In addition to capturing general details, such as size and type of the

data source, the strengths and limitations of each source are characterized and assessed across multiple *objective dimensions*:

Strengths:

- Large sample size
- Depth of subpopulation representation
- Representativeness across multiple factors
    - Diversity
    - Patient generated
    - Sensor generated
    - Provider types (academic, community)
    - Geographic/regional
    - Outcomes (clinical, genomics, PROs)
    - Longitudinal in scope
    - Comprehensiveness of care journey
        - Closed/Open system
- Linkage capability with other data sources
- Low complexity of administrative barriers
    - Little/no additional funding requirements
    - Minimal legal hurdles for formal partnerships
- Number of publications supported by the data sources

Limitations

- Small sample size
- Dated aspect of claims data
- Structured vs unstructured data (requiring AI skills and experience)
    - Depends on the depth of technical skills available at the respective data holder orgs
    - Need to be specific on understanding what technical skills are needed: difference between programmer skills vs. “administrative” data analysis skills

Commented [HKA1]: For discussion: Are the "administrative" data analysis skills for reporting only whereas the programmer skills are for predictive modeling?

Appendix

Table 1: Data source general information

Title	
Data holder	
Data type	
Total number of records	
Total number of records relevant to case study*	
Key focus areas	
Unique features	

** Example: for the advanced NSCLC case study, what is the number of patient records have been diagnosed with adv. NSCLC (as determined by the agreed-upon criteria)*

Table 2: Data source dimensions and characterization

Commented [HKA2]: For discussion:

1. What additional dimensions would we want to include here under the "Dimension" column?

2. For the "Additional context" column: what helpful information could/should be entered here?

Dimension	Strengths	Limitations	Add
Sample size			
Extent of subpopulation representation			
Representativeness of diversity factors			
Provider type (academic, community, etc.)			
Geographic/regional			
Outcomes (types including clinical, genomics, PROs, etc.)			
Longitudinal in scope			
Comprehensiveness of care journey			

Dimension	Strengths	Limitations	Additional context
Closed/Open system			
Age of data (current or outdated)			
Structured or unstructured			
Linkage capability with other data sources			
Complexity of administrative barriers for data access			
Need for additional funding requirements			
Legal requirements for formalizing partnerships			
Number of publications supported by the data sources			

Table 3: Data partner organizational information

Organization Name	
Type of Organization	
Size of Organization	
Areas of specialization	
Unique Features/Areas of Expertise	

Table 4: Data partner technical skill dimensions and characterization

Commented [HKA3]: For discussion: What additional dimensions would we want to include here under the "Dimension" column?

Dimension	Strengths	Limitations	Depth/Breadth of Expertise
Statistics/Statistical Modeling			
Healthcare data analysis (claims, EHR, Registries, etc.)			
Predictive Analytics			
AI & Machine Learning			
Data Science			
Computer Science			
Natural Language Processing			
Number of recent publications relevant to specific case study/focus area			