Beyond Illness-Death: Capturing Real-World Disease Progression with Multistate Models

Susie Bayarri Biostatistics Sessions, organised by the IGTP Program in Public Health and Primary Healthcare (CORE)

Guadalupe Gómez Melis[1] K. Langohr[1] I. Arostegui[2]

April 9, 2026

[1]Department of Statistics and Operations Research (UPC)
Institute for Research and Innovation in Health (IRIS)
[2]Department of Mathematics, University of the Basque Country (UPV/EHU)

Outline

Motivation: DIVINE and Basque Country project

Motivating Data: Hospital vs. Population Cohorts

To capture real-world disease dynamics, we analyze data from two complementary perspectives

Two Distinct Data Sources:

1. Hospital COVID-19 data from Catalunya: Detailed clinical events for each patient.
2. Nationwide COVID-19 population data from the Basque Country: Broad view of the entire population's health records.

Goal: Multistate Framework to handle both scales of information.

Dynamic evaluat**I**on of CO**VI**D-19 cli**N**ical stat**E**s and their prognostic factors to improve the intra-hospital patient management

DIVINE

Dynamic evaluatIon of coVId-19 cliNical statEs

Multidisciplinary Composition

A synergistic collaboration of 21 researchers between Clinical and Biostatistics teams

Clinical Experts

- **Infectious Diseases:** Front-line specialists managing COVID-19 patients.
- **Clinical Pharmacology:** Experts in drug interactions and trial safety.
- **Role:** Define clinical questions, design questionnaires, and ensure evidence-based measures.

Methodology Team: Biostatisticians

- Experts in survival analysis and complex modeling.
- Extensive experience in biomedical collaboration.
- **Role:** Implement state-of-the-art methodology and ongoing data analysis.

Team Diversity: $\approx$ 50/50 Clinical-Statistical Split
Gender Balance: 11 Women out of 21 Members (52%)

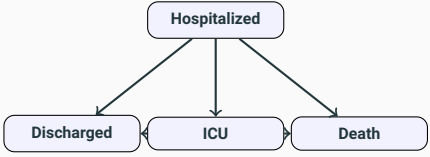
DIVINE's Goals

Primary Objective

Identify the most clinically relevant prognostic factors for respiratory distress, ICU admission, death, or discharge in a cohort of hospitalized adults with confirmed COVID-19.

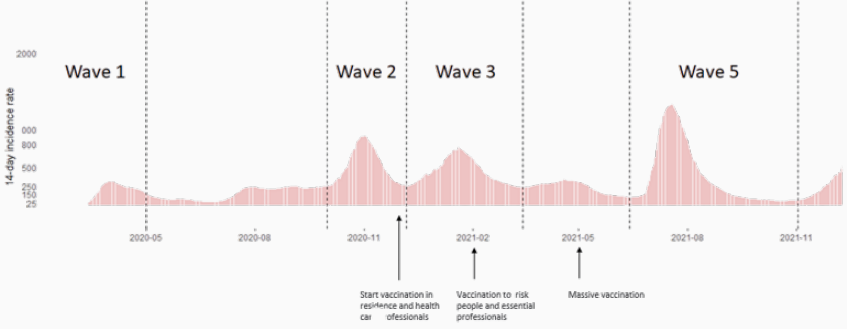
- **Model** disease course using **MSM** to account for intermediate clinical states.
- **Develop** a dynamic prediction tool for early identification of high-risk patients.
- **Estimate** incubation periods for the first time in a Catalan cohort.
- **Assess** the evolution of the disease and the patients' profile over time.

Clinical Pathways



DIVINE's Data: Electronic form in REDCap designed *ad hoc*

- Prospective multicenter cohort study of hospitalized adults with PCR-proven SARS-CoV-2 infection
- MetroSud: 5 hospitals from Barcelona Metropolitan area
- Age $\geq$ 18
- More than 5,000 hospitalized COVID-19 patients during the first five waves of the pandemic.



Multistate model: 6 states and 8 transitions

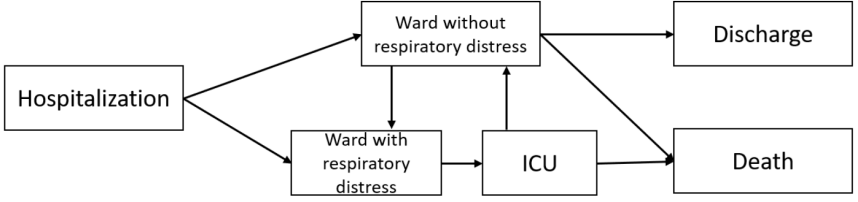
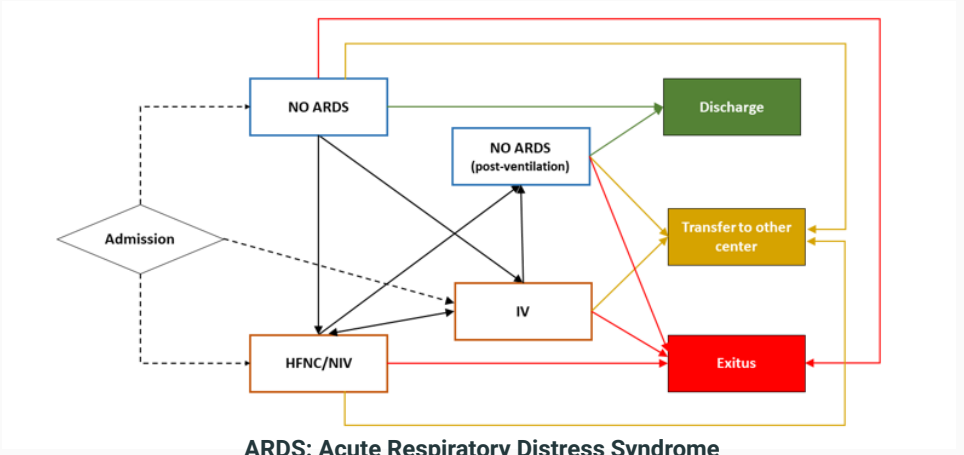


