Using statistical methods and reproducible tools to gain new insights from biomedical and public health data

Ariel Mundo Ortiz

Centre de Recherches Mathematiques, Université de Montréal

MfPH Next Generation Seminar Series

3/15/23

Introduction

- Data is the core of research. However, data is not information, as it needs to be processed before we can get information from it.

Introduction

- Data is the core of research. However, data is not information, as it needs to be processed before we can get information from it.
- This is specially true in the case of health research: public health, or biomedical data can be complex, and decisions along the analysis can result in different interpretations.

Introduction

- Data is the core of research. However, data is not information, as it needs to be processed before we can get information from it.
- This is specially true in the case of health research: public health, or biomedical data can be complex, and decisions along the analysis can result in different interpretations.
- In this talk I will focus on two examples that showcase how we can get more insight from data

The Case of Public Health Data

COVID-19

- COVID-19 vaccination has been an important component of public health strategies aimed at managing the pandemic.

[1] Nafilyan et al. 2021.
[2] Gerretsen et al. 2021.

COVID-19

- COVID-19 vaccination has been an important component of public health strategies aimed at managing the pandemic.
- However, COVID-19 vaccination has not been equal across different population segments.

[1] Nafilyan et al. 2021.
[2] Gerretsen et al. 2021.

COVID-19

- COVID-19 vaccination has been an important component of public health strategies aimed at managing the pandemic.
- However, COVID-19 vaccination has not been equal across different population segments.

[1] Nafilyan et al. 2021.
[2] Gerretsen et al. 2021.

COVID-19

- COVID-19 vaccination has been an important component of public health strategies aimed at managing the pandemic.
- However, COVID-19 vaccination has not been equal across different population segments.
- Individuals with lower income, and those belonging to a racial/ethnic minority have had lower vaccination uptake[1,2].

[1] Nafilyan et al. 2021.
[2] Gerretsen et al. 2021.

COVID-19: The Case of Ontario

- The Fields Institute collected some very nice data regarding COVID-19 vaccination in Ontario, the *Survey of COVID-19 related Behaviours and Attitudes*.

COVID-19: The Case of Ontario

- The Fields Institute collected some very nice data regarding COVID-19 vaccination in Ontario, the *Survey of COVID-19 related Behaviours and Attitudes*.
  - The survey ran between late 2021 and early 2022 and collected socio-demographic information along with self-reported vaccination status ("Have you received the first dose of the Covid vaccine?")

COVID-19: The Case of Ontario

Table 1: Selected socio-economic factors from the survey

Variable	Levels
Age group	16-34,35-54,55 and over
Income bracket (CAD)	under 25,000, 25,000-59,999, 60,000 and above
Race/ethnicity	Arab/Middle Eastern, Black, East Asian/Pacific Islander, Indigenous, Latin American, Mixed, South Asian, White Caucasian, Other

COVID-19: The Case of Ontario

- Other studies have analyzed the dependency on vaccination status using socio-economic data.

COVID-19: The Case of Ontario

- Other studies have analyzed the dependency on vaccination status using socio-economic data.
- We could do the same, but what other information could we get from this data?

COVID-19: The Case of Ontario

- Other studies have analyzed the dependency on vaccination status using socio-economic data.
- We could do the same, but what other information could we get from this data?
- From a Public Health Perspective, there have been some relatively recent developments in Ontario.

COVID-19: The Case of Ontario

- However, Ontario adopted in late 2019 the Health Regions for healthcare and phased out the Local Health Integration Network (LHIN) approach.

COVID-19: The Case of Ontario

- However, Ontario adopted in late 2019 the Health Regions for healthcare and phased out the Local Health Integration Network (LHIN) approach.
- The change is relatively new, and therefore, geographical data can be used to analyze data within the different Health Regions.

COVID-19: The Case of Ontario

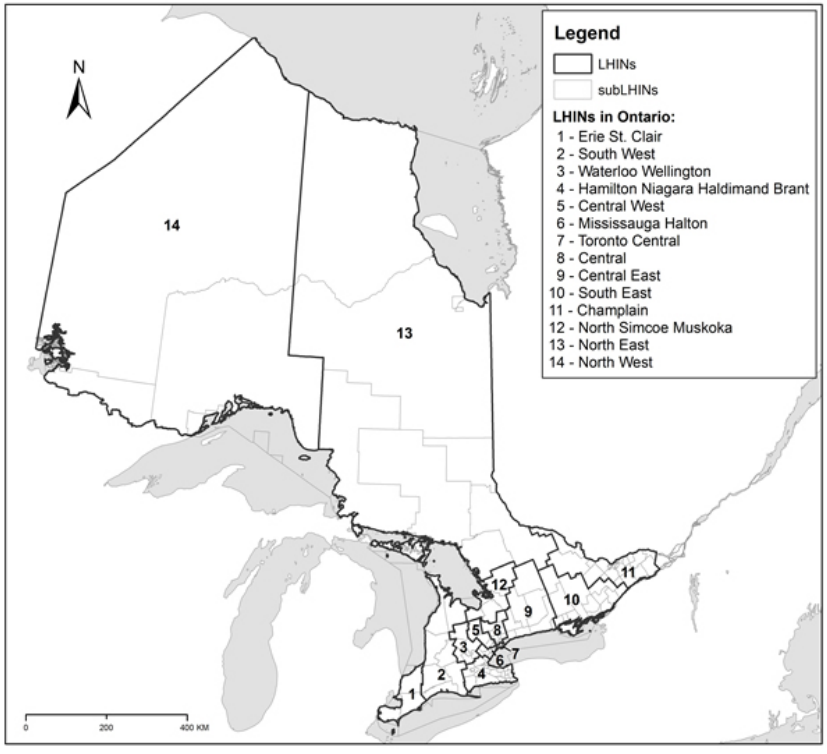

Figure 1: Ontario LHINs (Crighton et al. 2015)

