A National Academies of Sciences, Engineering, & Medicine Workshop: Addressing Resistance in the Development of Cancer Immune Modulator Therapeutics

Session 4: Current Challenges and Opportunities: The Role of Data & Computational Tools

AI, Data Science, and BigData Approaches to Accelerate/Expand Research and Evaluation

November 14-15, 2022
Washington, D.C.

Usama Fayyad
u.fayyad@northeastern.edu
Executive Director, Institute for Experiential AI
& Professor of the Practice
Khoury College of Computer Sciences

EAI The Institute for Experiential AI
Northeastern University

Disclosures

- I represent the Institute for Experiential AI at Northeastern University – a private not-for-profit institution
- The faculty members of Institute for EAI receive research funding from NIH, NSF, FDA, DARPA and many other public and private research funding agencies
- The Institute of EAI works with and seeks projects with companies (private and public) in the AI+Life Sciences area – we leverage such applied projects to drive new AI research as well as provide experiential learning opportunities to students from Northestern University and learners from partner organizations
- I am also affiliated as chairman of a company I founded in 2008: Open Insights. Historically Open Insights has worked with pharma and other tech and manufacturing companies on big data, data science and AI projects.

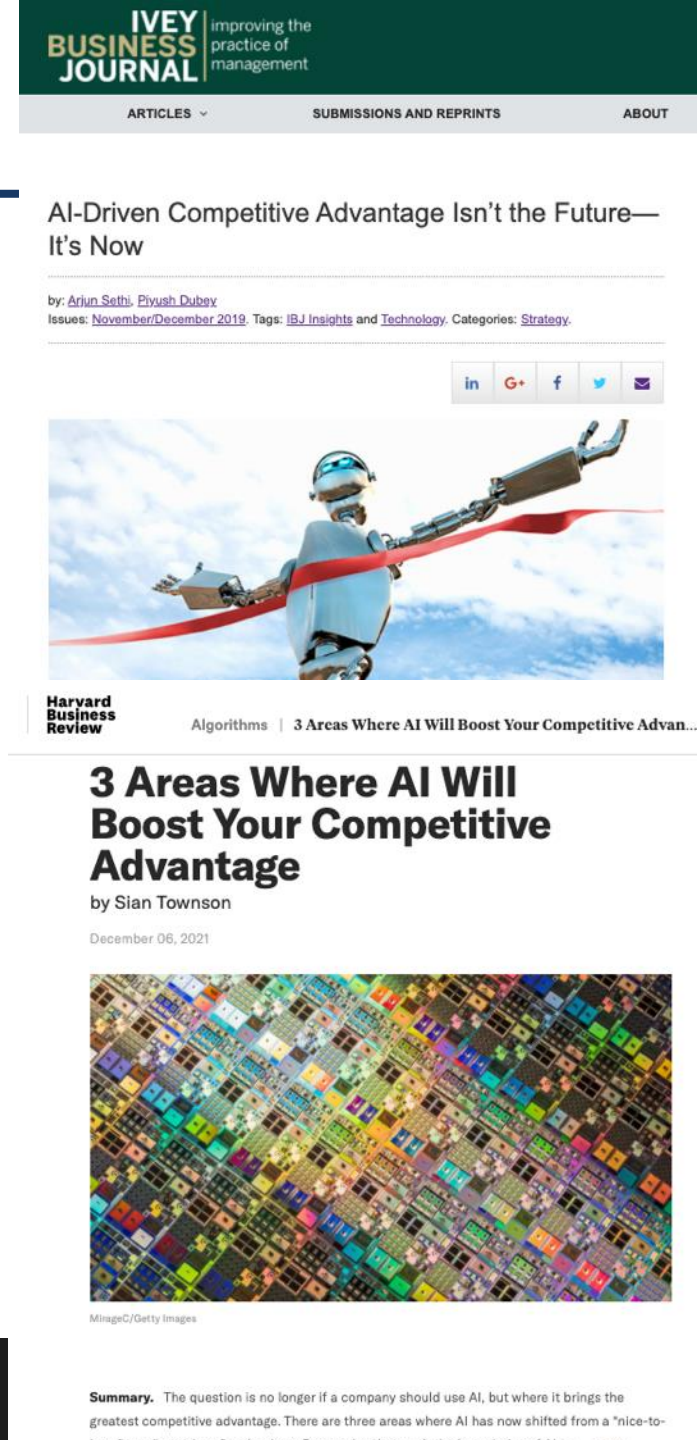
Overview of this talk

Making AI work correctly is one of the grand challenges facing us today in many fields

- Digitization has taken hold and has been greatly accelerated with COVID-19 pandemic
- AI has been a great and difficult challenge
- Machine learning is the dominant part of any working AI
- ML has a huge dependency on Data

What would be the impact of leveraging what AI, Data, and Data Science have to offer to Life Sciences and Cancer Immune Modulator Therapeutics research and development?

- Applications in Healthcare in general
- Applications in Life Science
- Applications in Cancer Immune Modulator Therapeutics

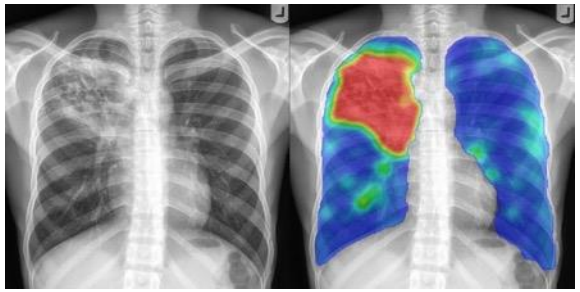


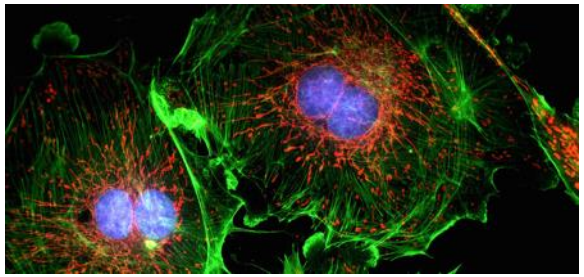
What Has Digital Transformation Looked Like in Health?