Figure 1. Multi-state model for our study

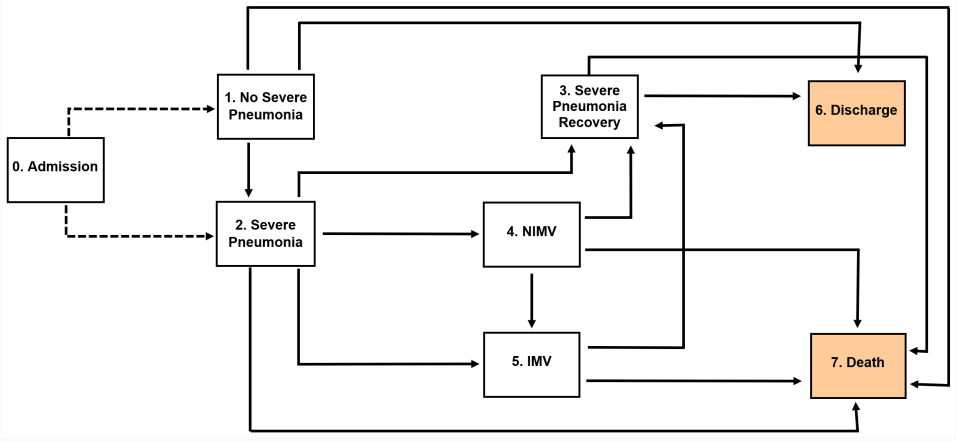
Multistate model: 8 states and 13 transitions



ARDS: Acute Respiratory Distress Syndrome

HFNC: High-Flow Nasal Cannula, IV: Invasive Ventilation

Complex Multistate model: 7 states and 14 transitions



4. NIMV: Non Invasive Mechanical Ventilation, 5. IMV: Invasive Mechanical Ventilation (ICU)

MSMpred Shiny app: Two main objectives

1. **Fit** a multistate model to specific data (Based on mstate R package[2])
2. **Predict** the evolution for a new individual

`https://www.grbio.eu/pubs/MSMpred/` and `https://www.grbio.eu/pubs/MSMpred2/`

[1] Garmendia Bergés, L., Cortés Martínez, J. and Gómez Melis, G. (2023). *MSMpred: Interactive modelling and prediction of individual evolution via multistate models*. BMC Med Res Methodol 23, 126

[2] de Wreede,L.C., Fiocco, M. and Putter, H. (2011). mstate: An R Package for the Analysis of Competing Risks and Multi-State Models. Journal of Statistical Software, 38(7), 1-30.

MSMpred: model specification

MSMpred helps to define the **paths** between states.

Define the transitions

From:

nopneum

To:

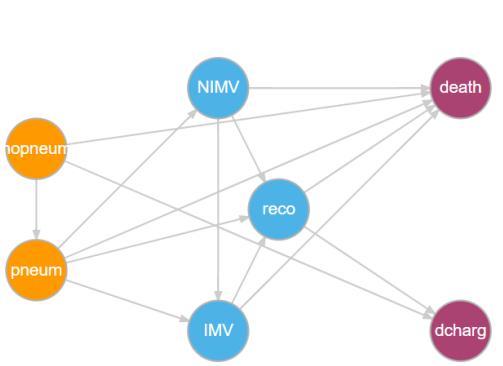
nopneum

Add Delete

Number of events for each transition

	nopneum	pneum	reco	NIMV	IMV	dcharg	death
nopneum	0	405	0	0	0	1386	43
pneum	0	0	213	217	163	0	26
reco	0	0	0	0	0	447	12
NIMV	0	0	109	0	98	0	10
IMV	0	0	137	0	0	0	124
dcharg	0	0	0	0	0	0	0

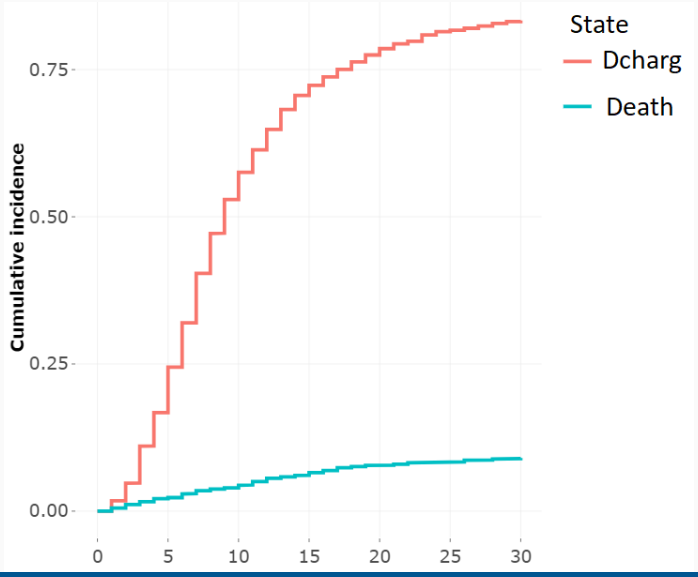
Multistate model diagram



MSMpred: model specification

	sex	age	psi	card_vasc	safi	charlson_fact	crprot	lympho
nopneum -> pneum	✔	✔	✔				✔	
nopneum -> dcharg	✔	✔				✔		
nopneum -> death	✔	✔		✔				
pneum -> reco	✔	✔	✔					
pneum -> NIMV	✔	✔	✔					
pneum -> IMV	✔	✔					✔	
pneum -> death	✔	✔				✔		
reco -> dcharg	✔	✔				✔		
reco -> death	✔	✔						
NIMV -> reco	✔	✔	✔					
NIMV -> IMV	✔	✔	✔					
NIMV -> death	✔	✔						
IMV -> reco	✔	✔	✔					
IMV -> death	✔	✔	✔	✔				

