Modeling COVID-19 high proportion death rates

Team 6: Alex Foster, Felipe Alvarado, Raman Hundal, Ankit Gubiligari

March 6, 2024

AGENDA

01 Overview

02 Data Background

03 Feature Selection

04 Modeling Decisions

05 Final Model Interpretation

06 Conclusion

Overview

Project Origin

Stakeholder

Government entities at both the state and federal levels are seeking to comprehend the interplay of policies and demographics that influence the highest survival rates amidst elevated proportions of pandemic-related deaths.

Challenge

- Data is not standardized across states and regions in the US
- Demographic information is collected sparsely and not entirely accurate.
- Modeling this data is inherently difficult with such volatile data.

Analytical Plans & Goals

1. Create a dataset that contains state level population and pandemic policy information together
2. Consider multiple modeling approaches to address high proportion pandemic caused deaths
3. Provide detailed analysis of which factors and policies are most important to consider during a pandemic from the data provided.

Data Preprocessing

Data Overview

Data was collected from a variety of sources:

01	Binned Age Population Projections (CDC)
02	Demographic Population Estimates 2020-2022 (Census.gov)
03	COVID-19 Policy Tracker Data (OxCGRT)
04	COVID-19 Deaths and Cases Data (New York Times)
05	COVID-19 Vaccine Tracker (CDC)

Data Preprocessing

There were several steps involved in combining our datasets

1. Combining Datasets:
   a. We integrated multiple datasets to streamline our analysis
2. Yearly Projection Data Processing:
   a. Conducted averaging to summarize yearly projection data, particularly focusing on population estimates
3. Aggregation by Month and State:
   a. Aggregated data by month and state to facilitate detailed analysis
4. Granularity Adjustment:
   a. Recognizing that state-level data might be overly granular, we grouped states into broader regions: West, Midwest, Southwest, and Northeast
5. Data Granularity Levels:
   a. The majority of our data was collected at the daily/weekly level
6. Death Proportion Calculation:
   a. Calculated the death proportion by dividing the number of deaths by the population estimate for a given year.
7. Death_Event Variable Creation:
   a. Introduced the 'death_event' variable, indicating whether the death proportion exceeded 0.8, serving as a critical threshold for analysis and decision-making.

Final Dataset Details:

1250 rows (covering 25 months and 50 states)

50 potential explanatory features

Target Variable: 'death_event' (binary)

Documentation: https://scikit-learn.org/stable/modules/generated/sklearn.impute.KNNImputer.html

COVID Policies/Variable Interpretation

- **Some examples of selected variables and their differing value keys:**

- **Vaccine Availability (V2A_Vaccine.Availability..summary.)**
  - 1 = Lowest tier - Least strict level of COVID-19 policies: vaccine availability
  - 2 = Next level to which COVID-19 policies allow for vaccine availability
  - 3 = Highest tier - Strictest level of COVID-19 policies: vaccine availability

- **School Closing (C1M_School.closing)**
  - 0 = Highest tier - Strictest Level of COVID-19 policies for closing a school.
  - Numbers in between = the level to which PCR test are required to enter school.
  - 3 = Lowest Tier - Least strict level of COVID-19 policies for closing a school.

- **International Travel Controls (C8EV_International.travel.controls)**
  - 1 = Highest tier - Strictest COVID-19 policies on international travel
  - 3 = Lowest tier - Least Strict COVID-19 policies on international travel

Feature Selection

RFE and Correlation Analysis for Logistic Regression

Selected Features
X20.24.years
X40.44.years
X45.49.years
C3M_Cancel.public.events
X50.54.years
X15.19.years
X25.29.years
X0.4.years
X60.64.years
H6M_Facial.Coverings

Variable	Correlation
doses_distributed_cumulative	0.1944324
num_months	0.1910613
X	0.1898049
doses_administered_cumulative	0.1869141
V2A_Vaccine.Availability..summary.	0.1761854
V1_Vaccine.Prioritisation..summary.	0.1731717
regionNortheast	0.1555786
X85..years	0.1449583
H7_Vaccination.policy	0.1410544
X80.84.years	0.1384333

Backward Elimination for Cox Proportional Hazard Model

Using this technique we removed the following features:

new_doses_distributed	new_doses_administered	0-4 years
5-9 years	80-84 years	Workplace closing
Close public transport	Public information campaigns	Contact Tracing

Modeling Interpretation

Logistic Regression

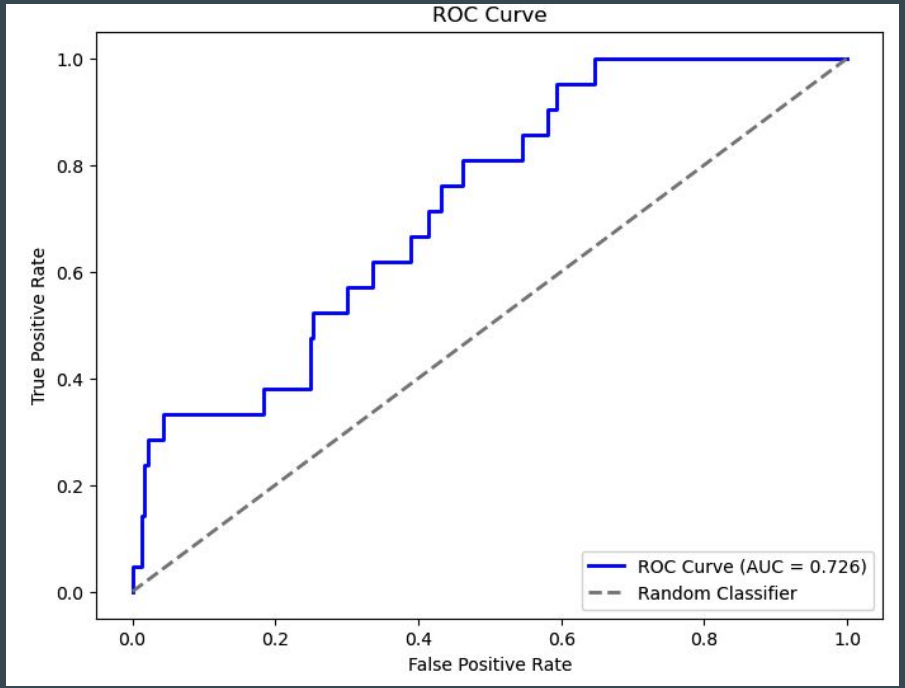
Coefficients of Note

- Odd of death event increase by a factor of 2.45 for each unit increase in 85+ years
- Odd of death event increase by a factor of 1.47 for each unit increase in Northeast Region
- Odd of death event increase by a factor of 3.44 for each unit increase in Doses Distributed Cumulative

