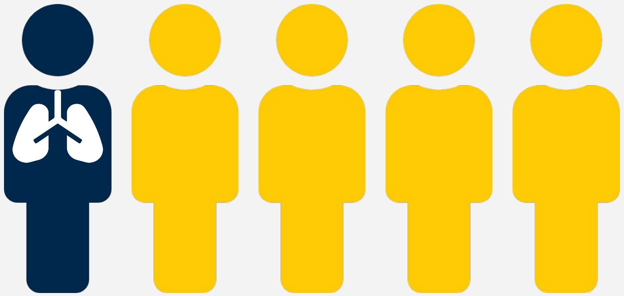
BIOSTATISTICS

Deep Learning for Semi-Competing Risks and Statistics in the Community:

Some Thoughts, Current Work, and Future Directions

Stephen Salerno
PhD Candidate, Biostatistics, University of Michigan

Approximately ***1 in 5*** cancer deaths are attributed to ***lung cancer***



5-year survival rate of ***1 in 5*** (Bade and Cruz, 2020), with prognosis depending on ***individualized risk factors*** (Ashworth et al., 2014)

World Health Organization, International Agency for Research on Cancer, Latest global cancer data: Cancer burden rises to 18.1 million new cases and 9.6 million cancer deaths in 2018.

Motivation comes from the ***Boston Lung Cancer Study*** (BLCS), a large cancer epidemiology cohort examining:

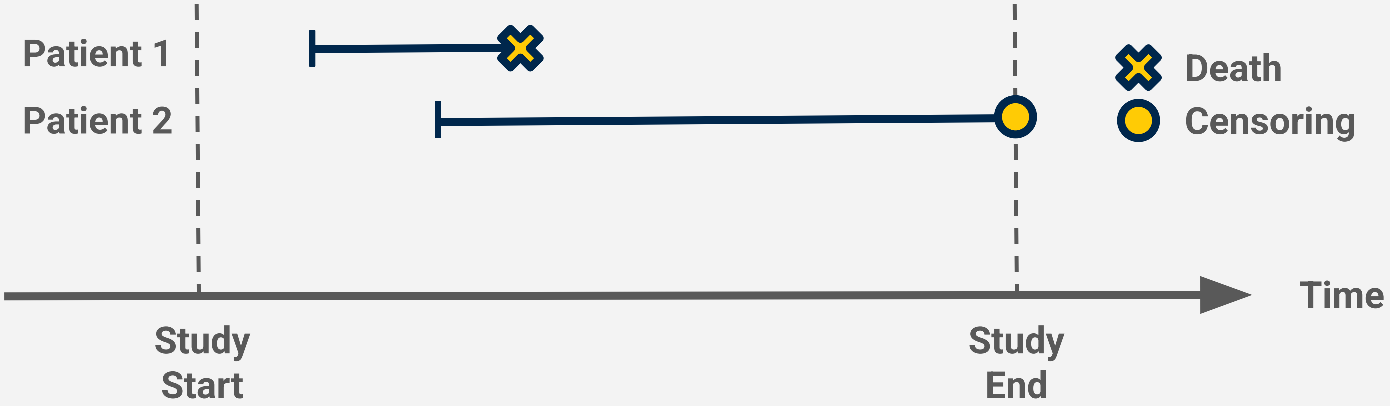
1. Complex mechanisms governing relationships between ***risk factors***

2. Efficacy of ***treatments***

3. Methods for accurately ***predicting survival***

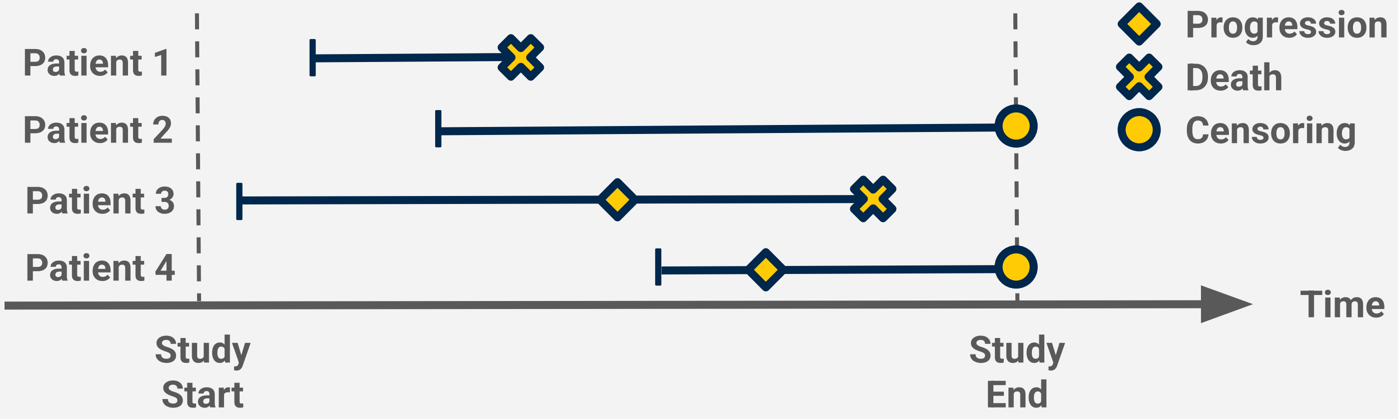
https://www.medicalnewstoday.com/articles/323434

Mortality is often the ***endpoint of choice*** for clinical trials and cohort studies



Survival analysis deals with ***time-to-event*** outcomes which may be ***censored***

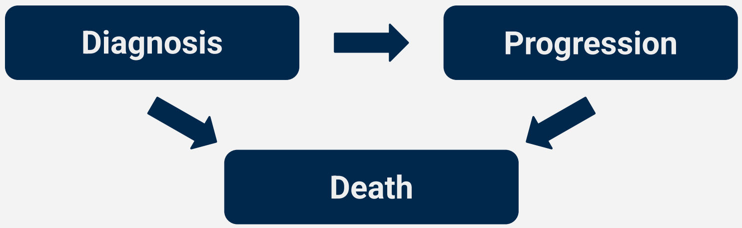
Non-fatal events such as recurrence, progression may occur ***prior to death***



Non-fatal and ***fatal*** events are ***semi-competing*** (Fine et al., 2001)

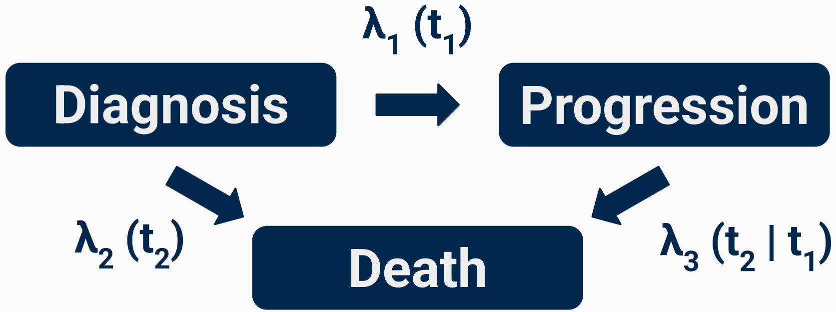
Many studies report on lung cancer ***outcomes***, however:

- ***Progression-free survival*** is used
- Mortality is considered ***without other events***



When progression and death ***do not correlate well***, this leads to ***biased results*** (Jazić et al., 2016)

Consider modeling the ***hazards*** of ***transitioning*** between ***states***:



