UPV EHU

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

A Multi-State Modeling Framework for Complex Disease Trajectories: A Population-Based Analysis of 400,000 COVID-19 Cases in the Basque Country

Klaus Langohr Inma Arostegui Guadalupe Gómez Melis

July 1, 2026

Outline

COVID-19 Data Set From the Basque Country

The Multi-State Model

Modelling Reinfection of COVID-19

Future Work and Conclusions

COVID-19 Data Set From the Basque Country

Basque Country Project: Nationwide COVID-19 Population Data

EUSKO JAURLARITZA
GOBIERNO VASCO
OSASUN SAILA
DEPARTAMENTO DE SALUD

Basque Country Project: Strategic Goals

Goal 1: Mapping Patient Trajectories (before Reinfection)

Analyze disease progression using nationwide population-based registry data.

- **Aim:** Demonstrate how MSMs capture the full **patient journey** better than standard survival models.

Goal 2: Modeling COVID-19 Reinfection

Quantify the risk of a new episode ($\geq$ 120 days post-primary infection).

- **Challenge:** Accounting for **landmark constraints** and **competing risk of death**.
- **Definition:** Reinfection as a distinct state transition within the longitudinal history.

Basque Country COVID-19 Cohort

Data set

Source: Basque Country Health Service (Osakidetza).

Population: Individuals with a 1st SARS-CoV-2 positive test.

- $n = 377\,285$ patients.
- Recruitment: 3/2020 – 1/2022.
- Follow-up: Until **April 9, 2022**.

Hospital: Admission within **15 days** (post+).

Follow-up:

- **Non-hospitalized:** 90 days from diagnosis.
- **Hospitalized:** 90 days from diagnosis or discharge (whichever occurred last).

Clinical Factors

Comorbidity: Charlson Index (CCI).

Treatments: Based on ATC classification.

Polymedication: ≥ 6 concurrent treatments.

Immunity status: Time-dependent covariate based on vaccination and diagnosis.

Study Periods (Waves)

P1 (Lockdown): *3/2020 – 6/2020*

P2 (Transition): *7/2020 – 12/2020*

P3 (Alpha/Delta): *1/2021 – 13/12/2021*

P4 (Omicron): *14/12/2021 – 9/1/2022*

Vaccination campaign started in Dec 2020.

Basque Country COVID-19 Cohort

Data set

Source: Basque Country Health Service (Osakidetza).

Population: Individuals with a 1st SARS-CoV-2 positive test.

- $n = 377\,285$ patients.
- Recruitment: 3/2020 – 1/2022.
- Follow-up: Until **April 9, 2022**.

Hospital: Admission within **15 days** (post+).

Follow-up:

- **Non-hospitalized:** 90 days from diagnosis.
- **Hospitalized:** 90 days from diagnosis or discharge (whichever occurred last).

Clinical Factors

Comorbidity: Charlson Index (CCI).

Treatments: Based on ATC classification.

Polymedication: ≥ 6 concurrent treatments.

Immunity status: Time-dependent covariate based on vaccination and diagnosis.

Study Periods (Waves)

P1 (Lockdown): *3/2020 – 6/2020*

P2 (Transition): *7/2020 – 12/2020*

P3 (Alpha/Delta): *1/2021 – 13/12/2021*

P4 (Omicron): *14/12/2021 – 9/1/2022*

Vaccination campaign started in Dec 2020.

Descriptive Analysis

Key Findings ($n = 377\,285$)

Gender:
Balanced cohort, but men overrepresented in ICUs (**68.5%**).

Age:

- Hospitalizations increase with higher age.
- Ages **60–80**: $\approx$ 58% of ICU cases.
- Patients **80+**: rarely sent to ICU.

Deprivation index:
Socioeconomic status was not a limiting factor for healthcare access.

Nursing homes:
6.2% hospitalized vs. only 0.6% ICU (Triage/Frailty effect).

Mortality Patterns ($n = 5\,149$ deaths [1.4%])

Gender:

- In-hospital deaths (**53%**): more men (**57.5%**).
- Out-of-hospital (**47%**): more women (**58.4%**).

Age:

- Mean age at death: > 80 **years**.
- Mortality below age **50** is residual.
- Patients **90+**: **37.2%** of out-of-hospital, **20.3%** of in-hospital deaths.

Deprivation index:
Remarkable equity, with deaths evenly distributed across quintiles ($\approx$20% each).