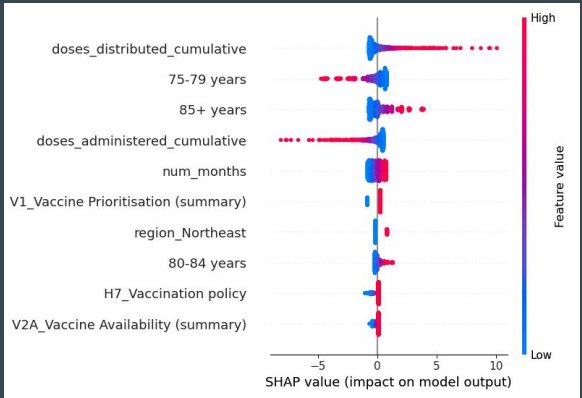
Feature	Coefficient	Odds Ratio
doses_administered_cumulative	-0.989984	0.371583
num_months	0.469462	1.599133
doses_distributed_cumulative	1.236876	3.444834
V2A_Vaccine Availability (summary)	0.197311	1.218123
V1_Vaccine Prioritisation (summary)	0.470216	1.600340
region_Northeast	0.385069	1.469715
H7_Vaccination policy	0.241765	1.273495
85+ years	0.895182	2.447782
80-84 years	0.291748	1.338765
75-79 years	-1.025651	0.358563

Examining Predictions

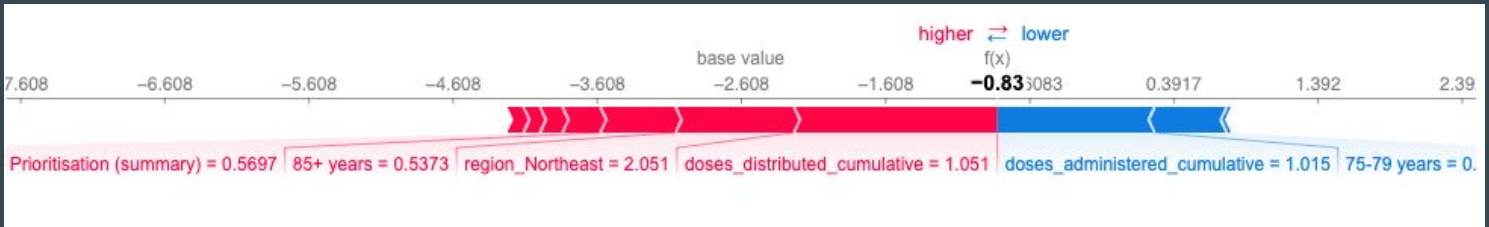
- Accuracy of 92%
- Recall (Sensitivity): 5%
- Precision: 50%



Explaining our Model's Decisions (shapley stuff)



Using SHAP values we can provide an overall explanation of our model and a breakdown of factors influenced specific model predictions.



Cox Proportional Hazard Model

Build CPH Model with majority of variables and examining Hazard Ratios to further prune our explanatory variable set.

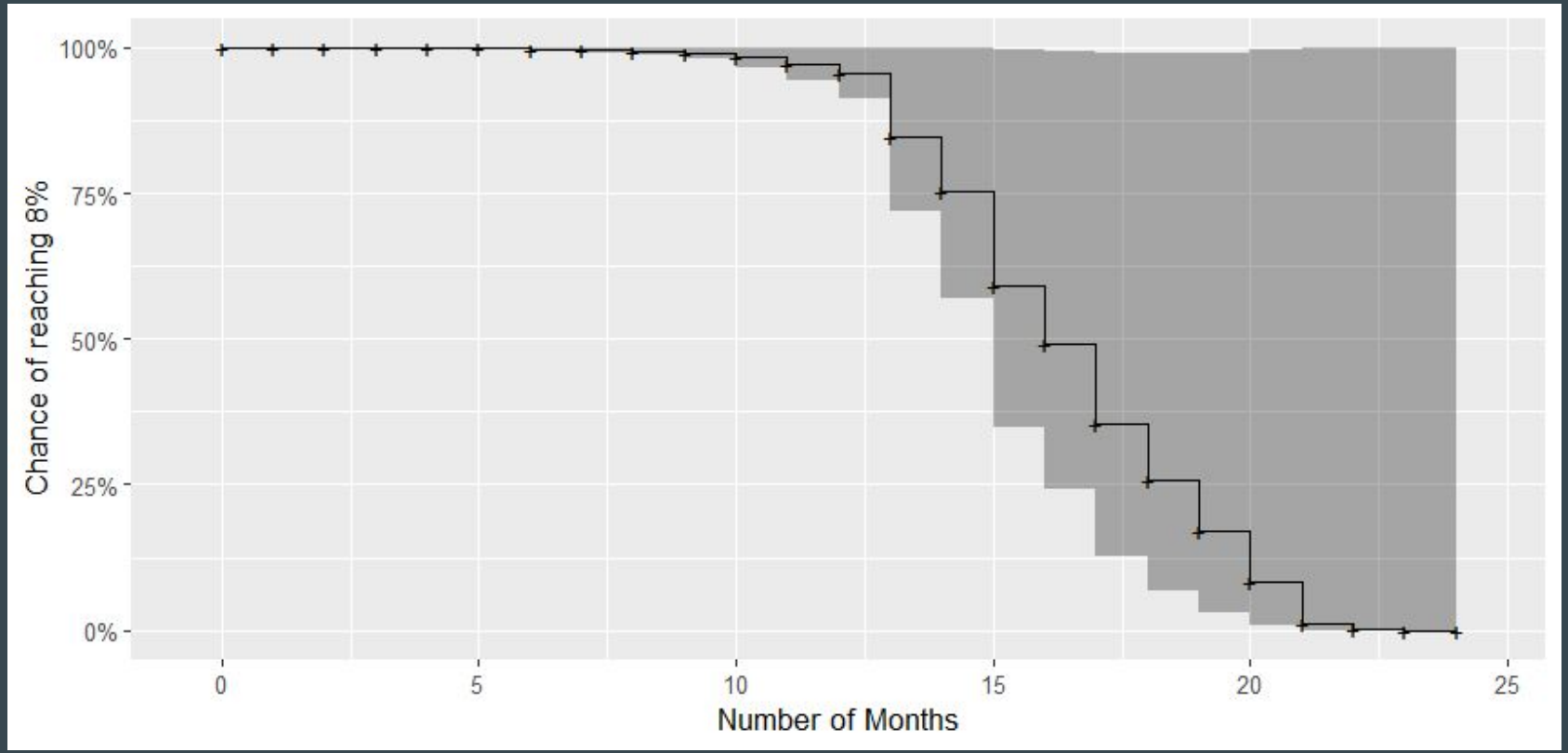
- All age demographic variables had values of 1.0 or within 0.02 of 1.0. Variables that remained were our COVID Policy variables