We can parameterize an ***illness-death model*** as:

$$\lambda_1(t_1 \mid \gamma_i, x_i) = \gamma_i \times \lambda_{01}(t_1) \times \exp\{h_1(x_i)\}; \quad t_1 > 0$$

$$\lambda_2(t_2 \mid \gamma_i, x_i) = \gamma_i \times \lambda_{02}(t_2) \times \exp\{h_2(x_i)\}; \quad t_2 > 0$$

$$\underbrace{\lambda_3(t_2 \mid t_1, \gamma_i, x_i)}_{\text{Hazard Function}} = \underbrace{\gamma_i}_{\text{Frailty}} \times \underbrace{\lambda_{03}(t_2 - t_1)}_{\text{Baseline Hazard}} \times \underbrace{\exp\{h_3(x_i)\}}_{\text{Risk Function}}; \quad t_2 > t_1 > 0$$

Hazard = Frailty × Baseline Hazard × Risk Function

The likelihood for the ***observed data***, D, is given by:

$$L(\psi;\mathcal{D}) = \prod_{i=1}^{n} \int_0^{\infty} \frac{\theta^{-\frac{1}{\theta}}}{\Gamma\left(\frac{1}{\theta}\right)} \times \gamma_i^{\frac{1}{\theta}-1} \times e^{-\frac{\gamma_i}{\theta}} \times \gamma_i^{\delta_{i1}+\delta_{i2}} \times \left[\lambda_{01}(Y_{i1})\boxed{e^{h_1(\mathbf{x}_i)}}\right]^{\delta_{i1}}$$

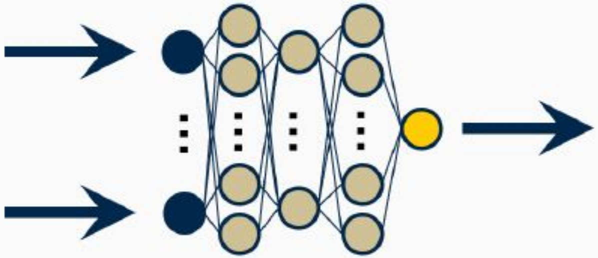
$$\times \left[\lambda_{02}(Y_{i2})\boxed{e^{h_2(\mathbf{x}_i)}}\right]^{(1-\delta_{i1})\delta_{i2}} \times \left[\lambda_{03}(Y_{i2}-Y_{i1})\boxed{e^{h_3(\mathbf{x}_i)}}\right]^{\delta_{i1}\delta_{i2}}$$

$$\times \exp\left\{-\gamma_i\left[\Lambda_{01}(Y_{i1})\boxed{e^{h_1(\mathbf{x}_i)}} + \Lambda_{02}(Y_{i1})\boxed{e^{h_2(\mathbf{x}_i)}} + \delta_{i1}\Lambda_{03}(Y_{i2}-Y_{i1})\boxed{e^{h_3(\mathbf{x}_i)}}\right]\right\} d\gamma_i$$

True ***risk functions*** governed by potentially ***complex relationships***

How to ***estimate/predict*** these risk functions ***more accurately***?

Deep learning has emerged as a powerful tool for ***survival prediction***, but no work has been done on semi-competing outcomes



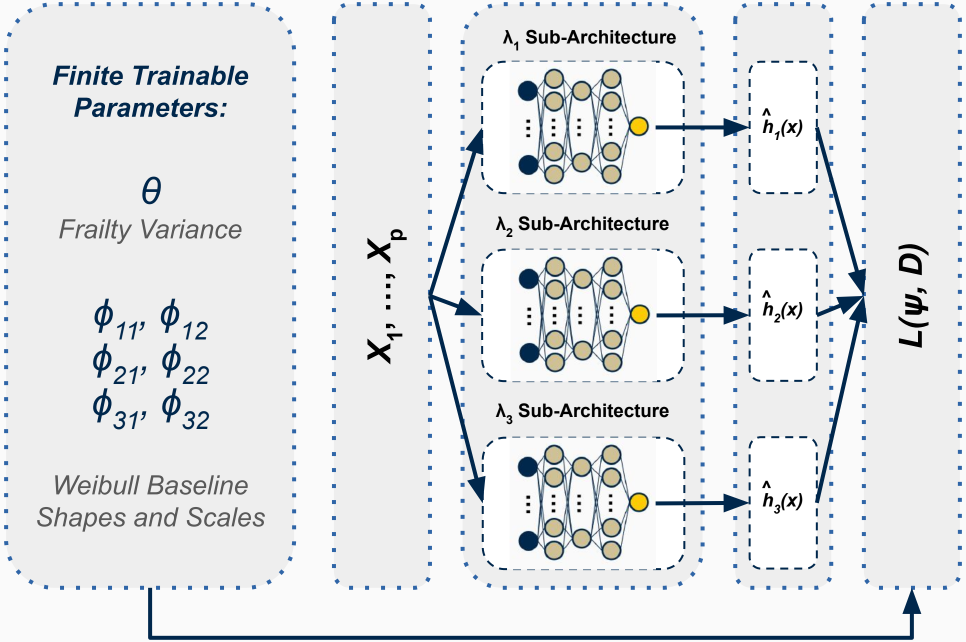
Artificial neural networks try to mirror how the human brain functions, with ***nodes*** connected through ***affine transformations***

We proposes a ***multi-task deep neural network*** with three risk-specific ***sub-networks***, corresponding to each state ***transition***

And a finite set of ***trainable parameters*** to specify the ***frailty variance*** (θ) and the ***baseline hazards*** $(\varphi_{g1}, \varphi_{g2})$:

$$\lambda_{0g}(s) = \varphi_{g1}\varphi_{g2}s^{\varphi_{g2}-1};\ g = 1, 2, 3$$

We propose the use of deep learning to estimate the ***risk functions*** for each ***hazard*** (i.e, state transition)