Nursing homes:
2.3% of the cohort, but account for **47.6%** of out-of-hospital deaths.

Descriptive Analysis

Key Findings ($n = 377\,285$)

Gender:

Balanced cohort, but men overrepresented in ICUs (**68.5%**).

Age:

- Hospitalizations increase with higher age.
- Ages **60–80**: $\approx$ 58% of ICU cases.
- Patients **80+**: rarely sent to ICU.

Deprivation index:

Socioeconomic status was not a limiting factor for healthcare access.

Nursing homes:

6.2% hospitalized vs. only 0.6% ICU (Triage/Frailty effect).

Mortality Patterns ($n = 5\,149$ deaths [1.4%])

Gender:

- In-hospital deaths (**53%**): more men (**57.5%**).
- Out-of-hospital (**47%**): more women (**58.4%**).

Age:

- Mean age at death: $>$ 80 **years**.
- Mortality below age **50** is residual.
- Patients **90+**: **37.2%** of out-of-hospital, **20.3%** of in-hospital deaths.

Deprivation index:

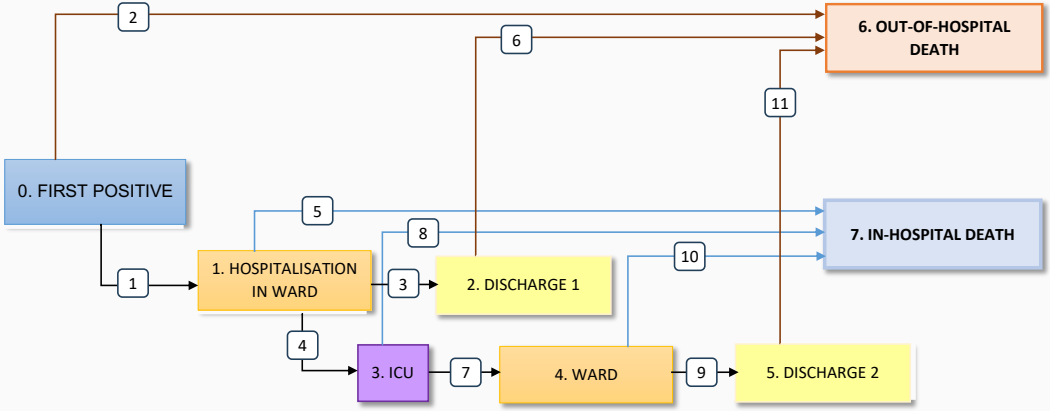
Remarkable equity, with deaths evenly distributed across quintiles ($\approx$20% each).

Nursing homes:

2.3% of the cohort, but account for **47.6%** of out-of-hospital deaths.

The Multi-State Model

MSM with 8 States and 11 Transitions



Journal of Statistical Software

January 2011, Volume 38, Issue 7. *http://www.jstatsoft.org/*

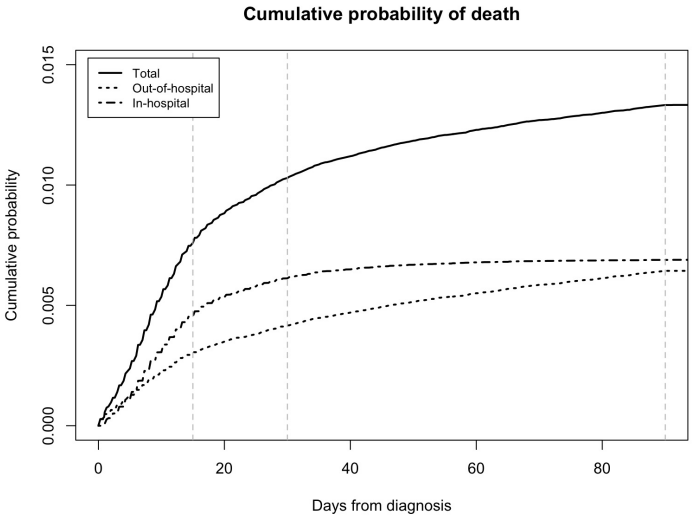
mstate: An R Package for the Analysis of Competing Risks and Multi-State Models

