Mahidol University
Faculty of Medicine Ramathibodi Hospital

C&B | Department of Clinical Epidemiology and Biostatistics

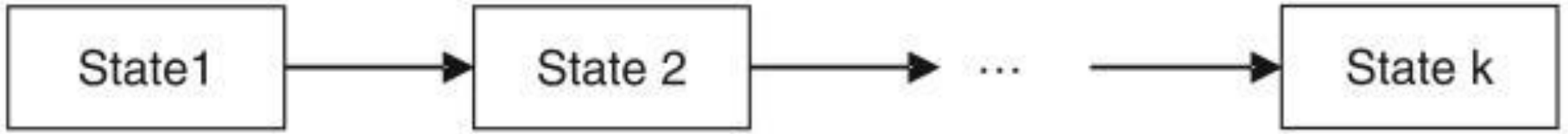
Mastering Data Preparation in Multi-State Models

Romen Samuel Wabina, MSc.
PhD candidate, Data Science for Healthcare and Clinical Informatics

Multi-State Models

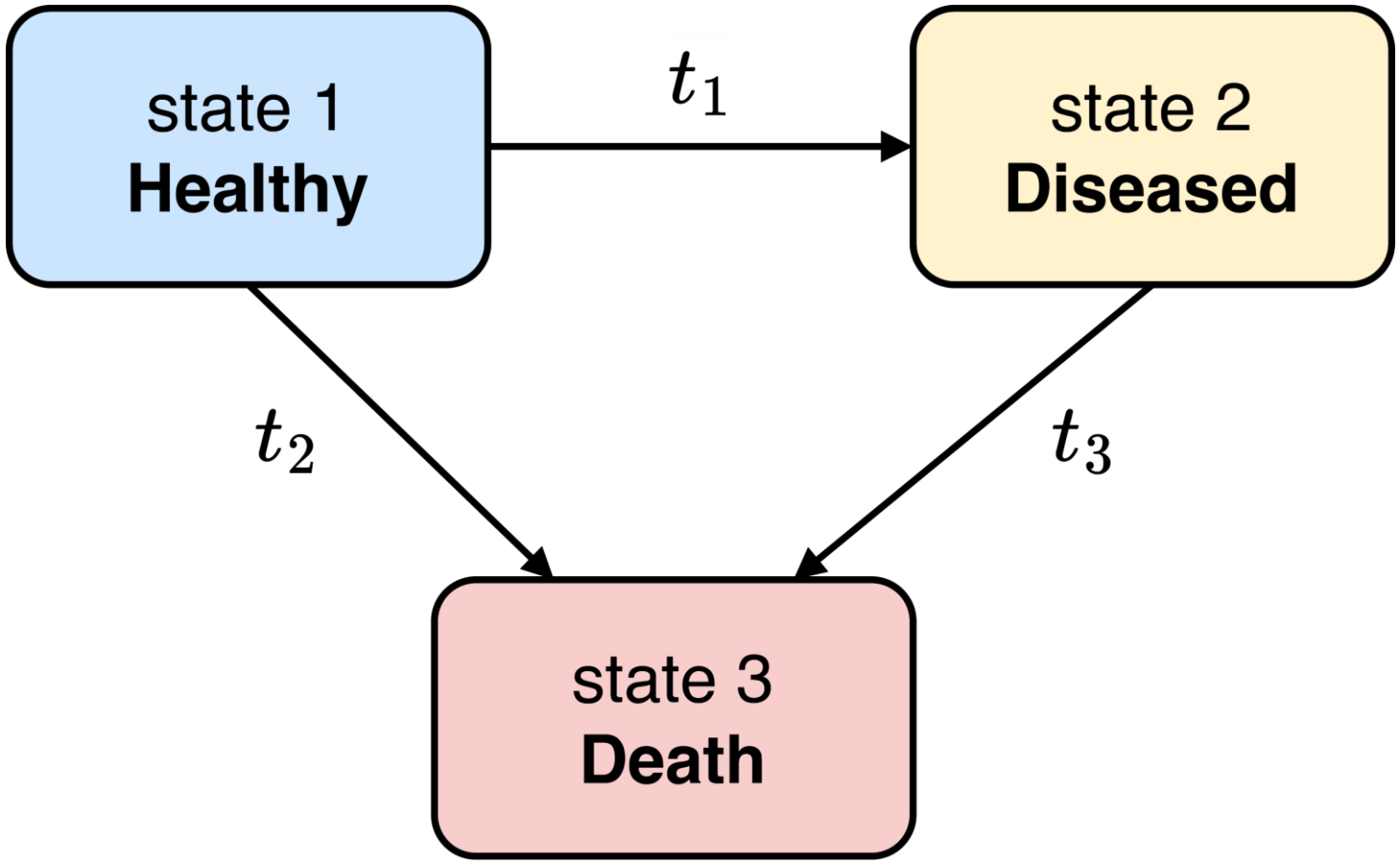
- In survival analysis, we often concentrate on the time to a single event of interest
- In practice, there are many clinical examples of where a patient may experience a variety of intermediate events

- Model a process where subjects transition from one state to the next
- Time-to-event analysis
  - Analyze the time until an event of interest occurs
  - Time since diagnosis of disease to death



- More than one interested events
  - **time since hypertension** to **myocardial infarction**, **death**?
  - **time since chronic kidney disease** (CKD) to **end-stage renal disease**, **death**?

Components of a Multi-State Model



Objectives in Multi-State Models

- Modeling disease progression and transitions between states
  - To estimate the probability of being in each state at distinct time
  - To estimate the transition hazards, which are effects of risk factors
    - Evaluate how comorbidities, medications, and laboratory tests influence transition rates

Concepts: Markov Model

- Transition rates $\lambda_{ij}(t)$ describe the probability of transitioning to state j, given that an individual is in state i at time t

$$\lambda_{ij}(t; \mathcal{F}_t) = \lim_{\delta t \to 0} \delta t^{-1} \mathbb{P}\{S(t + \delta t) = j | S(\delta t) = i, \mathcal{F}_t\}$$

- If $\lambda_{ij}(t; \mathcal{F}_t) = \lambda_{ij}(t)$ where t is time since start of process, i.e., no dependence on history $\mathcal{F}_t$:
  - time spent in the current state
  - states visited previously by the individual and time spent in them
- Then the process is Markov.

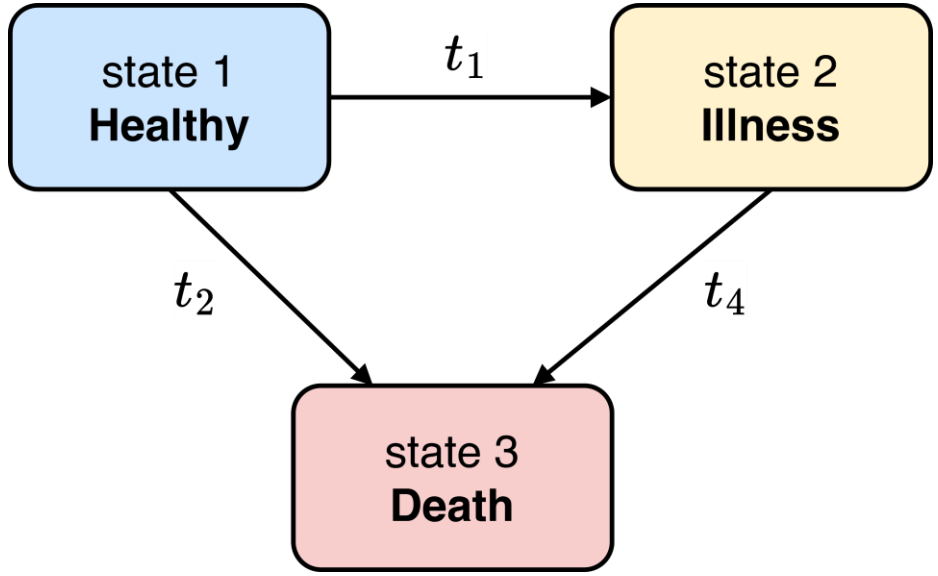
- Markov process: transition rates depend only on the current state i
  - Not the specific path taken to arrive at $S(t)$
    - Patient 1: $\mathrm{CKD3A} \rightarrow \mathrm{CKD3B} \rightarrow \mathrm{CKD4}$
    - Patient 2: $\mathrm{CKD3A} \rightarrow \mathrm{CKD4}$
  - Both patients have the same probability progressing to CKD5.

Concepts: Transition Matrix and Probabilities

- Consider an irreversible illness-death model with states $s = 3$
- Multi-state models can be represented by different matrices

1. Transition matrix $\mathbb{T} \in \mathbb{R}^{\{s \times s\}}$
   - illustrates the possible transitions a patient is at risk of experiencing from a given state.

	State 1	State 2	State 3
State 1		t_1	t_2
State 2			t_3
State 3			



2. Transition probability matrix $\mathbb{P}(t) \in \mathbb{R}^{\{s \times s\}}$
   - Each element in $\mathbb{P}$ represents the probability of transitioning from state i to state j

	State 1	State 2	State 3
State 1	$p(t_{ii})$	$p(t_1)$	$p(t_2)$
State 2	0	$p(t_{ii})$	$p(t_3)$
State 3	0	0	1

- Diagonal elements: rate of staying in state i
- Upper non-diagonal elements:
  - rate of moving from one state to another

Data Preparation

- Many survival studies have their data stored initially in a wide format.

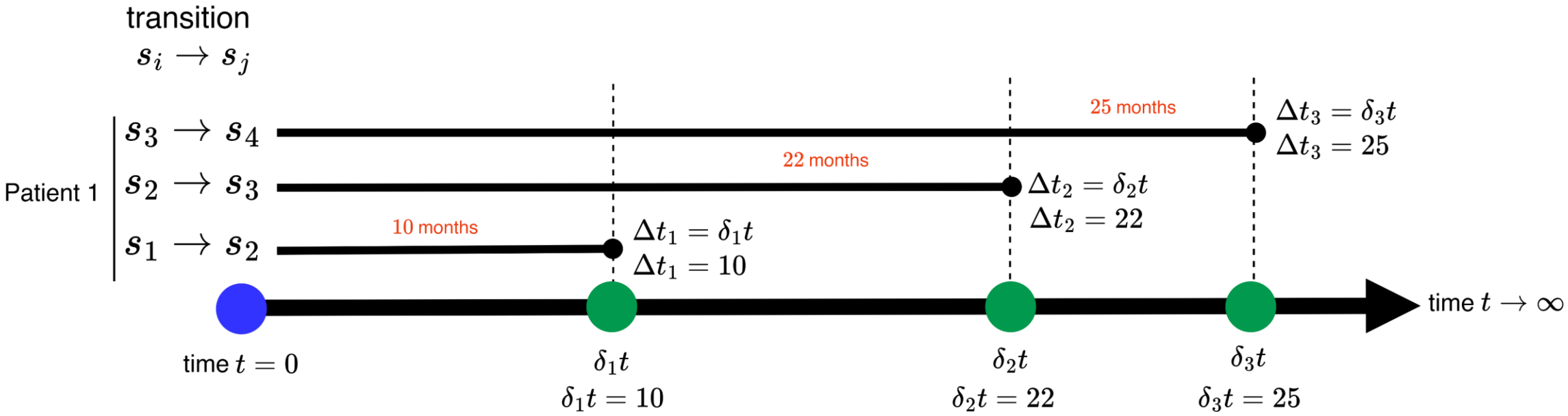
PID	time (s_1)	status (s_1)	time (s_2)	status (s_2)	time (s_3)	status (s_3)
1	0	1	5	1	15	1
2	0	1	20	1	30	0
3	0	1	12	0	12	1

- Long format allows the most flexibility for multi-state modelling
  - Competing risks – Each row represents one patient 'at risk' for one transition.

