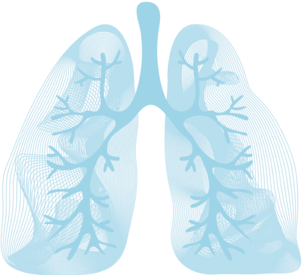
# Causal Framework for Subgroup Treatment Evaluation using Multivariate Generalized Mixed Effect Models with Longitudinal Data

Yizhen Xu
Advisor: Scott Zeger
yxu143@jhu.edu

# Motivation - Scleroderma

Scleroderma is a chronic autoimmune disease marked by hardening of the skin and internal organs. The severity of the illness varies and might cause pain, stiffness, and exhaustion. For the approximately 300,000 people in the United States diagnosed with scleroderma, there are no approved treatments currently. The two main subtypes of systematic sclerosis are limited cutaneous SSc (slow progression over a long time) and diffuse cutaneous SSc (affects larger skin area and progresses quickly to organs).

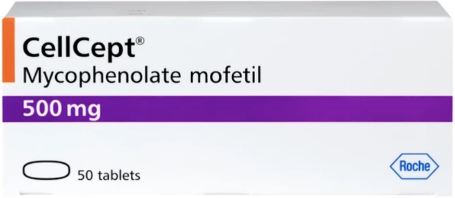


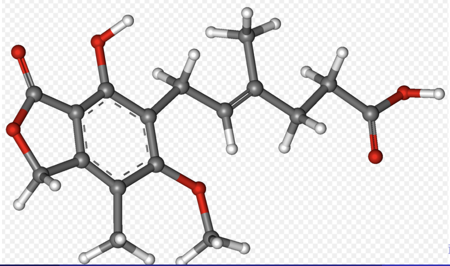
Sources: https://labblog.uofmhealth.org/industry-dx/systemic-scleroderma-treatments-where-are-we-now and https://sclerodermanews.com/social-clips/lung-involvement-fibrosis-scleroderma/

# Motivation - Scleroderma

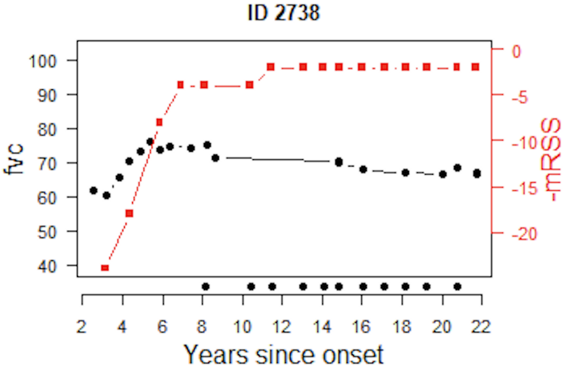
## Immunosuppressant medication for scleroderma

CellCept (mycophenolate mofetil) is an oral medication developed by Genentech (a member of the Roche group) that can improve lung function in people with scleroderma. In Scleroderma Lung Study II, MMF resulted in improvements in the modified Rodnan skin score (mRSS) among diffuse patients over 24 months.



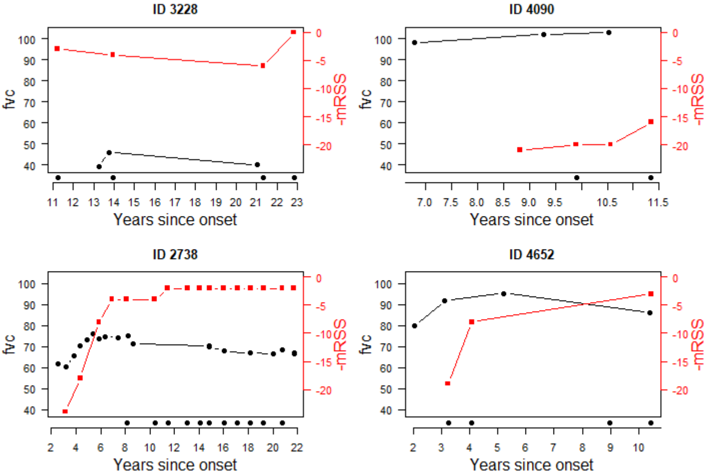


# Motivation - Scleroderma



- Data source: The Johns Hopkins Precision Medicine Analytics Platform (PMAP) Registry
- People who had the first treatment between 2010-01-01 and 2020-01-01, and the first medical visit happened within 5 years of disease onset