Liesbeth C. de Wreede
Leiden University
Medical Center

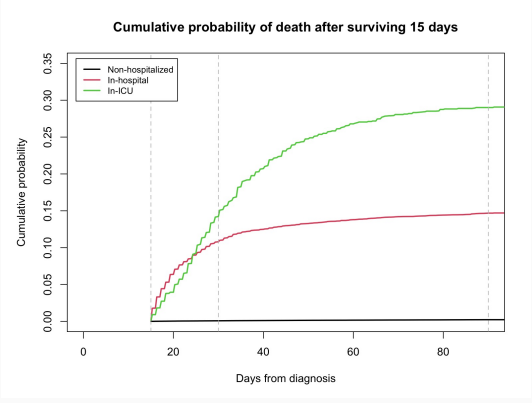
Marta Fiocco
Leiden University
Medical Center

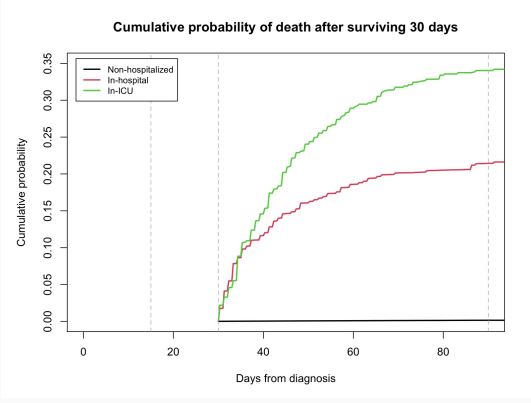
Hein Putter
Leiden University
Medical Center

Nonparametric Analysis: Evolution of Mortality Risk over Time



Nonparametric Analysis: Evolution of Mortality Risk over Time





Semiparametric Analysis: Summary of Transition-Specific Cox Models

	Transition		Covariate							
	Initial	Final	Sex: M vs. W	Age: 8 categories	NH: Yes vs. No	DI: Q1 to Q5	Period: 4 periods	Immunity	CCI: 5 categories	Polimed.: ≥6 vs. <6
1	Positive	Hospitalization	—(+)—>	—(+)—>	—(−)—>	—(+)—>	~~~	—(−)—>	—(+)—>	—(+)—>
2	Positive	Death (OUT)	—(+)—>	—(+)—>	—(+)—>	NS	~~~	—(−)—>	—(+)—>	NS
3	Hospitalization	Discharge	—(−)—>	—(−)—>	—(−)—>	—(−)—>	~~~	—(+)—>	—(−)—>	NS
4	Hospitalization	ICU	—(+)—>	~~~	—(−)—>	NS	~~~	—(−)—>	—(−)—>	NS
5	Hospitalization	Death (IN)	—(+)—>	—(+)—>	—(+)—>	NS	~~~	NS	—(+)—>	NS
6	Discharge	Death (OUT)	—(+)—>	—(+)—>	—(+)—>	—(−)—>	~~~	NS	—(+)—>	—(−)—>
7	ICU	Discharge ICU	—(−)—>	~~~	—(+)—>†	NS	~~~	NS	—(−)—>	NS
8	ICU	Death (IN)	NS	—(+)—>	NS	NS	~~~	NS	—(+)—>	NS
9	Discharge ICU	Discharge	NS	~~~	NS	—(−)—>	~~~	NS	—(−)—>	—(−)—>†
10	Discharge ICU	Death (IN)	NS	—(+)—>	NS	NS	~~~	—(+)—>	—(+)—>	—(+)—>†
11	Discharge	Death (OUT)	NS	—(+)—>	—(+)—>†	NS	NS	—(+)—>	—(+)—>	NS

NS: non-significant and †: p-values between 0.05 and 0.1

$-(+)\rightarrow$: positive linear relation, $-(-)\rightarrow$: negative linear relation, $\sim\!\sim$: non-linear relation