- **Slower** *digitization and digitalization* than other fields
- Great advances in device tech – from surgical to imaging
- Interesting examples of inconsistent adoption of digital & AI

Some sample examples of uses of AI:

	Routine Medical Diagnostics	Cell Imaging	The Omics
✓	Radiology: digitized and a lot of automation	Great advances in technology down to sub-single-cell	Great advances in Genomics, Proteomics, and Transcriptomics
✗	Pathology: still in the land of analog and little AI processing	Very few large-scale uses	Very little work on combining omics and advancing to metabolomics



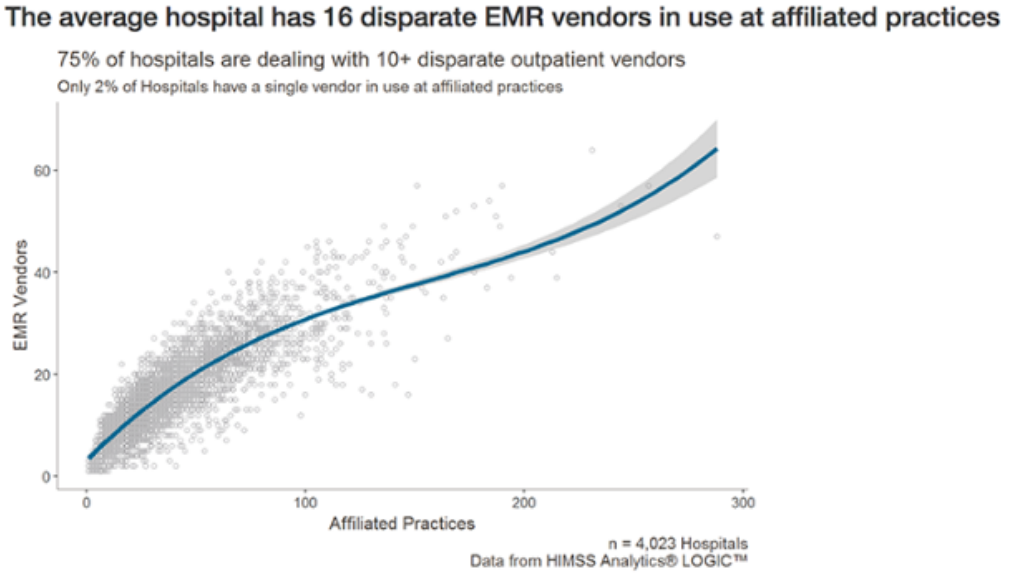


What About Electronic Medical Records?

- Mandated digital coding and medical notes
- Failure to standardize the Data
- No two EHR's look the same…
- Very fragmented space, and even the largest EHR systems do not inter-operate
- No real incentives to share data

This results in great difficulty to leverage AI/ML and automated analytics to help leverage the wealth of data

The average hospital has 16 disparate EMR vendors in use at affiliated practices Most hospitals have 10 EHR systems, and only 2% are down to only two EHR systems!



Meanwhile the platform is burning in the U.S.

- Cost of healthcare is growing out of control as share of GDP
- > $4T total medical spend in U.S. (CMS 2021 – 10% annual growth),
- It is estimated that > $400B of administrative waste
- We need serious help in figuring out how to scale healthcare in a more economically sustainable model
- AI approaches hold a lot of promise to help
  - In Healthcare in general
  - In Health Sciences
  - In Life Sciences and Drug Discovery
  - In understanding diseases better, faster
  - In evaluating therapy effectiveness better, faster

AI could help, but there is a serious catch…

- Working AI needs ML and Data
- Data requires digitization
- Data capture, representation, sharing, and management remains a largely unaddressed challenge in healthcare…