COVID-19: The Case of Ontario

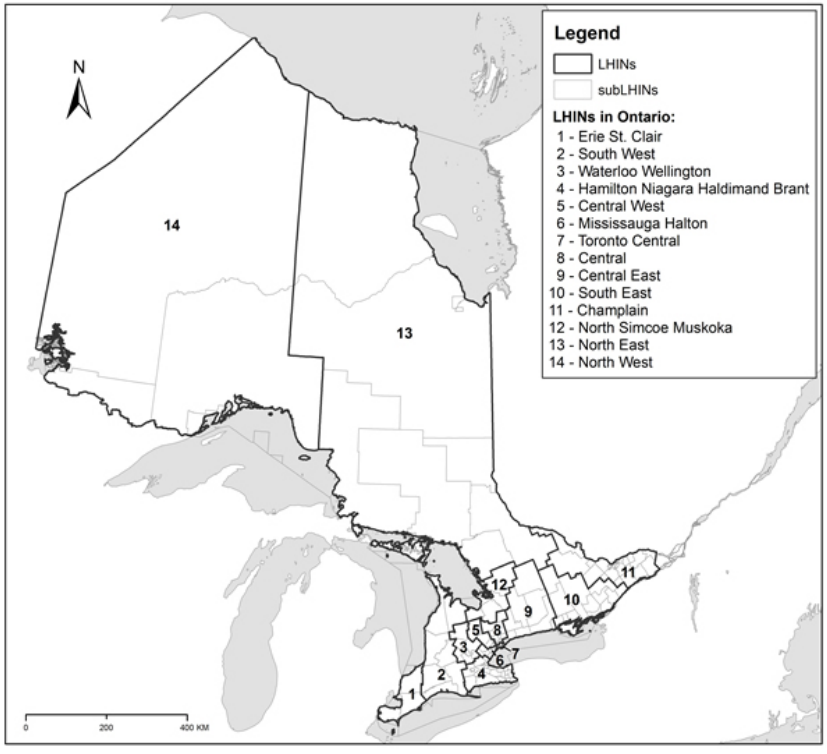

Figure 1: Ontario LHINs (Crighton et al. 2015)

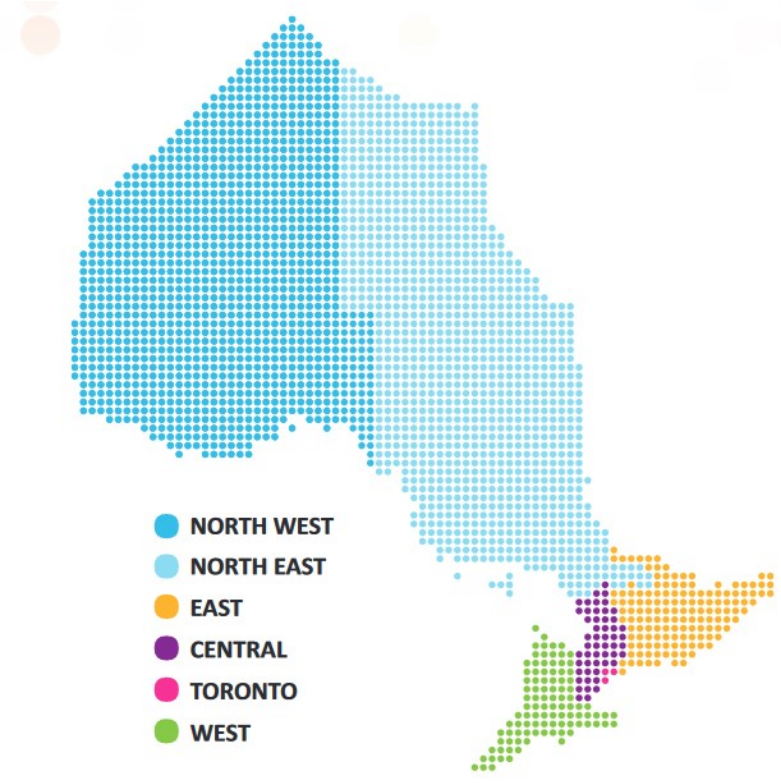

Figure 2: Ontario Health Regions (Ontario Business Health Plan 2022-2023)

COVID-19: The Case of Ontario

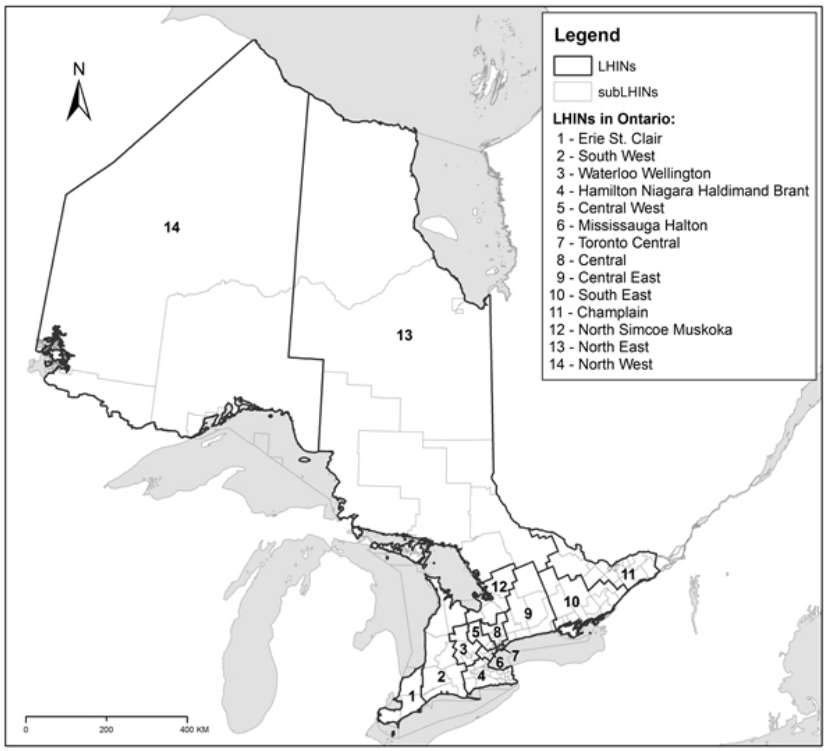

Figure 1: Ontario LHINs (Crighton et al. 2015)

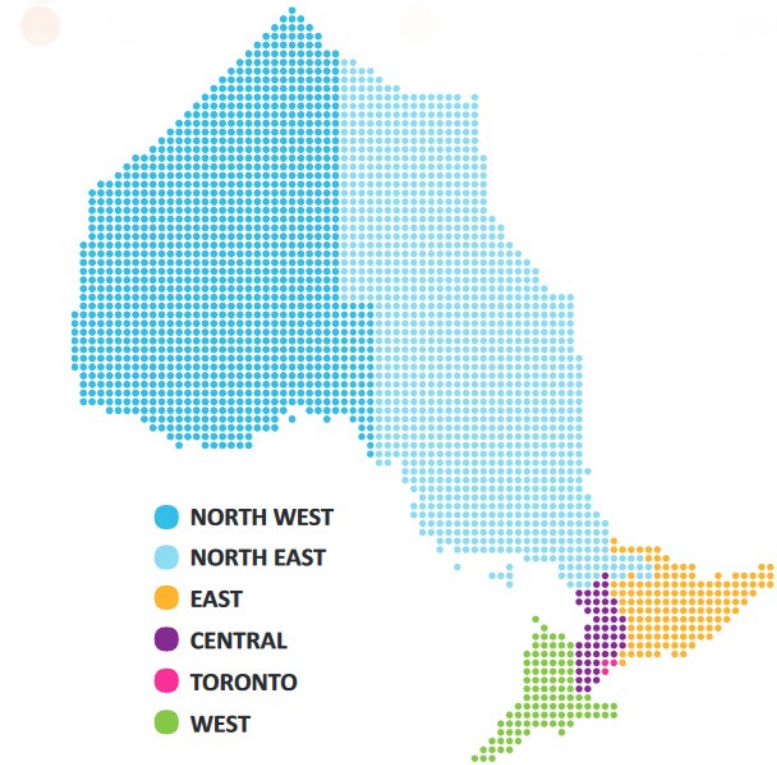

Figure 2: Ontario Health Regions (Ontario Business Health Plan 2022-2023)