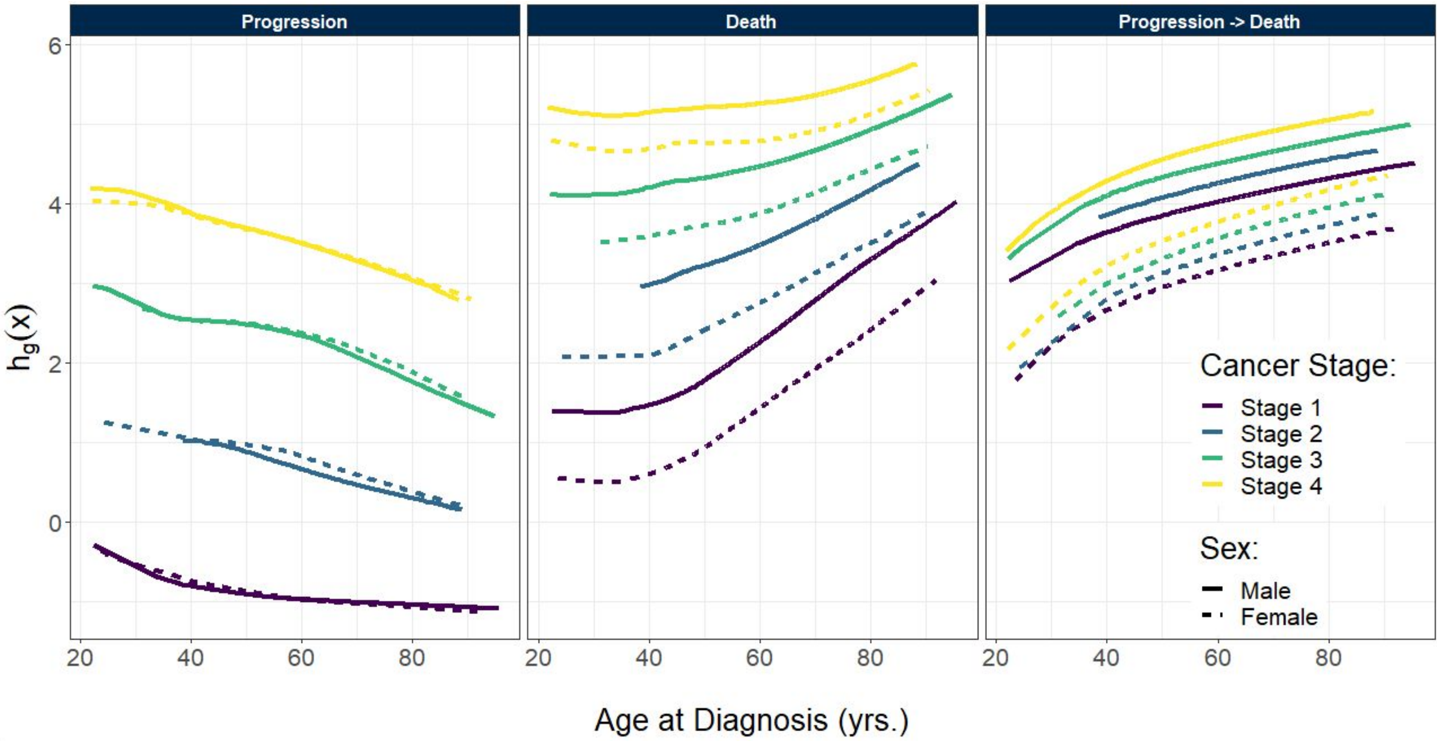
BIOSTATISTICS

In our ***first project***, focused on ***5,296 patients*** with non-small cell lung cancer, diagnosed between June 1983 and October 2021

Investigated time to ***disease progression*** and ***death***, where progression might be censored by death or the study endpoint

	Progression Observed	**Censored**
Death Observed	111 (2%)	1,916 (36%)
Censored	224 (4%)	3,045 (58%)

Log-risk functions of age at diagnosis on each state transition, stratified by sex (solid versus dashed lines) and initial cancer stage (line color); https://www.annualreviews.org/doi/abs/10.1146/annurev-statistics-032921-022127



Estimated ***frailty variance*** (θ) to be 3.15, suggesting ***progression*** is ***correlated with death***

Potential ***nonlinear effects*** of age that differ by event transition, ***interactions*** between cancer stage, and sex

We assumed ***parametric*** baseline hazards, ***optimized directly***

But …

We want a ***non-parametric*** model for both the ***baseline hazards*** and covariate ***risk functions*** to achieve greater flexibility and accuracy

But …

Direct maximization of the likelihood function is ***challenging***

Treating ***frailties*** as ***unobserved***, the ***complete data likelihood*** is:

$$L(\psi; \mathcal{D}, \gamma) = \prod_{i=1}^{n} \frac{\theta^{-\frac{1}{\theta}}}{\Gamma\left(\frac{1}{\theta}\right)} \times \gamma_i^{\frac{1}{\theta}-1} \times e^{-\frac{\gamma_i}{\theta}} \times \gamma_i^{\delta_{i1}+\delta_{i2}} \times \left[\lambda_{01}(Y_{i1}) e^{h_1(\mathbf{x}_i)}\right]^{\delta_{i1}}$$

$$\times \left[\lambda_{02}(Y_{i2}) e^{h_2(\mathbf{x}_i)}\right]^{(1-\delta_{i1})\delta_{i2}} \times \left[\lambda_{03}(Y_{i2} - Y_{i1}) e^{h_3(\mathbf{x}_i)}\right]^{\delta_{i1}\delta_{i2}}$$

$$\times \exp\left\{-\gamma_i \left[\Lambda_{01}(Y_{i1}) e^{h_1(\mathbf{x}_i)} + \Lambda_{02}(Y_{i1}) e^{h_2(\mathbf{x}_i)} + \delta_{i1}\Lambda_{03}(Y_{i2} - Y_{i1}) e^{h_3(\mathbf{x}_i)}\right]\right\}$$

EM algorithm provides a ***numerically stable*** approach to estimation

The ***expected log-complete data likelihood***, or ‘Q’ function, is

$$Q\left(\psi \mid \mathcal{D}, \psi^{(m)}\right) = Q_1 + Q_2 + Q_3 + Q_4,$$

where Q_1, Q_2, Q_3, and Q_4 are ***separable*** w.r.t the ***model parameters***

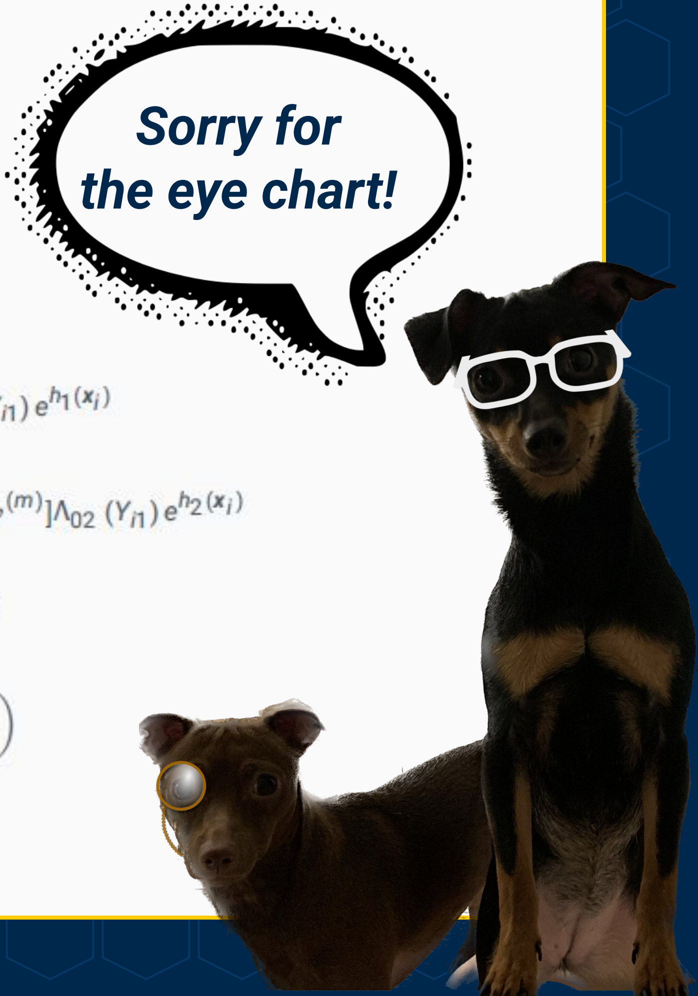
‘Q’ function components:

$$Q_1 = \sum_{i=1}^{n} \delta_{i1}\mathbb{E}[\log(\gamma_i)|\mathcal{D}, \psi^{(m)}] + \delta_{i1}\left\{\log\left[\lambda_{01}\left(Y_{i1}\right)\right] + h_1(\mathbf{x}_i)\right\} - \mathbb{E}[\gamma_i|\mathcal{D}, \psi^{(m)}]\Lambda_{01}\left(Y_{i1}\right)e^{h_1(\mathbf{x}_i)}$$

$$Q_2 = \sum_{i=1}^{n} \delta_{i2}\mathbb{E}[\log(\gamma_i)|\mathcal{D}, \psi^{(m)}] + (1-\delta_{i1})\,\delta_{i2}\left\{\log\left[\lambda_{02}\left(Y_{i2}\right)\right] + h_2(\mathbf{x}_i)\right\} - \mathbb{E}[\gamma_i|\mathcal{D}, \psi^{(m)}]\Lambda_{02}\left(Y_{i1}\right)e^{h_2(\mathbf{x}_i)}$$

$$Q_3 = \sum_{i=1}^{n} \delta_{i1}\delta_{i2}\left\{\log\left[\lambda_{03}\left(Y_{i2}\right)\right] + h_3(\mathbf{x}_i)\right\} - \mathbb{E}[\gamma_i|\mathcal{D}, \psi^{(m)}]\delta_{i1}\left(\Lambda_{03}\left(Y_{i2} - Y_{i1}\right)\right)e^{h_3(\mathbf{x}_i)}$$

$$Q_4 = \sum_{i=1}^{n} -\frac{1}{\theta}\log(\theta) + \left(\frac{1}{\theta} - 1\right)\mathbb{E}[\log(\gamma_i)|\mathcal{D}, \psi^{(m)}] - \frac{1}{\theta}\mathbb{E}[\gamma_i|\mathcal{D}, \psi^{(m)}] - \log\Gamma\left(\frac{1}{\theta}\right)$$