Semiparametric Analysis: Summary of Transition-Specific Cox Models

	Transition		Covariate							
	Initial	Final	Sex: M vs. W	Age: 8 categories	NH: Yes vs. No	DI: Q1 to Q5	Period: 4 periods	Immunity	CCI: 5 categories	Polimed.: ≥6 vs. <6
1	Positive	Hospitalization	—(+)—>	—(+)—>	—(−)—>	—(+)—>	~~~	—(−)—>	—(+)—>	—(+)—>
2	Positive	Death (OUT)	—(+)—>	—(+)—>	—(+)—>	NS	~~~	—(−)—>	—(+)—>	NS
3	Hospitalization	Discharge	—(−)—>	—(−)—>	—(−)—>	—(−)—>	~~~	—(+)—>	—(−)—>	NS
4	Hospitalization	ICU	—(+)—>	~~~	—(−)—>	NS	~~~	—(−)—>	—(−)—>	NS
5	Hospitalization	Death (IN)	—(+)—>	—(+)—>	—(+)—>	NS	~~~	NS	—(+)—>	NS
6	Discharge	Death (OUT)	—(+)—>	—(+)—>	—(+)—>	—(−)—>	~~~	NS	—(+)—>	—(−)—>
7	ICU	Discharge ICU	—(−)—>	~~~	—(+)—>†	NS	~~~	NS	—(−)—>	NS
8	ICU	Death (IN)	NS	—(+)—>	NS	NS	~~~	NS	—(+)—>	NS
9	Discharge ICU	Discharge	NS	~~~	NS	—(−)—>	~~~	NS	—(−)—>	—(−)—>†
10	Discharge ICU	Death (IN)	NS	—(+)—>	NS	NS	~~~	—(+)—>	—(+)—>	—(+)—>†
11	Discharge	Death (OUT)	NS	—(+)—>	—(+)—>†	NS	NS	—(+)—>	—(+)—>	NS

NS: non-significant and †: p-values between 0.05 and 0.1

$-(+)\rightarrow$: positive linear relation, $-(-)\rightarrow$: negative linear relation, $\sim\sim$: non-linear relation

Semiparametric Analysis: Summary of Transition-Specific Cox Models

Gender & Age

Gender:

Men face higher risk of hospitalization/ICU, but sex becomes **non-significant** once in the ICU.

Age & Mortality:

Older age is related to higher death rates in every state. Even after recovery, older adults still face greater health risks.

Advanced age is the primary predictor of **death after discharge**.

The ICU "Ceiling":

Risk of ICU entry drops for patients **80+**; triage shifts toward ward-based or palliative care.

Immunity, Frailty & Equity

Vaccination:

Reduces risk of hospitalization and death, as well as times to discharge.

Nursing Homes:

Strongest predictor of out-of-hospital death; less ICU stays due to ICU "ceiling".

Social Equity:

Deprivation index was **non-significant** for survival; healthcare access remained equitable.

Clinical Complexity:

CCI and polymedication are universal drivers of poor clinical trajectories.

Semiparametric Analysis: Summary of Transition-Specific Cox Models

Gender & Age

Gender:

Men face higher risk of hospitalization/ICU, but sex becomes **non-significant** once in the ICU.

Age & Mortality:

Older age is related to higher death rates in every state. Even after recovery, older adults still face greater health risks.

Advanced age is the primary predictor of **death after discharge**.

The ICU "Ceiling":

Risk of ICU entry drops for patients **80+**; triage shifts toward ward-based or palliative care.

Immunity, Frailty & Equity

Vaccination:

Reduces risk of hospitalization and death, as well as times to discharge.

Nursing Homes:

Strongest predictor of out-of-hospital death; less ICU stays due to ICU "ceiling".

Social Equity:

Deprivation index was **non-significant** for survival; healthcare access remained equitable.

Clinical Complexity:

CCI and polymedication are universal drivers of poor clinical trajectories.

Modelling Reinfection of COVID-19

Basque Country Project: Strategic Goals

Goal 1: Mapping Patient Trajectories (before Reinfection)

Analyze disease progression using nationwide population-based registry data.

- **Aim:** Demonstrate how MSMs capture the full **patient journey** better than standard survival models.