- Requires 6 or 7 columns depending on your time-scale approach
  - Patient identification column (e.g., `PID`)
  - Transition columns `FROM` and `TO`
  - `STATUS` indicates the state transition, specifying whether the patient has experienced the disease or was censored
  - Columns that depend on your time-scale approach:
    - Clock-forward: `TSTART` and `TSTOP`
    - Clock-reset: `TSTART`, `TSTOP`, and `TIME = TSTOP-TSTART`

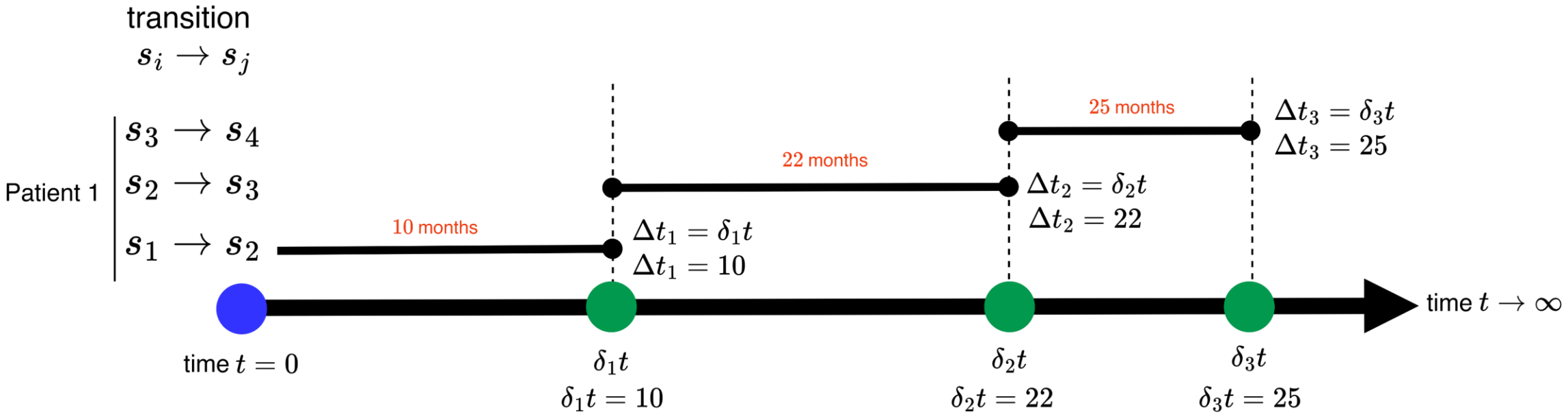
Time-Scales in Multi-State Models

- Clock-forward (i.e., Markov)
  - Time scale progresses continuously from a baseline point (i.e., initial state)
    - column TSTOP in data preparation
  - Does not track the time spent specifically in each intermediate state
  - Treats the total time since the baseline as sufficient information for calculating transition hazards.
    - Markov? Past states and their durations don't influence future transitions.



Time-Scales in Multi-State Models

- Clock-reset (i.e., semi-Markov)
  - Resets to zero each time a new state is entered
    - column TIME in data preparation
  - Transition hazards are based on the time spent in the current state rather than the overall time since the baseline event.
  - Why semi-Markov?
    - Time-scale depends on the time when the present state was reached



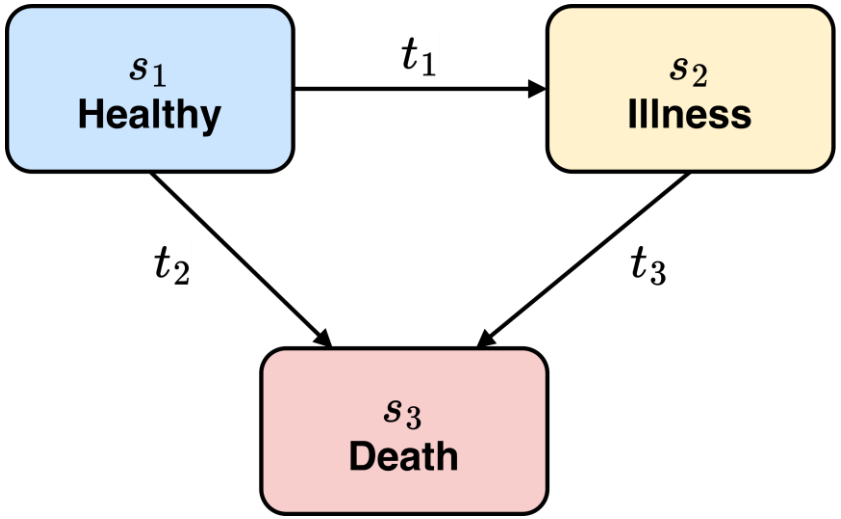
Data Preparation: Quick Overview

- Consider a unidirectional illness-death model with three states s_1, s_2 and s_3 with a wide format dataset.

PID	time (s_1)	status (s_1)	time (s_2)	status (s_2)	time (s_3)	status (s_3)
1	0	1	5	1	15	1
2	0	1	20	1	30	0
3	0	1	12	0	12	1

- How a multi-state dataset looks like:

PID	pathway	transition	from	To	tstart	tstop	status
1	$s_1 \to s_2 \to s_3$	1	s_1	s_2	0	5	1
1	$s_1 \to s_2 \to s_3$	2	s_1	s_3	0	10	0
1	$s_1 \to s_2 \to s_3$	3	s_2	s_3	5	10	1
2	$s_1 \to s_2$	1	s_1	s_2	0	20	1
2	$s_1 \to s_2$	2	s_1	s_3	20	30	0
2	$s_1 \to s_2$	3	s_2	s_3	20	30	0
3	$s_1 \to s_3$	1	s_1	s_2	0	12	0
3	$s_1 \to s_3$	2	s_1	s_3	0	12	1



Column checklist:

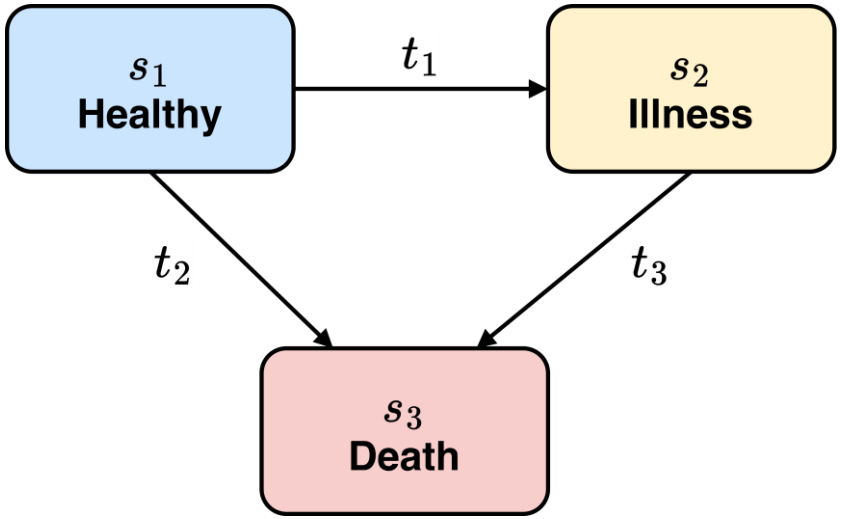
- ✓ PID
- ✓ transition (FROM and TO)
- ✓ STATUS
- ✓ PATHWAY (optional)
- ✓ TSTART and TSTOP
  TIME (clock-reset approach)

Data Preparation: Illness-death model

- Consider an irreversible illness-death model with three states s_1, s_2 and s_3

- Suppose we have a wide-format dataset with two variables (time and status) for each state

PID	pathway	time (s_1)	status (s_1)	time (s_2)	status (s_2)	time (s_3)	status (s_3)
1	$s_1 \rightarrow s_2 \rightarrow s_3$	0	1	5	1	15	1
2	$s_1 \rightarrow s_2$	0	1	20	1	30	0
3	$s_1 \rightarrow s_3$	0	1	12	0	12	1



	State 1	State 2	State 3
State 1		t_1	t_2
State 2			t_3

- Pathway represents the patients' *journey* since entry time until the end of study date. (optional)

- Let's try to prepare the data in a stepwise logical way!

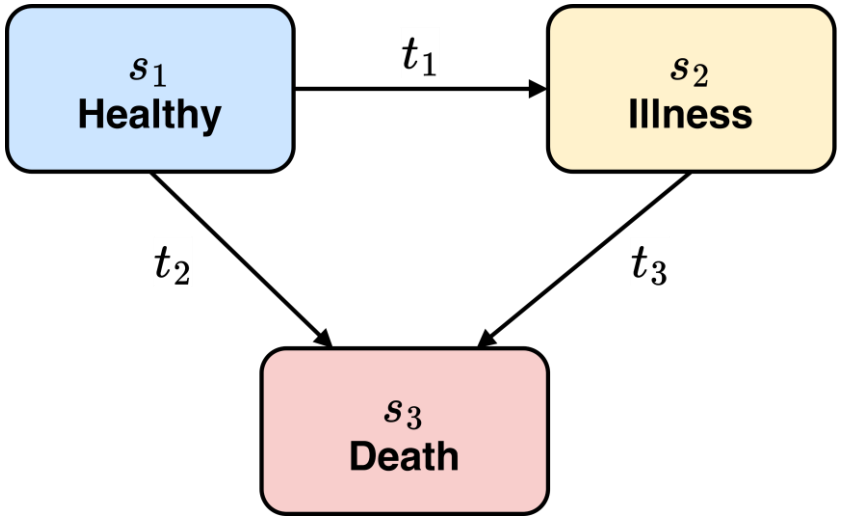
Objective

- **Reshape wide to long format**
- **Each patient should initially have three records depending on whether the patients have had experienced the event for each state**

Data Preparation: Illness-death model

- Suppose we have a wide-format dataset with two variables for each state (time and status)

PID	pathway	time (s_1)	status (s_1)	time (s_2)	status (s_2)	time (s_3)	status (s_3)
1	$s_1 \rightarrow s_2 \rightarrow s_3$	0	1	5	1	15	1
2	$s_1 \rightarrow s_2$	0	1	20	1	30	0
3	$s_1 \rightarrow s_3$	0	1	12	0	12	1



First, create your transition matrix!