We ***extend*** the EM algorithm to a ***hybrid multi-task deep learning*** approach for semi-competing risk prediction:

E-Step: Frailties ***imputed*** given data, current M-Step estimates

M-Step: Estimate non-parametric ***cumulative baseline hazard*** by non-decreasing step functions and ***frailty variance***

N-Step: Maximize the 'Q' function w.r.t. the ***deep neural network parameters*** for each risk function, $h_g(x_i)$; $g = 1, 2, 3$

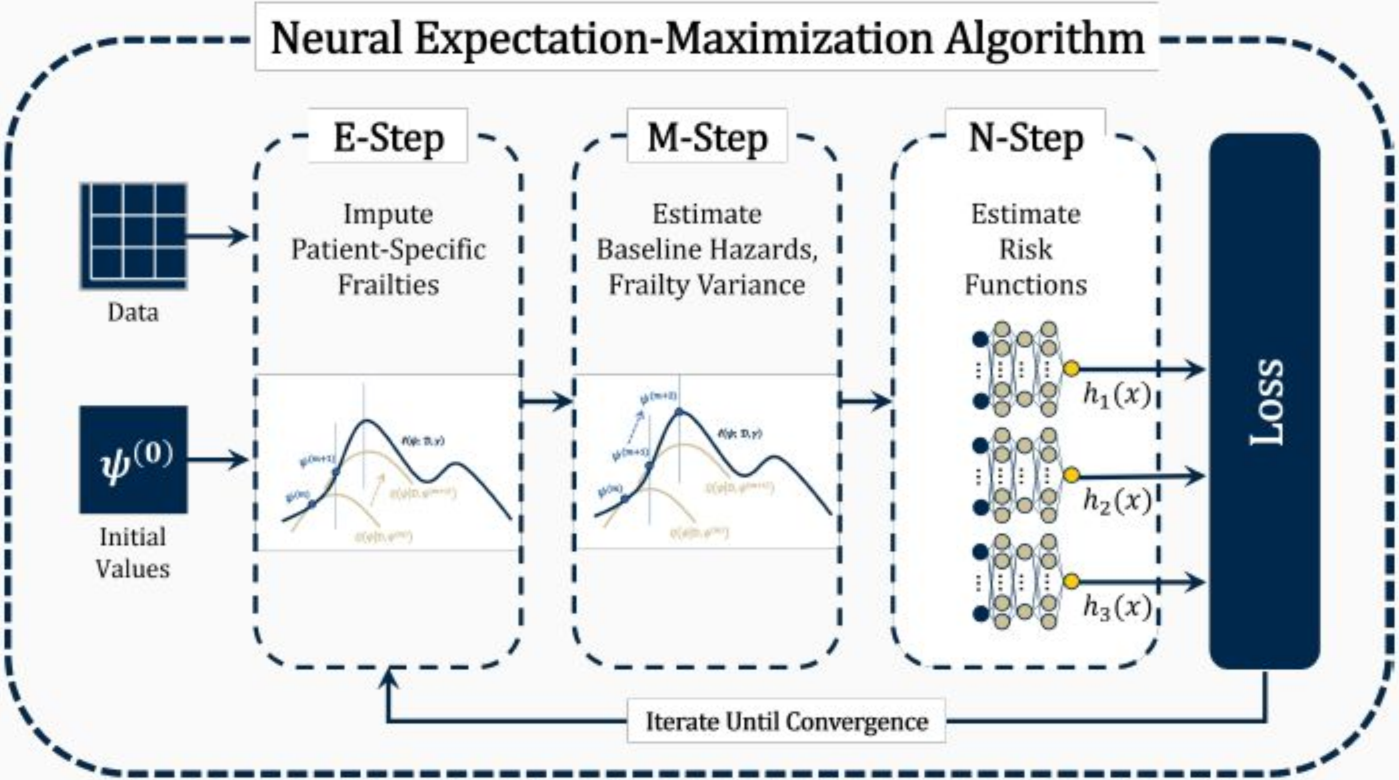


Figure: Overview of our proposed neural expectation-maximization algorithm

Further, no tailored metrics to assess ***predictive accuracy***. We propose a ***bivariate extension*** to the ***Brier Score*** [Brier et al., 1950]

$$
\begin{aligned}
BBS_c(t) = & \frac{\pi_i(t)^2 \cdot \mathbb{I}\{Y_{i1} \le t,\ \delta_{i1} = 1,\ Y_{i1} \le Y_{i2}\}}{\hat{G}_i(Y_{i1})} \\
& + \frac{\pi_i(t)^2 \cdot \mathbb{I}\{Y_{i1} \le t,\ Y_{i2} \le t,\ \delta_{i1} = 0,\ \delta_{i2} = 1,\ Y_{i1} \le Y_{i2}\}}{\hat{G}_i(Y_{i2})} \\
& + \frac{[1 - \pi_i(t)]^2 \cdot \mathbb{I}\{Y_{i1} > t,\ Y_{i2} > t\}}{\hat{G}_i(t)}
\end{aligned}
$$

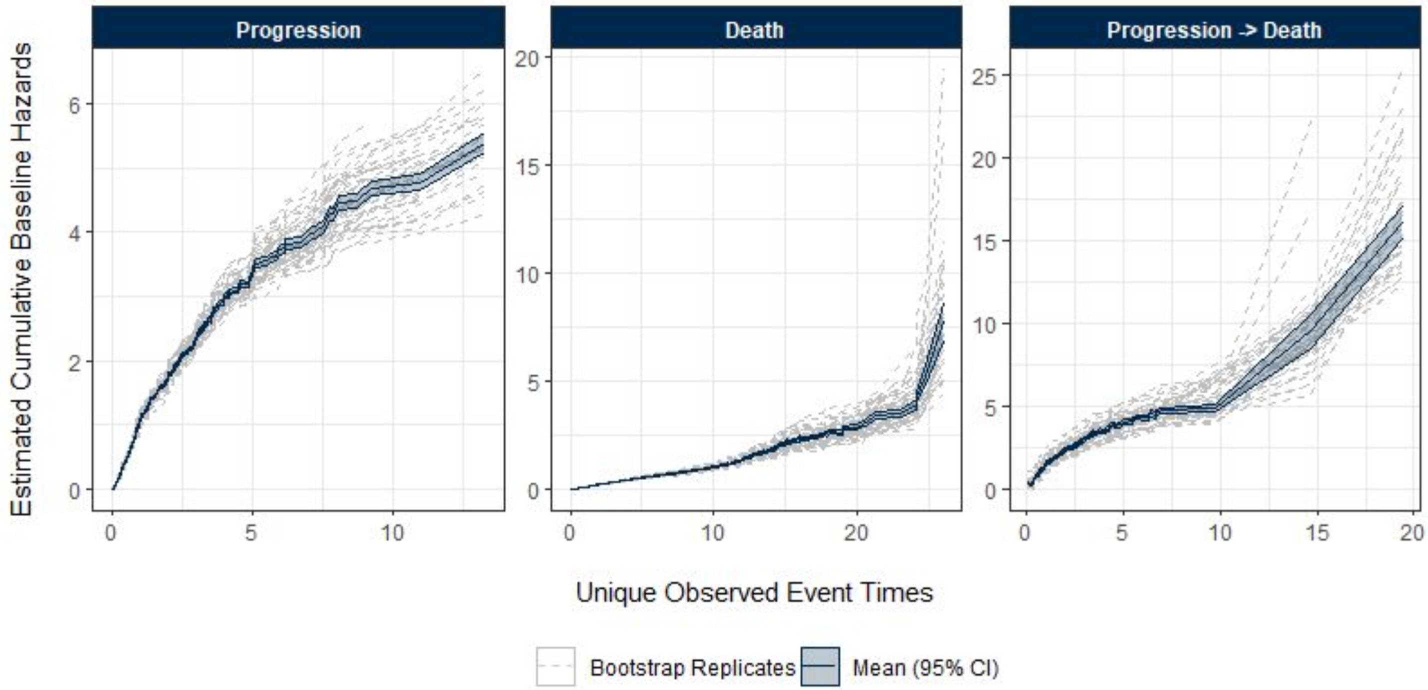
$\pi_i(t)$ is an ***estimate of $S_i(t)$*** $= Pr(T_{i1} > t, T_{i2} > t), G_i(t) = Pr(C_i > t) > 0$