- Data cleaning is **important**

COVID-19: The Case of Ontario

- Therefore, we decided to integrate the different Health Regions in our analysis to determine the odds of vaccination.

$$\log\left(\frac{p(\text{vac})}{1-p(\text{vac})}\right) = \beta_0 + \beta_1(\text{Age group}) + \beta_2\ \text{Race} + \beta_3\ \text{Health Region} + \beta_4\ \text{Income} + \beta_5(\text{Health Region} \times \text{Race}) + \beta_6\ (\text{Income} \times \text{Race}) \quad (1)$$

Results

Table 2: **Selected** Multivariable Regression Results

Characteristic	OR	95% CI	p-value
Income (CAD)			
60000 and above	—	—	
25000-59999	0.59	0.39, 0.89	0.011
under 25000	0.37	0.25, 0.56	<0.001
Race			
White/Caucasian	—	—	
Arab/Middle Eastern	0.31	0.14, 0.69	0.004
Black	0.32	0.17, 0.60	<0.001
East Asian/Pacific Islander	1.15	0.50, 2.66	0.7
Indigenous	0.44	0.19, 1.02	0.056
Latin Aamerican	0.28	0.11, 0.67	0.004
Mixed	0.64	0.25, 1.65	0.4
Other	0.22	0.12, 0.41	<0.001
South Asian	0.91	0.49, 1.69	0.8
Health Region			
Toronto	—	—	
Central	1.47	0.92, 2.35	0.11
East	1.42	0.90, 2.23	0.13
West	1.55	1.05, 2.30	0.029
Income and Race			
25000-59999 * Arab/Middle Eastern	1.79	0.67, 4.83	0.2
under 25000 * Arab/Middle Eastern	3.05	1.26, 7.39	0.013
25000-59999 * Black	1.34	0.59, 3.05	0.5
under 25000 * Black	3.19	1.45, 6.99	0.004
25000-59999 * East Asian/Pacific Islander	0.42	0.17, 1.05	0.062
under 25000 * East Asian/Pacific Islander	1.16	0.47, 2.86	0.8
25000-59999 * Indigenous	1.36	0.48, 3.89	0.6
under 25000 * Indigenous	1.45	0.55, 3.80	0.5
25000-59999 * Latin American	1.24	0.45, 3.43	0.7

Results

Characteristic	OR	95% CI	p-value
under 25000 * Latin American	2.80	1.04, 7.51	0.041
25000-59999 * Mixed	0.85	0.32, 2.26	0.7
under 25000 * Mixed	1.10	0.37, 3.27	0.9
25000-59999 * Other	6.93	2.65, 18.1	<0.001
under 25000 * Other	4.59	2.33, 9.05	<0.001
25000-59999 * South Asian	1.20	0.51, 2.85	0.7
under 25000 * South Asian	2.00	0.93, 4.30	0.077
Race and Health Region			
Arab/Middle Eastern * Central	0.66	0.26, 1.70	0.4
Black * Central	0.44	0.19, 0.98	0.046
East Asian/Pacific Islander * Central	0.98	0.38, 2.53	>0.9
Mixed * East	0.91	0.28, 3.03	0.9
other * East	1.05	0.39, 2.83	>0.9
South Asian * East	0.52	0.19, 1.45	0.2
Arab/Middle Eastern * West	1.00	0.37, 2.73	>0.9
Black * West	0.76	0.32, 1.80	0.5
East Asian/Pacific Islander * West	0.52	0.20, 1.34	0.2
Indigenous * West	0.39	0.14, 1.09	0.073
Latin American * West	0.94	0.32, 2.72	>0.9
Mixed * West	0.37	0.12, 1.16	0.089
Other * West	0.41	0.18, 0.93	0.032
South Asian * West	0.41	0.18, 0.95	0.037

[1] OR = Odds Ratio, CI = Confidence Interval

How do we interpret this?

- Our results show that there were disparities in vaccination uptake in Ontario.

[3] Hawkins 2020.

How do we interpret this?

- Our results show that there were disparities in vaccination uptake in Ontario.
- People in certain racial minority groups had lower odds of vaccination than White/Caucasian individuals.

[3] Hawkins 2020.

How do we interpret this?

- Our results show that there were disparities in vaccination uptake in Ontario.
- People in certain racial minority groups had lower odds of vaccination than White/Caucasian individuals.
- However, individuals that identified with a racial/ethnic minority and that were in a low household income bracket (<60k CAD) had higher odds of vaccination than individuals with a high household income.

[3] Hawkins 2020.

How do we interpret this?

- Our results show that there were disparities in vaccination uptake in Ontario.
- People in certain racial minority groups had lower odds of vaccination than White/Caucasian individuals.
- However, individuals that identified with a racial/ethnic minority and that were in a low household income bracket (<60k CAD) had higher odds of vaccination than individuals with a high household income.
- This is likely caused by the type of occupation: people in racial minorities, and those with a low household income work in essential occupations[3], and thus potentially got the vaccine to be able to work.

[3] Hawkins 2020.

How do we interpret this?

- But there are also intra-provincial differences in vaccine uptake within the Health Regions:

How do we interpret this?

- But there are also intra-provincial differences in vaccine uptake within the Health Regions:
  - For example, South Asian individuals in the West Health Region had lower odds of vaccination that in other Health Regions.

How do we interpret this?

- But there are also intra-provincial differences in vaccine uptake within the Health Regions:
  - For example, South Asian individuals in the West Health Region had lower odds of vaccination that in other Health Regions.
  - These results provide a more comprehensive assessment of COVID-19 vaccination rates within Ontario, as they showed that certain minority groups within specific income brackets and certain Health Regions had differences in vaccination.

The Case of Biomedical Data

Longitudinal Data

- Biomedical studies often collect longitudinal data to see the effect of an intervention over time:

Longitudinal Data

- Biomedical studies often collect longitudinal data to see the effect of an intervention over time:
  - How a chemotherapy treatment changes the metabolism of a tumor

Longitudinal Data

- Biomedical studies often collect longitudinal data to see the effect of an intervention over time:
  - How a chemotherapy treatment changes the metabolism of a tumor
  - How the concentration of a drug changes over time in the blood

Longitudinal Data

- Biomedical studies often collect longitudinal data to see the effect of an intervention over time:
  - How a chemotherapy treatment changes the metabolism of a tumor
  - How the concentration of a drug changes over time in the blood
- How is this data typically analyzed?