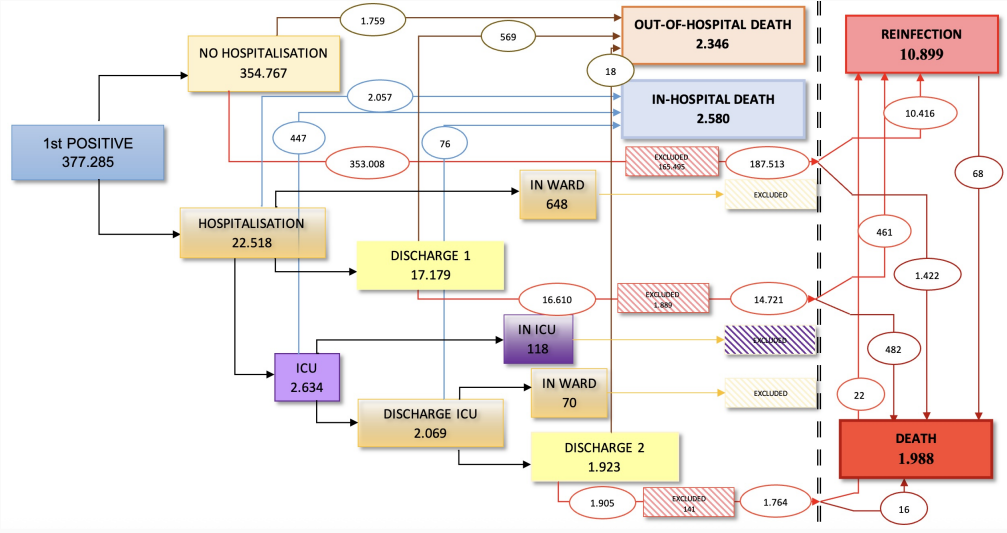
Arostegui, Langohr, Barrio, Larrea-Agirre, Quintana, Gómez-Melis. *Multistate modeling of disease progression for population-based registry data: Application to Covid-19. Submitted.*

Goal 2: Modeling COVID-19 Reinfection

Quantify the risk of a new episode ($\geq$ 120 days post-primary infection).

- **Challenge:** Accounting for **landmark constraints** and **competing risk of death**.
- **Definition:** Reinfection as a distinct state transition within the longitudinal history.

Complete flowchart with sample sizes and number of event



Reinfection Analysis: Understanding Disease Dynamics

COVID-19 Reinfection Definition: A new positive test occurring at least 120 days after the first infection in individuals not currently hospitalized.

Remark: Hospitalized patients were considered eligible for reinfection analysis only after discharge and recovery stabilization.

The Study Cohort

- **Total SARS-CoV-2 positive cases:** $n = 377\,285$
- **At risk of reinfection:** $n = 203\,998$ (54.1%)
- **Confirmed reinfections:** $n = 10\,899$ (5.34%)

Main Goal & Motivation

- Identify key factors linked to reinfection risk.
- Analyze how disease progression after the first infection relates to reinfection risk.

Hypothesis

Reinfection analysis estimates the risk of an individual experiencing a **repeated infection** after overcoming the primary infection, providing a deeper understanding of long-term disease dynamics.

Reinfection Analysis: Understanding Disease Dynamics

COVID-19 Reinfection Definition: A new positive test occurring at least 120 days after the first infection in individuals not currently hospitalized.

Remark: Hospitalized patients were considered eligible for reinfection analysis only after discharge and recovery stabilization.

The Study Cohort

- **Total SARS-CoV-2 positive cases:** $n = 377\,285$
- **At risk of reinfection:** $n = 203\,998$ (54.1%)
- **Confirmed reinfections:** $n = 10\,899$ (5.34%)

Main Goal & Motivation

- Identify key factors linked to reinfection risk.
- Analyze how disease progression after the first infection relates to reinfection risk.

Hypothesis

Reinfection analysis estimates the risk of an individual experiencing a **repeated infection** after overcoming the primary infection, providing a deeper understanding of long-term disease dynamics.

Methodological Approaches for Modeling Recurrence

$\boldsymbol{M_1}$: MSM based on all times to events prior to reinfection; $\boldsymbol{M_2}$: Reinfection sub-model

1. Stratified Approach

Joint or separate **competing risks models** for M_2 based on the **final state** reached in M_1.

- Simple & intuitive.
- Loses power due to fragmentation.

2. Sequential ($M_2|M_1$)

Augmented Covariate Models:

- Integration of baseline and clinical covariates (e.g., days in ICU).
- Inclusion of **interaction terms** with the reached state.

Integrated Risk Score (IRS):

- **Predicted Probability:** Uses information prior to landmark t_L to estimate the probability of reaching specific states
- **Linear Predictors:** Utilizes M_1 summary statistics to account for individual heterogeneity and refine estimates.