	State 1	State 2	State 3
State 1		t_1	t_2
State 2			t_3
State 3			

Data Preparation: Illness-death model

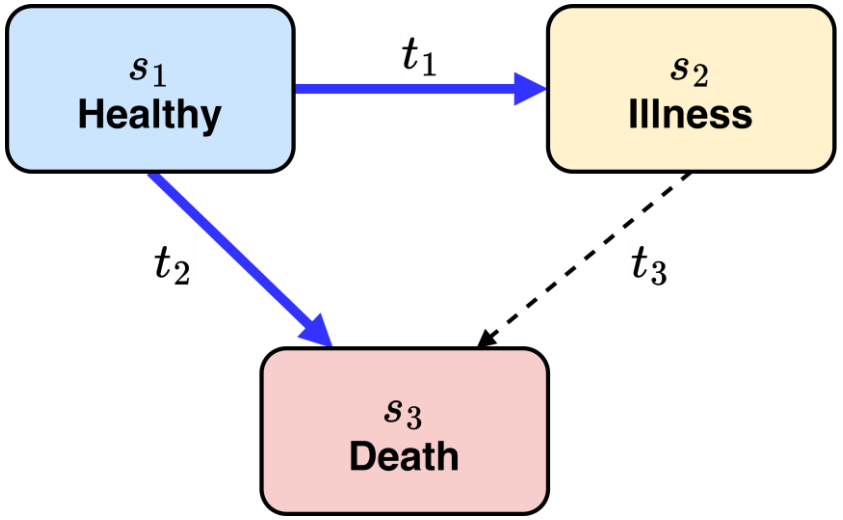
- Suppose we have a wide-format dataset with two variables for each state (time and status)

PID	pathway	time (s_1)	status (s_1)	time (s_2)	status (s_2)	time (s_3)	status (s_3)
1	$s_1 \rightarrow s_2 \rightarrow s_3$	0	1	5	1	15	1
2	$s_1 \rightarrow s_2$	0	1	20	1	30	0
3	$s_1 \rightarrow s_3$	0	1	12	0	12	1

- From s_1, patient 1 is at risk of going to s_2 or s_3 (transition matrix)

A

PID	from	to	transition	tstart	tstop	status
1	s_1	s_2	t_1	0	5	1
1	s_1	s_3	t_2	0	5	0



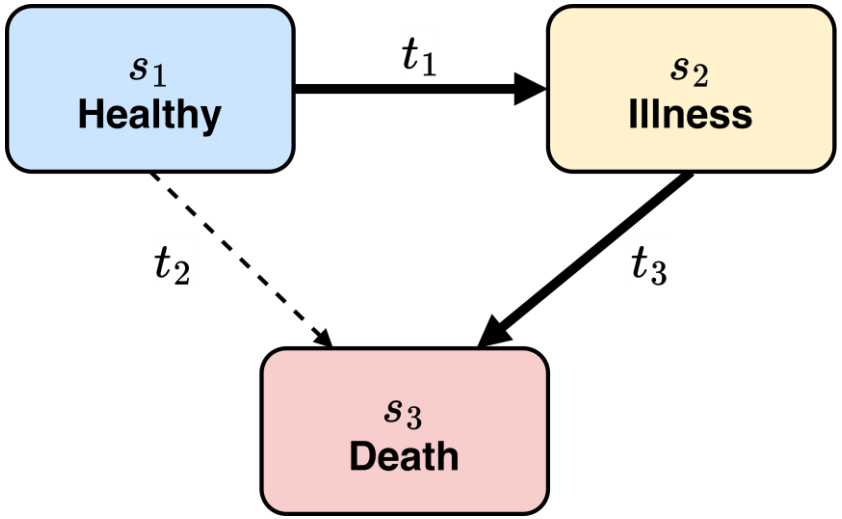
	State 1	State 2	State 3
State 1		t_1	t_2
State 2			t_3

patient-at-risk

Data Preparation: Illness-death model

- Suppose we have a wide-format dataset with two variables for each state (time and status)

PID	pathway	time (s_1)	status (s_1)	time (s_2)	status (s_2)	time (s_3)	status (s_3)
1	$s_1 \to s_2 \to s_3$	0	1	5	1	15	1
2	$s_1 \to s_2$	0	1	20	1	30	0
3	$s_1 \to s_3$	0	1	12	0	12	1



- From s_1, patient 1 is at risk of going to s_2 or s_3 (transition matrix)

A

PID	from	to	transition	tstart	tstop	status
1	s_1	s_2	t_1	0	5	1
1	s_1	s_3	t_2	0	5	0

- After 5 months, patient 1 entered s_2
- From s_2, patient is at risk of s_3

B

PID	from	to	transition	tstart	tstop	status
1	s_2	s_3	t_3	5	10	1

- Concatenate tables A and B

	State 1	State 2	State 3
State 1		t_1	t_2
State 2			t_3

patient-at-risk

Hence, patient 1 with pathway $s_1 \to s_2 \to s_3$ has 3 rows.

Data Preparation: Illness-death model

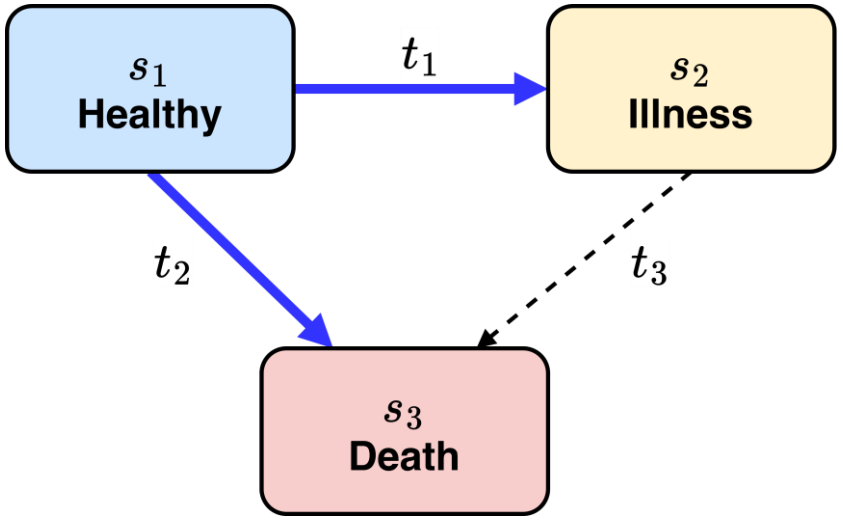
- How about patient 2 with pathway $s_1 \rightarrow s_2$?

PID	pathway	time (s_1)	status (s_1)	time (s_2)	status (s_2)	time (s_3)	status (s_3)
1	$s_1 \rightarrow s_2 \rightarrow s_3$	0	1	5	1	15	1
2	$s_1 \rightarrow s_2$	0	1	20	1	30	0
3	$s_1 \rightarrow s_3$	0	1	12	0	12	1

- Like patient 1, patient 2 is at risk of going to s_2 or s_3

A

PID	from	to	transition	tstart	tstop	status
2	s_1	s_2	t_1	0	20	1
2	s_1	s_3	t_2	0	20	0



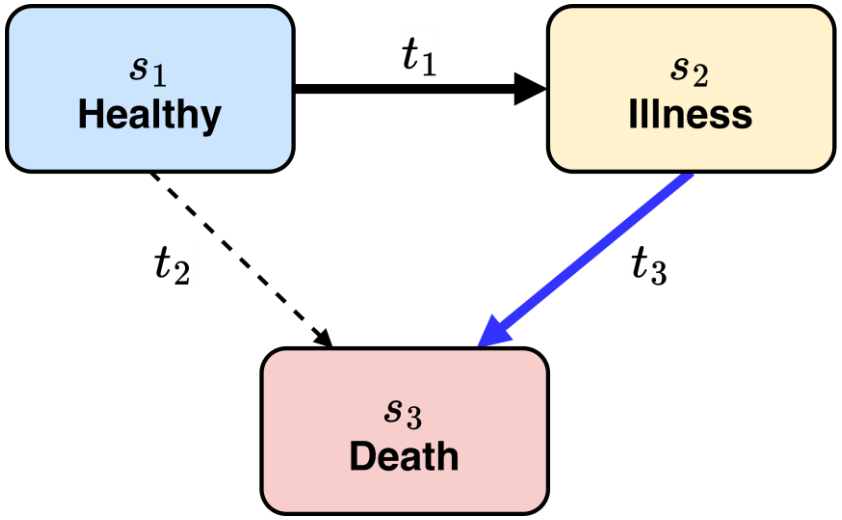
	State 1	State 2	State 3
State 1		t_1	t_2
State 2			t_3

patient-at-risk

Data Preparation: Illness-death model

- How about patient 2 with pathway $s_1 \rightarrow s_2$?

PID	pathway	time (s_1)	status (s_1)	time (s_2)	status (s_2)	time (s_3)	status (s_3)
1	$s_1 \rightarrow s_2 \rightarrow s_3$	0	1	5	1	15	1
2	$s_1 \rightarrow s_2$	0	1	20	1	30	0
3	$s_1 \rightarrow s_3$	0	1	12	0	12	1



- Like patient 1, patient 2 is at risk of going to s_2 or s_3

A

PID	from	to	transition	tstart	tstop	status
2	s_1	s_2	t_1	0	20	1
2	s_1	s_3	t_2	0	20	0

- Patient 2 entered s_2 and at risk of going to s_3
- Unlike patient 1, patient 2's status should be censored

B

PID	from	to	transition	tstart	tstop	status
2	s_2	s_3	t_3	20	30	0

	State 1	State 2	State 3
State 1		t_1	t_2
State 2			t_3

patient-at-risk

Hence, patient 2 with pathway $s_1 \rightarrow s_2$ has 3 rows.

Data Preparation: Illness-death model

- Now, for patient 3

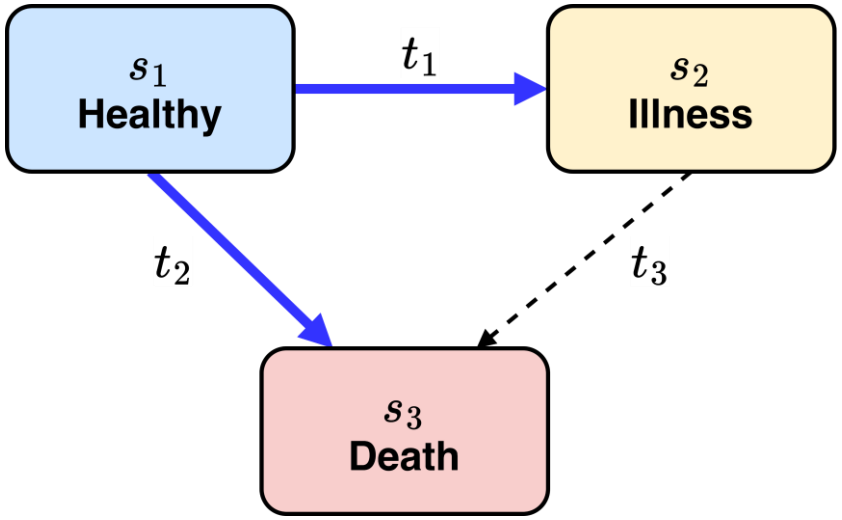
PID	pathway	time (s_1)	status (s_1)	time (s_2)	status (s_2)	time (s_3)	status (s_3)
1	$s_1 \to s_2 \to s_3$	0	1	5	1	15	1
2	$s_1 \to s_2$	0	1	20	1	30	0
3	$s_1 \to s_3$	0	1	12	0	12	1

- Like previous patients, patient 3 is at risk of going to s_2 and s_3
- But no risk from $s_2 \to s_3$ since patient never went to s_2

PID	from	to	transition	tstart	tstop	status
2	s_1	s_2	t_1	0	12	0
2	s_1	s_3	t_2	0	12	1

- The row with transition t_3 should not be in the dataset.