$E[BBS_c(t)]$ equals ***MSE of $\pi_i(t)$***, plus a ***constant*** w.r.t. $\pi_i(t)$

Returning to the Boston Lung Cancer Study, baseline hazards ***highest*** in ***sojourn time*** between progression and death

5-year ***iBBS*** for our method was 0.32 vs. 0.68 from a traditional model, suggesting that a linear model might not be ***as predictive***

Estimated ***cumulative baseline hazards*** and 95% bootstrap CI



BIOSTATISTICS

Our current work considers:

1. Efficiency
2. Interpretability
3. Causality

Some thoughts…

- Non-fatal events impact ***illness trajectories/treatment decisions***
- Interested in ***'true' effect*** of intervention/exposure on progression
- Progression is ***difficult to estimate*** and associations with risk factors/treatments are forgone despite being of clinical interest

Our proposal…

A three-stage approach for estimating the ***causal effect*** of treatment on a ***non-fatal outcome*** in the presence of ***dependent censoring***:

1. ***Derive*** the marginal, non-fatal ***survival function***

2. ***Impute*** our outcome using ***jackknife pseudo-values***

3. ***Estimate*** average treatment effect using ***causal deep learning***

1. ***Derive*** the marginal, non-fatal ***survival function***

$$S_1(t) = [S_*(t)^{1-\theta} - S_2(t)^{1-\theta} + 1]^{1/(1-\theta)}$$

Where $S_*(t)$ is the ***progression-free*** survival function, $S_2(t)$ is the marginal ***fatal survival*** function, and θ is the ***copula parameter***

Based on ***Clayton copula*** with connection to ***previous GFCMM***

1. ***Marginal distribution*** of non-fatal event time as a function of event-free survival and fatal event survival is ***always estimable***

Need to ***estimate θ***, the frailty variance (restricted GFCMM) and equivalent ***dependency parameter*** (Clayton Copula) “ad hoc”

→ Using extended ***concordance estimator*** of Oakes (1982) proposed in Fine et al. (2001)

2. Impute our outcome using ***jackknife pseudo-values***

$$\hat{S}_1^{i}(t) = n\,\hat{S}_1(t) - (n-1)\,\hat{S}_1^{-i}(t)$$

where $\hat{S}_1(t)$ and $\hat{S}_1^{-i}(t)$ are the ***estimates*** of $S_1(t)$ using ***all*** *n* subjects and ***excluding*** the *i*th subject, respectively.

This ***leave-one-out*** estimator for $S_1(t)$ represents the ***contribution*** of the *i*th individual in estimating $E[S_1(t)]$ in the sample

3. ***Estimate*** average treatment effect using ***causal deep learning***

The ***average causal risk difference*** is given by:

$$E[I(T_{i1}^{1} > t)] - E[I(T_{i1}^{0} > t)]$$

An estimate of the ***ATE*** for the average causal risk difference is:

$$A\hat{T}E = 1/n \sum_i \hat{S}_{i1}(t \mid X_i, Z = 1) - \hat{S}_{i1}(t \mid X_i, Z = 0)$$

For a causal variable of interest, ***Z***

3. Estimate average treatment effect using ***causal deep learning***

Predict ***potential outcomes*** and estimate the survival ATE by modeling pseudo-values conditional on ***risk factors*** in ***DNN***

Network output optimized under the common ***binary cross-entropy*** loss function

→ Faster ***learning rate/convergence*** than MSE due to ***steeper gradient*** when the predicted output is far from the true output

- ***Survival probabilities*** are more natural to interpret than hazards

- Because we have a ***consistent estimate*** of $S_1(t)$

1. $S_1^i(t)$ is ***approximately independent*** of $S_1^j(t)$ for $i \neq j$ as $n \to \infty$
2. $\lim_{n\to\infty} E[S_1^i(t) \mid Z_i, X_i] = S_1(t \mid Z, X)$

- With (1) and (2), these ***pseudo-values*** can be used as a ***response variables*** in our deep learning framework

- Imputed outcome ***more efficient*** for deep learning

Preliminary Simulations:

Example comparison ATE calculation for parametric vs. proposed method with ***linear*** vs. ***non-linear*** risks generated

Risk Function	Empirical ATE	Parametric		Proposed	
		Bias	**MSE**	**Bias**	**MSE**
Linear	0.3100	0.0183	0.0003	0.0157	0.0002
Non-Linear	0.3283	0.0590	0.0035	0.0343	0.0012

Preliminary Analysis of the BLCS Study:

Considered ***n = 4,700*** patients in the BLCS with ***NSCLC*** and ***stage 1-3a*** (considered operable) at diagnosis

Estimated ***average difference*** in probability of 5-year time-to-progression between ***first-line treatment*** options

→ ***Surgical resection*** vs. ***chemotherapy or radiation***, adjusting for socio-demographic and genetic risk factors

Preliminary Analysis of the BLCS Study:

Estimated survival conditional ***ATE of 0.156***, suggesting potential ***longer-term benefit*** of surgery, consistent with current literature

Estimated ***copula dependence*** between progression and survival to be $\boldsymbol{\theta = 4.38}$, corresponding to a ***Kendall's*** $\boldsymbol{\tau = 0.6865}$

Overall…

Clinical motivation for this work comes from the ***semi-competing*** nature of patient ***health events*** in the ***Boston Lung Cancer Study***

Statistical motivation comes from ***interest in methods*** for high-dimensional survival analysis, deep learning, causal inference

Many ***exciting opportunities*** for ***future development***

Some next steps…

Prediction intervals to quantify uncertainty

Extending these methods and our analysis to ***incorporate***

- ***Longitudinal*** risk factors over disease course
- ***High-dimensional*** covariate, such as ***CT imaging features***
- ***More comprehensive*** events (e.g. second primaries)


"And now for something completely different."
Monty Python

stat com

Data science is ubiquitous in big business and academic research

What about …

Ann Arbor Area Community Foundation

A ***community foundation*** allocating $18 million to improve ***quality of life*** services for seniors?

food for thought

A ***mobile food pantry*** determining optimal ***service locations*** in low-income areas?

ozone HOUSE safe place. real support.

A ***youth center*** predicting ***crisis calls*** from high-risk, runaway, and homeless youth?

Community organizations also stand to benefit from statistical insight …

… they may lack the **time**, **resources**, or **knowledge** to collect and analyze data

stat com | Statistics in the Community

stat com

A **community outreach program** that offers the expertise of graduate students, **free of charge**, to non-profit governmental and community partners in the areas of **data organization**, **analysis**, and **interpretation**.

2001: STATCOM founded at Purdue

2006: ASA Member Initiatives Grant

2006: 10 chapters chartered, including Michigan!