3. Joint Approach

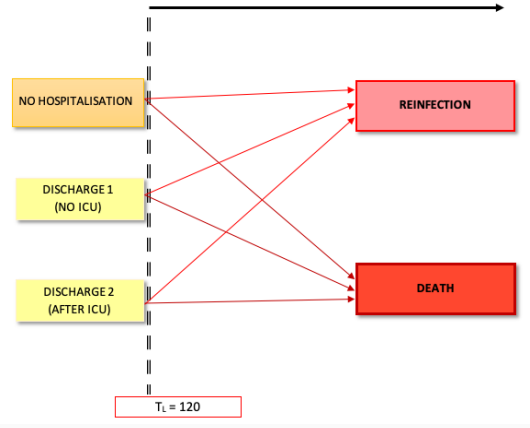
A global extension of the M_1 .

- Simultaneous estimation.
- Higher computational cost.

Stratified Approach: Competing Risk Model From Landmark Time $t_L = 120$

Main Objective:

Model risk of reinfection based on the prior course of events considering death as competing event.



Stratified Approach: Competing Risk Model From Landmark Time $t_L = 120$

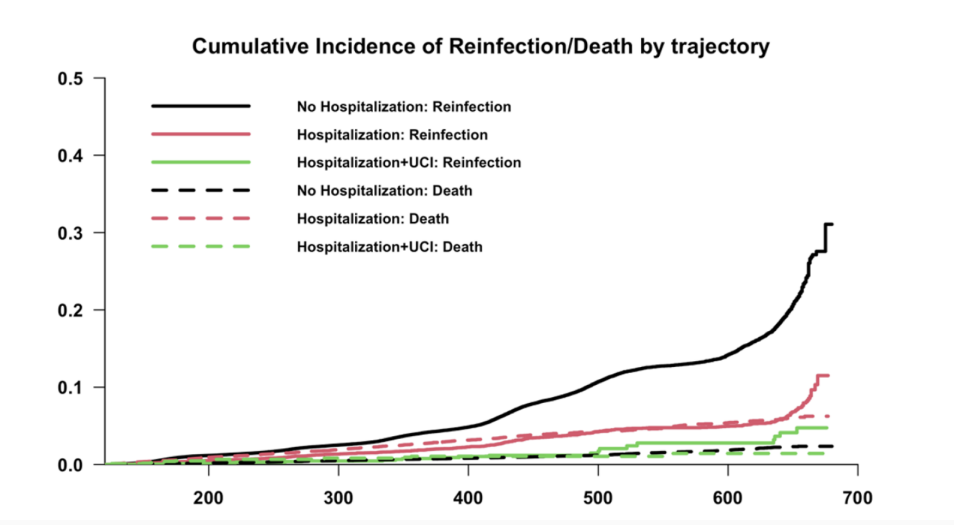
Statistical Framework

Nonparametric estimation of the cumulative incidence function and fit of **cause-specific Cox model**. Baseline and time-dependent covariates are introduced at the landmark time $t_L = 120$ days, taking into account the individual's prior clinical state.

Cohort at Risk ($n = 203\,998$ individuals at t_L):

		Reinfections		Deaths	
Prior State (at t_L)	**Count (n)**	***n***	**%**	***n***	**%**
1. Non-hospitalized	187 513	10 416	5.55%	1422	0.76%
2. Hospitalized (No ICU)	14 721	461	3.13%	482	3.27%
3. Hospitalized (ICU)	1 764	22	1.25%	16	0.91%
Total	**203 998**	**10 899**	**5.34%**	**1920**	**0.94%**

Stratified Approach: Nonparametric Estimation of the Cumulative Incidence Function



Stratified Approach: Stratified Cause-Specific Cox Models

Summary	Effect of covariates					
Outcome	Sex (M)	Age	CCI	NH	Period*Baseline immunity	Δ Immunity
Reinfection	–	–	+	+	+	–
Death	+	+	+	+	+/–	–

Note: +/- signs indicate the direction of effect on Reinfection or Death risk (CCI: Charlson Comorbidity Index, NH: Nursing Home).

Key Findings

Age: Older age groups show consistently lower reinfection risk compared with the reference group (< 30).

Temporal Dynamics: Later pandemic periods are associated with increased reinfection and mortality risks.

Basal Immunity: Prior immunity significantly attenuates mortality risk in later periods.

Effect Heterogeneity: Covariate effects differ substantially between reinfection and death.

Further Analyses

Cause-specific Cox models among people previously hospitalized (with and without ICU stay) do not reveal effect on reinfection risk of time in hospital and ICU, respectively, .

Laura Belmonte: *Development of multistate and competing-risk methodologies for characterizing the recurrence of events in population cohorts: application to COVID-19 data.* Trabajo final de grado.