Hence, patient 3 with pathway $s_1 \to s_3$ only has 2 rows.



	State 1	State 2	State 3
State 1		t_1	t_2
State 2			t_3

patient-at-risk

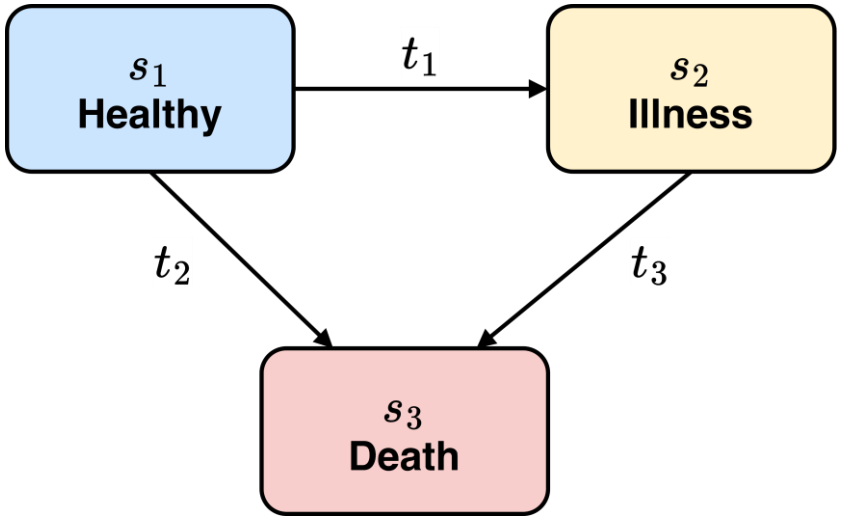
Data Preparation: Illness-death model

- So, what's the final data?

PID	pathway	time (s_1)	status (s_1)	time (s_2)	status (s_2)	time (s_3)	status (s_3)
1	$s_1 \rightarrow s_2 \rightarrow s_3$	0	1	5	1	15	1
2	$s_1 \rightarrow s_2$	0	1	20	1	30	0
3	$s_1 \rightarrow s_3$	0	1	12	0	12	1

- How the final dataset looks like?
  - Eight rows in your multi-state dataset

PID	from	to	transition	tstart	tstop	status
1	s_1	s_2	t_1	0	5	1
1	s_1	s_3	t_2	0	5	0
1	s_2	s_3	t_3	5	10	1
2	s_1	s_2	t_1	0	20	1
2	s_1	s_3	t_2	0	20	0
2	s_2	s_3	t_3	20	30	0
3	s_1	s_2	t_1	0	12	0
3	s_1	s_3	t_2	0	12	1

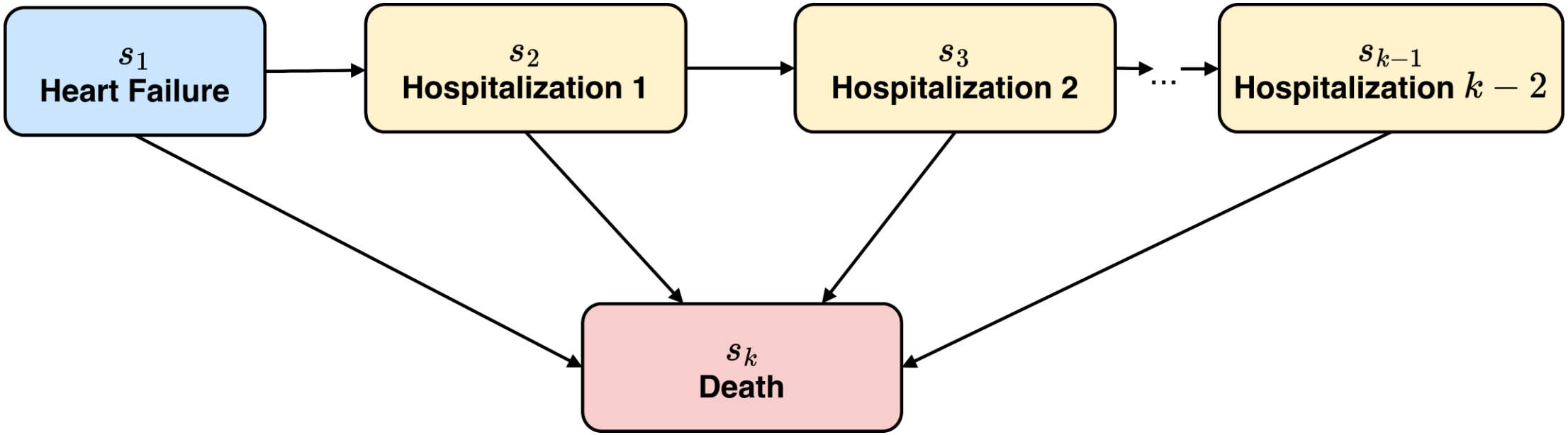


	State 1	State 2	State 3
State 1		t_1	t_2
State 2			t_3

Recurrent events

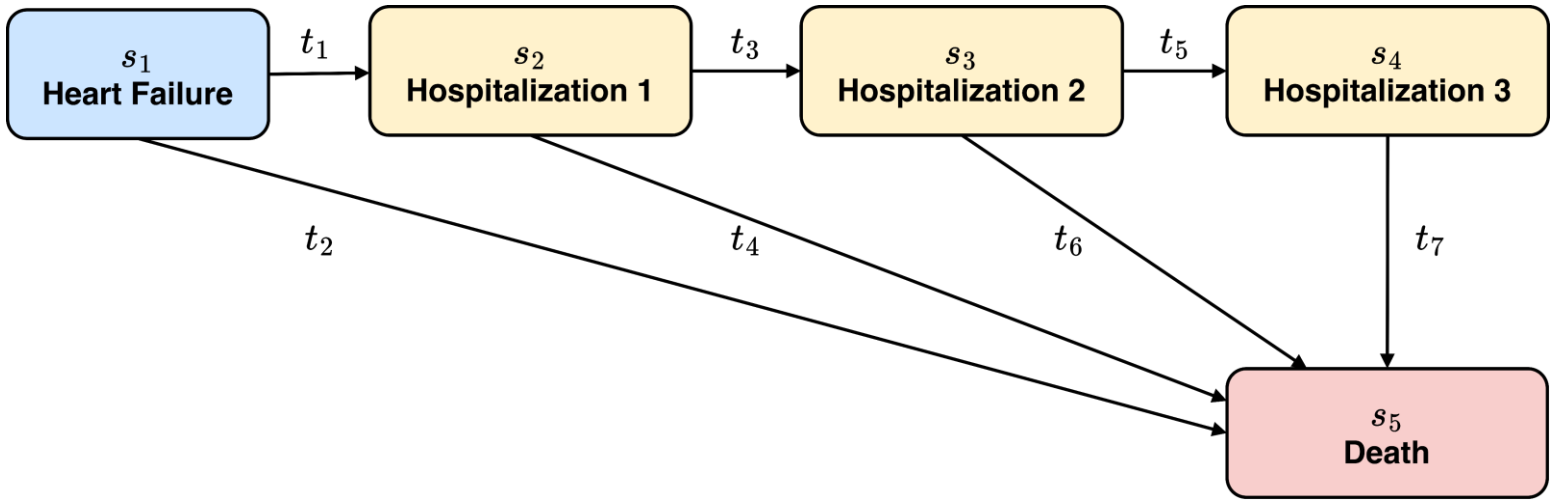
- Another type of multi-state is the possibility that the event can occur multiple times
  - Cancer recurrence, infections, hospitalization, relapses for drug abuse

- A key decision when modeling recurrent events is whether and how to condition on previous recurrences.



Heart Failure Multi-State Model

- Consider this multi-state model with five states and seven transitions
- Assume patients can have at most three hospitalizations



Transition Matrix

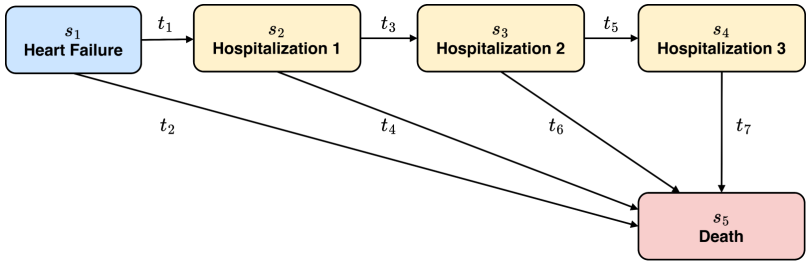
	s_1	s_2	s_3	s_4	s_5
s_1		t_1			t_2
s_2			t_3		t_4
s_3				t_5	t_6
s_4					t_7

Data Preparation: HF Multi-State Model

- Consider this multi-state model with five states and seven transitions
- Assume patients can have at most three hospitalizations

- Suppose we have a wide-format dataset for three patients

PID	time(s_1)	stat(s_1)	time(s_2)	stat(s_2)	time(s_3)	stat(s_3)	time(s_4)	stat(s_4)	time(s_5)	stat(s_5)
1	0.00	1	8.13	1	28.43	1	29.47	0	29.47	1
2	0.00	1	22.33	1	32.53	0	32.53	0	32.53	1
3	0.00	1	8.13	1	18.33	1	20.30	1	30.50	1



- Patient 1 progressed from $s_1 \rightarrow s_2 \rightarrow s_3 \rightarrow s_5$
- Patient 2 progressed from $s_1 \rightarrow s_2 \rightarrow s_5$
- Patient 3 progressed from $s_1 \rightarrow s_2 \rightarrow s_3 \rightarrow s_4 \rightarrow s_5$

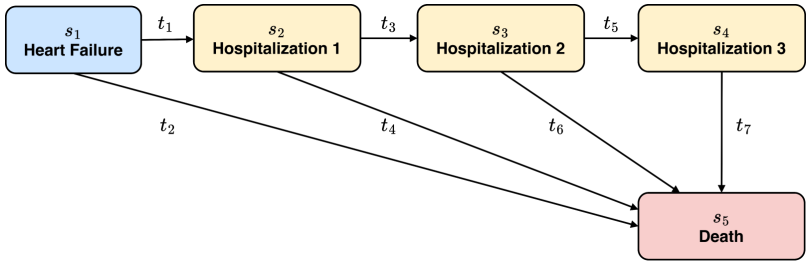
Transition Matrix

	s_1	s_2	s_3	s_4	s_5
s_1		t_1			t_2
s_2			t_3		t_4
s_3				t_5	t_6
s_4					t_7

Data Preparation: HF Multi-State Model

- Let's use the wide-format to prepare multi-state data

PID	time(s_1)	stat(s_1)	time(s_2)	stat(s_2)	time(s_3)	stat(s_3)	time(s_4)	stat(s_4)	time(s_5)	stat(s_5)
1	0.00	1	8.13	1	28.43	1	29.47	0	29.47	1
2	0.00	1	22.33	1	32.53	0	32.53	0	32.53	1
3	0.00	1	8.13	1	18.33	1	20.30	1	30.50	1