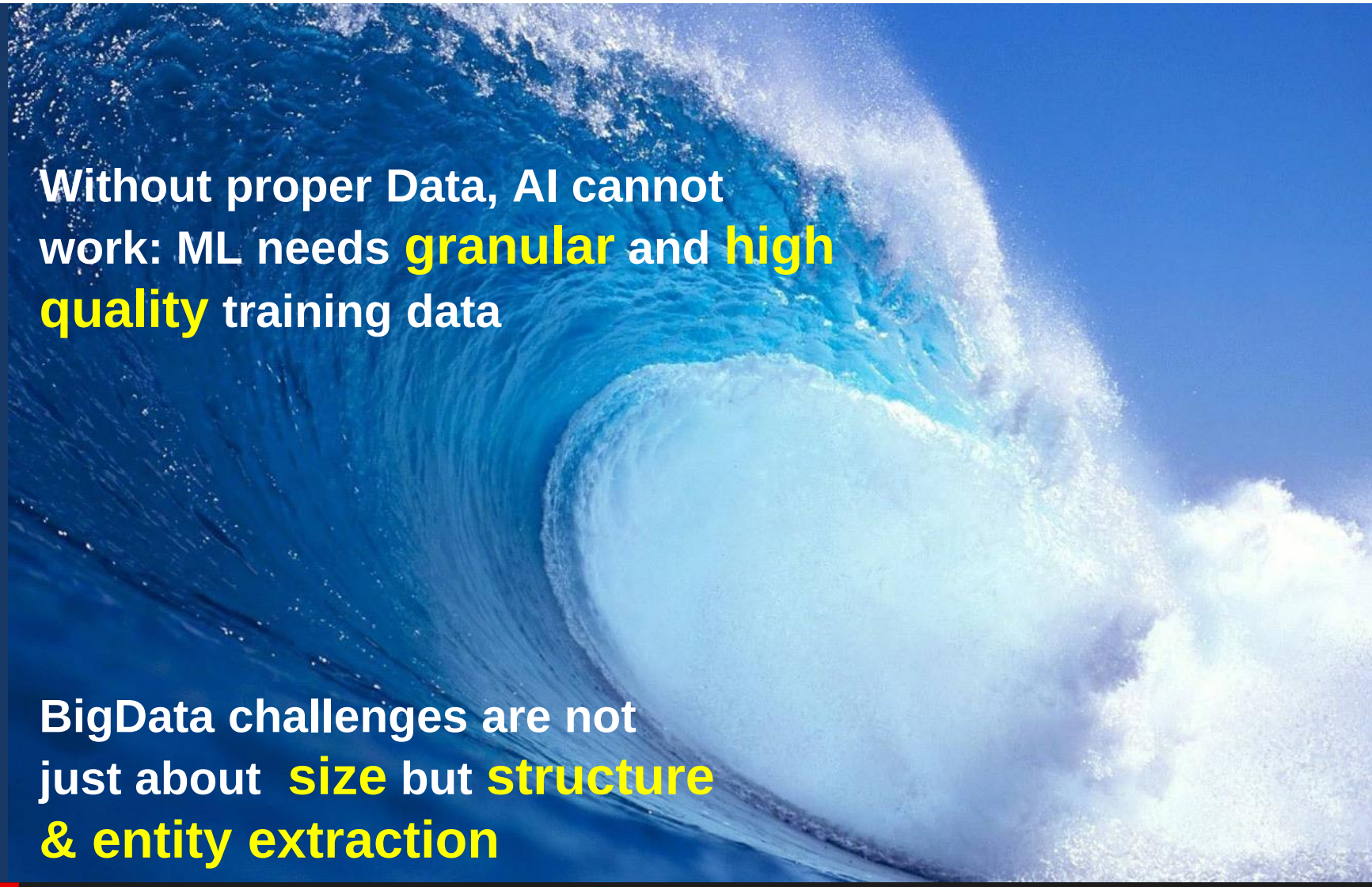
Digitization Produces Much More Data

But most organizations are not equipped to effectively manage data as an asset

How do we make this Data a working asset?

New economy of Interactions is rich with unstructured data

90% of Data in any organization is UNSTRUCTURED



Without proper Data, AI cannot work: ML needs granular and high quality training data

BigData challenges are not just about size but structure & entity extraction

Areas where AI Can Help

- NLP to leverage unstructured text data – LLM (large language models) and other open source methods for image and TS analysis
- Image analysis tools to leverage and retrieve related image data (*query by example, pattern recognition, etc.*)
- Graph-based and network representations
- Network Science models for understanding multi-factor interactions
- Multi-omics approaches to counter the single-omics traditions

Why Multi-omics?

- **Detect and understand interactions between different omics**
  - Genome or Exome
  - Proteomics
  - Transcriptomics
  - Metabolomics
- **Most work appears to focus on working within one "silo"**
  - Genomics and transcriptomics are "mature"
  - Metabolomics evolving and difficult but is critical in immuno-response and cell-response analysis
- **Example:** Exome or Genome "similar" patients have different responses to the same cancer therapeutic:
  - Respond well vs. no response?
  - Respond well but then stop responding after some time?

The resistance is also “cultural” and “social”

- **Clinical protocols are hard to change** (for good reason): from a Data Science/AI technologist perspective:
  - Typically simple
  - Typically ”outdated”
  - Typically do not leverage latest technology, science, or math (probabilistic modeling)
- **Consider an Example in Single-Cell Analytics**
  - Single cell metabolomics or in-vivo imaging/video
  - Can actually observe direct effect of therapy at cell-level – e.g. are tumor cells being attacked by immune cells?
  - Can evaluate effectiveness/impact in days (not weeks or months)
  - **How do you get uptake of this new evidence by a clinician?**
  - Typical outcome: follow the “weeks to months” observation cycle instead of adjusting therapy
    - ***What do we need to trust the micro and cell-level responses as much as we trust the “phenotype” observables?***

Example Problems Where AI and Data Can Help in the Health & Life Sciences

A sampling of a very large space in Health and Life Sciences

AI in Healthcare through Data

- **AI to understand social determinants of health:**
  - conditions in the environment where people are born, live, learn, work, play, worship, and age that affect a wide range of health, functioning, and quality of life outcomes and risks.
  - Economic stability, educational access and quality, healthcare access and quality, neighborhood and built environment, social and community context (https://health.gov/healthypeople/priority-areas/social-determinants-health)
- **AI to reduce medical errors:**
  - Medical errors are 3rd leading cause of death (251K deaths annually) and costs the US billions of dollars (https://www.marsdd.com/news/saving-lives-with-ai-machine-learning-reduced-hospital-mortality-rates-by-20-percent/)
- **Closing the loop – Digital Health in the Home:**
  - Tracking and feedback to healthcare providers on what happens outside the clinical setting- leveraging of wearables, nearables, and IOT devices in general
- **Therapeutics and Protocol effectiveness evaluation** through population health data crossed with other data sources (shopping, exercise, etc.)