MSMpred: Cumulative incidence of absorbing states



MSMpred: Hazard ratios for the fitted model

Transition-specific Cox models: $\alpha_{hj}(t|\boldsymbol{Z}) = \alpha_{hj,0}(t)\exp(\beta'\boldsymbol{Z})$

MSMpred allows us to obtain the risks associated with the transitions by means of the **Hazard Ratios**: $\text{HR}_{Z_l} = \exp(\beta_l)$

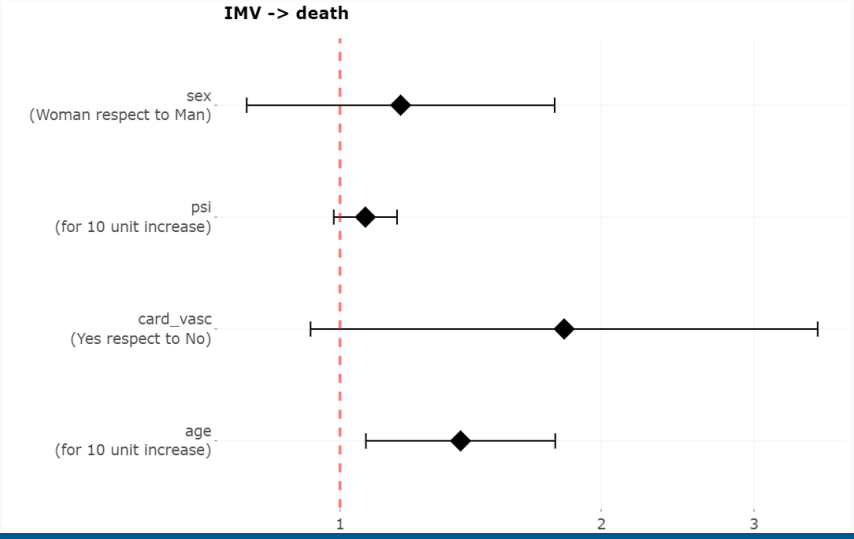
Table of model coefficients:

Show 10 entries Search: (IMV -> death)

	coef	HR (95%CI)	p-value
sex (IMV -> death)	0.189	1.21 (0.80, 1.83)	0.372
age (IMV -> death)	0.029	1.03 (1.00, 1.06)	0.027
psi (IMV -> death)	0.006	1.01 (1.00, 1.02)	0.153
card_vasc (IMV -> death)	0.612	1.84 (0.94, 3.62)	0.075

Showing 1 to 4 of 4 entries (filtered from 45 total entries) Previous 1 Next

MSMpred: forest plot for hazard ratios



MSMpred: predictions

Predictions

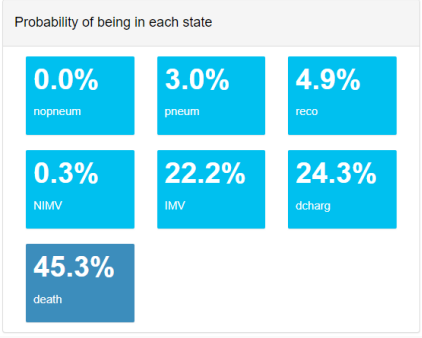
MSMpred provides the **probabilities** of being in a specific state for a new hospitalized patient.

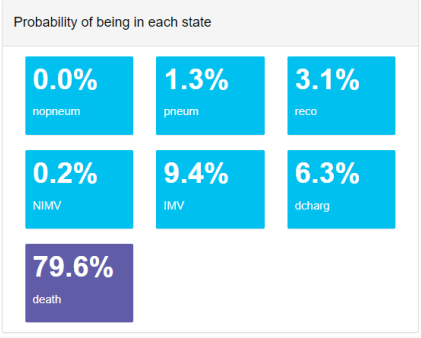
Initial state: severe pneumonia

Individual 1: man, 80 y.,
psi = 50, no CDV, low CCI.

Follow-up time: 30 days

Individual 2: man, 80 y.,
psi = 110, CDV, very high CCI.





DIVINE's Related References:

2020 *Modeling the Coronavirus Disease 2019 Incubation Period: Impact on Quarantine Policy* Pak; Langoh; Ning; Cortés; Gómez; Shen. Mathematics.

2022 *SARS-Cov-2 incubation period according to vaccination status during the fifth COVID-19 wave in a tertiary-care center in Spain: a cohort study* Cortés, J.; Pak, D.; Abelenda-Alonso, G.; Langohr, K.; Ning, J.; Rombauts, A.; Colom, M.; Shen, Y.; Gómez Melis. BMC infectious diseases.

2023 *MSMpred: Interactive modelling and prediction of individual evolution via multistate models*. Garmendia Cortés and Gómez Melis. BMC Medical Research Methodology.

2023 *Characteristics and Outcomes by Ceiling of Care of Subjects Hospitalized with COVID-19 During Four Waves of the Pandemic in a Metropolitan Area: A Multicenter Cohort Study.* Pallarès et al. Infect Dis Ther.

2024 *Second-Order Markov multistate models.* Besalú and Gómez Melis. SORT.

2025 *Semi-Markov Multistate Modeling Approaches for Multicohort Event History Data*. Piulachs et al. Biometrical Journal.

