RESEARCH ON TAP

Health Data Science

Wednesday, November 29, 2023

bu.edu/research/events
Agenda

- Welcome Remarks
- Presentations
  - Kevin Lane
  - Neha Gondal
  - Jonathan Huggins
  - Kayoko Shioda
  - Shariq Mohammed
  - Archana Venkataraman
  - Diane Joseph-McCarthy
  - Pawel Przytycki
  - Heather Hsu
  - Judith J. Lok
  - Vijaya B. Kolachalama
- Closing Remarks
Health Data Science

Debbie Cheng

Professor of Biostatistics
Director, Population Health Data Science Program
Department of Biostatistics, School of Public Health
DATA SCIENCE
SHAPING GLOBAL POPULATION HEALTH

The Population Health Data Science Program promotes the development and quality of health data science research. We aim to create new interdisciplinary research synergies between health data science researchers, provide opportunities for education and training in data science, and ultimately facilitate the generation of new insights from large data sources to advance population health worldwide.
Goals for Today

- Strengthen our BU health data science community
- Engage new scholars in health data science
- Foster new collaborations across the university
- Provide seed funding opportunity to advance new research
Population Health Data Science
Seed Funding Awards

- $5000 for 1 year period
- 1-Page Proposal (background, aims, approach)
- PI ROT attendee (faculty, postdoc or student)
- Team must include SPH member
- Deadline: 12/13/23

Details: https://sites.bu.edu/ph-datascience/
Cloud Processing and Machine Learning for Local, National and Global Public Health Research

Kevin Lane, PhD MA
Assistant Professor
Chief Data Officer at BUSPH Center for Climate Health
Department of Environmental Health, Boston University School of Public Health

Boston University Office of Research
Community measurements of aviation-related air pollution near Logan International airport

Sean Mueller
EH Doctoral Candidate

Prasad Patil PhD,
Biostatistics Assistant Professor

ML Source Quantification Methods

PNC
1-sec aggregated to hourly

ML Model
1. Correlated Data
2. Pre-processing
3. Circular Predictors
4. Cross-Validation
5. Missing Data
6. Sample Weighting
7. Outliers
8. ML Method
9. Hyperparameter Tuning

Model Equation
\[ \text{PNC} \sim \text{Flight} + \text{Traffic} + \text{Met} + \text{Temporal} \]

Model Evaluation
\[ \text{RMSE} = \sqrt{\frac{\sum (y - \hat{y})^2}{n}} \quad \text{MAE} = \frac{\sum |y - \hat{y}|}{n} \quad R^2 = \text{(Adj. R-sq.)} \]

Interpretation
Shapley Additive Explanations (SHAP)

Sean Mueller
EH Doctoral Candidate

Prasad Patil PhD,
Biostatistics Assistant Professor

XGBoost Model
\[ \text{RMSE} = 7530 \quad \text{MAE} = 4590 \quad R^2 = 0.64 \]
National Studies of Environmental and Community Social Determinants of Health and Well-Being: Measuring Geospatial Access to Resources

Figure 1: Illustration of Methods for Estimating Access to the Nearest Grocery Store

Potential for Measurement Error in Environmental Variables Indicating Access to Resources, by Distance Measurement Method: Amruta Nori-Sarma, Keith Spangler, Nina Cesare, Biqi Wang, Kimberly Dukes, Kevin Lane

Nori-Sarma et al, JESEE 2022

Spangler et al. (2023), https://doi.org/10.1016/j.sste.2023.100606

Keith Spangler, PhD
Research Scientist EH

Dennis Milechin, P.E., GISP
Scientific Programmer/Analyst IS&T
CHAIR Study Goals: Develop nationwide 1 km x 1 km daily PM2.5 and Temperature models from 2008 – 2020 for all of India.

BUSPH Researchers: provide technical and structural support for researchers across the globe via SCC OnDemand
Modeling Health Behaviors using Social Network Data

Neha Gondal

Assistant Professor
Department of Sociology and Faculty of Computing & Data Sciences
Social Networks and Health

- Health behaviors tend to cluster within social networks. E.g.,
  - Smoking and drinking (McMillan et al. 2018, Cheadle at al 2013))
  - Oral health (Heaton and Gondal 2023; Maupome et al. 2016)
  - Obesity and exercise (Centola and Rijt 2015, Shoham et al 2012)
  - Eating behavior (Fletcher et al 2011)
  - Mental health (Schaefer et al. 2011)

- Two ways to model health outcomes using social network analysis techniques.
ERGMs for Sociometric Data

- What are the micro-level structures that concatenate to produce the overall structure of the network.

- E.g., is there homophily on health behaviors like smoking?
ABMs

- Agent-based models are useful for testing theory when data are unavailable, infeasible, or unethical.

- Agents interact in social networks.