Approach 2: Integrated Risk Score (IRS)

Motivation

Summarize an individual's complex clinical history from M_1 into a single, statistically powerful value: the **Predicted State Occupation Probability**.

Probability at Landmark Time t_L: For an individual i in risk group $k \in \{1, 2, 3\}$, the IRS is:

$$Z_i^k = P(X_{i,t=t_L} = k \mid X_{i,t_0} = 0),$$

based on **all** transition-specific Cox models from M_1:

$$\alpha_{jk}(t \mid \mathbf{X}_i) = \lambda_{jk0}(t) \exp\left(\beta_k^\top \mathbf{X}_i(t)\right).$$

Idea: This score captures the **cumulative risk footprint** of an individual before they even reach the landmark "at-risk" state for reinfection.

Approach 2: Preliminary Results & Current Challenges

Preliminary Findings ($n = 300$)

Previously hospitalized individuals show a **lower reinfection risk** (possible immunity effect).

Higher occupation probabilities (Z_i^k) are associated with **higher reinfection risk**.

Covariates included in M_1: sex, age, CCI, and infection period.

Methodological Strengths

Automatically captures interactions and non-linearities from model M_1.

Bridges multi-state history with competing-risk prediction.

Computational Challenges

Extremely high computational cost.

$>$ 1 hour for only $n = 300$ individuals.

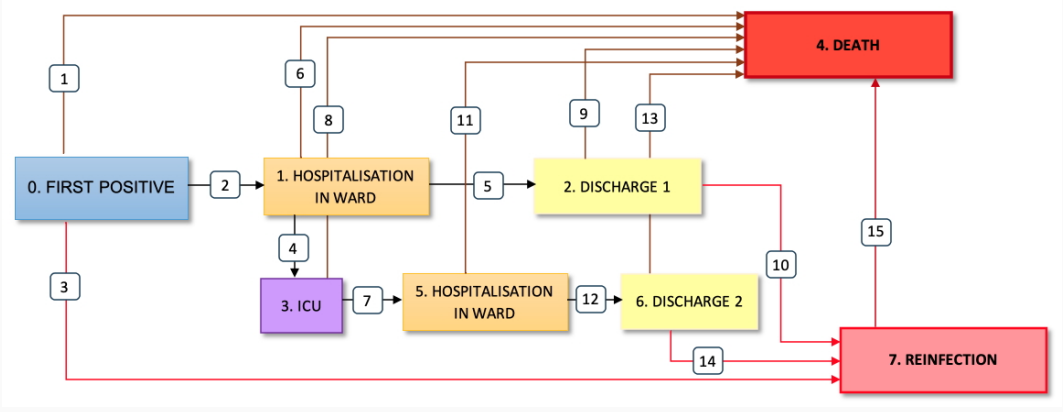
Population-scale analyses ($n \approx 400,000$) become computationally prohibitive without massive parallelization.

Inference Challenges

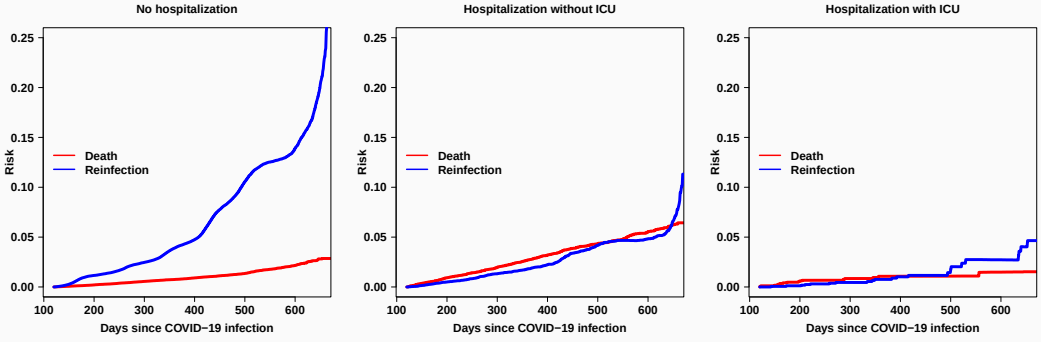
Because Z_i^k is a predicted quantity, analytical standard errors for M_2 are not directly available. **Current requirement:** Multi-stage bootstrap procedures, increasing computational burden.