- Patient 1 progressed from $s_1 \rightarrow s_2 \rightarrow s_3 \rightarrow s_5$
  - Admitted twice in the hospital

row	PID	from	to	tstart	tstops	status
1	1	s_1	s_2	0.00	8.13	1
2	1	s_1	s_5	0.00	8.13	0
3	1	s_2	s_3	8.13	28.43	1
4	1	s_2	s_5	8.13	28.43	0
5	1	s_3	s_4	28.43	29.47	0
6	1	s_3	s_5	28.43	29.47	1

Transition Matrix

	s_1	s_2	s_3	s_4	s_5
s_1		t_1			t_2
s_2			t_3		t_4
s_3				t_5	t_6
s_4					t_7

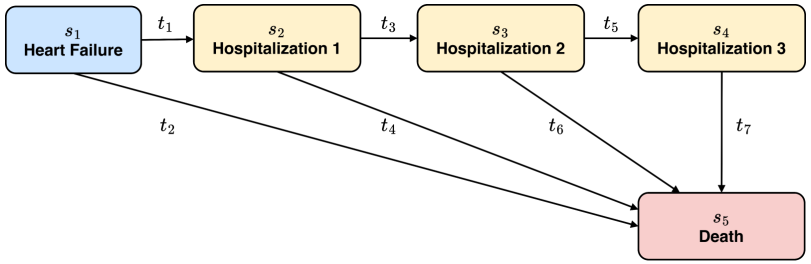
patient-at-risk

- Patient 1 ($s_1 \rightarrow s_2 \rightarrow s_3 \rightarrow s_5$) should have six rows
- Observe that each state is at risk to two states (two rows)

Data Preparation: HF Multi-State Model

- Let's use the wide-format to prepare multi-state data

PID	time(s_1)	stat(s_1)	time(s_2)	stat(s_2)	time(s_3)	stat(s_3)	time(s_4)	stat(s_4)	time(s_5)	stat(s_5)
1	0.00	1	8.13	1	28.43	1	29.47	0	29.47	1
2	0.00	1	22.33	1	32.53	0	32.53	0	32.53	1
3	0.00	1	8.13	1	18.33	1	20.30	1	30.50	1



- Patient 2 progressed from $s_1 \rightarrow s_2 \rightarrow s_5$
  - Admitted once to the hospital prior to death

row	PID	from	to	tstart	tstops	status
1	2	s_1	s_2	0.00	22.33	1
2	2	s_1	s_5	0.00	22.33	0
3	2	s_2	s_3	22.33	32.53	0
4	2	s_2	s_5	22.33	32.53	1

- Patient 2 ($s_1 \rightarrow s_2 \rightarrow s_5$) should have four rows
- Like in patient 1, each state is at risk to two states (two rows)

Transition Matrix

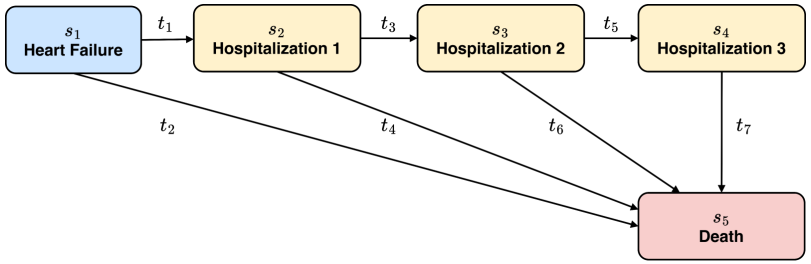
	s_1	s_2	s_3	s_4	s_5
s_1		t_1			t_2
s_2			t_3		t_4
s_3				t_5	t_6
s_4					t_7

patient-at-risk

Data Preparation: HF Multi-State Model

- Let's use the wide-format to prepare multi-state data

PID	time(s_1)	stat(s_1)	time(s_2)	stat(s_2)	time(s_3)	stat(s_3)	time(s_4)	stat(s_4)	time(s_5)	stat(s_5)
1	0.00	1	8.13	1	28.43	1	29.47	0	29.47	1
2	0.00	1	22.33	1	32.53	0	32.53	0	32.53	1
3	0.00	1	8.13	1	18.33	1	20.30	1	30.50	1



- Patient 3 progressed from $s_1 \rightarrow s_2 \rightarrow s_3 \rightarrow s_4 \rightarrow s_5$
  - Admitted thrice to the hospital prior to death

row	PID	from	to	tstart	tstops	status
1	3	s_1	s_2	0.00	8.13	1
2	3	s_1	s_5	0.00	8.13	0
3	3	s_2	s_3	8.13	18.33	1
4	3	s_2	s_5	8.13	18.33	0
5	3	s_3	s_4	18.33	20.30	1
6	3	s_3	s_5	18.33	20.30	0
7	3	s_4	s_5	20.30	30.50	1

Transition Matrix

	s_1	s_2	s_3	s_4	s_5
s_1		t_1			t_2
s_2			t_3		t_4
s_3				t_5	t_6
s_4					t_7

patient-at-risk

- Patient 3 ($s_1 \rightarrow s_2 \rightarrow s_3 \rightarrow s_4 \rightarrow s_5$) should have 7 rows
- Each state is at risk to two states (two rows) except s_4

Data Preparation: HF Multi-State Model

- Checking the number of rows for all patient is impossible

- How to check the overall data?
  - HF multi-state model has $k = 5$ states
  - Number of states of rehospitalization $k - 2$

- Suppose patient i has n hospitalizations, then patient i has rows r_i with R as the total number of rows

$$r_i = \begin{cases} 2n + 1; \text{if } s_{k-1} \in \mathbb{P} \\ 2n + 2; \text{otherwise} \end{cases} \qquad R = \sum_i r_i$$

- Patient 1 has 2 hospitalizations $r_1 = 2n + 2 = 6$
- Patient 2 has 1 hospitalization $r_2 = 2n + 2 = 4$

- Patient 3 has 3 hospitalizations and reached s_4

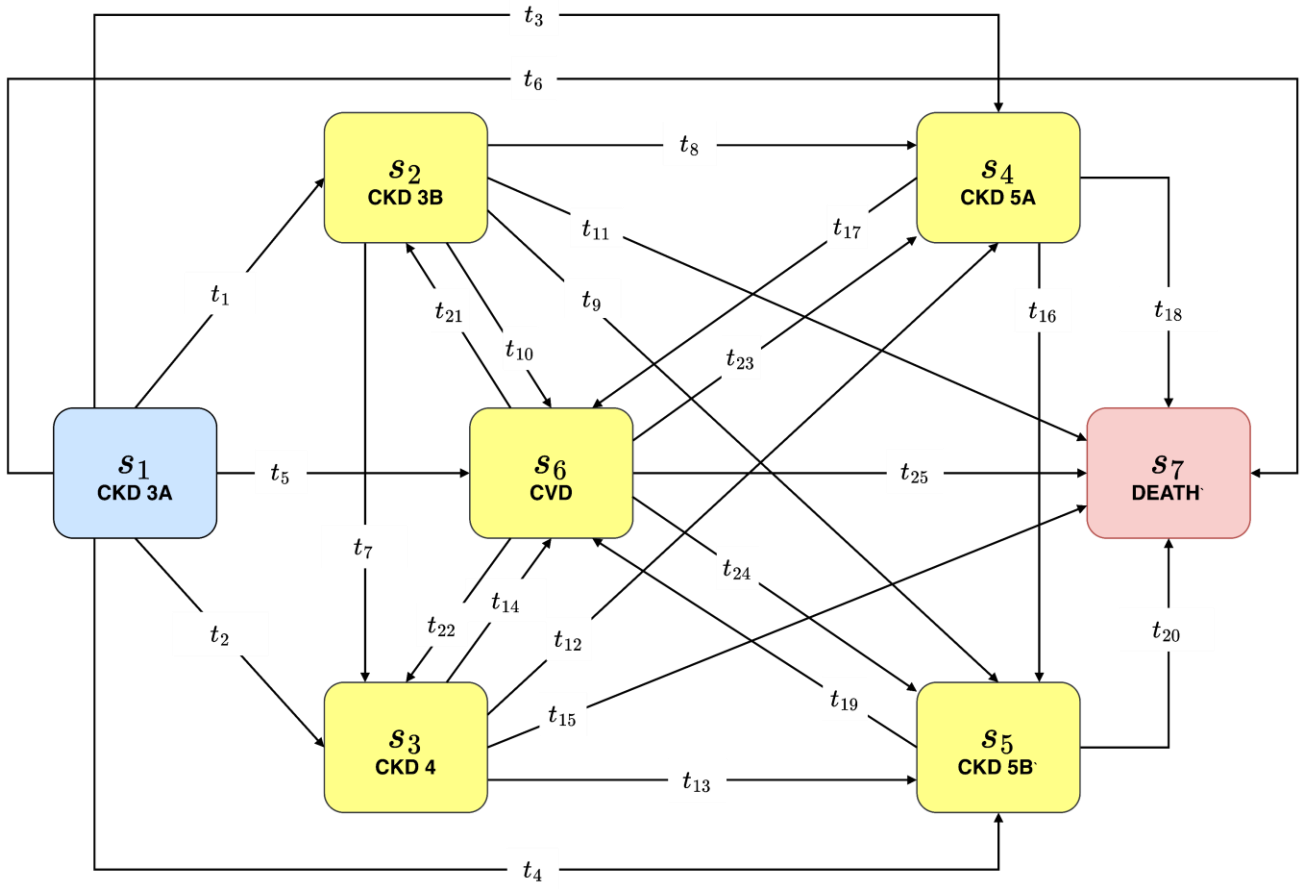
$$r_3 = 2n + 1 = 7$$

- $R = r_1 + r_2 + r_3 = 17$

- Therefore, 17 rows in total.

row	PID	from	to	tstart	tstops	status
1	1	s_1	s_2	0.00	8.13	1
2	1	s_1	s_5	0.00	8.13	0
3	1	s_2	s_3	8.13	28.43	1
4	1	s_2	s_5	8.13	28.43	0
5	1	s_3	s_4	28.43	29.47	0
6	1	s_3	s_5	28.43	29.47	1
7	2	s_1	s_2	0.00	22.33	1
8	2	s_1	s_5	0.00	22.33	0
9	2	s_2	s_3	22.33	32.53	0
10	2	s_2	s_5	22.33	32.53	1
11	3	s_1	s_2	0.00	8.13	1
12	3	s_1	s_5	0.00	8.13	0
13	3	s_2	s_3	8.13	18.33	1
14	3	s_2	s_5	8.13	18.33	0
15	3	s_3	s_4	18.33	20.30	1
16	3	s_3	s_5	18.33	20.30	0
17	3	s_4	s_5	20.30	30.50	1

CKD Multi-State Model



- Consider this multi-state model with seven states and 25 transitions
- Allows reversible transitions from CVD

Transition Matrix

	CKD3A	CKD3B	CKD4	CKD5A	CKD5B	CVD	DEATH
CKD3A		t_1	t_2	t_3	t_4	t_5	t_6
CKD3B			t_7	t_8	t_9	t_{10}	t_{11}
CKD4				t_{12}	t_{13}	t_{14}	t_{15}
CKD5A					t_{16}	t_{17}	t_{18}
CKD5B						t_{19}	t_{20}
CVD		t_{21}	t_{22}	t_{23}	t_{24}		t_{25}

Assumptions in CKD Multi-State Model

1. Linear pathway assumption
   - $\mathrm{CKD3A} \rightarrow \mathrm{CKD3B} \rightarrow \mathrm{CKD4} \rightarrow \mathrm{CKD5A} \rightarrow \mathrm{CKD5B} \rightarrow \mathrm{CVD} \rightarrow \mathrm{DEATH}$
   - CKD is characterized by a progressive, irreversible loss of kidney function over time

2. Mutual exclusivity assumption
   - Each state in the model is distinct and mutually exclusive
     - e.g., if patient is in one state, they cannot be in another state at the same time.