- AMBs can be used to test public health interventions (e.g., should intervention target individual or a network of connected persons) and combined with ERGMs to generate realistic models and outcomes.
Thank You!

gondal@bu.edu
More Data, More Problems?

Jonathan Huggins

Assistant Professor
Department of Mathematics & Statistics, CAS
+ Faculty of Computing & Data Sciences
Models: turning data into knowledge

**Model** = family of possible data-generating processes
- Designed based on scientific / other prior knowledge
- Interpretable

![Phylogenetic tree](https://www.researchgate.net/figure/Phylogenetic-position-of-peacock-with-respect-to-other-bird-genomes-The-phylogenetic_fig2_327750618)
More data, more problems

Models approximate reality
- With only a little data, can’t tell that model is wrong
- More data = more detail

Georges Seurat, Parade de cirque (1889)
Example 1: Phylogenetic tree inference

Problem: If two trees are equally good, then standard methods declares one of them is best

- We show: **bagged posterior** resolves the issue
- Example: infer phylogeny of 13 whale species

  ➢ **Standard method:** × 0 trees the same
  ➢ **Bagged posterior:** ✓ 22 trees the same
Example 2: Cell type discovery

**Problem:** model imperfect + using standard methods: more data ⇒ more cell types spuriously “discovered”

- Solution we developed **provably** solves spurious discovery problem

### Clustering accuracy (F-measure)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>standard approach</td>
<td>0.53</td>
<td>0.48</td>
<td>0.62</td>
<td>0.45</td>
<td>0.54</td>
<td>0.59</td>
</tr>
<tr>
<td>our solution</td>
<td>0.63</td>
<td>0.92</td>
<td>0.94</td>
<td>0.99</td>
<td>0.99</td>
<td>0.98</td>
</tr>
<tr>
<td>alternative solution</td>
<td>0.67</td>
<td>0.88</td>
<td>0.93</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
</tr>
</tbody>
</table>
Data Science to Guide Vaccine Policies
& One Health Approach

Kayoko Shioda
Assistant Professor
Global Health, School of Public Health
Practical & methodological challenges with data and analytic approaches

Develop new data stream and analytic approaches

Robust scientific evidence to guide vaccine policies
Estimating the cumulative incidence of SARS-CoV-2 based on seroprevalence, accounting for antibody waning

How many people have been infected by SARS-CoV-2?

Serosurvey data

Mortality data

Bayesian model

Estimated % population ever infected

New York City

Estimated cumulative incidence

CDC seroprevalence

Estimated seroprevalence

Boston University Office of Research

(Shioda, et al. Epidemiology 2021; Shioda, et al. JID 2022)
What is the **optimal timing** to administer the 2\textsuperscript{nd} dose of mRNA COVID-19 vaccines?

---

**Findings:**

- Delaying the 2\textsuperscript{nd} dose by ~10 days would provide **stronger long-term** protection.
- But longer delay would increase the risk.

---

**Plans:** PCV, rotavirus vaccines

(Shioda, *et al.* medRxiv 2023)
Zoonotic pathogens spread via water, food, soil and animal contacts. How many infections are coming from each pathway?

Transmission dynamic modeling

**Findings:**
If the **foodborne** pathway is reduced by 50%, the total *Campylobacter* infection would decrease by **30%** (95% CrI: 24-36%)

**Future Plans:**
How does this change by weather & climate? What’s the optimal climate-resilient control measure?

(Shioda, *et al*. EHP 2023)
Statistical Methods for Complex Health Data

Shariq Mohammed

Assistant Professor, Department of Biostatistics
Rafik B. Hariri Junior Faculty Fellow
shariqm@bu.edu
Imaging-genomic associations in gliomas

- Detect imaging markers of tumor heterogeneity associated with molecular characteristics
- Using imaging markers to predict clinical outcomes

**Statistical methods**
- Bayesian modeling
- Variable selection
- Functional data analysis

*Mohammed et. al. (2023)*, *Med Image Anal*
*Mohammed et. al. (2022)*, *Am J Neuroradiol*
*Mohammed et. al. (2021)*, *Ann Appl Stat*
Understanding tumor micro-environment

- Offer localized insights on gene expression changes across the tissue
- Quantify interaction between cell types to predict response to treatment

**Statistical methods**
- Bayesian modeling
- Spatial statistics
- Differential geometry

*Krishnan, Mohammed et. al. (2022), Sci Rep*
Cognitive assessment using digital neuropsychological tests

- Quantify performance on digital tests – high frequency pen movement data
- Distinguish participant profiles based on novel performance measures
- Potential non-invasive diagnostic tools for early detection

**Statistical methods**
- Functional data analysis
- Prediction models

Framingham Heart Study Brain Aging Program Pilot Award
Community well-being indices

- Rank zip-codes for overall well-being and its subscales using individual-level survey data
  - Zip-code neighborhood using driving times
Engineering Solutions to Brain Dysfunction