# Motivation - Scleroderma



# Motivation

$$\begin{tikzcd}
Y_{i0} \arrow[r] \arrow[dr] & Y_{i1} \arrow[r] \arrow[dr] & Y_{i2} \\
 & A_{i1} \arrow[u] \arrow[r] & A_{i2} \arrow[u]
\end{tikzcd}$$

What is the causal effect of $\bar{a}_2 = (1, 1)$ versus $\bar{a}'_2 = (0, 0)$?

For patient $i$ at day $t$,

- $Y_{it}$: outcomes
- $A_{it}$: binary treatment status

# Challenges

Patient heterogeneity regarding treatment assignment and biomarker progression is the primary factor for treatment effect evaluation.

- Clinician's treatment decisions are highly related to unmeasured factors
  e.g. a physician may not ventilate a patient who is too physically weak
- Unmeasured factors may strongly influence biomarkers progression. Standard approaches in longitudinal causal inference may lead to biased estimates. (Yang and Lok, 2018)

# Challenges

Patient heterogeneity regarding treatment assignment and biomarker progression is the primary factor for treatment effect evaluation.

- Clinician's treatment decisions are highly related to unmeasured factors
- Unmeasured factors may strongly influence biomarkers progression. Standard approaches in longitudinal causal inference may lead to biased estimates. (Yang and Lok, 2018)

$$\begin{tikzcd}
 & & b_i^Y \arrow[dl] \arrow[d] \arrow[dr] & \\
Y_{i0} \arrow[r] \arrow[dr] & Y_{i1} \arrow[r] \arrow[dr] & Y_{i2} \arrow[r] \arrow[dr] & Y_{i3} \\
 & A_{i1} \arrow[r] \arrow[u] & A_{i2} \arrow[r] \arrow[u] & A_{i3} \arrow[u] \\
 & & b_i^A \arrow[ul] \arrow[u] \arrow[ur] &
\end{tikzcd}$$

$(b_i^A, b_i^Y) \sim N(0, G)$

# Proposal

- Goal:
  - We want to account for patient heterogeneity in treatment assignment and biomarkers progression for longitudinal causal inference
- Solution:
  - Include random effects to partially identify time-invariant unmeasured confounders
  - Adopt the potential outcomes framework, combining g-computation and generalized mixed models (GMM)

# Proposal

- Goal:
  - We want to account for patient heterogeneity in treatment assignment and biomarkers progression for longitudinal causal inference
- Solution:
  - Include random effects to partially identify time-invariant unmeasured confounders
  - Adopt the potential outcomes framework, combining g-computation and generalized mixed models (GMM)

# What's new about our method

- Existing work on longitudinal causal inference with GMM usually expresses a causal estimand as a function of the treatment sequence, covariates, and fixed effect coefficients

- Random effects are included in our calculation of the causal estimand to better address patient heterogeneity

# What’s new about our method

- Existing work on longitudinal causal inference with GMM usually expresses a causal estimand as a function of the treatment sequence, covariates, and fixed effect coefficients
- Random effects are included in our calculation of the causal estimand to better address patient heterogeneity
- The cause must precede the effect in time

# What's new about our method

- Existing work on longitudinal causal inference with GMM usually expresses a causal estimand as a function of the treatment sequence, covariates, and fixed effect coefficients
- Random effects are included in our calculation of the causal estimand to better address patient heterogeneity
- The cause must precede the effect in time
- We propose to dynamically estimate the subject-specific random effects as subject-level observations accumulate over time

# What's new about our method

- Existing work on longitudinal causal inference with GMM usually expresses a causal estimand as a function of the treatment sequence, covariates, and fixed effect coefficients