AI in the Life Sciences – some examples

- Drug discovery and drug design
- Therapeutics Effectiveness: cell biology and chemistry
- AI for High-throughput processing: e.g. metabolomics, processing of imaging, interpreting simulation results
- Combining multiple-omics to improve prediction and enhance understanding – enable new models
- AI for drug validation: Computational models of Interactomics (the omics of cell interactions):
  - temporal dynamics, cell kinetics, cell types
  - prediction models of human response to therapy to save a “trial-and-error” approach on patients.
- Disease spread tracking and hyperlocal to global models

AI/ML – Data Science for Immune Modulation Therapy

- Multi-dimensional predictive models
- Drug/agent combination models
- Network science models to account for large variable spaces
- Knowledge graphs to “find similar” and match against research studies – to reduce large search spaces
- What about long-term tracking of treatment/patients?
  - Very possible but requires standardization of data
  - Stable, extensible, and backward-compatible data standards
  - Requires policy and governance to implement and maintain (think EMR’s with *focus on standardization)*

Northeastern University
Observational Health Data Sciences and Informatics Center

OHDSI

A global community with worldwide impact
Working together across six continents and 80 countries, OHDSI (Observational Health Data Sciences and Informatics) —pronounced "odyssey"— members transform patient care using transparent research methods and standardized, open-source data.

What is the OHDSI Center?

Founded in 2021, the OHDSI Center at Northeastern's Roux Institute equips learners and researchers with the training, credentials, and resources they need to extract value from observational health data.

Through collaborations with academia, industry, and government worldwide, we generate study results that serve as accurate, reliable guides for healthcare decision making.

ohdsi.northeastern.edu

Advance research. Access data.

The OHDSI Lab enables population-based research on standardized health data maintained by the global OHDSI community for 12% of the world's patient population.

Support for research programs.

See for example:

The Proceedings of the National Academy of Sciences 113– 27
Characterizing treatment pathways at scale using the OHDSI network: https://www.pnas.org/doi/10.1073/pnas.1510502113.

Partner with us.

Collaborations in research to advance patient care and quantitative methods. Deep expertise in novel statistical and epidemiologic methods. Co-design learning programs.

Northeastern University
Observational Health Data
Sciences and Informatics Center

Towards Common Data Models for RWE.

- Access open-source data
- Workforce development in RWE
- Sponsored research
- Collaborative research
- RWE consultancy services
- Chart your own post-doc journey
- Make the most of your sabbatical

Backed by World Class Experts in Large Scale Observational Data Science

David Madigan
Special Advisor
Provost

Christian Reich
Executive Director

Kristin Kostka
Director

Asieh Golozar
Director of Clinical
Research

Justin Manjourides
Curriculum Lead

Louisa Smith
Faculty

Brianne Olivieri-Mui
Faculty

Stephen Flaherty
Faculty

ohdsi.northeastern.edu

FDAAA calls for establishing Risk Identification and Analysis System

SEC. 905. ACTIVE POSTMARKET RISK IDENTIFICATION AND ANALYSIS.

(a) IN GENERAL.—Subsection (k) of section 505 of the Federal Food, Drug, and Cosmetic Act (21 U.S.C. 355) is amended by adding at the end the following:

"(3) ACTIVE POSTMARKET RISK IDENTIFICATION.—

"(A) DEFINITION.—In this paragraph, the term 'data' refers to information with respect to a drug approved under this section or under section 351 of the Public Health Service Act, including claims data, patient survey data, standardized analytic files that allow for the pooling and analysis of data from disparate data environments, and any other data deemed appropriate by the Secretary.

"(B) DEVELOPMENT OF POSTMARKET RISK IDENTIFICATION AND ANALYSIS METHODS.—The Secretary shall, not later than 2 years after the date of the enactment of the Food and Drug Administration Amendments Act of 2007, in collaboration with public, academic, and private entities—

"(i) develop methods to obtain access to disparate data sources including the data sources specified in subparagraph (C);

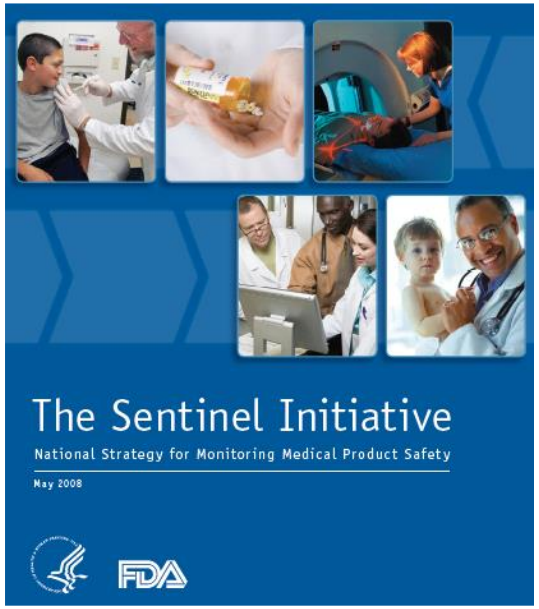
"(ii) develop validated methods for the establishment of a postmarket risk identification and analysis system to link and analyze safety data from multiple sources, with the goals of including, in aggregate—

"(I) at least 25,000,000 patients by July 1, 2010; and

"(II) at least 100,000,000 patients by July 1, 2012; and

"(iii) convene a committee of experts, including individuals who are recognized in the field of protecting data privacy and security, to make recommendations to the Secretary on the development of tools and methods for the ethical and scientific uses for, and communication of, postmarketing data specified under subparagraph (C), including recommendations on the development of effective research methods for the study of drug safety questions.