Linear Models

$$y_{ijt} = \beta_0 + \beta_1 \times treatment_j + \beta_2 \times time_t + \beta_3 \times time_t \times treatment_j + \varepsilon_{ijt} \quad (2)$$

where,

y_{ijt}: is the response for subject i in treatment group j at time t

Linear Models

$$y_{ijt} = \beta_0 + \beta_1 \times treatment_j + \beta_2 \times time_t + \beta_3 \times time_t \times treatment_j + \varepsilon_{ijt} \quad (2)$$

where,

y_{ijt}: is the response for subject i in treatment group j at time t

β_0: the mean group value

Linear Models

$$y_{ijt} = \beta_0 + \beta_1 \times treatment_j + \beta_2 \times time_t + \beta_3 \times time_t \times treatment_j + \varepsilon_{ijt} \quad (2)$$

where,

y_{ijt}: is the response for subject i in treatment group j at time t

β_0: the mean group value

$time_t$, $treatment_j$: fixed effects

Linear Models

$$y_{ijt} = \beta_0 + \beta_1 \times treatment_j + \beta_2 \times time_t + \beta_3 \times time_t \times treatment_j + \varepsilon_{ijt} \quad (2)$$

where,

y_{ijt}: is the response for subject i in treatment group j at time t

β_0: the mean group value

$time_t$, $treatment_j$: fixed effects

β_1, β_2 and β_3: linear slopes of the fixed effects.

Linear Models

$$y_{ijt} = \beta_0 + \beta_1 \times treatment_j + \beta_2 \times time_t + \beta_3 \times time_t \times treatment_j + \varepsilon_{ijt} \quad (2)$$

where,

y_{ijt}: is the response for subject i in treatment group j at time t

β_0: the mean group value

$time_t$, $treatment_j$: fixed effects

β_1, β_2 and β_3: linear slopes of the fixed effects.

ε_{ijt}: error, assumed to be $\sim N(0, \sigma^2)$

Linear Models

$$y_{ijt} = \beta_0 + \beta_1 \times treatment_j + \beta_2 \times time_t + \beta_3 \times time_t \times treatment_j + \varepsilon_{ijt} \quad (2)$$

where,

y_{ijt}: is the response for subject i in treatment group j at time t

β_0: the mean group value

$time_t$, $treatment_j$: fixed effects

β_1, β_2 and β_3: linear slopes of the fixed effects.

ε_{ijt}: error, assumed to be $\sim N(0, \sigma^2)$

A LMEM follows the same exact structure, only incorporates a random effect α_{ij}, which allows for different intercepts.

Trends Over Time

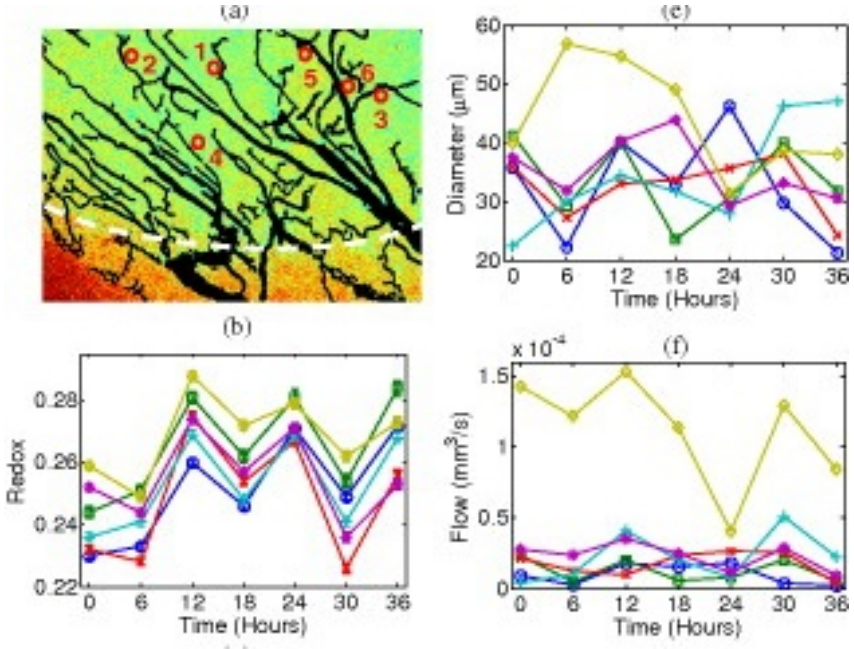

Figure 3: Tumor imaging data (Skala et al. 2010)

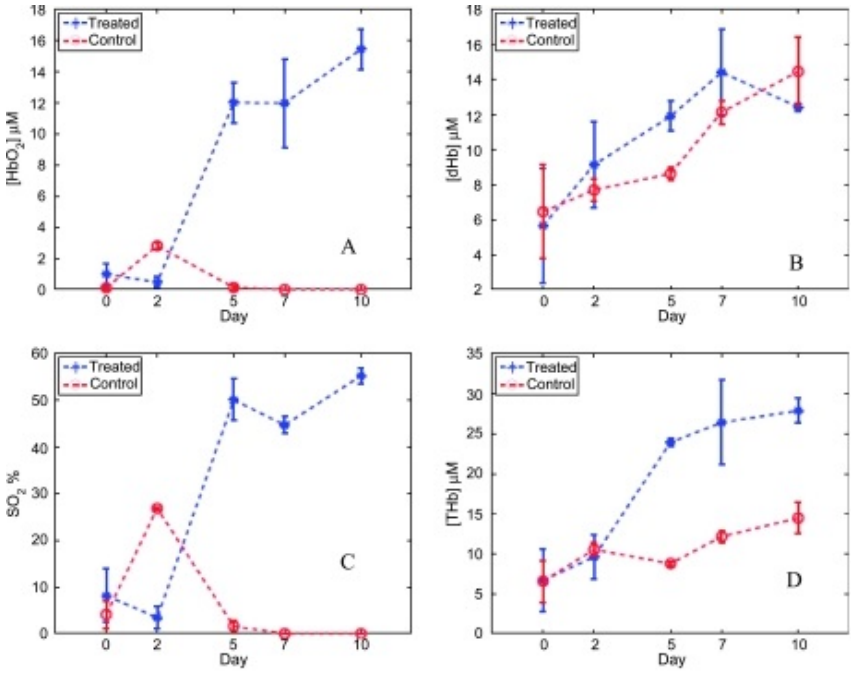

Figure 4: Tumor oxygenation data (Vishwanath et al. 2009)

Trends Over Time

- The issue in those data is that the trends are not linear, and therefore, a linear model will miss changes in the signal where some metabolic or physiological relevant change is taking place.

[4] Beck and Jackman 1998.