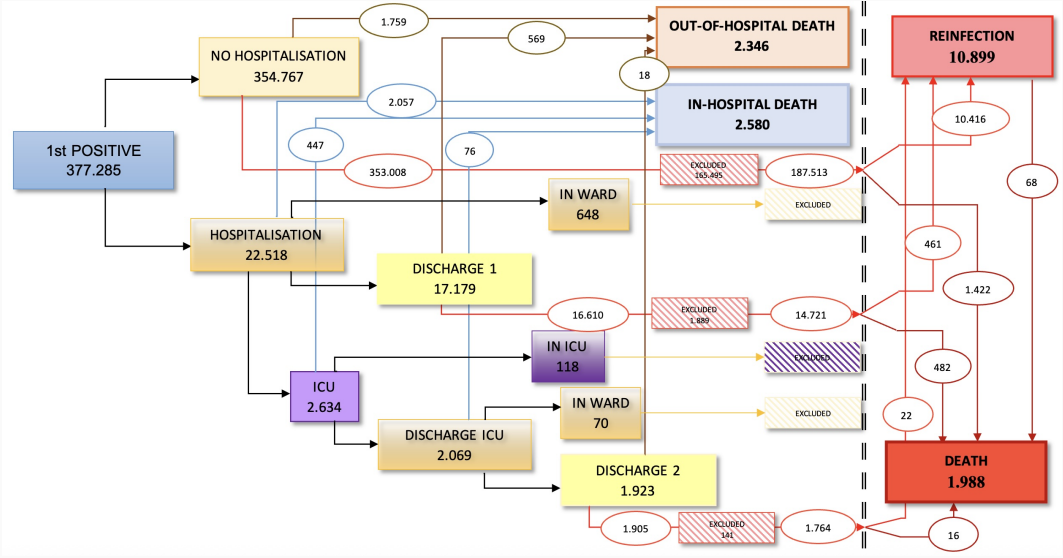
2024 *Clinical Characteristics and Predictors of complications and mortality in hospitalized octogenarian patients with COVID-19: an ambispective study*. Arroyo-Huidobro, M. et al *European Geriatric Medicine*.

2026 *Clinical Effectiveness of Intensive Care Unit Admission in Older Adults with COVID-19 during the First Pandemic Wave in Spain* . Arroyo-Huidobro, M. et al *BMC Geriatrics.*

Basque Country Project: Nationwide COVID-19 population data

EUSKO JAURLARITZA
GOBIERNO VASCO
OSASUN SAILA
DEPARTAMENTO DE SALUD

Motivating Data: Flowchart of subjects and trajectories



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.
- *Ref: Arostegui et al. Multistate modeling of disease progression for population-based registry data: Application to Covid-19. International Journal in Medical Informatics (under review)*

Goal 2: Modeling COVID-19 Reinfection

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

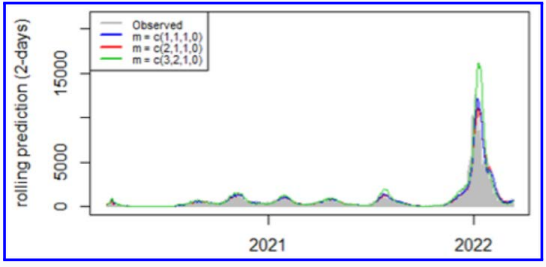
- **Challenge:** Accounting for **landmark constraints** and **competing risk of death**.
- **Definition:** Reinfection as a distinct state transition within the longitudinal history.

Basque Country COVID-19 Cohort

Data

Analyze disease progression using nationwide population-based registry data.

- **Source:** EHR from the Basque Country Health Service (Osakidetza).
- **Population:** Individuals with a 1st SARS-CoV-2 positive test.
  - $n = 377,285$ patients.
  - Recruitment: Mar 2020 – Jan 2022.
  - Follow-up: Until **April 9, 2022**.
- **Hospital:** Admission within **15 days** (post+)
- **Follow-up Duration:**
  - **Non-hospitalized:** 90 days from diagnosis.
  - **Hospitalized:** 90 from diagnosis or discharge (whichever occurred last).



Positive cases prediction.
Larrea et al. Int J Med Inform, 2025.

Data reflects real-world disease progression across multiple waves.

370,000+ individuals: Massive statistical power from Osakidetza for robust multistate modeling.

Covariates and Temporal Framework

Clinical Baseline Factors

- **Comorbidity:** Charlson Index (CCI).
- **Treatments:** Based on ATC classification.
- **Polymedication:** ≥ 6 concurrent baseline treatments.

Immunity Status

Time-dependent based on vaccination (t_i) and diagnosis (t_0):

- Effective immunity reached **14 days** after the relevant event:
  - Post-1st dose: $t \geq t_1 + 14$
  - Re-infection: $t \geq t_0 + 14$
  - Post-2nd dose: $t \geq t_2 + 14$

Study Periods (Waves)

P1 **Lockdown:**
Mar – Jun 2020

P2 **Transition:**
Jul – Dec 2020

P3 **Alpha/Delta:**
Jan – Dec 13, 2021

P4 **Omicron:**
Dec 14, 2021 – Jan 9, 2022

Vaccination campaign started in Dec 2020.

Methodological Approach: From Data to MSM

1. Statistical Strategy

- **Framework:** Cox-based Multistate Model.
- **Model structure:** 8 states and 11 transitions.
- **Estimation:** Nelson-Aalen for non-parametric transition probabilities:

$$P(T_k \leq t | T_k \geq s)$$

2. Transition-Specific Cox Models

- **Covariates:** Baseline + Time-dependent (Immunity).
- **Selection:** Only significant factors retained (forced Age/Gender).
- **Effect Measure:** Hazard Ratios (HR) as instantaneous risk.

Descriptive Analysis: Key Findings ($n = 377,285$)

Demographic Trends

- **Gender Distribution:** Balanced cohort (47% Male), but significant overrepresentation in ICUs (**68.5%**).

  Implication: Gender must be evaluated as a primary covariate for transition intensities.

- **Age Gradient:**
  - Mean age increases from **46** (total) to **65** in hospital admissions.
  - Peak ICU demand: Ages **60–80** (comprising $\approx$ 58% of ICU cases).
  - Patients 80+ frequently hospitalized but rarely admitted to the ICU

Clinical & Social Indicators

- **Deprivation Index (DI):**
  - High **homogeneity** across all quintiles (approx. 20% each).
  - Socioeconomic status was not a limiting factor for healthcare access.
- **Nursing Homes:** 6.2% hospitalized vs. only 0.6% ICU (Triage/Frailty effect).