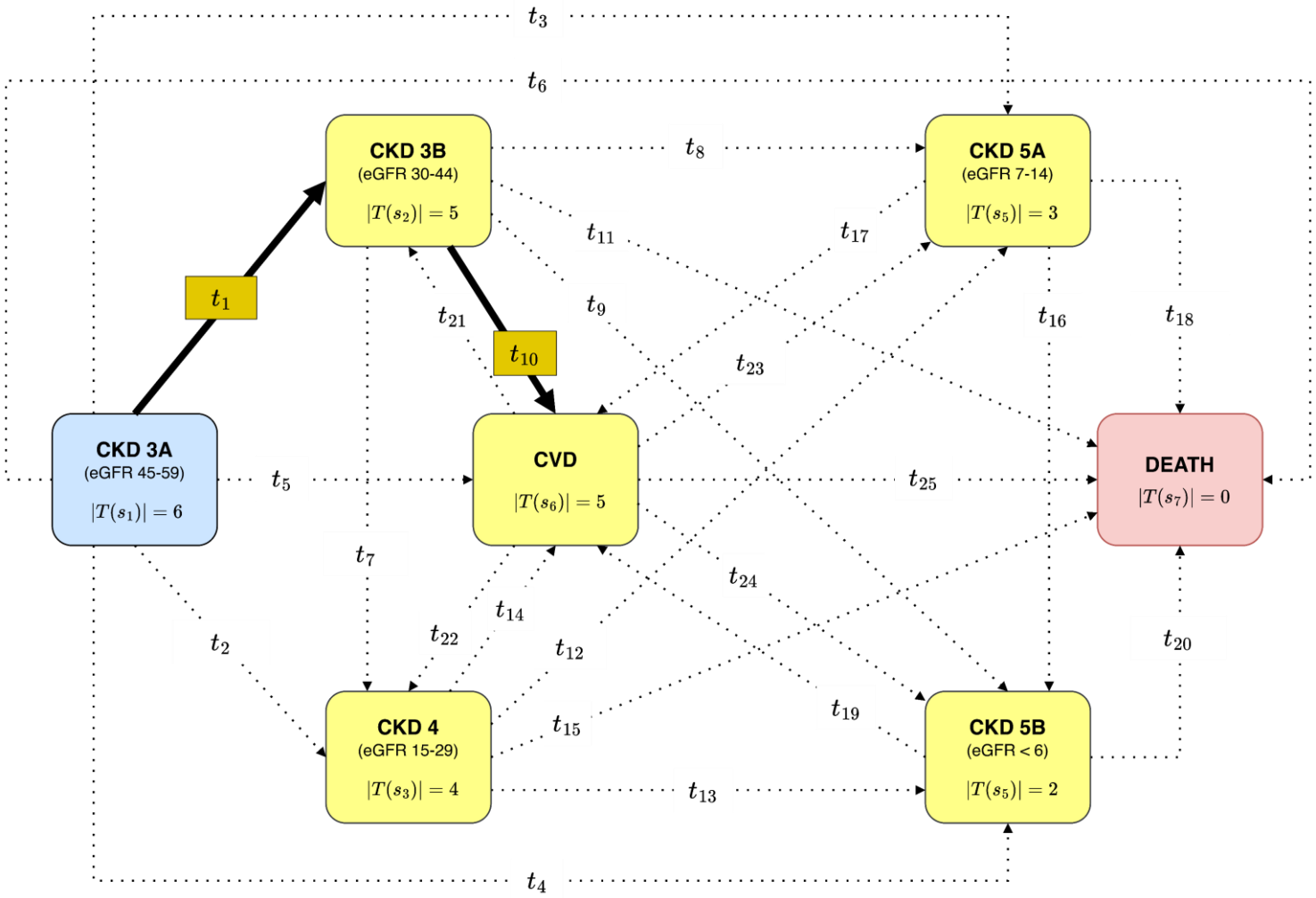
3. Patients cannot transition back to lower complication state after the subject enters a more severe complication state
   - Incorrect: $\mathrm{CKD3A} \rightarrow \mathrm{CKD3B} \rightarrow \mathrm{CKD3A}$

4. Transition is allowed from CVD state to other CKD complication state

5. Each patient will be followed until get into absorbing state (death) and cannot be move out from this state

Data Preparation: CKD Multi-State Model

Example 1. Suppose a patient has the following disease pathway: **CKD3A → CKD3B → CVD**



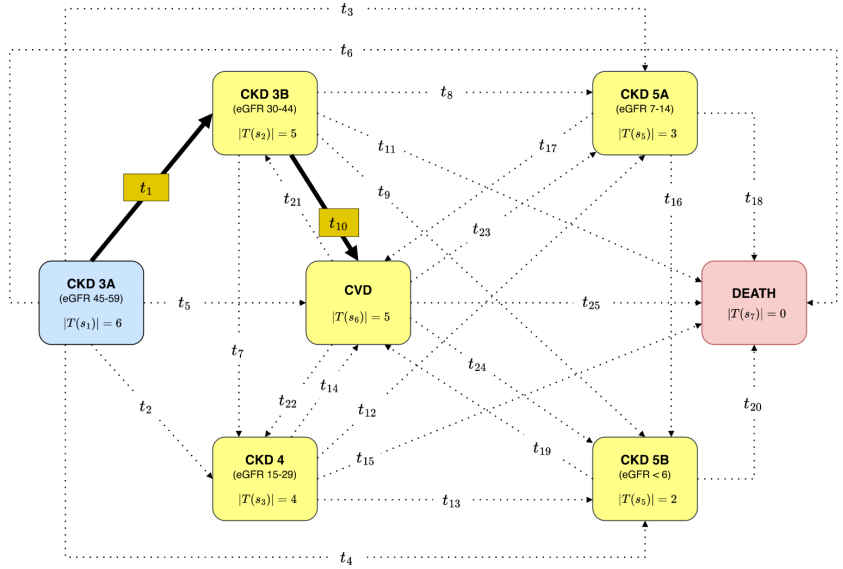
Data Preparation: CKD Multi-State Model

Suppose a patient has the following disease pathway: **CKD3A → CKD3B → CVD**

- At CKD3A, patient is at risk to six different states

row	PID	from	to	tstart	tstops	status
1	1	CKD3A	CKD3B	0.00	4.64	1
2	1	CKD3A	CKD4	0.00	4.64	0
3	1	CKD3A	CKD5A	0.00	4.64	0
4	1	CKD3A	CKD5B	0.00	4.64	0
5	1	CKD3A	CVD	0.00	4.64	0
6	1	CKD3A	DEAD	0.00	4.64	0

- Patient entered CKD3B at 4.64th month
- From CKD3A, this patient should have 6 rows



	CKD3A	CKD3B	CKD4	CKD5A	CKD5B	CVD	DEATH
CKD3A		t_1	t_2	t_3	t_4	t_5	t_6
CKD3B			t_7	t_8	t_9	t_{10}	t_{11}
CKD4				t_{12}	t_{13}	t_{14}	t_{15}
CKD5A					t_{16}	t_{17}	t_{18}
CKD5B						t_{19}	t_{20}
CVD		t_{21}	t_{22}	t_{23}	t_{24}		t_{25}

Data Preparation: CKD Multi-State Model

Suppose a patient has the following disease pathway: **CKD3A → CKD3B → CVD**

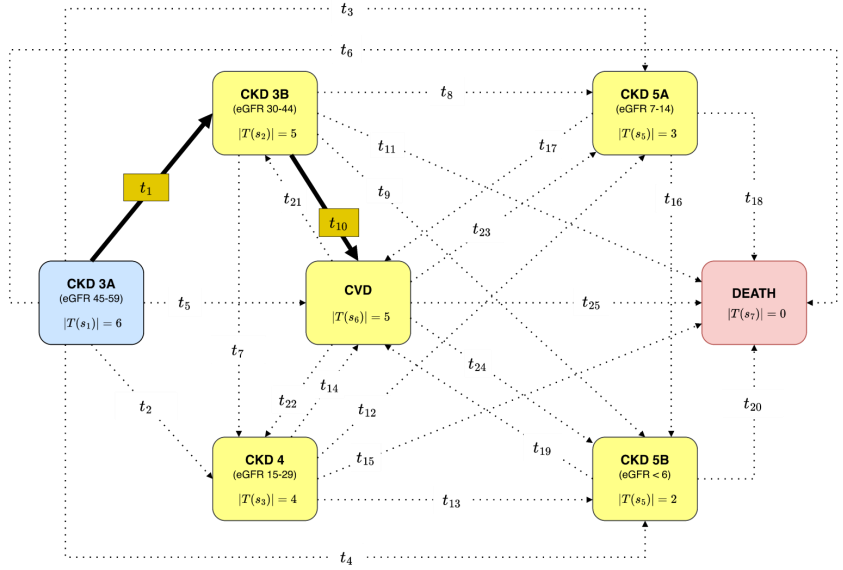
- At CKD3A, patient is at risk to six different states

row	PID	from	to	tstart	tstops	status
1	1	CKD3A	CKD3B	0.00	4.64	1
2	1	CKD3A	CKD4	0.00	4.64	0
3	1	CKD3A	CKD5A	0.00	4.64	0
4	1	CKD3A	CKD5B	0.00	4.64	0
5	1	CKD3A	CVD	0.00	4.64	0
6	1	CKD3A	DEAD	0.00	4.64	0