2016: Most defunct, Michigan growing

2023: Michigan STATCOM is thriving

STATCOM'S ENGAGEMENT

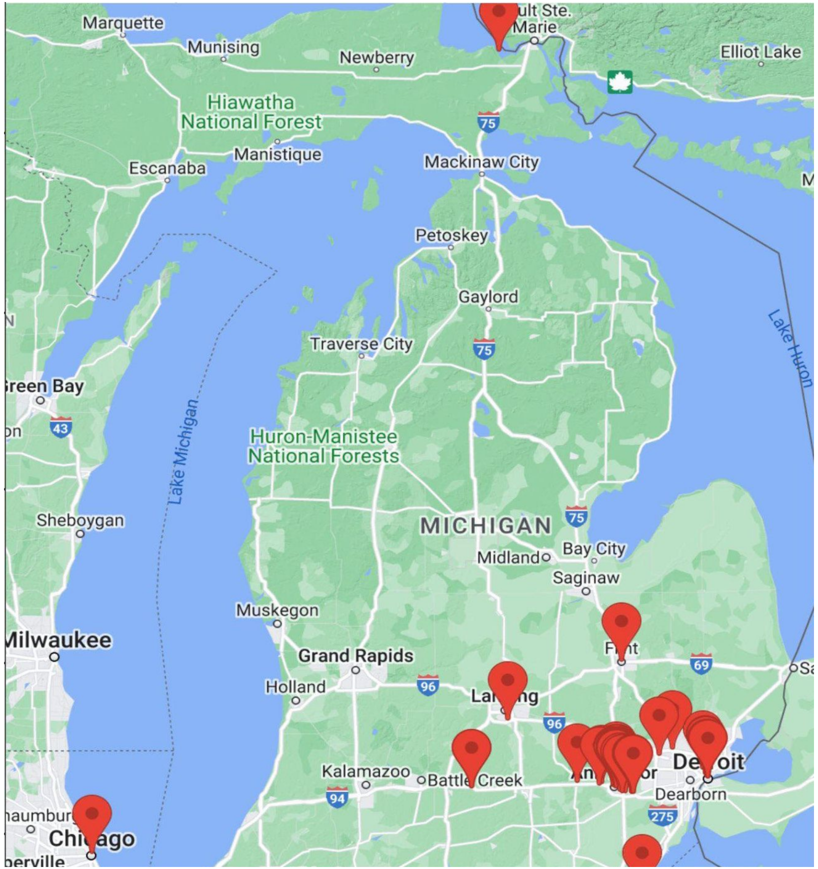
TAKEN FROM WHEN I STARTED IN 2016 UNTIL TODAY

62
PROJECTS COMPLETED

255
MEMBERS ENGAGED

10
DEPARTMENTS REPRESENTED

statcom



HANNAN CENTER
LEARN | CONNECT | CREATE

Ann Arbor Area
Community Foundation

CATHOLIC
SOCIAL
SERVICES
OF WASHTENAW COUNTY

OSHER
LIFELONG
LEARNING
INSTITUTE

Stand WITH Trans
ADVOCATE
CELEBRATE
EDUCATE

food
for thought

ACC&D
Alliance for Contraception
in Cats & Dogs

MICHIGAN
BROADBAND
JUMPSTART

STARR
COMMONWEALTH

food gatherers
fighting hunger where we live

WASHTENAW COUNTY
WATER RESOURCES COMMISSIONER

M
PUBLIC
HEALTH

Area
Agency on
Aging 1-B
Answers you can trust

DCAN
DETROIT
COLLEGE
ACCESS
NETWORK

Genesee County
Health Department
Your Health. Our Work.

ozone HOUSE
safe place. real support.

the
CHILDREN'S
Center

60+ Survey of Washtenaw County

statcom

The AAACF received $18 million dollars to help improve the quality of life of older adults aging in place within the county, especially those with lower life expectancy and socioeconomic status

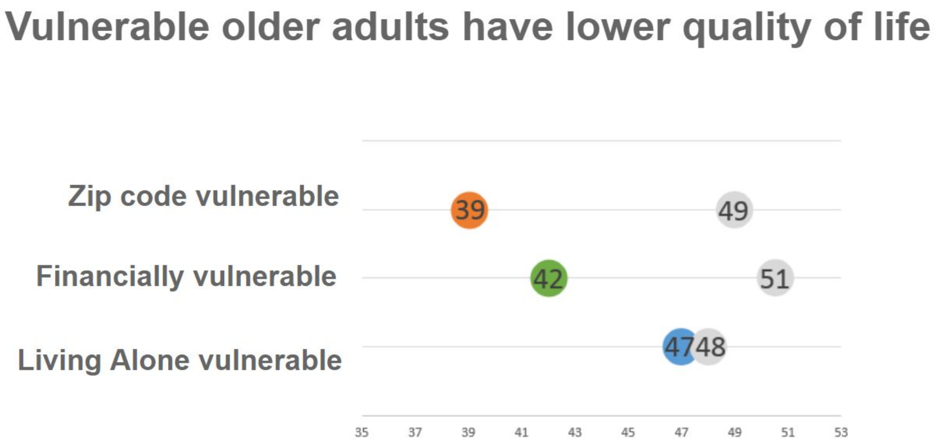
Project Overview

AAACF partnered with STATCOM and the Ginsberg Center at UM to write, distribute, and analyze a survey assessing the quality of life of older adults (60+) aging in place in Washtenaw County

Zip Code	Financial	Living Alone
48197 48198	Rent Assistance Medicaid Not Enough Money	Living Alone

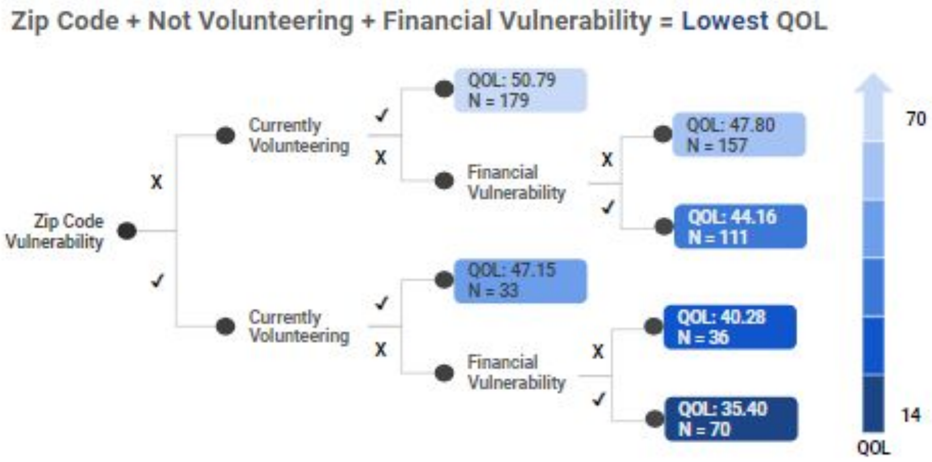
Why This Matters

Results are being used to allocate funding for services to older adult populations within the county, and community reports have informed decision making for local governments in smaller regions of the county



Our Results

Tree-based modeling showed that respondents living in vulnerable zip codes, who were not volunteering, and who were financially vulnerable comprised the lowest average quality of life group.



Why was this a good project?

1. STATCOM involved from conception to conclusion
2. Leveraged community partnerships for success
3. Findings were used to inform policy decisions
4. Students directly impacted this population

statcom

Our Current Projects (mine in yellow):

- Starr Commonwealth
- The Konnection
- Stand with Trans
- Michigan Center for Youth Justice
- Poverty Solutions

It’s **incredibly beneficial** having highly trained & knowledgeable STATCOM representatives help us analyze our data on recent graduate outcomes. They helped us get to the **‘aha’ moment** of understanding the story about what the data was telling us. I am so **thankful for the help** I received at STATCOM.

Shelagh Saenz,
STATCOM Community Partner

stat com

What has contributed to our growth?