Disease severity in this cohort is primarily driven by **Age** and **Male gender**, suggesting age-dependent care pathways. Socioeconomic factors show a remarkably equitable distribution.

Descriptive Analysis: Mortality Patterns ($n = 5,149$ deaths)

Location & Gender

- **In-hospital (53%):** Predominance of **Males** (57.5%).
- **Out-of-hospital (47%):** Higher proportion of **Females** (58.4%).
- **Clinical progression to death varies significantly by sex and setting.**

Age Impact

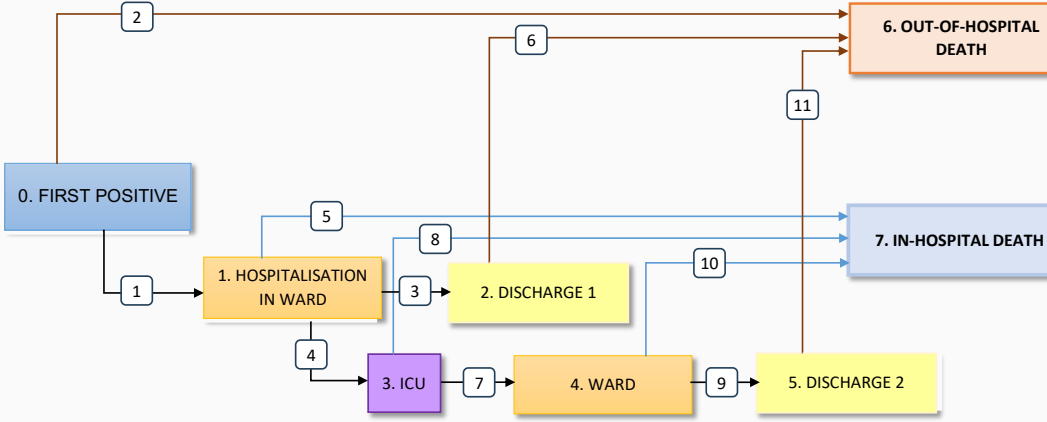
- Mean age at death is $>$ 80 **years**
- Death under 50 is Residual
- **The $\geq$ 90 group:** Represents 37.2% of out-of-hospital deaths vs. 20.3% in-hospital.

Nursing Homes & Social Index

- **Nursing Homes:**
  - Only 2.3% of the total cohort.
  - **Critical impact:** Account for **47.6%** of all out-of-hospital deaths.

  The most critical driver of non-hospital mortality pathways in the model.
- **Deprivation Index (DI):**
  - **Remarkable equity**.
  - Deaths are distributed evenly across quintiles ($\approx$ 20% each).

MSM with 8 States and 11 Transitions



Summary of transition-specific Cox estimates

	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

Figure 1: NH: Nursing home, DI: deprivation index, Immunity is time-dependent, Charlson comorbidity index (CCI)

- NS: non-significant and †: p-values between 0.05 and 0.1
- $-(+)\rightarrow$: positive and linear relation, $-(-)\rightarrow$: negative and linear relation, $\sim\sim$: non-linear relation

Impact of Gender and Age on Clinical Transitions

Gender:

- **Male Risk:** Increases hospitalization, ICU entry, and mortality.
- **ICU Threshold:** Gender becomes **Non-Significant** once in ICU; physiology outweighs sex for survival.

Age:

- **Mortality:** Constant positive linear correlation across all states.
- **The ICU "Ceiling":** Risk of ICU entry drops for ages ≥ 80 (ward-based care instead of ICU).

	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

- **Post-Discharge Risk:** Higher age significantly predicts **Death after Discharge**.
- "Recovery" does not mean safety for the elderly.

Gender drives initial severity; **Age** dictates the entire trajectory and determines ICU "triage" ceilings.

Determinants of Transitions: Immunity, Frailty, and Equity

Nursing Homes:

- **Mortality:** Strongest predictor of out-of-hospital death.
- Lower ICU risk (palliative/frailty protocols).

Vaccination & Immunity:

- **Protection:** Robust barrier to hospitalization/death.
- **Recovery:** Significantly **faster discharge** (shorter, less severe clinical courses).

Equity (Deprivation Index):

- **Impact:** ICU entry and Death are **Non-Significant (NS)**.
- **Conclusion:** Equitable health system outcomes regardless of socioeconomic status.

	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

Transition-specific Cox estimates ($n = 377,285$).

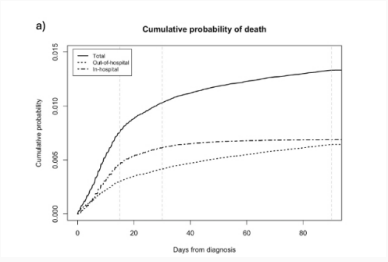
Clinical Complexity:

- **CCI:** Universal driver of poor outcomes.
- **Polymedication:** Predicts initial hospitalization and post-ICU mortality.

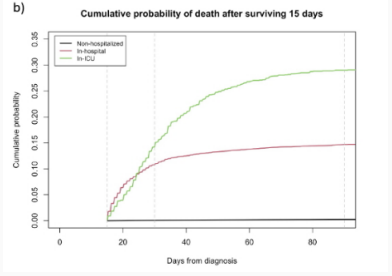
Immunity and **Nursing Homes** modulate trajectories, while **Social Deprivation** had no impact on survival equity.

Evolution of Mortality Risk over Time

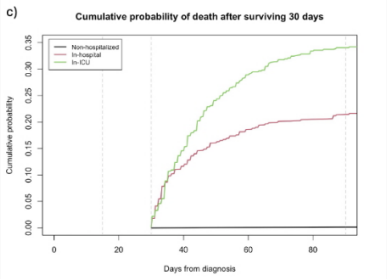
From Diagnosis ($s = 0$)



Surviving 15 days ($s = 15$)



Surviving 30 days ($s = 30$)



Remaining risk is highly dependent on the **current clinical setting**. While ICU status carries the highest intensity, the significant mortality in the hospital ward (14.7%) reflects the impact of frailty and direct progression in elderly populations.