Hazard Ratios for Survival Model

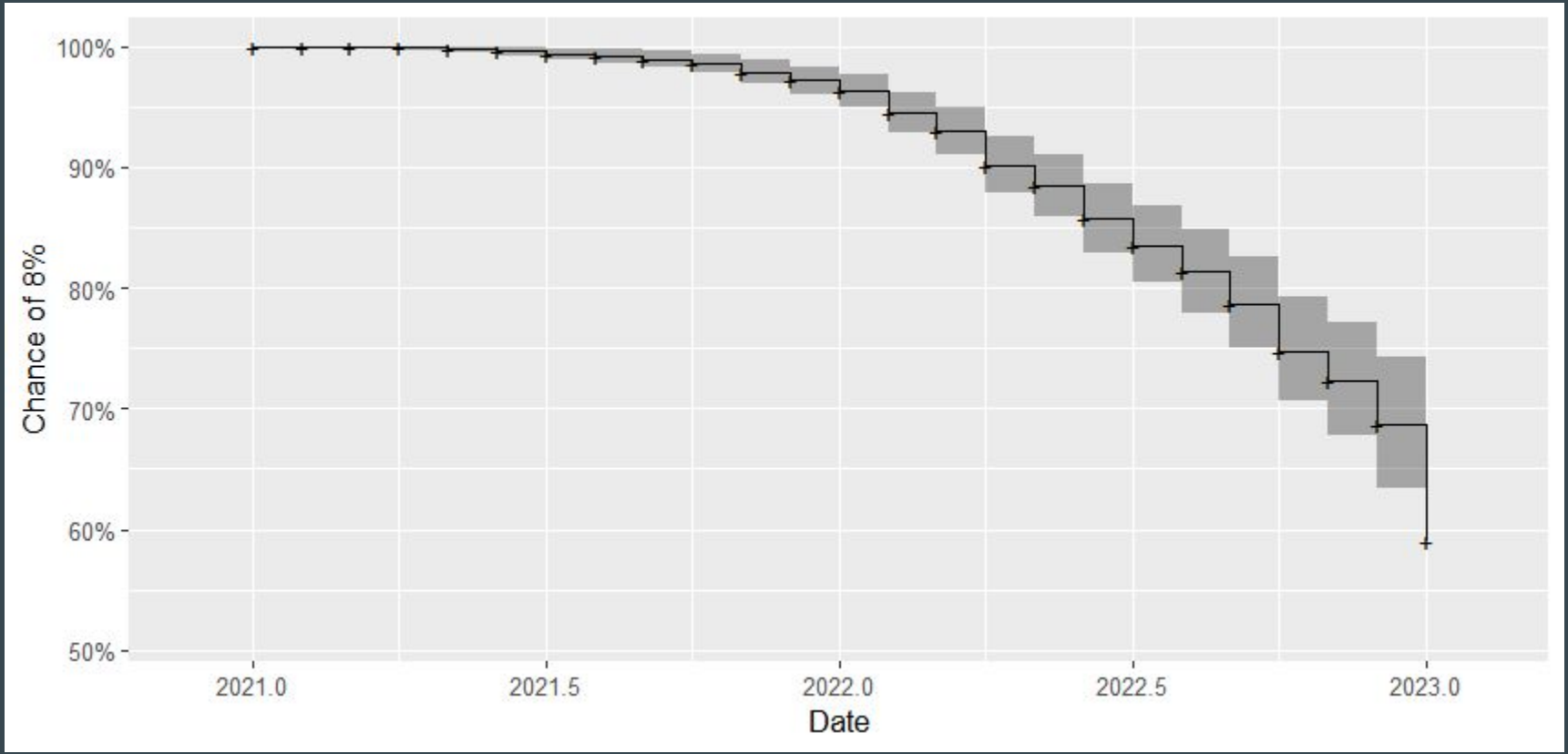
Variable	Hazard_Ratio
regionNortheast	4.747779e+00
regionSoutheast	5.378767e+00
regionSouthwest	4.264578e+00
regionWest	4.541234e-01
V2A_Vaccine.Availability..summary.	1.164012e+00
H8M_Protection.of.elderly.people	4.550751e+00
H7_Vaccination.policy	4.588620e-02
C1M_School.closing	1.271140e+00
C2M_Workplace.closing	1.756992e+00
C3M_Cancel.public.events	1.028338e+00
C5M_Close.public.transport	5.861179e-01
C6M_Stay.at.home.requirements	1.707789e-01
C7M_Restrictions.on.internal.movement	1.538952e-01
C8EV_International.travel.controls	4.309050e+00
E1_Income.support	5.145160e-01
E2_Debt.contract.relief	2.357136e+00
H2_Testing.policy	1.117800e+06
H3_Contact.tracing	1.424549e+00
H6M_Facial.Coverings	1.157846e+00

Documentation:

Cox Proportional Hazard Model



Survival Curves

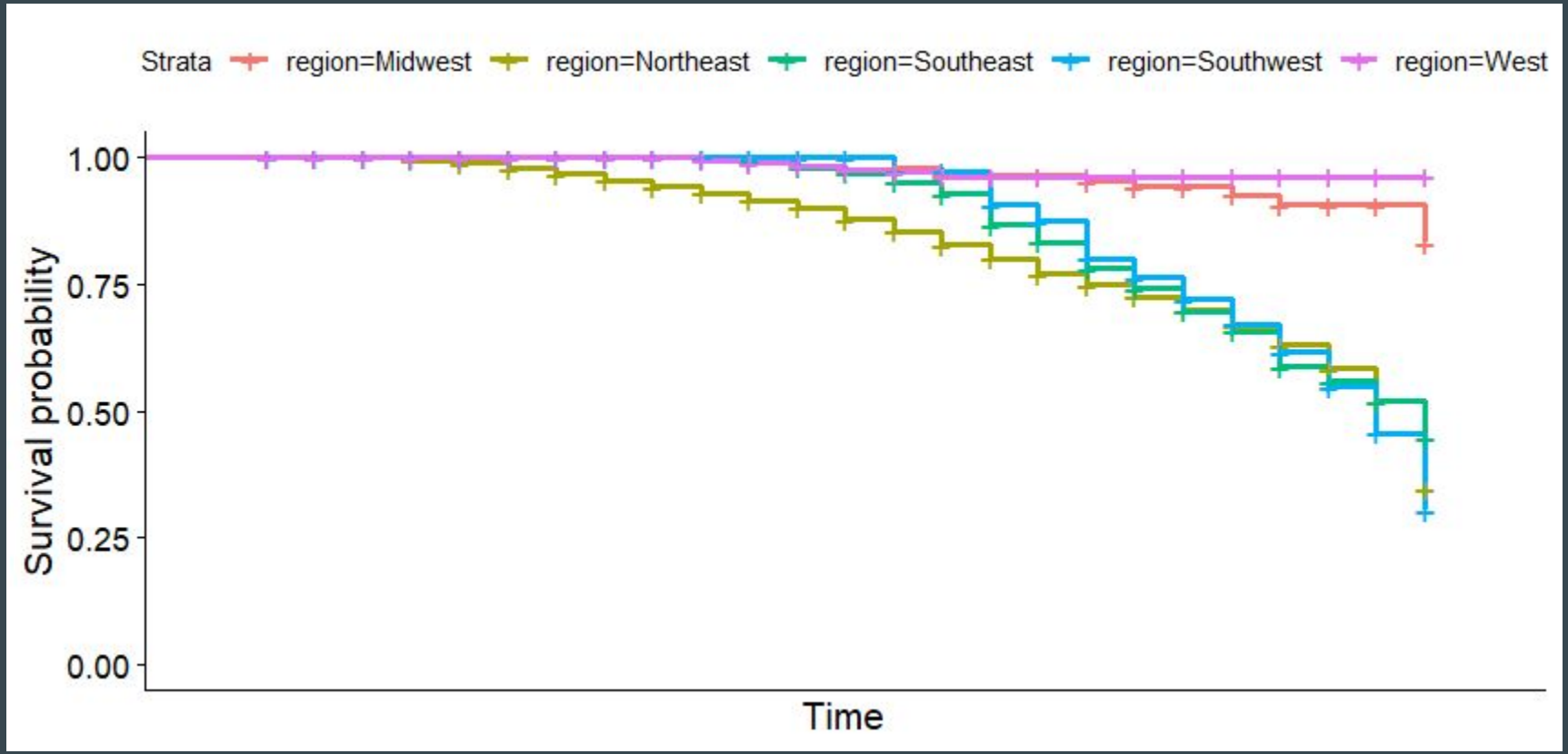


Predictive Metrics

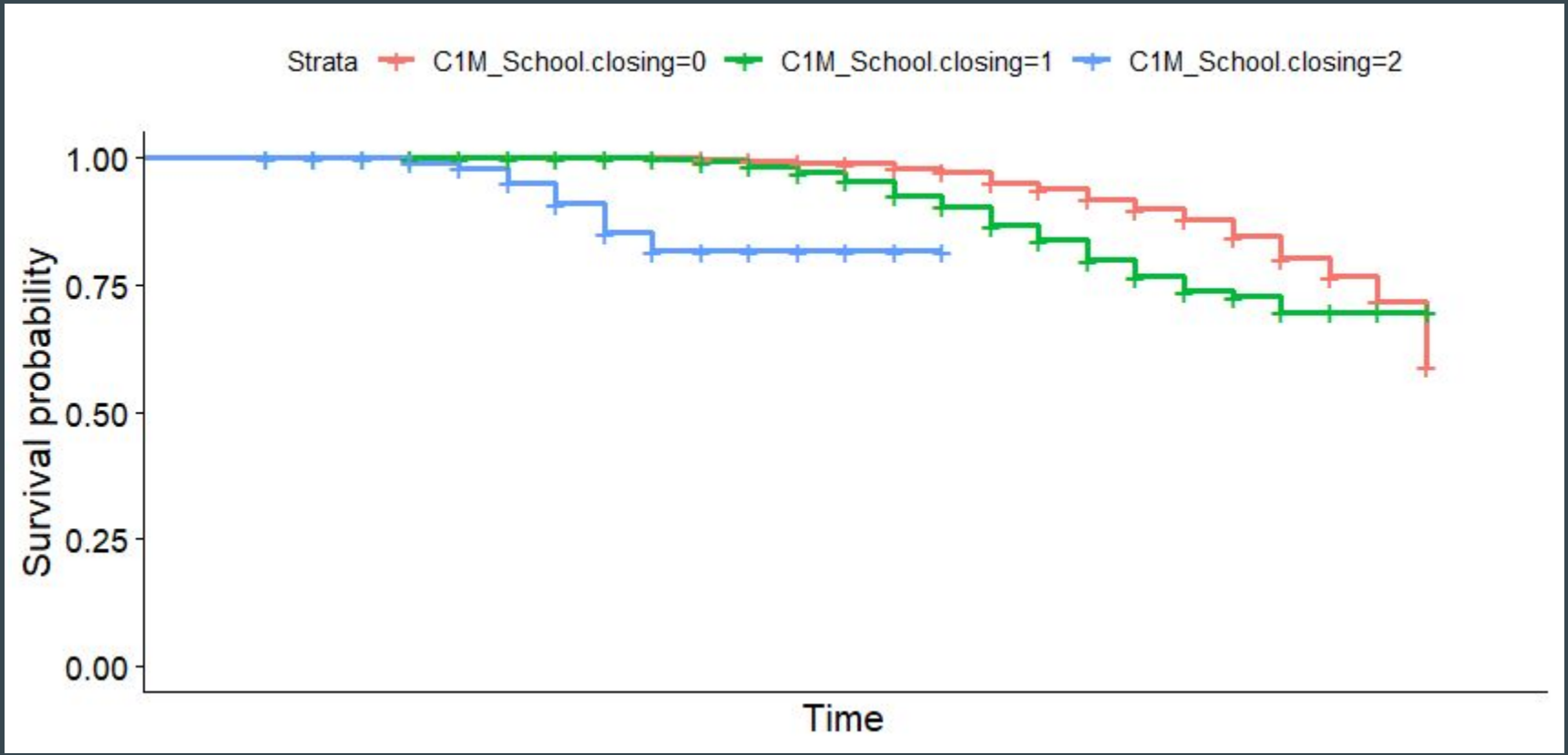
C Index	0.424302
Dxy	-0.1514
S.D.	0.042429
n	1250
missing	0
uncensored	1250
Relevant Pairs	304962
Concordant	129396
Uncertain	0