- Patient entered CKD3B at 4.64th month
- From CKD3A, this patient should have 6 rows
- At CKD3B, patient is at risk to five states

row	PID	from	to	tstart	tstops	status
1	1	CKD3B	CKD4	4.64	141.11	0
2	1	CKD3B	CKD5A	4.64	141.11	0
3	1	CKD3B	CKD5B	4.64	141.11	0
4	1	CKD3B	CVD	4.64	141.11	1
5	1	CKD3B	DEAD	4.64	141.11	0

- From CKD3B, patient should have 5 rows



	CKD3A	CKD3B	CKD4	CKD5A	CKD5B	CVD	DEATH
CKD3A		t_1	t_2	t_3	t_4	t_5	t_6
CKD3B			t_7	t_8	t_9	t_{10}	t_{11}
CKD4				t_{12}	t_{13}	t_{14}	t_{15}
CKD5A					t_{16}	t_{17}	t_{18}
CKD5B						t_{19}	t_{20}
CVD		t_{21}	t_{22}	t_{23}	t_{24}		t_{25}

Data Preparation: CKD Multi-State Model

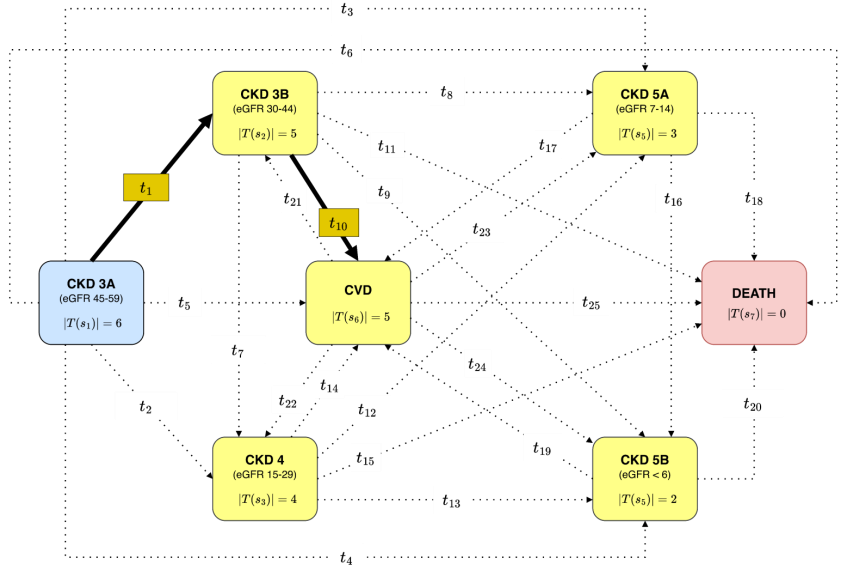
Suppose a patient has the following disease pathway: **CKD3A → CKD3B → CVD**

- At 141.11th month, patient moved from CKD3B to CVD
- A CVD patient is at risk of transitioning to 5 states

- Patient already diagnosed with CKD3B
- Hence, patient cannot transition back from CKD3B
  - Remove CVD to CKD3B

row	PID	from	to	tstart	tstops	status
~~1~~	~~1~~	~~CVD~~	~~CKD3B~~	~~141.11~~	~~149.10~~	~~0~~
2	1	CVD	CKD4	141.11	149.10	0
3	1	CVD	CKD5A	141.11	149.10	0
4	1	CVD	CKD5B	141.11	149.10	0
5	1	CVD	DEAD	141.11	149.10	0

- From CVD, this patient should only have 4 rows



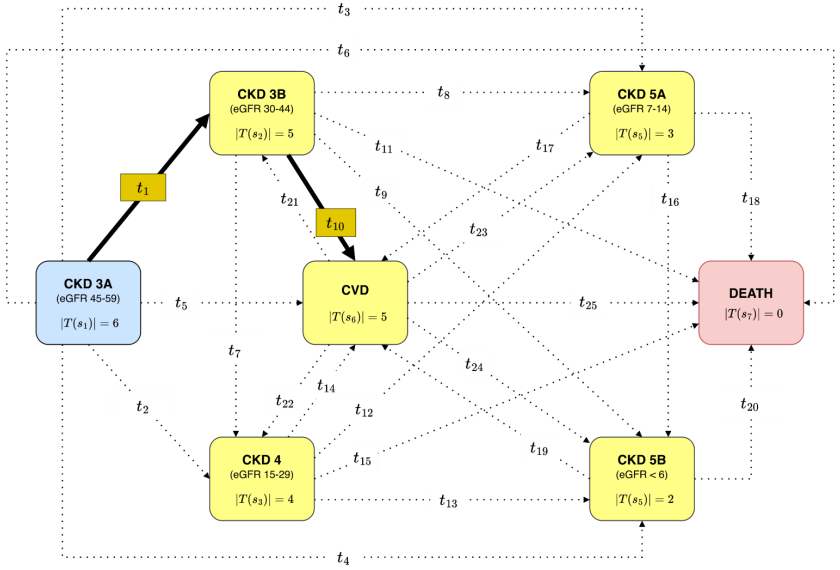
	CKD3A	CKD3B	CKD4	CKD5A	CKD5B	CVD	DEATH
CKD3A		t_1	t_2	t_3	t_4	t_5	t_6
CKD3B			t_7	t_8	t_9	t_{10}	t_{11}
CKD4				t_{12}	t_{13}	t_{14}	t_{15}
CKD5A					t_{16}	t_{17}	t_{18}
CKD5B						t_{19}	t_{20}
CVD		t_{21}	t_{22}	t_{23}	t_{24}		t_{25}

Data Preparation: CKD Multi-State Model

Suppose a patient has the following disease pathway: **CKD3A → CKD3B → CVD**

- CKD3A → CKD3B → CVD pathway should have 15 rows

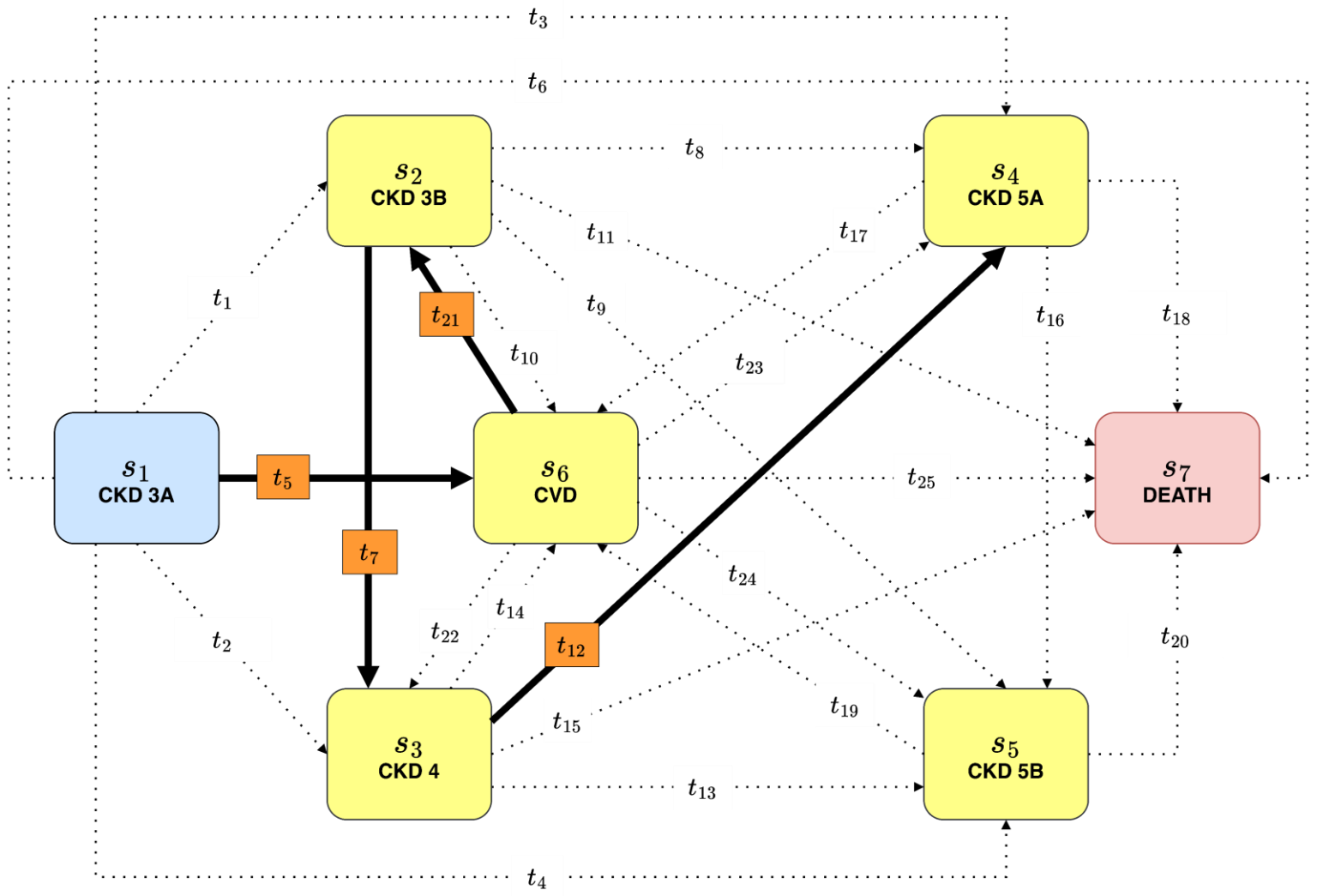
row	PID	from	to	tstart	tstops	status
1	1	CKD3A	CKD3B	0.00	4.64	1
2	1	CKD3A	CKD4	0.00	4.64	0
3	1	CKD3A	CKD5A	0.00	4.64	0
4	1	CKD3A	CKD5B	0.00	4.64	0
5	1	CKD3A	CVD	0.00	4.64	0
6	1	CKD3A	DEAD	0.00	4.64	0
7	1	CKD3B	CKD4	4.64	141.11	0
8	1	CKD3B	CKD5A	4.64	141.11	0
9	1	CKD3B	CKD5B	4.64	141.11	0
10	1	CKD3B	CVD	4.64	141.11	1
11	1	CKD3B	DEAD	4.64	141.11	0
12	1	CVD	CKD4	141.11	149.10	0
13	1	CVD	CKD5A	141.11	149.10	0
14	1	CVD	CKD5B	141.11	149.10	0
15	1	CVD	DEAD	141.11	149.10	0



	CKD3A	CKD3B	CKD4	CKD5A	CKD5B	CVD	DEATH
CKD3A		t_1	t_2	t_3	t_4	t_5	t_6
CKD3B			t_7	t_8	t_9	t_{10}	t_{11}
CKD4				t_{12}	t_{13}	t_{14}	t_{15}
CKD5A					t_{16}	t_{17}	t_{18}
CKD5B						t_{19}	t_{20}
CVD		t_{21}	t_{22}	t_{23}	t_{24}		t_{25}