*The beauty of the Multistate Model is that it is **dynamic.** While age and gender are crucial at the beginning, as the disease progresses, the patient's clinical trajectory, specifically whether they require ICU care or remain at home, becomes the dominant factor in predicting their next outcome.*

Conditional cumulative incidence of death up to day t

Time s	State k	$P(D \leq t \mid D \geq s) \cdot 100$ Time t		
		15	30	90
0	0 (Positive)	0.76	1.04	1.33
15	0 (Positive)	–	0.08	0.22
	1 (Hospital)	–	10.79	14.68
	3 (ICU)	–	14.23	29.01
30	0 (Positive)	–	–	0.15
	1 (Hospital)	–	–	21.44
	3 (ICU)	–	–	34.03

Estimated cumulative probability of death up to day t conditional on being alive in state k at day s ($s \leq t$) (in percentage). D is time to death, in-hospital (T_7) or out-of-hospital (T_6)

The cumulative probability of death up to day 90 from diagnosis for a subject who was in hospital (non-ICU) at day 15 is 0.1468 (14.68%)

Recurrence Analysis

Basque Country Project's Goals

Goal 1: Mapping Patient Trajectories

Analyze disease progression using nationwide population-based registry data.

- **Aim:** Demonstrate how MSMs capture the full "patient journey" better than standard survival models.
- *Ref: Arostegui et al. Multistate modeling of disease progression for population-based registry data: Application to Covid-19. International Journal in Medical Informatics (under review)*

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.

Recurrence Analysis: Understanding Disease Dynamics

Hypothesis

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

Main Goal & Motivation

- Identify robust **predictors of recurrence**.
- Map complete clinical trajectories based on population-level data.

The Study Cohort

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

COVID-19 Reinfection Definition: New positive test $\geq$ 120 days after primary infection in non-hospitalized individuals.

Remark: Hospitalized patients are not "at risk" of reinfection until post-discharge stability.

Recurrence in the MSM Framework: Two models: M1 and M2

M1: MSM based on all times to events prior to recurrence

- Captures complete trajectories of the full cohort.
- Includes all covariate effects on every transition.

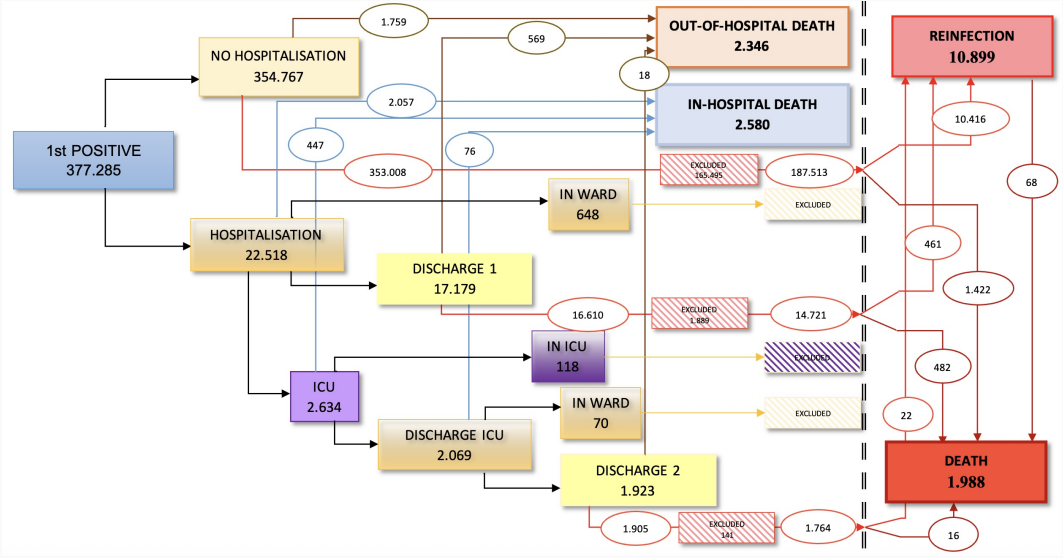
M2: Reinfection Sub-Model

Transitions from **M1** that can lead to re-infection.

- Population: Alive and non-hospitalized at Day 120.
- Accounts for **competing risks** (Reinfection vs. Death).

Methodological Challenge: The **landmark constraint** must be strictly incorporated to ensure that only individuals "at risk" (alive at day 120) enter the recurrence analysis.

Complete flowchart with sample sizes and number of event



Methodological Approaches for Modeling Recurrence

1. Stratified Approach

Separate 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.
- **High-dimensional:** 10 states and 18 transitions.
- Most computationally intensive.

Trade-off: Moving from (1) to (3) increases statistical efficiency but adds significant computational and interpretive complexity.

Common Modeling Framework: The Landmark Setting

Main Objective:

Characterize reinfection based on the prior course of events.

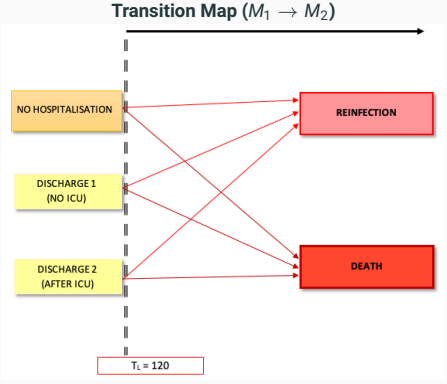
Statistical Foundation

- **Model:** Cox cause-specific hazards.
- **Baseline:** Landmark time $t_L = 120$ days.
- **Covariates:** Baseline and time-dependent variables introduced at landmark time t_L.

Individuals at Risk of reinfection ($t \geq t_L$):

1. **Risk Group 1 (General Pop):** Non-hospitalized patients.
2. **Risk Group (Post-Ward):** Hospitalized and discharged from Ward by t_L.
3. **Risk Group 3 (Post-ICU):** Hospitalized and discharged from ICU by t_L.