- Random effects are included in our calculation of the causal estimand to better address patient heterogeneity

- The cause must precede the effect in time

- We propose to dynamically estimate the subject-specific random effects as subject-level observations accumulate over time

# What’s new about our method

- A way to handle time-invariant unmeasured confounders in Bayesian causal inference

- Has the potential to be extended to guide the inclusion of latent variable models in Bayesian causal inference

# Model Specification

For patient *i* at day *t*,

- $Y_{it}$: outcomes
- $A_{it}$: binary treatment status
- $V_i$: baseline covariates

Generalized mixed model of time-varying components

$$
\begin{aligned}
Y_{it} &= f_Y(V_i, A_{i,t}, Y_{i,t-1}; b_i^Y, \theta^Y) \\
A_{it} &= f_A(V_i, A_{i,t-1}, Y_{i,t-1}; b_i^A, \theta^A) \\
b_i &= (b_i^Y, b_i^A) \sim MVN(0, G)
\end{aligned}
$$

# Advantages - GMM of outcome and exposure

- The random effects $(b_i^Y, b_i^A)$ captures a vector of time-invariant unobserved confounders in a flexible manner.
  - Repeated measurement makes it possible to partially identify individual heterogeneity that arises from unmeasured factors
  - Use random effect parameter(s) as an automatic uncertainty quantity for sensitivity analysis

# Advantages - GMM of modeling outcome and exposure

- Incorporate propensity score in the Bayesian causal inference
  - A major debate: the role of propensity score (PS) in Bayesian causal inference (Li et al. 2022)
  - PS model is ignorable in Bayesian inference of population ATE and mixed ATE
  - Existing ways: specify outcomes distribution based on PS, shared parameters/priors between PS and outcome models, posterior-based IPW or DR estimators

# Information Flow Diagram

$$\begin{tikzcd}
\boxed{b^A} \arrow[r, dash, dashed] & G \arrow[r, dash, dashed] & b_i^Y \arrow[dl] & \\
Y_{i0} \arrow[r] \arrow[dr] & Y_{i1} \arrow[r] \arrow[dr] & Y_{i2} \arrow[r] \arrow[dr] & Y_{i3} \\
 & A_{i1} \arrow[r] \arrow[u] & A_{i2} \arrow[r] \arrow[u] & A_{i3} \arrow[u]
\end{tikzcd}$$

# Information Flow Diagram

$$\begin{tikzcd}
\boxed{b^A} \arrow[rr, dash, dashed, "G" description] & & b_i^Y \arrow[d, blue] & \\
Y_{i0} \arrow[r] \arrow[dr] & Y_{i1} \arrow[r] \arrow[dr] \arrow[ur, blue] & Y_{i2} \arrow[r] \arrow[dr] & Y_{i3} \\
 & A_{i1} \arrow[r] \arrow[u] & A_{i2} \arrow[r] \arrow[u] & A_{i3} \arrow[u]
\end{tikzcd}$$

# Information Flow Diagram

$$\begin{tikzcd}
\boxed{b^A} \arrow[rr, dash, dashed, "G" description] & & b_i^Y \arrow[dr] & \\
Y_{i0} \arrow[r] \arrow[dr] & Y_{i1} \arrow[r] \arrow[ur] \arrow[dr] & Y_{i2} \arrow[r] \arrow[u] \arrow[dr] & Y_{i3} \\
 & A_{i1} \arrow[r] \arrow[u] & A_{i2} \arrow[r] \arrow[u] & A_{i3} \arrow[u]
\end{tikzcd}$$