Archana Venkataraman

Associate Professor
Department of Electrical & Computer Engineering
College of Engineering
Bringing AI into Clinical Neuroscience

### Challenges:
- Small Datasets – $N \sim O(10)$ or $N \sim O(100)$
- Patient Heterogeneity
- Interpretability is Paramount
- Noisy/Missing/Inadequate Training Labels
Addressing the Technical (and Data) Challenges

Deep-Generative Hybrids

Biologically-Informed Models

Robust and Adaptable Optimization

Boston University Office of Research
Deep Imaging-Genetics for Mental Health

Ghosal et al., NeuroImage 2021, SPIE 2021, ICLR, 2022
Multimodal Influences of Schizophrenia

Pathway Embedding

- Neuronal Action Potential, Extracellular Matrix Secretion
- Immunological Synapse, Macrophage Derived Foam Cell Differentiation
- Epithelial to Mesenchymal Transition, Cardiac Septum/Atrium Morphogenesis
- Calcium Ion Transport, Calcium Ion into Cytosol
- Histone Methylation, Meiotic Cell Cycle
- Suppression by Virus of Host STAT/Molecular Activity
- Lipid Storage, Sequestering of Triglyceride
- Endoplasmic Reticulum Stress Induces Dephosphorylation
- Cholesterol/Sterol Biosynthetic Process
- Stimulus Involved in Sensory Perception, Pseudouridine Synthesis

Boston University Office of Research
Pandemic Mitigation and Response Strategies

Diane Joseph-McCarthy

Executive Director, Bioengineering Technology & Entrepreneurship Center
Professor of the Practice
Department of Biomedical Engineering, College of Engineering
Map of current readiness with respect to targets, therapeutics, and vaccines

- Enhanced literature review using Large Language Models (LLMs) to reduce timelines
- Novel automated pipeline using prompt engineering strategies to identify drug targets
  - GPT-4 LLM by OpenAI with Zero-Shot-CoT prompting and ensembling top performer
- Applied to SARS-CoV-2 and Nipah Virus
  - Accuracy/F1-score/sensitivity/specificity: 92.87%/88.43%/83.38%/97.82% for SARS-CoV-2
  - 87.40%/73.90%/74.72%/91.36% for Nipah
Identify druggable sites on therapeutic target proteins from structures or homology models

- 10 SARS-CoV-2 targets
- FTMove to assess druggable sites rapidly
- Develop ranking scheme for druggable sites including allosteric/cryptic sites
- AlphaFold for proteins without structure
- Future drug repurposing efforts

AlphaFold Models Sufficient for Hot Spot Mapping

<table>
<thead>
<tr>
<th></th>
<th>Any hot spot, %</th>
<th>Top hot spot, %</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>13+, 50%</td>
</tr>
<tr>
<td>AlphaFold</td>
<td>62</td>
<td>79.0</td>
</tr>
<tr>
<td>X-ray structure</td>
<td>Bound</td>
<td>62</td>
</tr>
<tr>
<td></td>
<td>Unbound</td>
<td>48</td>
</tr>
</tbody>
</table>

Percentages of proteins with any hot spot or top hot spot with 13+ or 16+ probe clusters and at least 50% or 80% overlap with known fragment binder
Beta-Lactoglobulin – AlphaFold Generated Ensemble

- Creating a publicly available resource

@BostonUResearch | #researchontap

M. Lazou, A. Bekar-Cesaretli, D. Kozakov, D. Joseph-McCarthy, S. Vajda
Deciphering the Role of Cell Types (and States) in Disease

Pawel Przytycki
Assistant Professor
Faculty of Computing & Data Sciences
Waddington’s Epigenetic Landscape

- Coding Mutation
- Noncoding Mutation

@BostonUResearch | #researchontap
Multi-modal data

In individual cells…

Satpathy et al, Nat Biotech 2019
Capturing dynamic changes in cell type
Interpreting Genetic Variants

Autism Spectrum Disorder

Alzheimer's Disease

Boston University Office of Research
Optimizing Value-Based Care for Health Equity

Heather Hsu, MD MPH

Assistant Professor
Department of Pediatrics, Chobanian & Avedisian School of Medicine
Value-Based Care

Aims to improve healthcare value by linking provider reimbursement to measures of healthcare quality.