Analysis is conditional on surviving and being on Risk Groups 1, 2, or 3 at t_L. This handles the **immortal time bias** of the initial infection period. To enter M_2, a patient **must survive** the first 120 days, creating a "selected" population of survivors.



Stratified transitions at t_L.

Approach 1: Stratification by Risk of Reinfection Subcohort

Statistical Framework

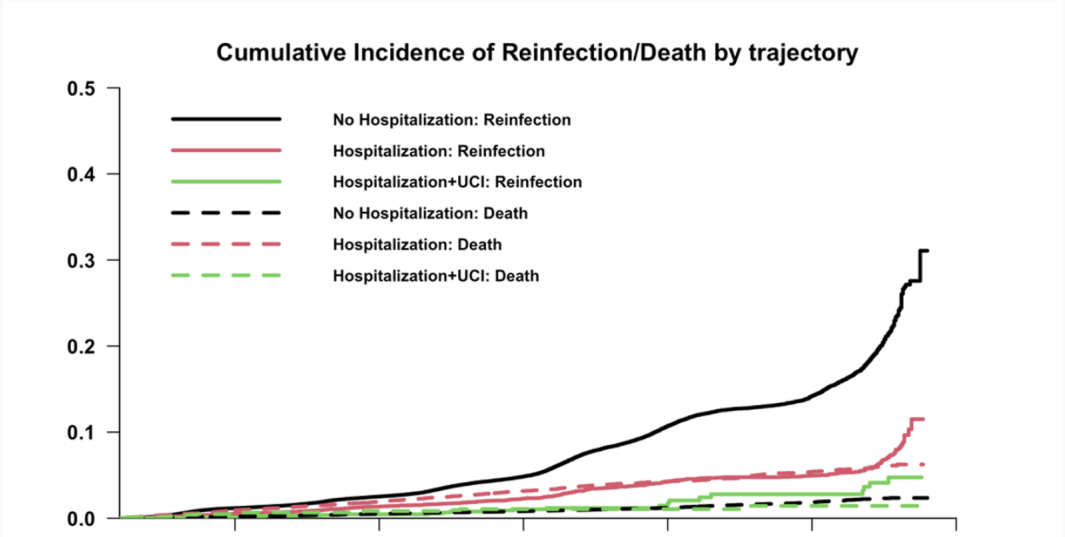
Maximize the likelihood $L(\theta)$ using a **Cox cause-specific hazards model**. All baseline and time-dependent covariates are introduced at the landmark time t_L, stratifying by the individual's prior clinical state.

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

Prior State (at t_L)	Count (n)	Reinfections	Rate (%)
1. Non-hospitalized	187,513	10,416	5.55%
2. Hospitalized (No ICU)	14,721	461	3.13%
3. Hospitalized (ICU)	1,764	22	1.25%
Total	**203,998**	**10,899**	**5.34%**

Note: Stratification allows for state-specific baseline hazards, accounting for the different recovery paths.

Results: Stratification approach I



Results: Stratification approach II

Stratified Cause-Specific Cox Models

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

Effect of significant covariates on Reinfection vs. Death (CCI: Charlson Comorbidity Index, NH: Nursing Home). +/- sign stands for + or - effect on the risk of reinfection or death.

Key Findings:

- **Age:** Non-linear effect observed; however, the risk of reinfection remains consistently lower relative to the reference group (< 30).
- **Temporal Interaction:** Later pandemic periods correlate with increased risk of reinfection and death, though **basal immunity** significantly mitigates this effect.

Methodological Insight:

- This heterogeneity justifies the need for an **Integrated Risk Score** to simplify implementation and interpretation.

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, NH: Nursing Home).

Key Findings:

- **Age:** Risk of reinfection is consistently **lower** in older groups vs. the reference (< 30).
- **Time & Immunity:** Later periods increase reinfection risk, but **basal immunity** significantly mitigates mortality.
- **Heterogeneity:** Significant differences in covariate effects justify the **Integrated Risk Score (IRS)**.

From Complexity to Utility: Moving from transition-specific HRs to a unified score simplifies clinical implementation and interpretation in `MSMpred`.

Approach 2: Integrated Risk Score (IRS)

Translating History into a Metric

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:

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

Encoding the Multi-State Dynamics: Z_i^k is not a raw variable; it integrates all baseline hazards (λ_{jk0}) and covariate effects (β) from the primary infection phase:

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

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

IRS in Competing Risk Models

Model Structure for Reinfection: The cause-specific Cox model for M_2 incorporates two key elements from M_1

Dual-Factor Risk Structure

- **Categorical (*Where they come from*):**
  The risk group reached at t_L:
  - $k = 1$: Non-hospitalized.
  - $k = 2$: Hospital (Ward).
  - $k = 3$: Hospital (ICU).

- **Continuous (*How influenced*):**
  The predicted probability IRS (Z_i^k) accounts for the effect of the covariates on the **intensity** of the path taken to reach that state.

Preliminary Findings ($n = 300$)

History Effect: Previously hospitalized individuals show a **lower risk** of reinfection (possible higher immunity?).

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

Probability Effect: Higher predicted occupation probabilities (Z_i^k) correlate with a **higher risk** of reinfection.

The IRS helps us distinguish between a **patient's current health** and their underlying **vulnerability**, leading to a much more accurate prediction of reinfection. The IRS filters out the **background noise** of their general frailty to see their true risk of catching COVID again.

Challenges of the IRS Approach

Advantages

- Exceptional **interpretive value**.
- Captures non-linearities and interactions from M_1 automatically.

Disadvantages

- Extreme **computational cost**.
- Complex statistical inference.

Computational Wall

Calculating probabilities for an 11-transition model is demanding:

- **Benchmark:** $>$ 1 hour for $n = 300$ individuals.
- **Population Scale:** At $n \approx 400,000$, the calculation becomes **computationally prohibitive** without massive parallelization.