Trends Over Time

- The issue in those data is that the trends are not linear, and therefore, a linear model will miss changes in the signal where some metabolic or physiological relevant change is taking place.
- Polynomial effects can be used, but they create biases at the boundaries of the covariates[4].

[4] Beck and Jackman 1998.

Generalized Additive Models (GAMs)

$$y_{ijt} = \beta_0 + \beta_1 \times treatment_j + f(time_t \mid \beta_j) + \varepsilon_{ijt} \qquad (3)$$

- The change of y_{ijt} over time is represented by the *smooth function* $f(time_t \mid \beta_j)$ with inputs as the covariates $time_t$ and parameters β_j.

Generalized Additive Models (GAMs)

$$y_{ijt} = \beta_0 + \beta_1 \times treatment_j + f(time_t \mid \beta_j) + \varepsilon_{ijt} \qquad (3)$$

- The change of y_{ijt} over time is represented by the *smooth function* $f(time_t \mid \beta_j)$ with inputs as the covariates $time_t$ and parameters β_j.

Generalized Additive Models (GAMs)

$$y_{ijt} = \beta_0 + \beta_1 \times treatment_j + f(time_t \mid \beta_j) + \varepsilon_{ijt} \qquad (3)$$

- The change of y_{ijt} over time is represented by the *smooth function* $f(time_t \mid \beta_j)$ with inputs as the covariates $time_t$ and parameters β_j.
- We can use a *basis function* to estimate the smooth function.

Generalized Additive Models (GAMs)

$$y_{ijt} = \beta_0 + \beta_1 \times treatment_j + f(time_t \mid \beta_j) + \varepsilon_{ijt} \qquad (3)$$

- The change of y_{ijt} over time is represented by the *smooth function* $f(time_t \mid \beta_j)$ with inputs as the covariates $time_t$ and parameters β_j.
- We can use a *basis function* to estimate the smooth function.

Generalized Additive Models (GAMs)

$$y_{ijt} = \beta_0 + \beta_1 \times treatment_j + f(time_t \mid \beta_j) + \varepsilon_{ijt} \qquad (3)$$

- The change of y_{ijt} over time is represented by the *smooth function* $f(time_t \mid \beta_j)$ with inputs as the covariates $time_t$ and parameters β_j.
- We can use a *basis function* to estimate the smooth function.
- Splines are helpful as basis functions: Thin plate regression splines (TPRS) are computationally efficient, and the underlying principle is that of polynomial pieces "joined" together

How GAMs work

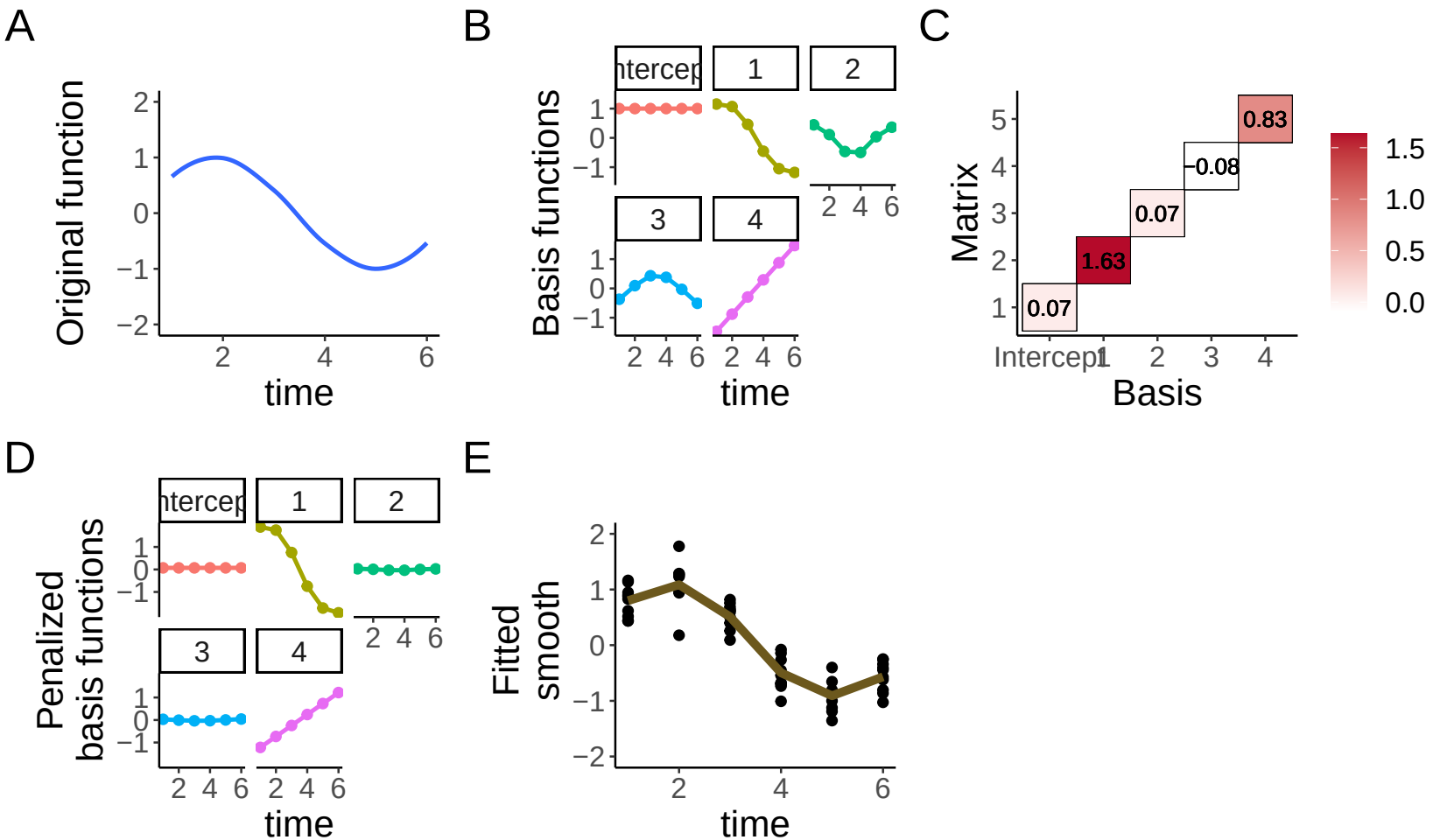

Figure 5: Fitting process of a GAM.

An Example

- Simulated data from a study on radiotherapy in a mouse model of melanoma[5].