You can’t fix what you don’t measure

Measurement is a cornerstone of value-based care.
MassHealth Value-Based Care Transformation

✓ Medical complexity
✓ Social complexity
BMC Health System & the MassHealth Value-Based Care Transformation

- Assess impact on quality of care & adequacy of payment
- Identify opportunities for payment model improvement
- Address potential measurement bias

40% enrollees are managed by BMC’s WellSense Plan
Opportunities for Value-Based Care Optimization

To improve population health & healthcare equity, we must reverse the trend of:

↓ Bonus payments
↑ Financial penalties
for the healthcare safety-net
Causal Mediation Analysis and the Promise of a Potential COVID-19 Treatment

Judith J. Lok

Associate Professor, Mathematics and Statistics
jjlok@bu.edu
Harvard’s Radcliffe Institute for Advanced Studies 2023-2024

Collaborators:
Victoria Herrera, Nicholas Bosch, Nelson Ruiz-Opazo, Allan Walkey
Boston University Chobanian and Avedisian School of Medicine
Motivation: direct and indirect effects

Direct and indirect effects decompose the effect of a treatment $A$ on outcome $Y$ into:

- Part that is mediated through covariate $M$ (indirect effect).
- Part that is not (direct effect).

Example:

A: blood pressure lowering medication, Outcome $Y$: heart attack.

How much of the effect of the blood pressure lowering medication $A$ is mediated by the effect of $A$ on blood pressure $M$, and how much (if any) works through other pathways?

Example:

In COVID patients admitted to the ICU: What is the indirect effect of a treatment $A$ eliminating Despr+ neutrophil nets $M$ on the final SOFA score $Y$?

SOFA score: describes the degree of multi-organ failure.
Causal diagram: ICU COVID data; exposure: SOFA-1.

\[ M: \text{Despr}+ \text{ neutrophil nets.} \]
\[ \text{SOFA-1: SOFA score at time 1, close to ICU admission.} \]
\[ \text{SOFA-2: SOFA score at time 2, the day before ICU-discharge or death.} \]
Causal diagram: ICU COVID data

\[ A \rightarrow M \rightarrow \text{SOFA-2} \]
\[ \text{SOFA-1} \]

\( M \): Despr+ neutrophil nets.
Putative exposure \( A \): eliminates Despr+ neutrophil nets.
SOFA-1: SOFA score at time 1, close to ICU admission.
SOFA-2: SOFA score at time 2, the day before ICU-discharge or death.
The organic indirect effect of $A$ that eliminates Despr+ neutrophil nets equals $\int_c E[Y|M = 0, C = c, A = 0] f_C(c)dc - EY^{(0)}$, where $C$ is SOFA-1. We can estimate this by

$$\frac{1}{n} \sum_{i=1}^{n} \hat{E}[Y_i|M_i = 0, C_i = c_i, A_i = 0] - \frac{1}{n} \sum_{i=1}^{n} Y_i :$$

the average expected outcome $Y$ had all $M_i$ been equal to 0 (if no direct effect), minus the average expected outcome without any treatment/intervention.

**No on-treatment outcome data needed** for organic indirect effect/pure indirect effect.

Indirect effect estimate: a decrease in time 2 SOFA score of 0.71.

ICU doctors are mostly interested in decreasing the SOFA score of patients with a somewhat high SOFA score at ICU admission (SOFA-1 $\geq$ 2). In those patients, the estimated indirect effect was more promising: 0.98 (95% CI: 0.27-2.13).
Screening Tools for Dementia

Vijaya B. Kolachalama, PhD, FAHA

Associate Professor
Department of Medicine
Department of Computer Science
Faculty of Computing & Data Sciences
Global burden of dementia

- More than 55 million people have dementia worldwide
- Every year, there are nearly 10 million new cases
- In 2019, dementia cost economies globally 1.3 trillion US dollars
- Multiple forms of dementia
  - Alzheimer’s disease (60-70% cases)
  - Lewy body dementia
  - Vascular dementia
  - Frontotemporal degeneration
  - Many more…
Global action plan on the public health response to dementia 2017-2025

1. Dementia as a public health priority
2. Dementia awareness and friendliness
3. Dementia risk reduction
4. Dementia diagnosis, treatment, care and support
5. Support for dementia carers
6. Information systems for dementia
7. Dementia research and innovation
Some hope…

Alzheimer's Drug May Benefit Some Patients, New Data Shows

The drug, lecanemab, made by Eisai and Biogen, also carried risks of brain swelling and bleeding and should be studied further, a report of the findings said.

Alzheimer's drug lecanemab hailed as momentous breakthrough

New Federal Decisions Make Alzheimer's Drug Leqembi Widely Accessible

The F.D.A. gave full approval to the drug, but added a black-box warning about safety risks. Medicare said it would cover most of the high cost.
Our research

- Using routinely collected clinical data, we are developing AI-based methods that can perform differential diagnosis of dementia.
- Such tools can assist neurologists and drug trials.
- We are performing pilot studies to evaluate our tools.

Funding

- National Institutes of Health (NIH)
  - National Heart, Lung, and Blood Institute
  - National Institute of Diabetes and Digestive and Kidney Diseases
  - National Institute on Aging
  - National Cancer Institute
  - American Heart Association
  - Gates Ventures
  - Johnson & Johnson