Inference & Uncertainty

Standard errors for M_2 cannot be calculated analytically because Z_i^k is a **predicted value**.
Requirement: Multi-stage bootstrap methods, further increasing the computational burden by 100 or 1000.

Approach 3: Joint M_1M_2 Modeling

The Full Disease Lifecycle

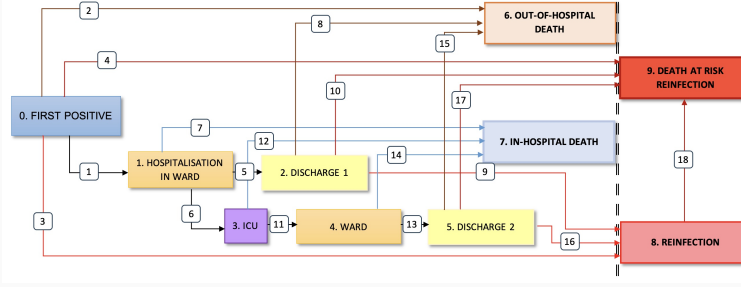
We extend the primary model (M_1) by integrating recurrence and subsequent mortality into a single stochastic process.

Model:

- **10 States**
- **18 Transitions**

New Absorption States:

- **State 8:** Reinfection.
- **State 9:** Death following a reinfection risk period.



Unified Flowchart: From primary infection to recurrence.

Approach 3 (Joint Model): Enforcing Temporal Constraints

Non-Standard Likelihood Modification

Simultaneous estimation of the 10-state process using a **single time origin** (diagnosis) while enforcing a 120-day "protection" phase via **mathematical latency**.

The Piecewise Hazard Solution: We manually constrain hazards to zero for $t \leq 120$:

- **Direct Reinfection (State 0 $\rightarrow$ 8):**

$$\lambda_{h8}(t \mid \mathbf{X}) = \lambda_{h8_0}(t) \exp(\beta^\top \mathbf{X})$$

  where $\lambda_{h8_0}(t) = 0$ **for** $t \leq 120$

- **Post-Discharge Reinfection (States $h \in \{2, 5\} \rightarrow 8$):**

$$\lambda_{h8}(t \mid \mathbf{X}) = \lambda_{h8_0}(t) \exp(\beta^\top \mathbf{X} + \gamma T)$$

  where T accounts for sojourn time and $\lambda_{h8_0}(t) = 0$ *for* $t \leq 120$.

Implications for Inference:

- **Structural Zero:** $P(\text{Reinfection} \mid t \leq 120) = 0$.
- **Immortal Period:** No information is contributed to reinfection parameters β_{h8} during the first 4 months.

Why "Latent"?

The transition exists in the state-space from $t = 0$, but the **stochastic gate** only opens at $t = 120$.

- Prevents misclassifying **persistent viral shedding** as reinfection.
- Integrates clinical criteria directly into the likelihood structure.

Mortality before $t = 120$ *acts as a competing risk that removes the individual before the transition activates.*

Joint Model: Implementation

PROs

- **Unified Inference:** Uses all population data in a single likelihood. Hence, there is no loss of information.
- **Sojourn Effects:** Naturally incorporates history via transition-specific hazards and captures the impact of hospital stay duration on reinfection.

CONs

- **Non-standard Implementation:** Manually setting hazards to zero until $t = 120$ is computationally complex. Extremely difficult to implement in standard software
- **Modeling options** and its implications.
  - **Full Model:** 10 states / 18 trans. Distinguishes death by prior clinical path. There are three states of death and two states of discharge
  - **Intermediate:** 8 states / 15 trans. Aggregates mortality states and preserves two discharge states.
  - **Simplified:** 6 states / 11 trans. Aggregates late-stage outcomes (State 5) and death (States 6, 7 and 9) to reduce dimensionality.

Future Work and Conclusions

Future Work: Methodology & Scaling

Inferential Validity

- **Bias Correction:** Formalize conditioning for the 120-day landmark.
- **Uncertainty:** Variance corrections for M_2 estimates (Bootstrapping).
- **Joint vs. IRS:** Accuracy vs. computational cost trade-off.

Sensitivity & Theory

- **The 120-day Gate:** Test sensitivity on global intensities.
- **Semi-Markov:** Incorporate duration-dependence (sojourn time).

Computational Scaling

- Parallelize for $n \approx 400,000$.
- Break the *1-hour/300-patient* bottleneck for real-time use.

`MSMpred` Integration

- Embed the **IRS framework** into the web tool.
- Scalable architecture for large datasets.

Clinical Horizons

- **Waning Immunity:** Model vaccination dynamics over time.
- **Generalization:** Apply to Cancer, Malaria, and Seasonal Flu.

Conclusions & Final Remarks

Methodology

MSMs move **beyond Illness-Death** models, successfully capturing real-world complexity and disease evolution from routine data.

Prediction: `MSMpred`

Bridges theory and practice, translating complex math into **individualized clinical tools** via an interactive platform.

Reinfection Risk

Our framework will help identify key predictors and clinical trajectories that characterize **long-term immunity** and reinfection dynamics.

Public Health Impact

Demonstrates the power of **population registries** to provide evidence-based support for policies and clinical management.

- **High Interpretability:** Despite model complexity, results are directly applicable to epidemiological surveillance.
- **Global Adaptability:** Methodology is **scalable beyond COVID-19**, enabling early interventions in any registry-based system.

Moltes gràcies per ser-hi avui compartint la nostra recerca!!!

Acknowledgements

Sponsored by 2021 SGR 01421, AGAUR, Catalunya.

GRBIO: Research Group in Biostatistics and Bioinformatics

10 years celebration

Grants: MICIU/AEI /10.13039/501100011033 and FEDER, UE: PID2023-148033OB-C21 (Statistics for Health Sciences: Advances in Survival Analysis and Clinical Trials) and PID2024-156800OB-I00 (Advances in Statistical Modelling for Health Sciences).