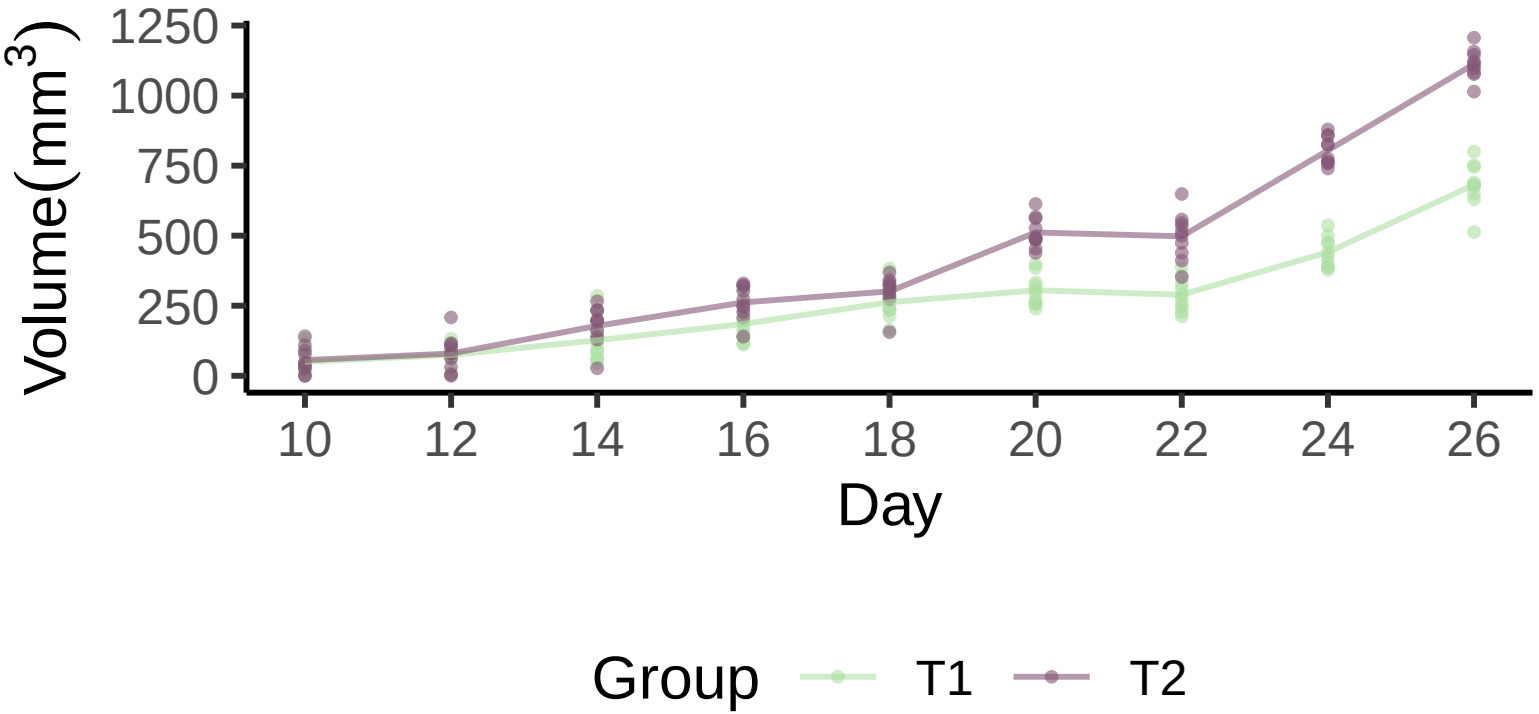

Figure 6: Tumor volume in two groups of tumors under radiotherapy

[5]Sen et al. 2011.