Partial Effects Charts

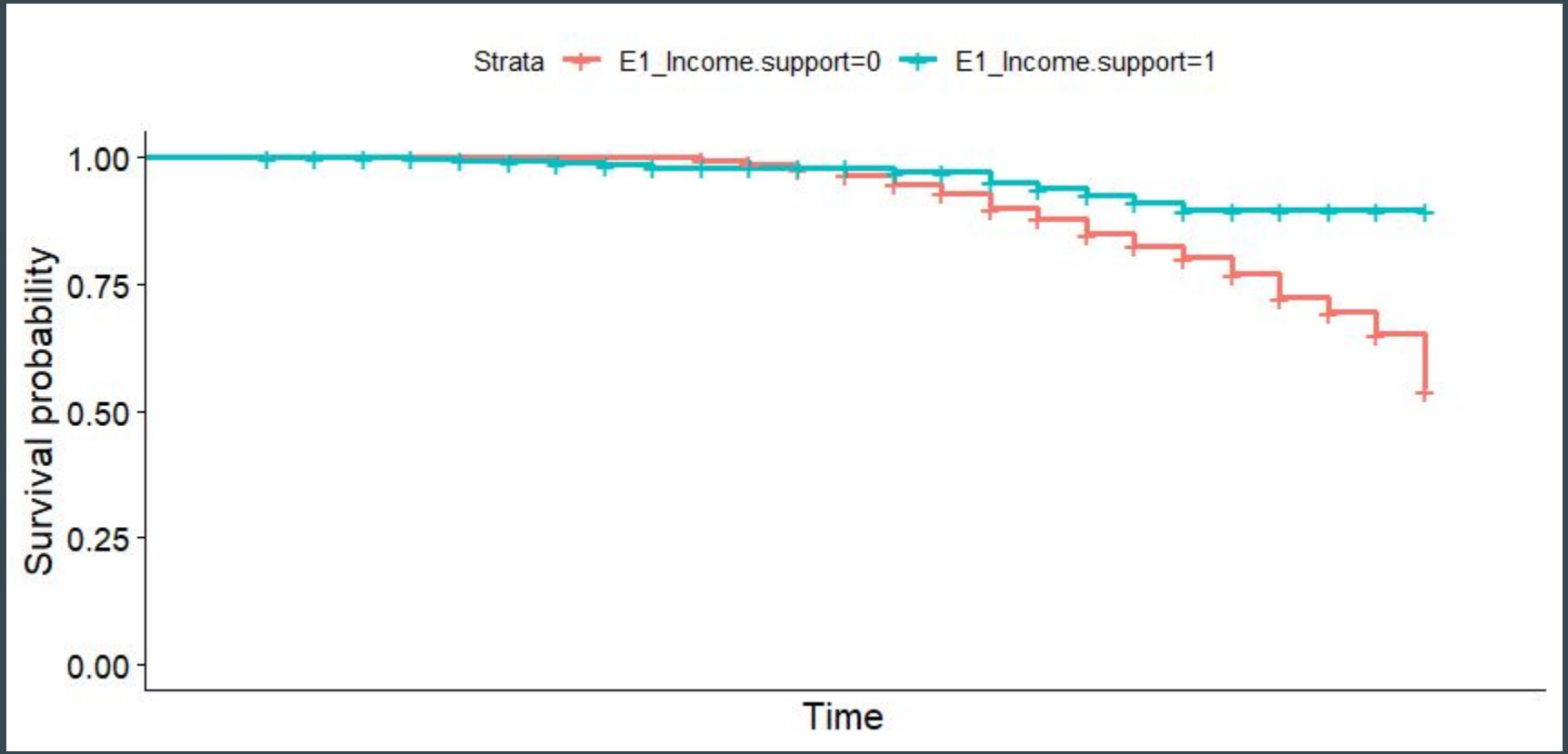
Region


Strata
region=Midwest
region=Northeast
region=Southeast
region=Southwest
region=West
1.00
0.75
0.50
0.25
0.00
Survival probability
Time