"(C) ESTABLISHMENT OF THE POSTMARKET RISK IDENTIFICATION AND ANALYSIS SYSTEM.—

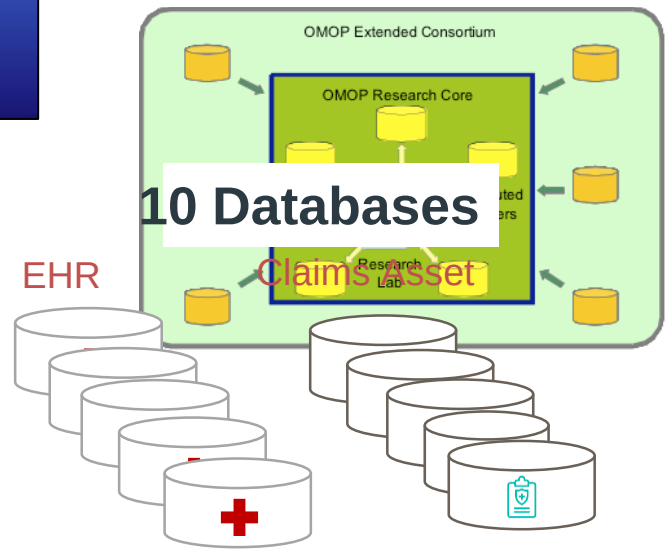


Risk Identification and Analysis System:

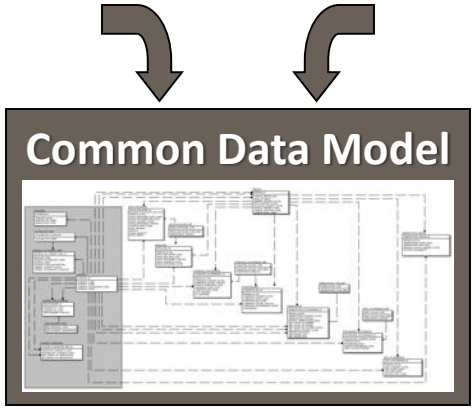
a systematic and reproducible process to efficiently generate evidence to support the characterization of the potential effects of medical products from across a network of disparate observational healthcare data sources

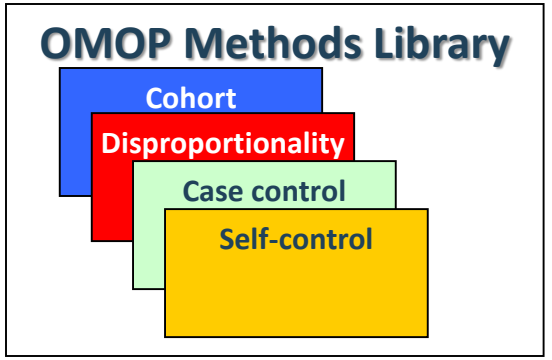
Observational Medical Outcomes Partnership

OMOP Experiment 2 (2011-2012)



- Open-source
- Standards-based





- 14 methods * 70 settings = **1,000 SAS scripts**

10 Drugs

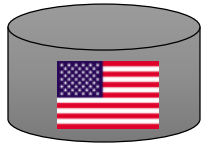
10 Outcomes

	ACE Inhibitors	**Amphotericin B**	**Antibiotics:** erythromycins, sulfonamides, tetracyclines	**Antiepileptics:** carbamazepine, phenytoin	**Benzodiazepines**	**Beta blockers**	**Bisphosphonates:** alendronate	**Tricyclic antidepressants**	**Typical antipsychotics**	**Warfarin**
Angioedema										
Aplastic Anemia										
Acute Liver Injury										
Bleeding										
Hip Fracture										
Hospitalization										
Myocardial Infarction										
Mortality after MI										
Renal Failure										
GI Ulcer Hospitalization										

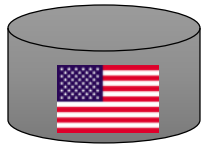
OMOP Experiment 2 (2011-2012)

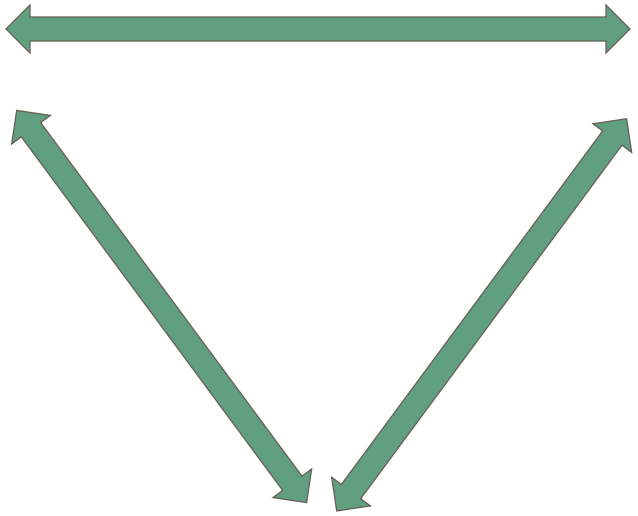
Observational Data

4 claims databases



1 ambulatory EMR





Methods

- Case-Control
- New User Cohort
- Disproportionality methods
- ICTPD
- LGPS
- Self-Controlled Cohort
- SCCS

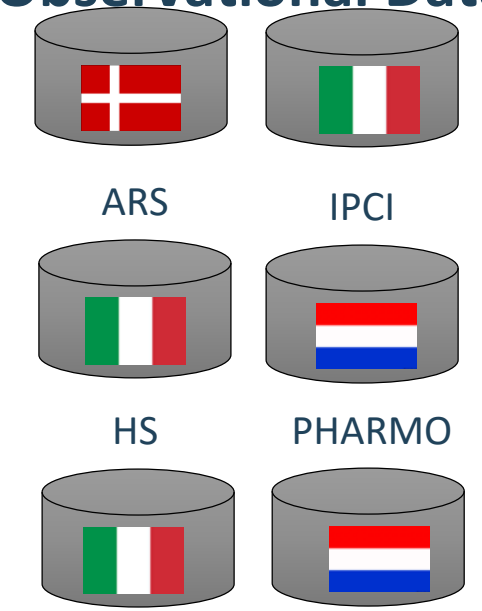
Drug-outcome pairs

	Positives	Negatives
Total	165	234
Myocardial Infarction	36	66
Upper GI Bleed	24	67
Acute Liver Injury	81	37
Acute Renal Failure	24	64