Fitting a GAM

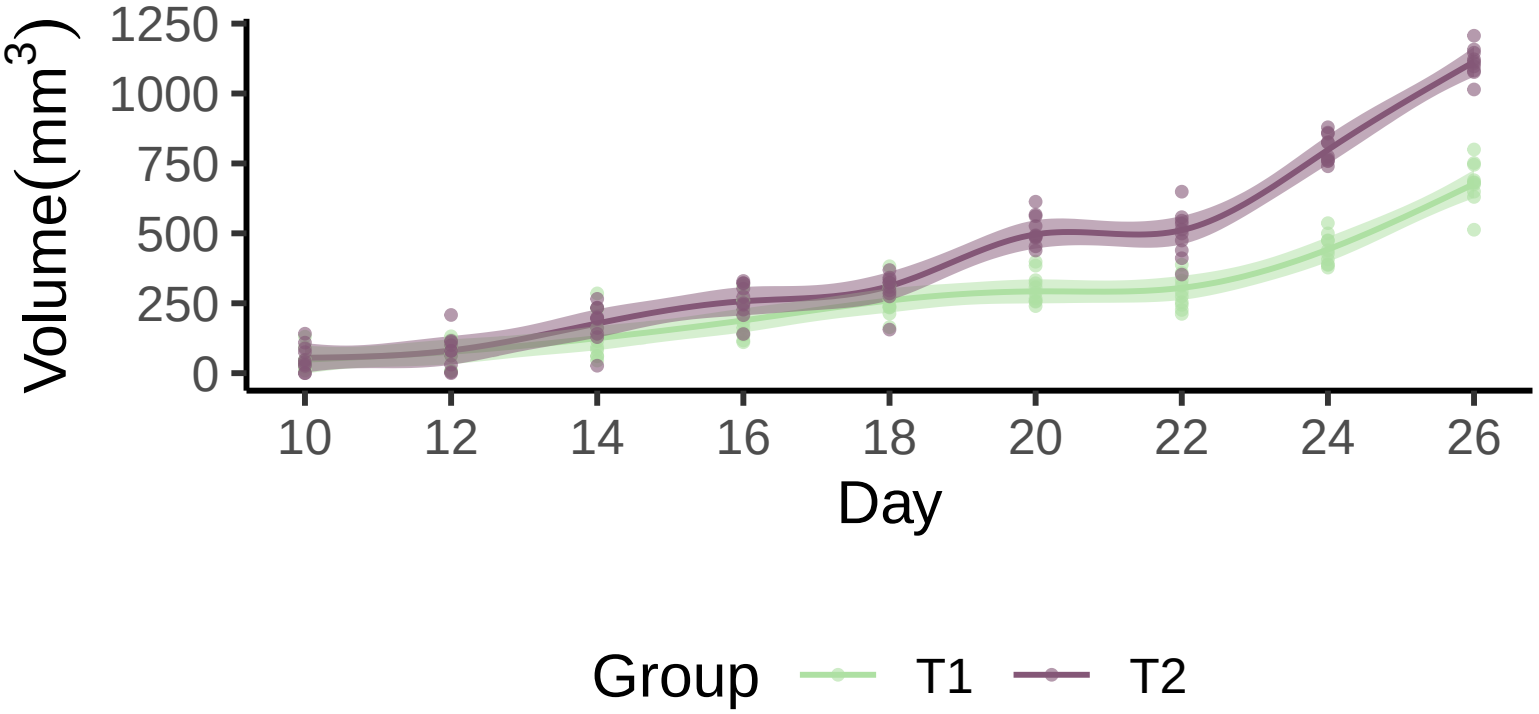

Figure 7: GAM fitted to simulated data

- The model captures the trend of the data

Fitting a GAM

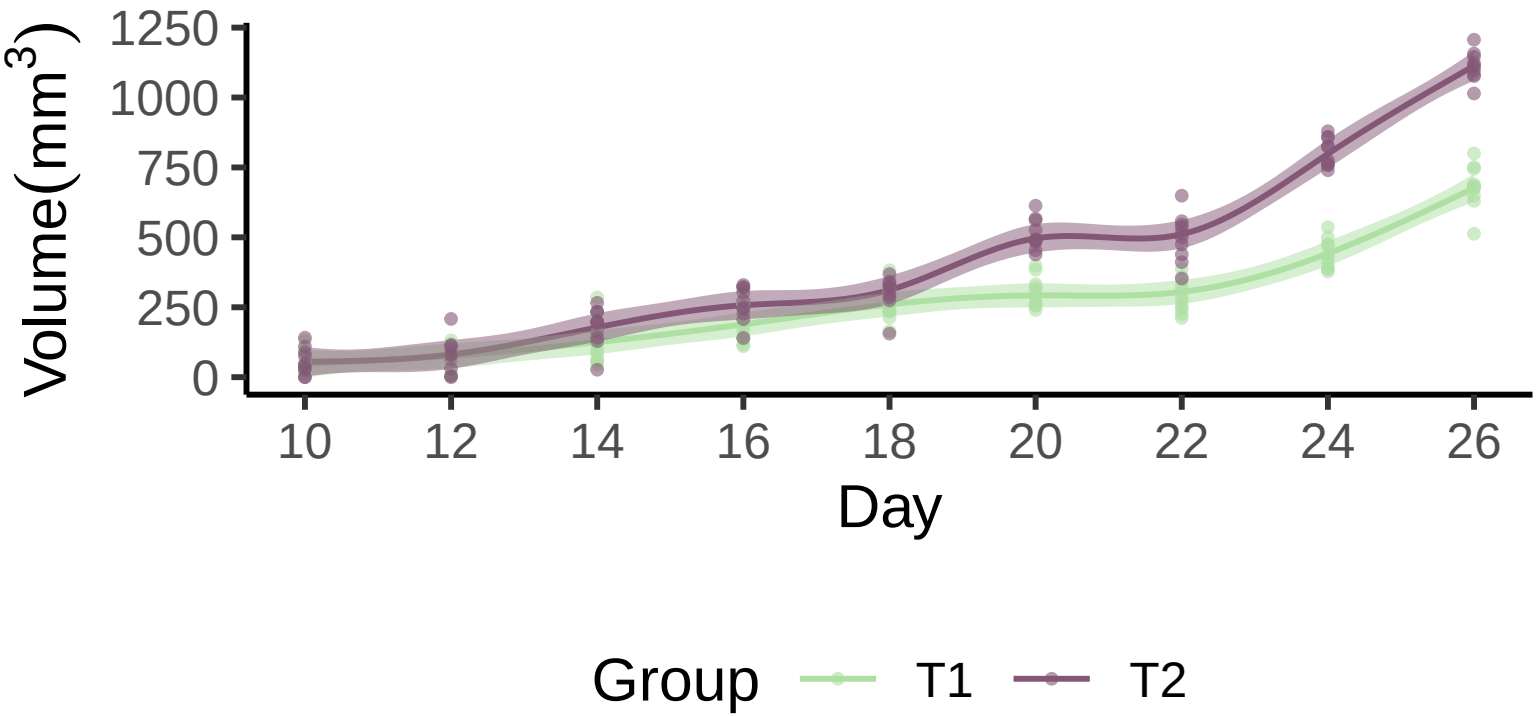

Figure 7: GAM fitted to simulated data

- The model captures the trend of the data
- We can furthermore compare the trends.

Differences

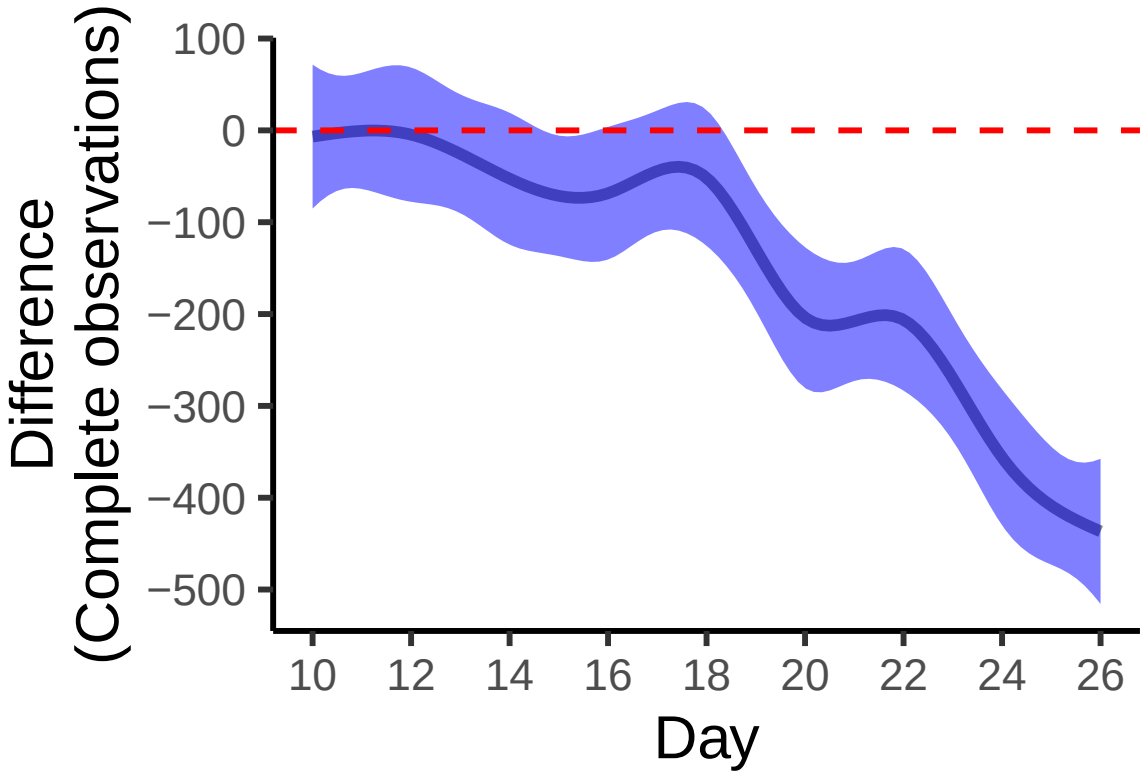

Figure 8: Pairwise comparisons between smooths

- We can compare the smooths for each group. Here, we see that T2 is significantly higher after day 18.