Ongoing Research: The IRS framework is under active methodological development. Current results are preliminary, but the approach shows strong potential for integrating complex clinical histories into interpretable reinfection risk prediction.

Approach 3: Joint M_1M_2 Modeling



Joint Model: Nonparametric Estimation of Reinfection and Death Risks



Note: Similar to competing risk analysis, but considers also death after reinfection.

Joint Model: Semiparametric Analysis and Comments

Preliminar results

Semiparametric Cox models adjusted for sex, comorbidity, and deprivation index also show higher reinfection risk among people without prior hospitalization.

PROs

Unified Inference:
Uses all population data within a single likelihood framework, avoiding information loss from separate analyses.

Sojourn Effects:
Allows to incorporate clinical history through transition-specific hazards and captures the effect of hospitalization duration on reinfection risk.

CONs

Higher computational cost:
Fit of complete model with $n = 377\,285$ to estimate reinfection risk for subset requires more computation than necessary.

Modeling Complexity:
Might be more difficult to implement if temporal constraints have to be considered.

Joint Model: Semiparametric Analysis and Comments

Preliminar results

Semiparametric Cox models adjusted for sex, comorbidity, and deprivation index also show higher reinfection risk among people without prior hospitalization.

PROs

Unified Inference:
Uses all population data within a single likelihood framework, avoiding information loss from separate analyses.

Sojourn Effects:
Allows to incorporate clinical history through transition-specific hazards and captures the effect of hospitalization duration on reinfection risk.

CONs

Higher computational cost:
Fit of complete model with $n = 377\,285$ to estimate reinfection risk for subset requires more computation than necessary.

Modeling Complexity:
Might be more difficult to implement if temporal constraints have to be considered.

Future Work and Conclusions

Conclusions & Future Research

Main Conclusions & Reinfection Modeling

Multi-state models provide a unified framework to reconstruct the complete **COVID-19 clinical trajectory** from diagnosis to recovery, death, and reinfection.

Reinfection risk depends not only on baseline characteristics, but also on the individual's previous disease pathway.
The proposed approaches showed complementary strengths:

- Stratified competing-risk models: interpretable and efficient.
- IRS framework: clinically informative summaries.
- Joint MSMs: statistically coherent global modeling.

Challenges & Future Research

Extending MSMs to nationwide longitudinal cohorts introduces **substantial challenges**:

- Landmark and temporal constraints.
- Competing risks and recurrent events.
- Massive computational burden.

Current work focuses on:

- Computation and interpretation of IRS.
- Dynamic prediction of reinfection risk.
- Extension to other chronic and recurrent diseases using registry data.

Conclusions & Future Research

Main Conclusions & Reinfection Modeling

Multi-state models provide a unified framework to reconstruct the complete **COVID-19 clinical trajectory** from diagnosis to recovery, death, and reinfection.

Reinfection risk depends not only on baseline characteristics, but also on the individual's previous disease pathway.
The proposed approaches showed complementary strengths:

- Stratified competing-risk models: interpretable and efficient.
- IRS framework: clinically informative summaries.
- Joint MSMs: statistically coherent global modeling.

Challenges & Future Research

Extending MSMs to nationwide longitudinal cohorts introduces **substantial challenges**:

- Landmark and temporal constraints.
- Competing risks and recurrent events.
- Massive computational burden.

Current work focuses on:

- Computation and interpretation of IRS.
- Dynamic prediction of reinfection risk.
- Extension to other chronic and recurrent diseases using registry data.

Acknowledgements

Grants:

- PID2024-156800OB-I00: Advances in Statistical Modelling for Health Sciences. Financiado por MICIU/AEI and FEDER, UE.
- RED2024-153680-T: Biostatnet 2025: Fortaleciendo la Investigación de Excelencia en Bioestadística a Nivel Nacional e Internacional Ministerio de Ciencia, Innovación y Universidades.
- PID2023-148033OB-C21: Statistics for Health Sciences: Advances in Survival Analysis and Clinical Trials. Financiado por MICIU/AEI /10.13039/501100011033 and FEDER, UE.
- 2021 SGR 01421: GRBIO: Grup de Recerca en Bioestadística i Bioinformàtica. Agència de Gestió d'Ajuts Universitaris i de Recerca.

¡Muchas gracias!

Moltes gràcies!

Eskerrik asko!

Thank you!