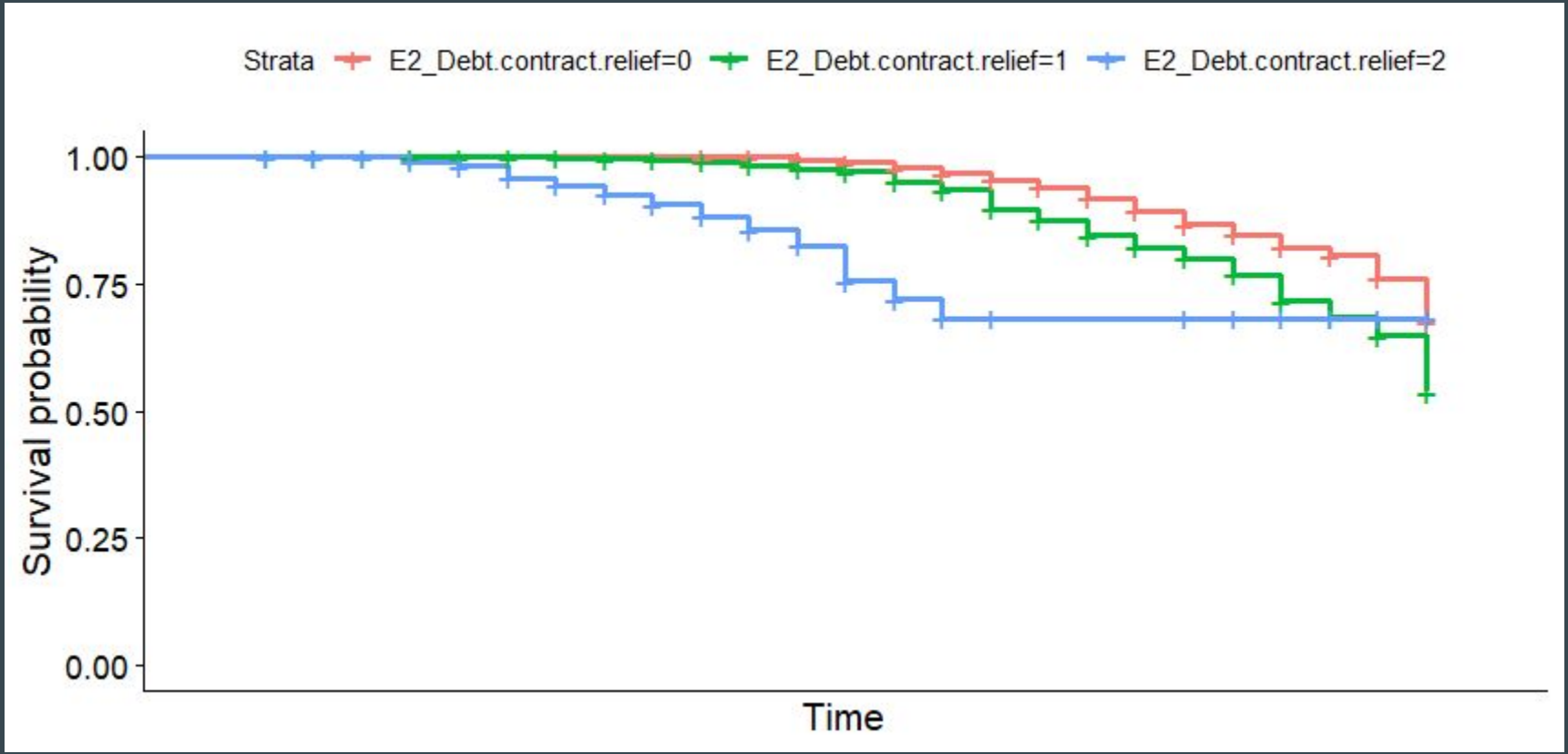
School Closing and COVID Policies



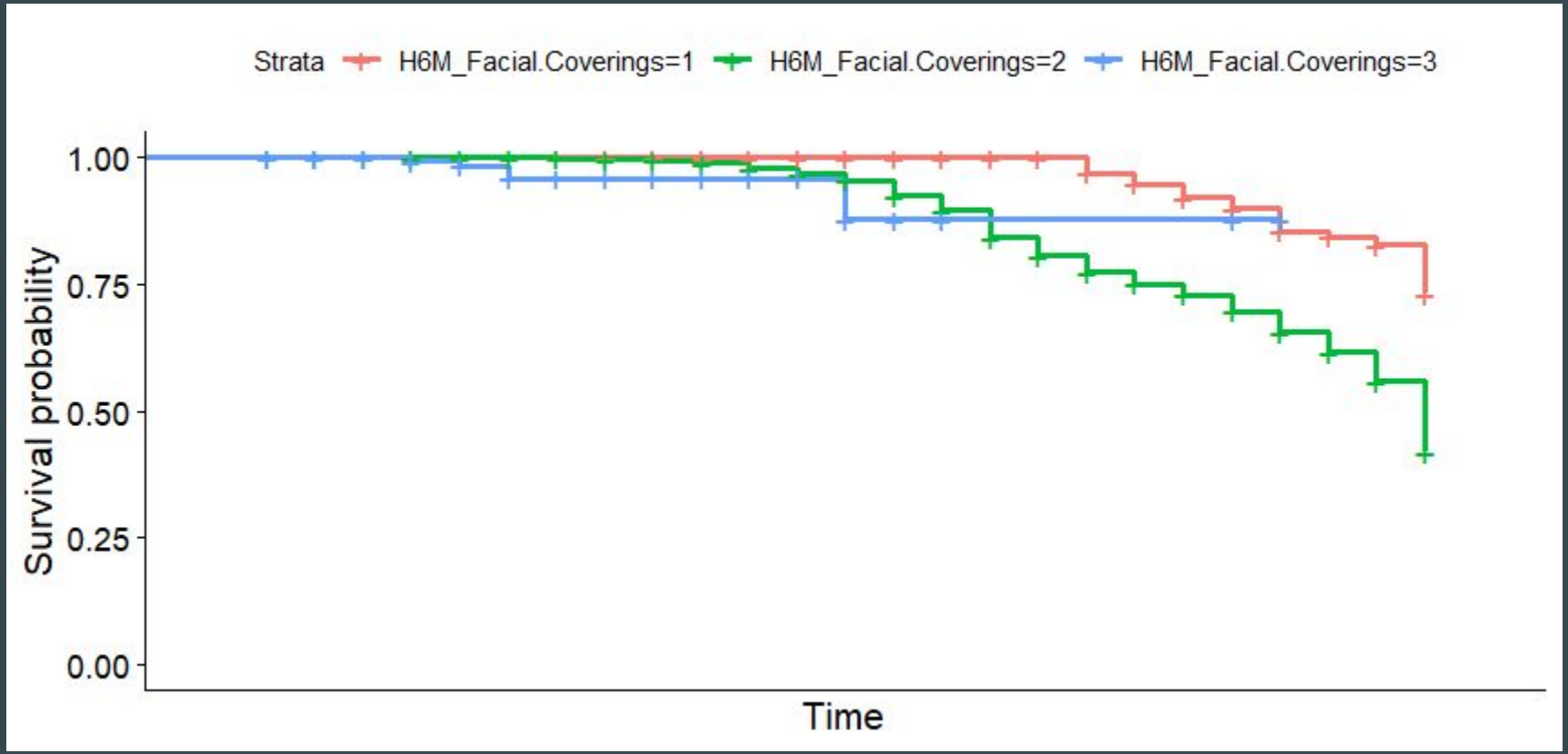
Income Support



Debt Contract Relief



Facial Coverings



Conclusion

Benefits of Dual Model Analysis

Cox Proportional Hazard Model

- Policy oriented model
- Provides insight into what policies implemented at the state level affect survival rate in terms of our event variable
- By utilizing partial effects we can understand exactly how influential features, policies in particular, are in reducing or increasing the chance of a high proportion death event in a given state.

Logistic Regression

- High predictive power
- Mix of demographic and policy features
- Provides predictive insight into when a catastrophic proportion of deaths may occur.
- Using Explainable AI we are able to understand exactly what features directed the decision making in our model

Limitations & Areas for Further Exploration

- Variability in Standardization: Data across states and regions in the U.S lack standardization in their values, posing challenges when performing comparative analysis or survival analysis.