European OMOP Experiment

eu-adr

Observational Data



Methods

- Case-Control
- New User Cohort
- Disproportionality methods
- ICTPD
- LGPS
- Self-Controlled Cohort
- SCCS

Drug-outcome pairs

	Positives	Negatives
Total	165	234
Myocardial Infarction	36	66
Upper GI Bleed	24	67
Acute Liver Injury	81	37
Acute Renal Failure	24	64

Results

Main findings in OMOP experiment

- Heterogeneity in estimates due to the choice of database
- Heterogeneity in estimates due to analysis choices
- Except little heterogeneity due to outcome definitions
- Good performance (AUC > 0.7) in distinguishing positive from negative controls for optimal methods when stratifying by outcome and restricting to powered test cases
- Self controlled methods perform best for all outcomes

OHDSI*/OMOP** Open Research Network

Fast access to open source observational health data

Open Source
Common Data Model, Vocabularies, Tools, Methods, Libraries of phenotypes

Community
Multiple Stakeholders: Academia, Government, Health System, Technology, Patient, Pharmaceutical, Payer

Data
Assets from Hospital EHRs, University Medical Centers, Specialty institutions for Oncology, Pediatrics, Immunology, Psychiatry. Countries that ban data distribution

Why Choose OHDSI/OMOP:

- ✓ **Faster, more reliable** studies across a series of datasets and data types
- ✓ **Reduced cost of ownership** including understanding coding schemes, writing statistical programs across databases or developing software
- ✓ **Expanded data access** via the OHDSI network and remote multi-center database studies

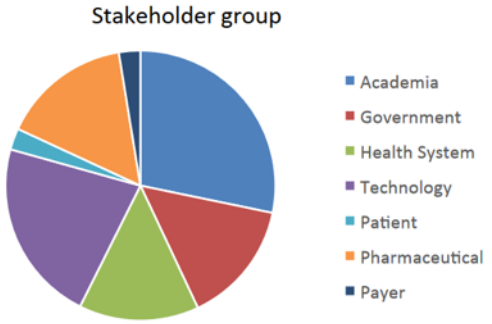
***OHDSI (Observational Health Data Science and Informatics)** is a public initiative independent of IQVIA.
****OMOP (Observational Medical Outcomes Partnership)** is no longer an active project. OMOP Common Data Model and Standardized Vocabularies are maintained and developed by OHDSI.

OHDSI

OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

What OHDSI is:

- ✓ **Open Source**
- ✓ **Community**
- ✓ **Data**



Why Choose OHDSI/OMOP:

- ✓ **Fast, reliable** studies across a series of datasets and data types
- ✓ **Reduced cost of ownership** including understanding coding schemes, writing statistical programs across databases or developing software
- ✓ **Expanded data access** via the OHDSI network and remote multi-center database studies

OHDSI Collaborators:

- 2,770 users
- 25 workgroups
- 18,700 posts on 3,250 topics

OHDSI Network:

- >150+ databases
- 21 countries
- 2.1B patient records, 369M ex-US

Our Learnings & Approach for DaaS Implementation

	Our learnings from prior experience	Our philosophy to change the game
	**Breaking the Budget** Setting up a Big Data Environment is expensive and takes a long time	**Environment as Code & Open Source** Accelerator Blueprints, environment up & running in days instead of weeks or months
	**Garbage In, Garbage Out** Getting quality data to business in timely manner is more difficult than it seems	**Built-In Governance & Processing** Proven framework, real time data collection & processing - Data Quality Processes
	**Structured Data Only** Lack of architecture and technology capabilities to leverage unstructured data	**BigData & Data Science Framework** Data science framework for unstructured data processing and entity extraction
	**Recipe for Failure** Long term Technology migration projects rather than iterative business deliveries	**Blueprint Strategic Architecture** Incremental build to reference architecture ***delivering quick business benefits***
	**Lack of Talent** Lack of data speciality results in generic skills doing data and coming up the learning curve	**Data Specialists & Data Academy** Skilled engineers who have implemented DaaS several times at large organizations

Themes in this talk:

1. AI is becoming an imperative in the digital age - yet challenging to make work
2. No Data ⇒ no working AI
3. Digitization ⇒ more complex & unstructured data - good news for AI?
4. Getting the data story right is the key enabler
5. There is a rational approach to getting to data assets - requires **Data Standards and Strategy**, **DaaS** with built-in **Data Governance**, and **incremental** build to reference **Data Architecture**

Where do efforts fail?

1. Lack of standards
2. Lack of pragmatic data governance enabling access & sharing
3. Lack of talent
    - Lack of skills in AI/ML/DS in Life Sciences
    - Lack of experience in the “art” of data science
    - Life Sciences + Data Science = Unicorn

Recommendation: address the issue of producing productive talent at these critical overlaps our

- How to train new graduates?
- How to upskill existing researchers and employees?
    - Since higher ed cannot address the gap in demand for talent

Talent Development at Institute for EAI

- our approach at the Institute for Experiential AI is “experiential education”
  - create the equivalent of a “medical residency” education – what it takes to **practice the art** as opposed to learn the principles of Data Science & AI
  - Requires working on real projects with real data and real constraints
- **Institute for EAI – 3 of the 4 focus areas:**

Thank you!
Any Questions?

USAMA M. FAYYAD

U.Fayyad@northeastern.edu

Assistant: R.Alshami@Northeastern.edu

AI.northeastern.edu

IEAI-NU
ufayyad

@Experiential_ai
@UsamaF

Institute for Experiential AI
Open Insights

3 Sample Case Studies

By Faculty at the Institutue for EAI at Northeastern University

Case Study:
Machine Learning in Medical Image Analysis

Professor Jennifer Dy
AI Faculty Director @ the Institute for Experiential AI

Skin Cancer Diagnosis

Jennifer Dy, Professor, ECE, IEAI

Reflectance Confocal Microscope (RCM) is **non-invasive** allowing imaging of nuclear morphology.