# Shardell and Ferrucci (2017)

G-formula under $(A_{i1}, A_{i2}) = (a_1, a_2)$, given subgroup information $V_i$ and heterogeneity effect $b_i = (b_i^A, b_i^Y) \sim N(0, G)$,

$$\begin{aligned}
& g(\overline{a}_2 | V_i, b_i) \\
=& \mathbb{E}\{Y_{i2} | V_i, \overline{A}_{i2} = \overline{a}_2, b_i\} \\
=& \int_{(Y_{i0}, Y_{i1})} \mathbb{E}\{Y_{i2} | V_i, \overline{Y}_{i1}, \overline{A}_{i2} = \overline{a}_2, b_i\} \\
& \qquad f(Y_{i1} | V_i, Y_{i0}, A_{i1} = a_1, b_i) f(Y_{i0}) dY_{i0} dY_{i1}
\end{aligned}$$

$$g(\overline{a}_2 | V_i) = \int_{b_i} g(\overline{a}_2 | V_i, b_i) f(b_i) db_i$$

# Proposal

$$
\begin{aligned}
& g(\overline{a}_2 | V_i, b_i^A) \\
=& \mathbb{E}\{Y_{i2} | V_i, \overline{A}_{i2} = \overline{a}_2, b_i^A\} \\
=& \int_{(Y_{i0}, Y_{i1})} \mathbb{E}\{Y_{i2} | V_i, \overline{Y}_{i1}, \overline{A}_{i2} = \overline{a}_2, b_i^A\} \\
& \qquad f(Y_{i1} | V_i, Y_{i0}, A_{i1} = a_1, b_i^A) f(Y_{i0}) dY_{i0} dY_{i1}
\end{aligned}
$$

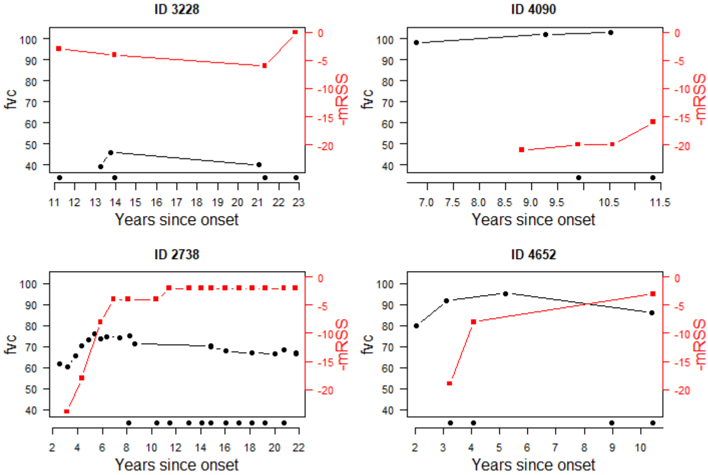
$$
g(\overline{a}_2 | V_i) = \int_{b_i^A} g(\overline{a}_2 | V_i, b_i^A) f(b_i^A) db_i^A
$$

# Proposal

$$
\begin{aligned}
& f(Y_{i1} | V_i, Y_{i0}, A_{i1} = a_1, b_i^A) \\
&= \int_{u_0} f(Y_{i1} | V_i, Y_{i0}, A_{i1} = a_1, b_i^A, b_i^Y = u_0) \\
&\qquad f(b_i^Y = u_0 | b_i^A) du_0
\end{aligned}
$$

$$
\begin{aligned}
& \mathbb{E}\{Y_{i2} | V_i, \overline{Y}_{i1}, \overline{A}_{i2} = \overline{a}_2, b_i^A\} \\
&= \int_{u_1} \mathbb{E}\{Y_{i2} | V_i, \overline{Y}_{i1}, \overline{A}_{i2} = \overline{a}_2, b_i^A, b_i^Y = u_1\} \\
&\qquad f(b_i^Y = u_1 | V_i, \overline{Y}_{i1}, A_{i1} = a_1, b_i^A) du_1
\end{aligned}
$$