- Departmental Support/Encouragement
- Dedicated Student Volunteers
- Strong University Partnerships

statcom

Our department supports experiential learning

STATCOM is a natural way for students to apply what they have learned in class while working in a team-based paradigm

It develops important skills all students should have when pursuing a graduate degree while giving back to the community

STATCOM provides a **great opportunity** to use data to help organizations focused on **public good**. It allows me to use my skills to help these organizations run more efficiently and **answer important questions**.

Tim NeCamp
Former STATCOM President

University Partnerships

SCHOOL OF PUBLIC HEALTH
BIOSTATISTICS
UNIVERSITY OF MICHIGAN

STUDENT LIFE
EDWARD GINSBERG CENTER
UNIVERSITY OF MICHIGAN

CEDER CENTER FOR EDUCATION DESIGN, EVALUATION, & RESEARCH
UNIVERSITY OF MICHIGAN

RACKHAM
STUDENT GOVERNMENT
UNIVERSITY OF MICHIGAN

SCHOOL OF
PUBLIC HEALTH
UNIVERSITY OF MICHIGAN

MIDAS MICHIGAN INSTITUTE FOR DATA SCIENCE
UNIVERSITY OF MICHIGAN

LSA COLLEGE OF LITERATURE, SCIENCE, AND THE ARTS
UNIVERSITY OF MICHIGAN

SCHOOL OF
SOCIAL WORK
UNIVERSITY OF MICHIGAN

Broader Initiatives

stat com

National Outreach

AMSTATNEWS

The Membership Magazine of the American Statistical Association

HOME ABOUT PDF ARCHIVES ADVERTISE STATISTICIANS IN HISTORY PODCAST

Home » Additional Features

STATCOM: Revitalization of Statistica Community Service at Universities

1 APRIL 2018 2,938 VIEWS ONE COMMENT

Evan Reynolds and Timothy NeCamp

Universities provide valuable resources for providing pro-bono statistical services, incluc connections and many statistically inclined students eager to apply their skills. STATCO of Michigan is leveraging these resources to increase its benefit to both the community a While several universities with graduate statistics and biostatistics programs founded ST the 2000s, much of their activity has declined since then. In contrast, STATCOM is large than ever at the University of Michigan.

NCSU-STATCOM

People

STATCOM
Statistics in the Community
At Columbia University's Mailman School of Public Health

Our Mission

y outreach program provided to New York City by graduate students in the Department of Biostatistics at Mailman e offer professional statistical consulting, free of charge, to non-profit community and local governmental groups in the areas of data organization, analysis, and interpretation.

Our History

r was founded in 2001 at Purdue University's Department of Statistics with support from a Member Initiatives Grant ical Association (ASA). Since then a network of STATCOM programs has been established and active chapters are nd Biostatistics departments throughout the country. The Columbia University chapter was founded in 2021 by laddie Stoms, Charly Fowler, Steven Lawrence, and Muhire Kwizera with the help of faculty advisor Cody Chiuzan.

+ Add to My Program Thu, 8/6/2020, 3:00 PM - 4:50 PM

566 * !

Teaching Data Science for Good: How University-Based Initiatives Are Shaping Future Statisticians'

Section on Statistical Consulting, Social Statistics Section, Section on Statistics and Data Science Educ

Organizer(s): Emily L Morris, Department of Biostatistics, University of Michigan

Chair(s): Emily L Morris, Department of Biostatistics, University of Michigan

3:05 PM Data for Good in Your Neighborhood: How Graduate Students and Local Communities Benefit from Collaborative Partnerships Presentation
Stephen Salerno, University of Michigan

3:35 PM "Data for Good" at Columbia's Data Science Institute
Tian Zheng, Columbia University

4:05 PM Data Science Education as an Economic and Public Health Intervention: How (Bio)Statisticians Can Lead Change in the World
Jeff Leek, Johns Hopkins Bloomberg School of Public Health

4:35 PM Floor Discussion

Fill out evaluation Thu, Jun 4, 1:20 PM - 2:55 PM

Organizer(s): Leah Jager, Johns Hopkins Bloomberg School of Public Health

Chair(s): Leah Jager, Johns Hopkins Bloomberg School of Public Health

1:25 PM Can Data Science Education Be Used as a Tool for Upward Mobility?
Presentation Aboozar Hadavand, Johns Hopkins University, Bloomberg School of Public Health

1:55 PM Incorporating Community-Based Learning Into the Classroom
Presentation Lynne Steuerle Schofield, Swarthmore College

2:25 PM Statistics in the Community: Community-University Partnerships Fostering Data Science Education
Presentation Stephen Salerno, Department of Biostatistics, University of Michigan

stat com

Bloomberg D4GX (2018)

Data for Good In Your Neighborhood: A case study on how data can benefit your local community

Timothy NeCamp
University of Michigan
Ann Arbor, MI
tnecamp@umich.edu

Evan Reynolds
University of Michigan
Ann Arbor, MI
evanlr@umich.edu

Emily Morris
University of Michigan
Ann Arbor, MI
emorrisl@umich.edu

Stephen Salerno
University of Michigan
Ann Arbor, MI
salernos@umich.edu

ABSTRACT

Data are beneficial resources for big businesses, national and state governments, and large nonprofits. In this article we describe how the benefits of data can be similarly reaped by local community organizations. We exemplar this benefit through Statistics in the Community (STATCOM) at the University of Michigan. STATCOM is an organization which connects statistics graduate students to local community organizations to provide free assistance around data organization, analysis, and interpretation. We describe three specific projects where STATCOM members partnered with a food assistance program, a crisis center, and a community foundation. For each project, we provide an overview of the community partner, the data available, the statistical methods used, the findings, and discuss how the analyses impacted the partner and the local community.

1. INTRODUCTION

A food assistance program needs to determine the optimal placement of mobile food pantries in low-income and underserved urban areas. A crisis call center hopes to accurately predict the number of calls from runaway, homeless, and high-risk youth in order to appropriately staff its counselors. A local community foundation wants to decide how to best allocate 18 million dollars in funds to improve quality of life services for their senior population.

Each of these organizations strives to provide public good directly to its surrounding communities but is hindered by these needs. Often, local community efforts hope to answer similar questions but do not have the means to do so [12]. What they sometimes do have, however, are data. The food assistance program, Food for Thought Toledo (FFT) [9], has data on levels of poverty, nutritional stress, food deserts, and other key neighborhood characteristics. The crisis center, Ozone House of Ann Arbor [16], has daily call logs for the past 10 years. The local community foundation, Ann Arbor Area Community Foundation (AAACF) [1] has responses to a survey instrument designed to assess quality of life in older adults.

In many academic fields, answering a question is synonymous with collecting and analyzing data. It is then through principled statistical methods that one can utilize the power of these data to make inferences and predictions. For local community organizations, data can be similarly harnessed to evaluate need, allocate resources, broaden impact, and assess efficacy. Unfortunately, the power of data often remains untapped due to a lack of resources and statistical support, especially at smaller local community organizations. Statistics in the Community (STATCOM) at the University of Michigan [17] aims to fix this problem by offering, free of charge, the expertise of statistics graduate students to nonprofit governmental and local community organizations in the areas of data organization, analysis, and interpretation. This includes schools, government agencies, health centers, homeless shelters, advocacy groups, libraries, adult learning centers, and other agencies and organizations.

STATCOM, which is directed and staffed primarily by graduate students, separates itself from other data for good organizations by working with small local community partners. Many other organizations, such as Data Science for Social Good [4], Statistics without Borders [18], DataKind [5], tend to work with either large institutions or international/national organizations (e.g., American Red Cross, UNICEF in Sierra Leone, and The Municipality of Rotterdam). STATCOM is currently undertaking 10 projects across Southeast Michigan and Northern Ohio, including organizations such as The Children's Center of Detroit [19], Genesee County Health Department [11], Detroit College Access Network [6], and Michigan Broadband Cooperative [13]. By bringing statistics into our community, STATCOM directly touches the people it helps, leaving the greatest impact on those in its direct proximity.