3-D view
Need guidance to focus narrow field of view on likely cancer spots
Mosaic increases field of view

Enface slices
at increasing depth

Mosaic

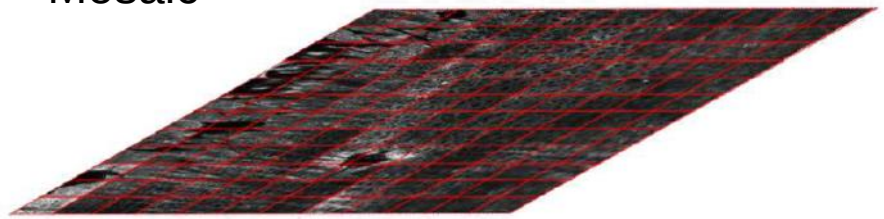
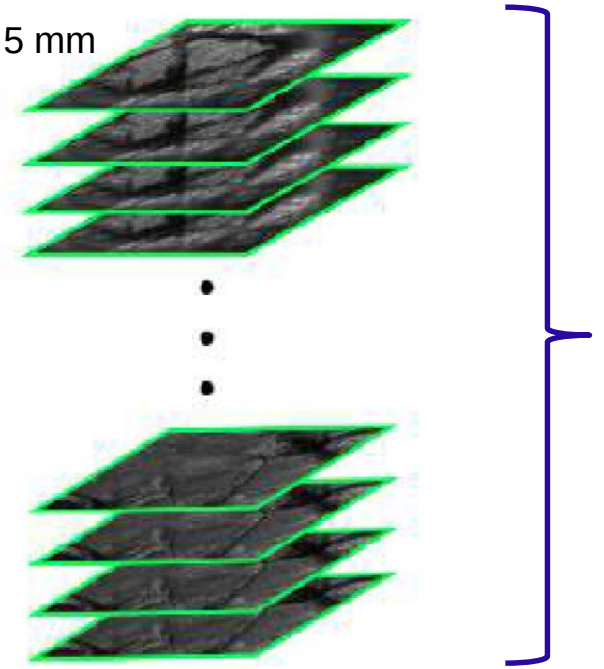
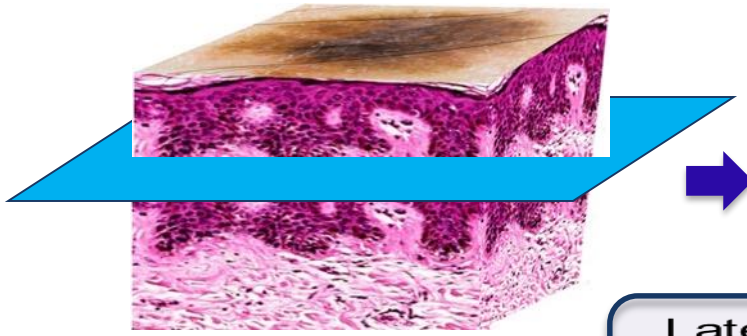
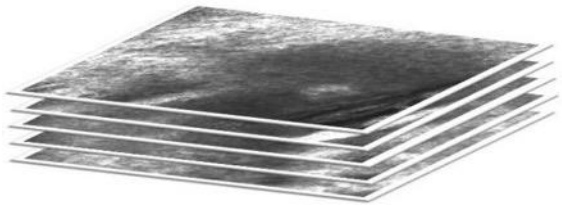


Image Stack





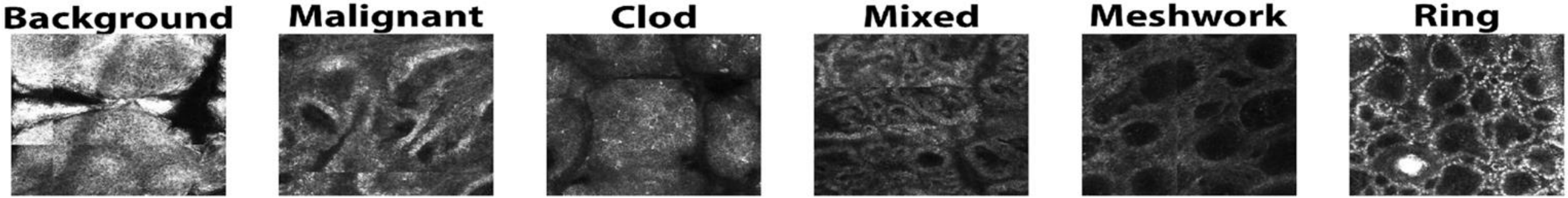


Lateral resolution: 0.5 μm
Section thickness: 2μm
Each RCM stack typically contains 50-80 slices