# Scleroderma Application



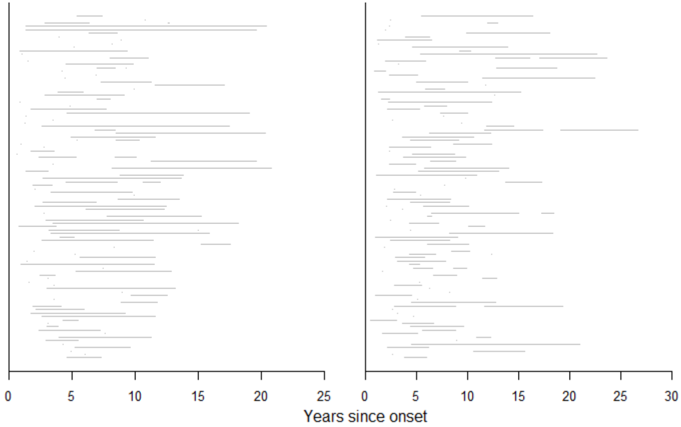
ID 3228

ID 4090

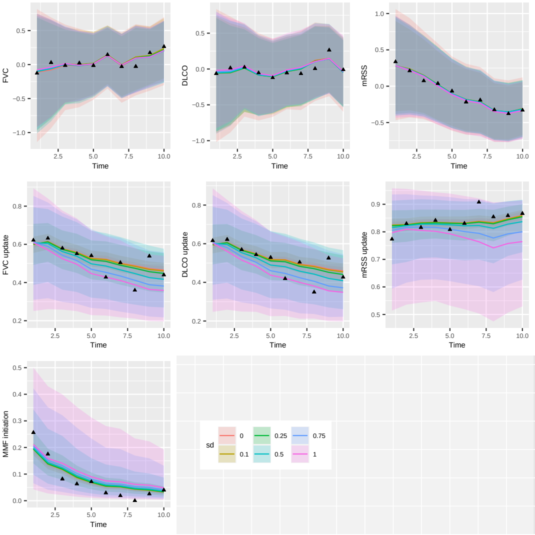
ID 2738

ID 4652

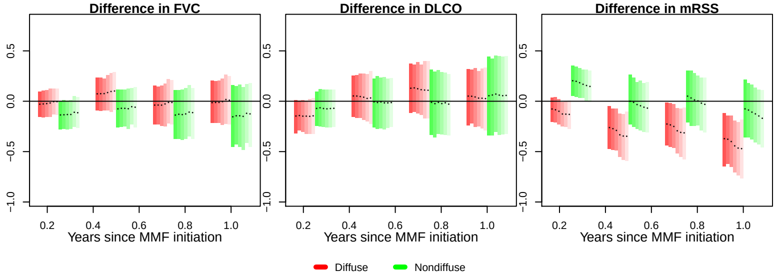
# Scleroderma Application



# Scleroderma Application

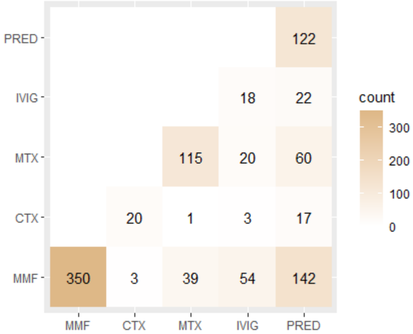


# Scleroderma Application

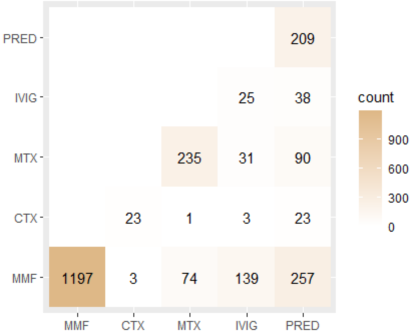


# Scleroderma Application

Drug Combinations by Person



Drug Combinations by 6 mo. Time Intervals



# Contributions

- Naturally includes time-invariant unmeasured confounders as model parameters that serve as automatic uncertainty quantities for sensitivity analysis
- A new way of incorporating propensity score in Bayesian causal inference under the potential outcomes framework
- Enables the combination of causal inference and latent variable models for dynamic decision making
- Integrates population-level knowledge and the dynamic accumulation of subject-level data into causal analysis

## End

# Thank you

Collaborators:

- Zeger, Scott - Johns Hopkins University
- Shah, Ami - Johns Hopkins University
- Kim, Jisoo - Johns Hopkins University