Differences

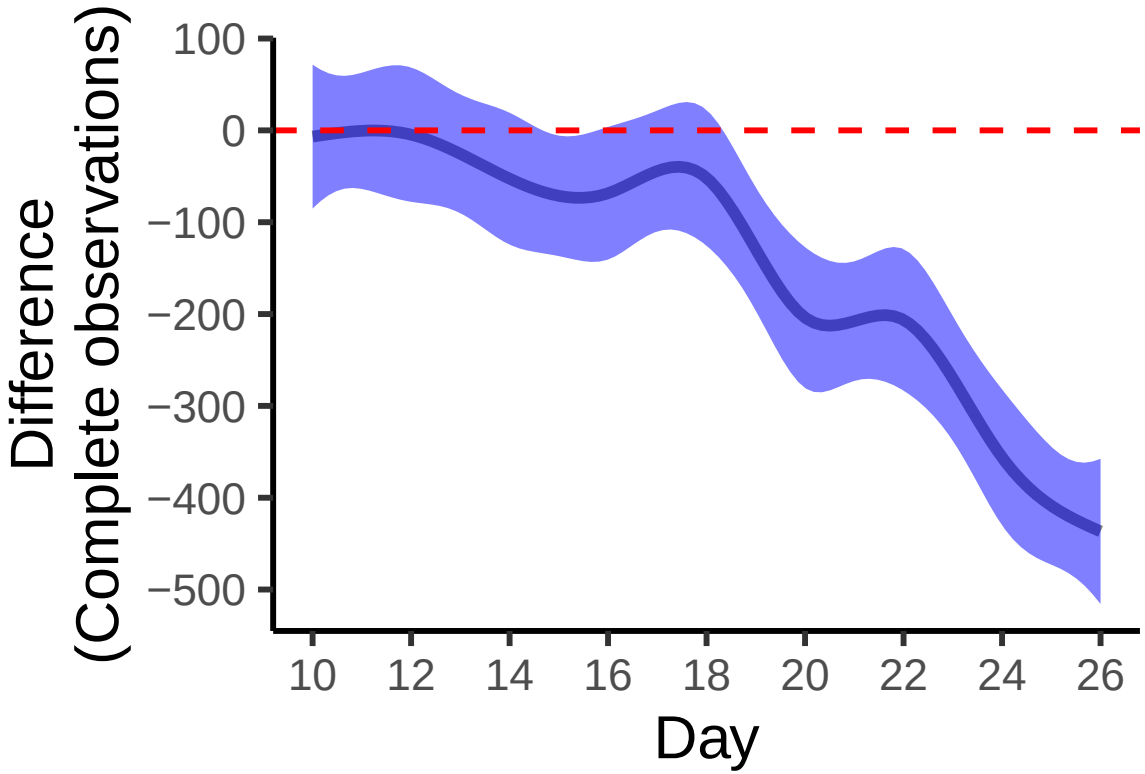

Figure 8: Pairwise comparisons between smooths

- We can compare the smooths for each group. Here, we see that T2 is significantly higher after day 18.
- This can give an idea of further explorations of biological

Addressing Reproducibility

- There is an ongoing need of making papers reproducible.

Addressing Reproducibility

- There is an ongoing need of making papers reproducible.
- This is specially important in the case of data/methods of health research.

Addressing Reproducibility

- There is an ongoing need of making papers reproducible.
- This is specially important in the case of data/methods of health research.
  - Otherwise, tools cannot be used by others.

Addressing Reproducibility

- There is an ongoing need of making papers reproducible.
- This is specially important in the case of data/methods of health research.
  - Otherwise, tools cannot be used by others.
- How are we addressing this in our research?

Addressing Reproducibility

- Using GitHub to share:

Addressing Reproducibility

- Using GitHub to share:
  - Data: Making publicly available the datasets used

Addressing Reproducibility

- Using GitHub to share:
  - Data: Making publicly available the datasets used
  - Methods: Sharing the code used for statistical analyses

Addressing Reproducibility

- Using GitHub to share:
  - Data: Making publicly available the datasets used
  - Methods: Sharing the code used for statistical analyses
- In synthesis, sharing all the information used to create a paper such that anyone can re-create the analysis, results, and the paper itself from the files provided.

Addressing Reproducibility

- For GAMs
  https://github.com/aimundo/GAMs-biomedical-research

Addressing Reproducibility

- For GAMs https://github.com/aimundo/GAMs-biomedical-research
- COVID-19: Work is ongoing, but repository will be ready when paper is submitted

Conclusion

- There is an ongoing need of analyzing public health data to address important disparities in areas such as vaccination.

Conclusion

- There is an ongoing need of analyzing public health data to address important disparities in areas such as vaccination.
- Semi-parametric statistical to analyze biomedical/public health longitudinal data, such as GAMs can provide better insight on periods where important biological changes might occur.

Acknowledgements

- The Nasri Lab (Université de Montréal)
  - Bouchra Nasri, PhD (PI)
  - Idriss Sekkak, PhD
  - Rado Ramasy
  - Fatima El-Mousawi
  - Rawda Berkat
- The Muldoon Lab (University of Arkansas)
  - Timothy J. Muldoon (PI)
- John R. Tipton (Los Alamos National Laboratory)

Beck, Nathaniel and Simon Jackman (Apr. 1998). "Beyond Linearity by Default: Generalized Additive Models". In: *American Journal of Political Science* 42.2, p. 596. DOI: 10.2307/2991772. URL: https://doi.org/10.2307/2991772.

Gerretsen, Philip et al. (Nov. 2021). "Individual determinants of COVID-19 vaccine hesitancy". In: *PLOS ONE* 16.11. Ed. by Leeberk Raja Inbaraj, e0258462. DOI: 10.1371/journal.pone.0258462. URL: https://doi.org/10.1371/journal.pone.0258462.

Hawkins, Devan (June 2020). "Differential occupational risk for COVID-19 and other infection exposure according to race and ethnicity". In: *American Journal of Industrial Medicine* 63.9, pp. 817–820. DOI: 10.1002/ajim.23145. URL: https://doi.org/10.1002/ajim.23145.

 Nafilyan, Vahe et al. (July 2021). "Sociodemographic inequality in COVID-19 vaccination coverage among elderly adults in England: a national linked data study". In: *BMJ Open* 11.7, e053402. DOI: 10.1136/bmjopen-2021-053402. URL: https://doi.org/10.1136/bmjopen-2021-053402.

 Sen, Arindam et al. (May 2011). "Mild Elevation of Body Temperature Reduces Tumor Interstitial Fluid Pressure and Hypoxia and Enhances Efficacy of Radiotherapy in Murine Tumor Models". In: *Cancer Research* 71.11, pp. 3872–3880. DOI: 10.1158/0008-5472.can-10-4482. URL: https://doi.org/10.1158/0008-5472.can-10-4482.