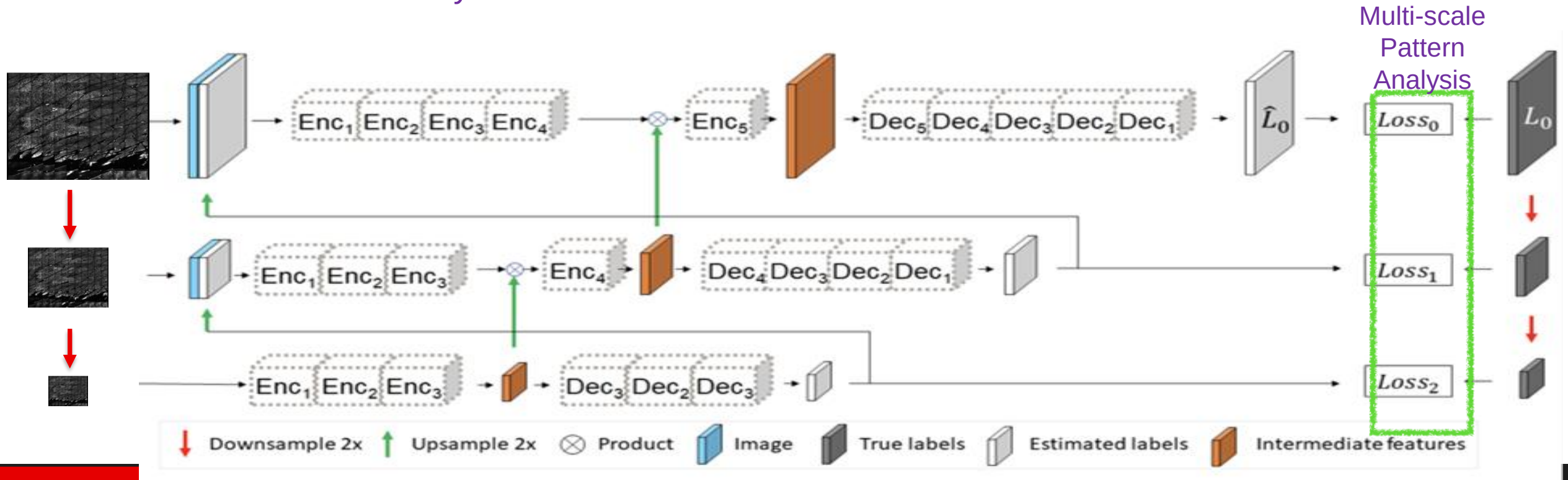
RCM Mosaic Segmentation

Patterns of interest are typically texture and are highly variable

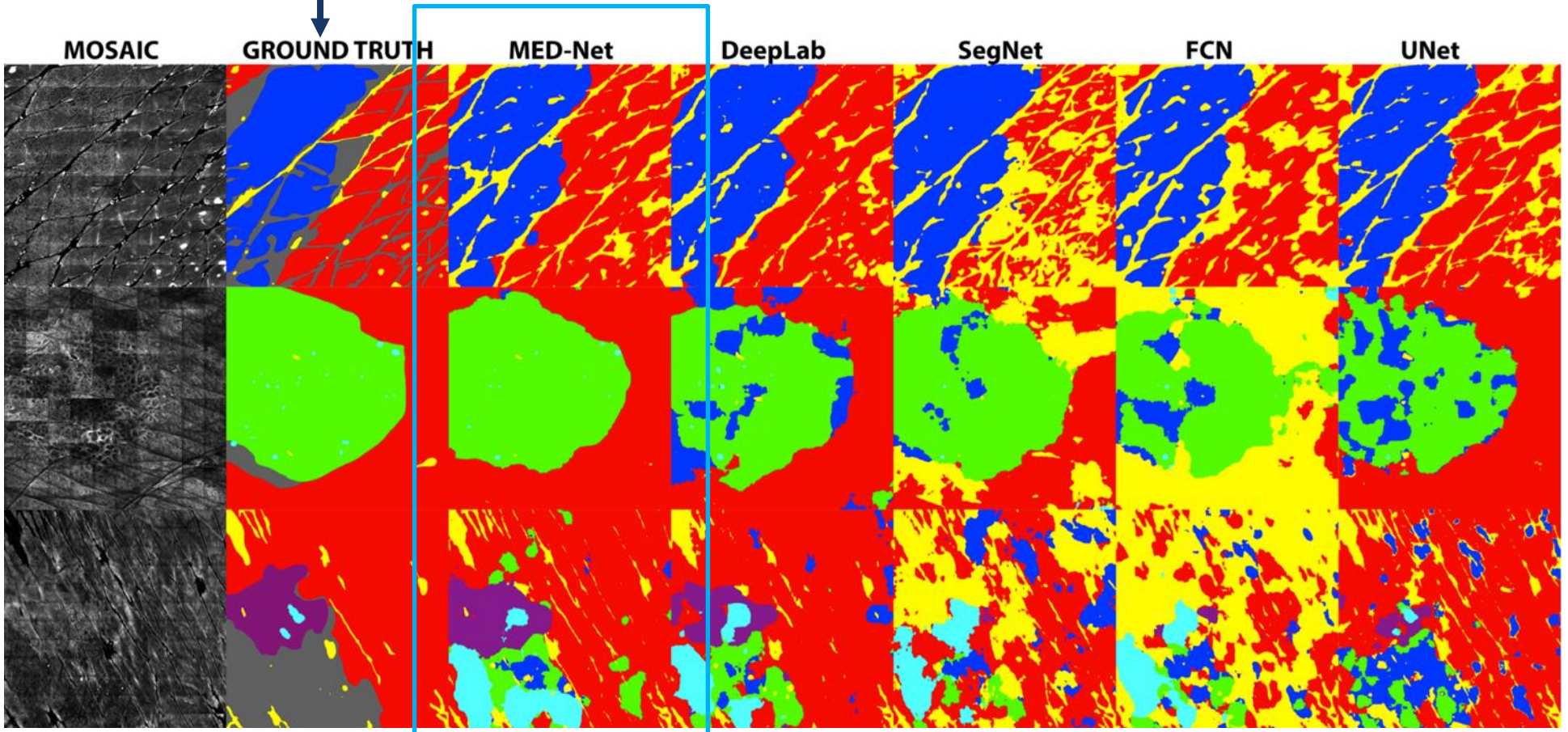
Kose et al., Medical Image Analysis 2019



Solution: Multiresolution Analysis



RESULTS



Benefits:

- Non-intrusive
- More accurate
- More informative
- in-vivo vs. biopsy
- fusing multi-resolution

Non_Lesion Artifact Meshwork Nested/Clod Ring Aspecific Not Labelled (Ignore)