- Sparse Demographic Information: Demographic data is collected somewhat sparsely and may not not completely and accurately reflect population/policy characteristics.

- Aggregating data by month and state might lead to unwanted effects, such as the oversimplification of temporal and spatial patterns, and potentially overlooking localized trends within shorter timeframes.

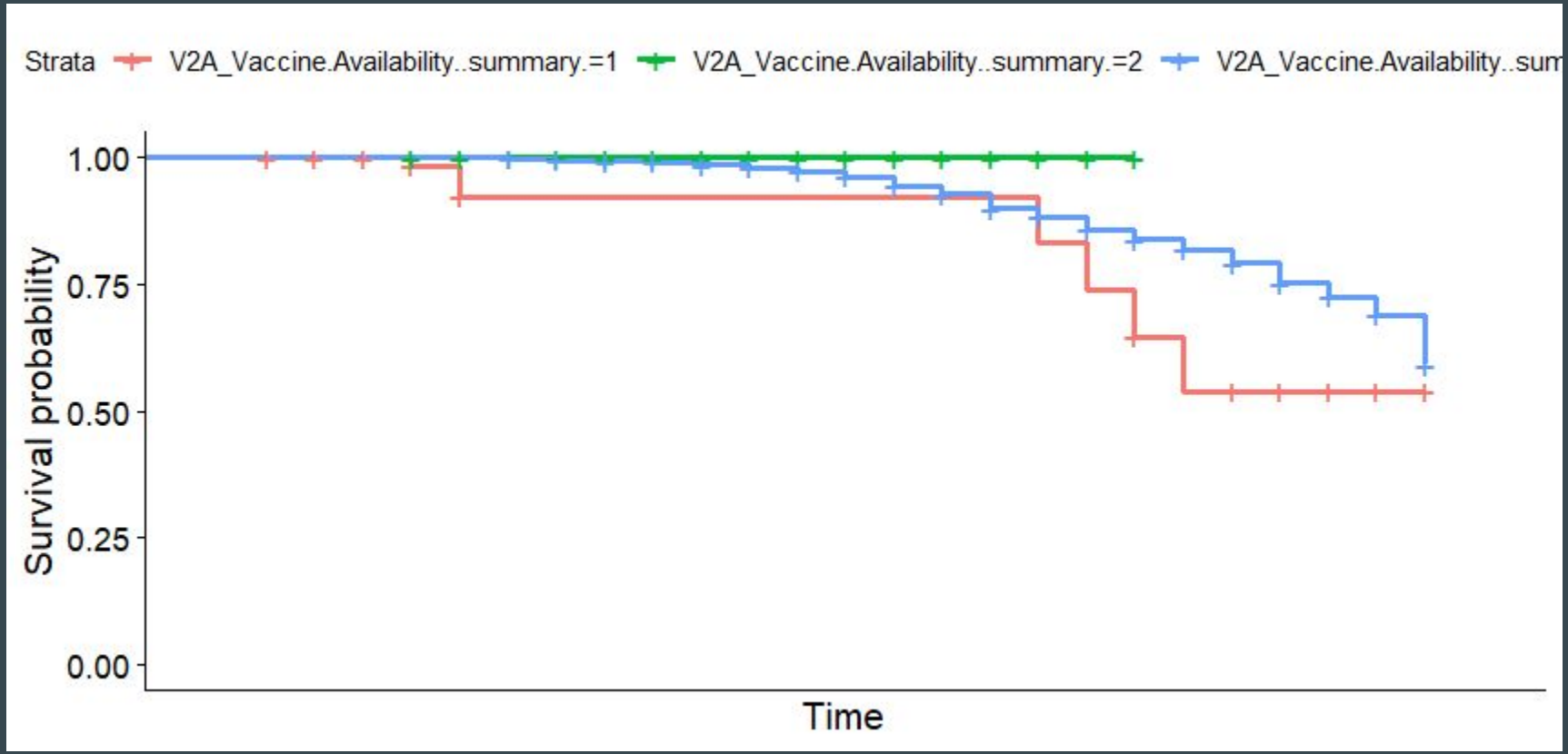
- Future research and data collection efforts should focus primarily on enhancing the standardization of values, improving demographic accuracy, and developing robust modeling techniques to mitigate the impact of the limitations above.