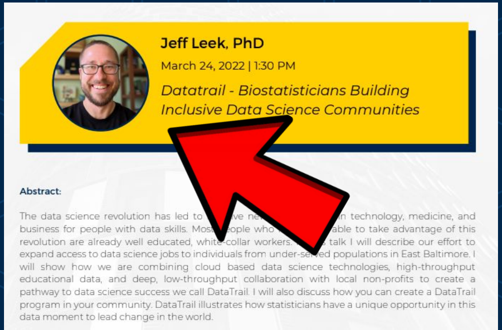
In this paper, we describe three specific STATCOM projects working with (1) FFT of Toledo, Ohio, (2) Ozone House of Ann Arbor, Michigan, and (3) the AAACF of Washtenaw County, Michigan. We highlight the problems at hand, the data being used to solve those problems, and interesting results and findings. We also emphasize the community impact and how each community partner is using STATCOM's work.

Bloomberg Data for Good Exchange Conference.
16-Sep-2018, New York City, NY, USA.

statcom

Data for Public Good Symposium

STATCOM, CTAC, CEDER, and MIDAS host a symposium to showcase research efforts and community-based partnerships that improve humanity by using data for good


Jeff Leek, PhD

March 24, 2022 | 1:30 PM

Datatrail - Biostatisticians Building Inclusive Data Science Communities

Abstract:

The data science revolution has led to [illegible]ive ne[illegible] in technology, medicine, and business for people with data skills. Most [illegible]eople who [illegible] able to take advantage of this revolution are already well educated, white-collar workers. [illegible] talk I will describe our effort to expand access to data science jobs to individuals from under-se[illegible]ed populations in East Baltimore. I will show how we are combining cloud based data science technologies, high-throughput educational data, and deep, low-throughput collaboration with local non-profits to create a pathway to data science success we call DataTrail. I will also discuss how you can create a DataTrail program in your community. DataTrail illustrates how statisticians have a unique opportunity in this data moment to lead change in the world.

stat com

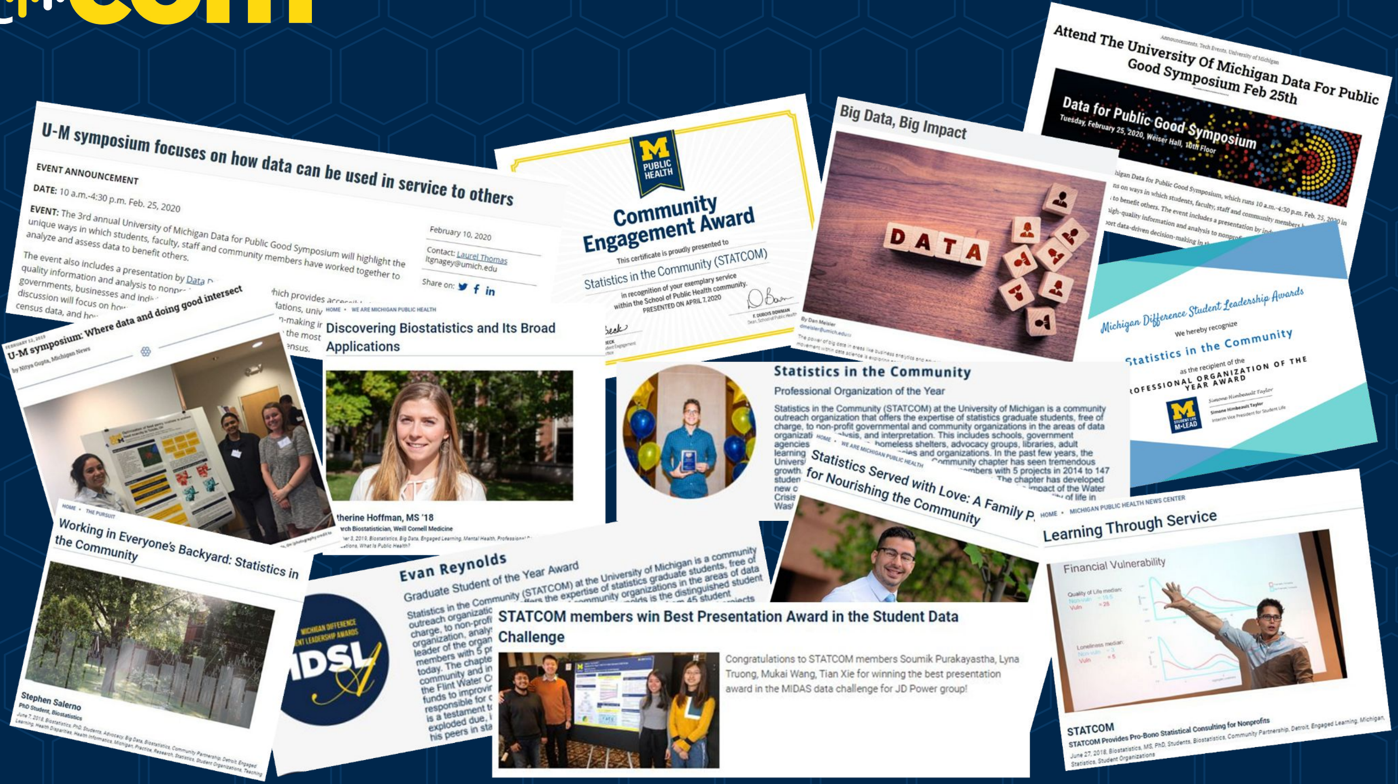
STATCOM hosts a series of workshops in R syntax, exploratory data analysis, data visualization, and reproducible research with RMarkdown/Projects:

stat com

Current Priorities/Goals:

- Engage more community partners and students
- Create sustainable processes and documentation
- Continue national outreach efforts
- Fund a student coordinator in our department

stat com


U-M symposium focuses on how data can be used in service to others
EVENT ANNOUNCEMENT
DATE: 10 a.m.-4:30 p.m. Feb. 25, 2020
EVENT: The 3rd annual University of Michigan Data for Public Good Symposium will highlight the unique ways in which students, faculty, staff and community members have worked together to analyze and assess data to benefit others.
February 10, 2020
Contact: Laurel Thomas
ltgnagey@umich.edu
Share on:
M PUBLIC HEALTH
Community Engagement Award
This certificate is proudly presented to
Statistics in the Community (STATCOM)
in recognition of your exemplary service within the School of Public Health community.
PRESENTED ON APRIL 7, 2020
Big Data, Big Impact
DATA
Attend The University Of Michigan Data For Public Good Symposium Feb 25th
Data for Public Good Symposium
Tuesday, February 25, 2020, Weiser Hall, 10th Floor
Michigan Difference Student Leadership Awards
We hereby recognize
Statistics in the Community
as the recipient of the
PROFESSIONAL ORGANIZATION OF THE YEAR AWARD
U-M symposium: Where data and doing good intersect
Discovering Biostatistics and Its Broad Applications
Statistics in the Community
Professional Organization of the Year
Working in Everyone's Backyard: Statistics in the Community
Stephen Salerno
PhD Student, Biostatistics
Evan Reynolds
Graduate Student of the Year Award
Statistics Served with Love: A Family P...
for Nourishing the Community
Learning Through Service
Financial Vulnerability
STATCOM
STATCOM Provides Pro-Bono Statistical Consulting for Nonprofits
STATCOM members win Best Presentation Award in the Student Data Challenge
Congratulations to STATCOM members Soumik Purakayastha, Lyna Truong, Mukai Wang, Tian Xie for winning the best presentation award in the MIDAS data challenge for JD Power group!

Many professional thanks …








… and personal ones, too!

BIOSTATISTICS

statcom

Contacting Me:

salernos@umich.edu

arxiv.org/abs/2212.12028

@StephenSalerno4

Contacting STATCOM:

statcom-um@umich.edu

sph.umich.edu/biostat/statcom/

@statcom_um