Appendix

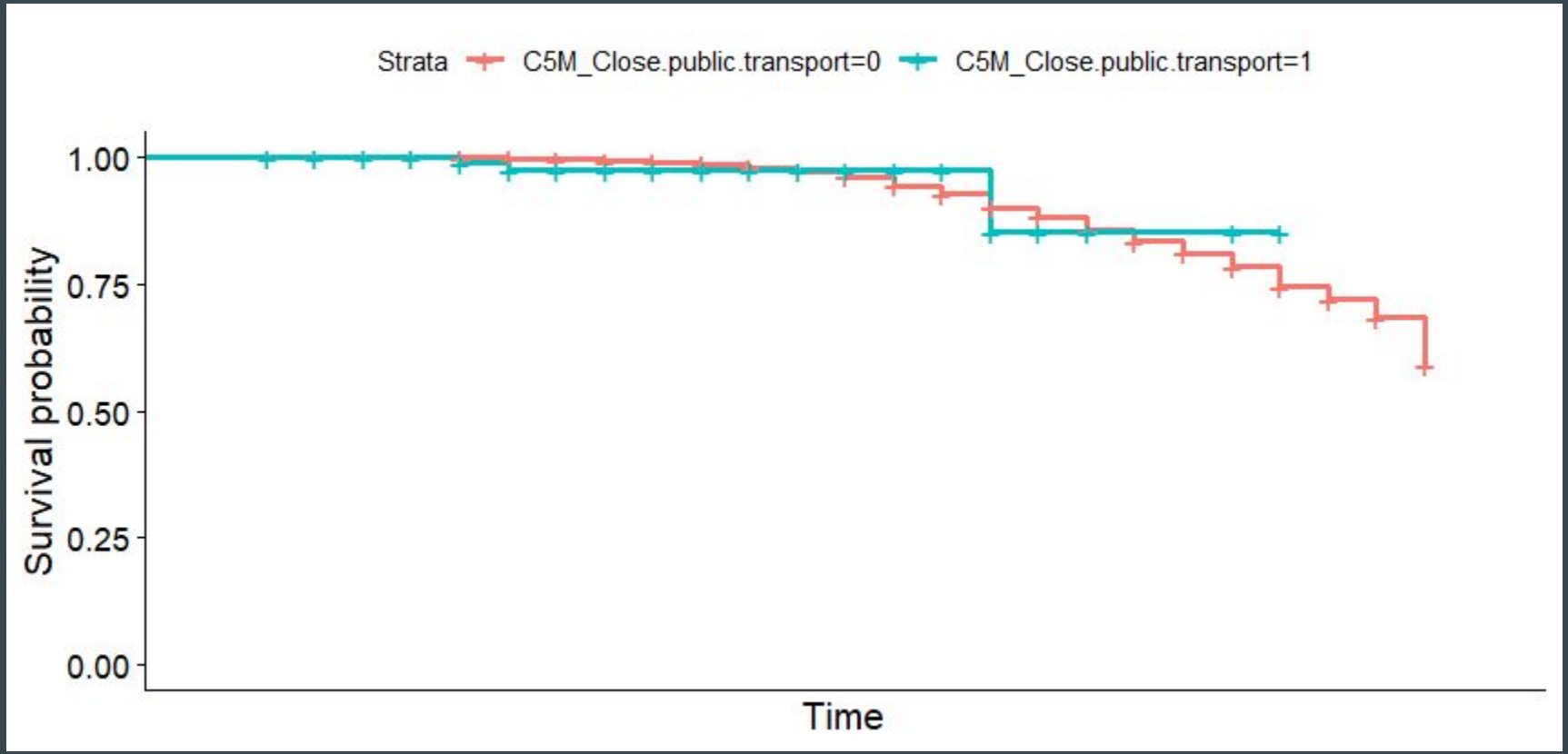
Data Sources

1. Demographic Age Bin Population Projections: https://wonder.cdc.gov/population-projections.html
2. Demographic Population Estimates: https://www.census.gov/data/tables/time-series/demo/popest/2020s-state-total.html
3. COVID Policy Tracker Data: https://github.com/OxCGRT/covid-policy-tracker/blob/master/data/United%20States/OxCGRT_USA_latest.csv
4. COVID Vaccine Tracker: https://www.kaggle.com/datasets/thedevastator/cdc-covid-19-vaccine-tracker
5. COVID-19 Dataset: https://www.kaggle.com/datasets/kalilurrahman/new-york-times-covid19-dataset

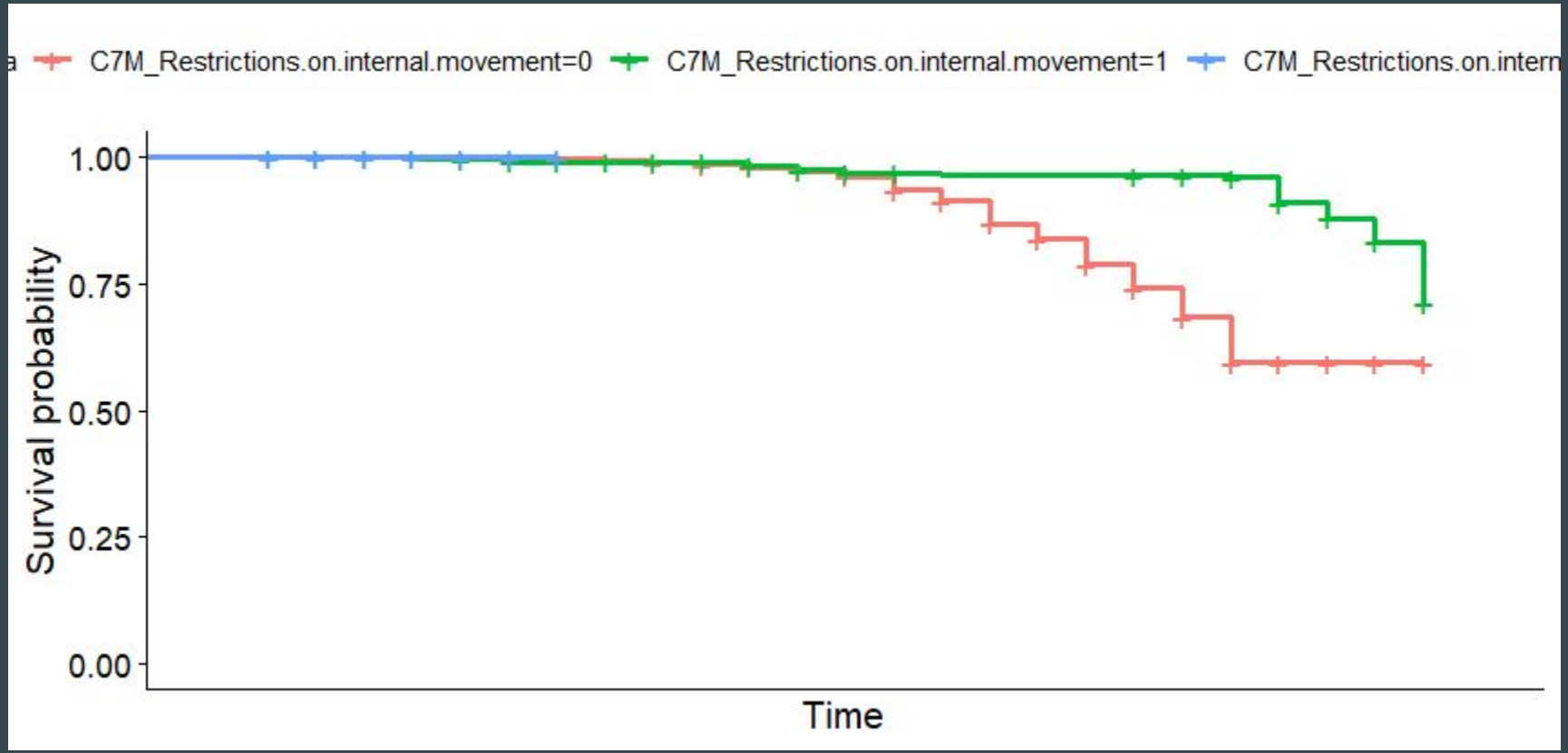
Partial Dependence Plots (Felipe)



Partial Dependence Plots (Felipe)



Partial Dependence Plots (Felipe)



Partial Dependence Plots (Felipe)