Data Preparation: CKD Multi-State Model

Example 2: **CKD3A → CVD → CKD3B → CKD4 → CKD5A**



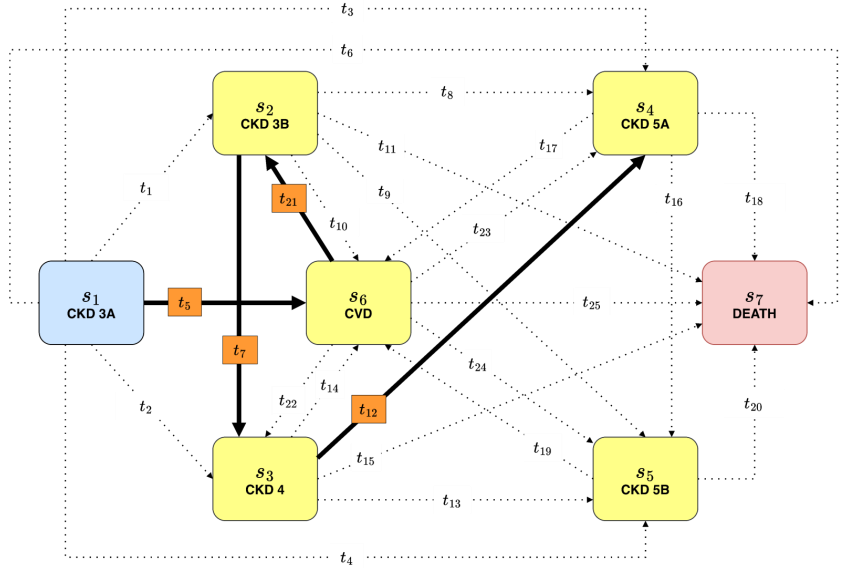
Data Preparation: CKD Multi-State Model

Suppose a patient has the following disease pathway: **CKD3A → CVD → CKD3B → CKD4 → CKD5A**

- A CKD3A patient is at risk to six different states

row	PID	trans	from	to	status	tstart	tstops	time
1	2	1	CKD3A	CKD3B	0	0.00	47.05	47.05
2	2	2	CKD3A	CKD4	0	0.00	47.05	47.05
3	2	3	CKD3A	CKD5A	0	0.00	47.05	47.05
4	2	4	CKD3A	CKD5B	0	0.00	47.05	47.05
5	2	5	CKD3A	CVD	1	0.00	47.05	47.05
6	2	6	CKD3A	DEAD	0	0.00	47.05	47.05

- Hence, CKD3A should have 6 rows



	CKD3A	CKD3B	CKD4	CKD5A	CKD5B	CVD	DEATH
CKD3A		t_1	t_2	t_3	t_4	t_5	t_6
CKD3B			t_7	t_8	t_9	t_{10}	t_{11}
CKD4				t_{12}	t_{13}	t_{14}	t_{15}
CKD5A					t_{16}	t_{17}	t_{18}
CKD5B						t_{19}	t_{20}
CVD		t_{21}	t_{22}	t_{23}	t_{24}		t_{25}

Data Preparation: CKD Multi-State Model

Suppose a patient has the following disease pathway: **CKD3A → CVD → CKD3B → CKD4 → CKD5A**

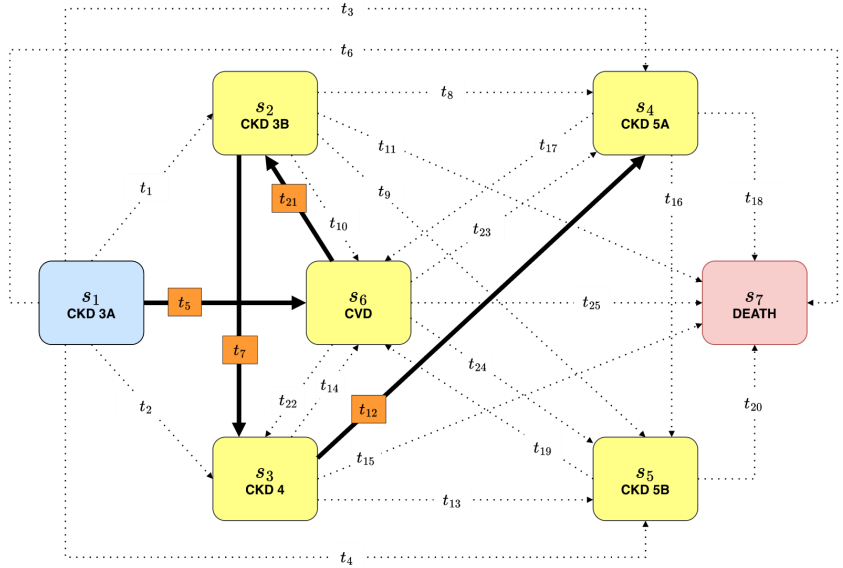
- A CKD3A patient is at risk to six different states

row	PID	trans	from	to	status	tstart	tstops	time
1	2	1	CKD3A	CKD3B	0	0.00	47.05	47.05
2	2	2	CKD3A	CKD4	0	0.00	47.05	47.05
3	2	3	CKD3A	CKD5A	0	0.00	47.05	47.05
4	2	4	CKD3A	CKD5B	0	0.00	47.05	47.05
5	2	5	CKD3A	CVD	1	0.00	47.05	47.05
6	2	6	CKD3A	DEAD	0	0.00	47.05	47.05

- Hence, CKD3A should have 6 rows
- At 47.05th month, CKD3A patient entered CVD which is at risk of entering to five states

row	PID	trans	from	to	status	tstart	tstops	time
1	2	21	CVD	CKD3B	1	47.05	49.55	2.50
2	2	22	CVD	CKD4	0	47.05	49.55	2.50
3	2	23	CVD	CKD5A	0	47.05	49.55	2.50
4	2	24	CVD	CKD5B	0	47.05	49.55	2.50
5	2	25	CVD	DEAD	0	47.05	49.55	2.50

- Patient should have 5 rows from CVD state.



	CKD3A	CKD3B	CKD4	CKD5A	CKD5B	CVD	DEATH
CKD3A		t_1	t_2	t_3	t_4	t_5	t_6
CKD3B			t_7	t_8	t_9	t_{10}	t_{11}
CKD4				t_{12}	t_{13}	t_{14}	t_{15}
CKD5A					t_{16}	t_{17}	t_{18}
CKD5B						t_{19}	t_{20}
CVD		t_{21}	t_{22}	t_{23}	t_{24}		t_{25}

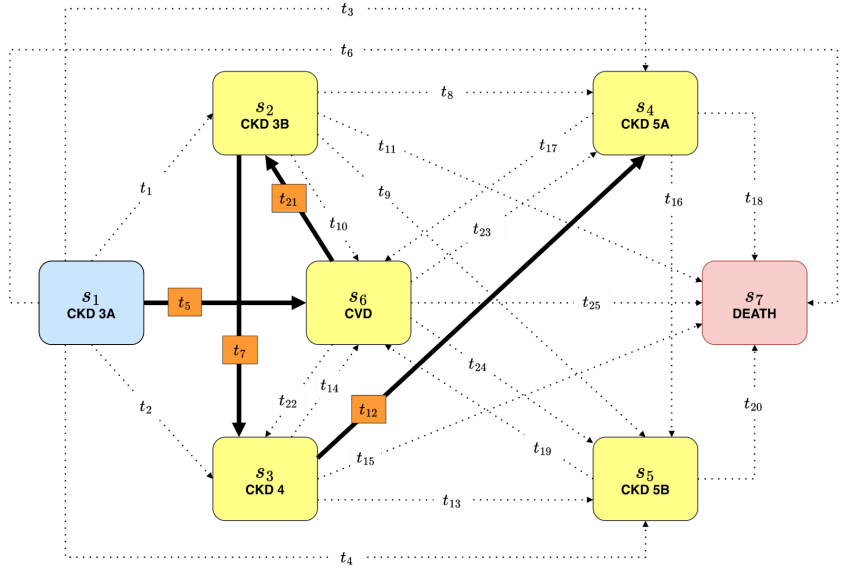
Data Preparation: CKD Multi-State Model

Suppose a patient has the following disease pathway: **CKD3A → CVD → CKD3B → CKD4 → CKD5A**

- When CVD patient entered CKD3B at 49.55th month,

row	PID	trans	from	to	status	tstart	tstops	time
1	2	7	CKD3B	CKD4	1	49.55	53.95	4.41
2	2	8	CKD3B	CKD5A	0	49.55	53.95	4.41
3	2	9	CKD3B	CKD5B	0	49.55	53.95	4.41
4	2	11	CKD3B	DEAD	0	49.55	53.95	4.41

patient is at risk of entering to 4 states (not 5)



	CKD3A	CKD3B	CKD4	CKD5A	CKD5B	CVD	DEATH
CKD3A		t_1	t_2	t_3	t_4	t_5	t_6
CKD3B			t_7	t_8	t_9	t_{10}	t_{11}
CKD4				t_{12}	t_{13}	t_{14}	t_{15}
CKD5A					t_{16}	t_{17}	t_{18}
CKD5B						t_{19}	t_{20}
CVD		t_{21}	t_{22}	t_{23}	t_{24}		t_{25}

Data Preparation: CKD Multi-State Model

Suppose a patient has the following disease pathway: **CKD3A → CVD → CKD3B → CKD4 → CKD5A**

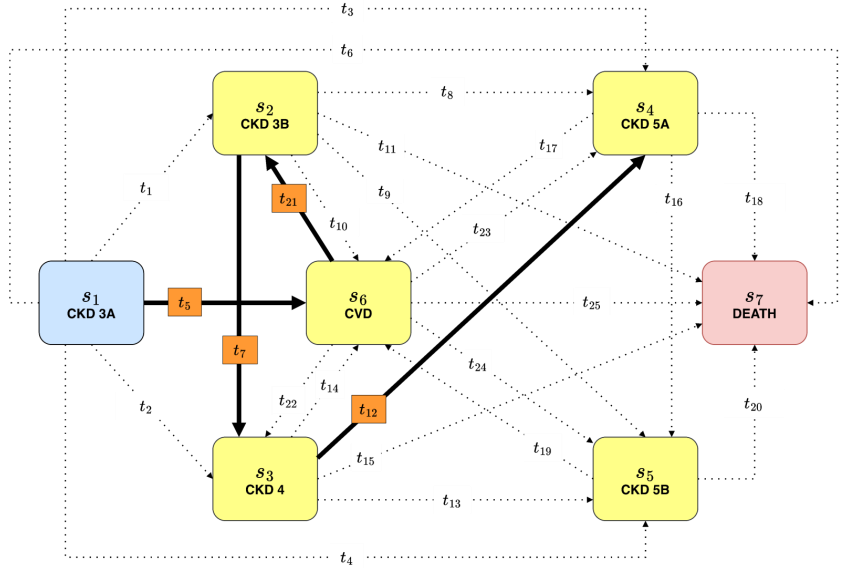
- When CVD patient entered CKD3B at 49.55th month,

row	PID	trans	from	to	status	tstart	tstops	time
1	2	7	CKD3B	CKD4	1	49.55	53.95	4.41
2	2	8	CKD3B	CKD5A	0	49.55	53.95	4.41
3	2	9	CKD3B	CKD5B	0	49.55	53.95	4.41
4	2	11	CKD3B	DEAD	0	49.55	53.95	4.41

patient is at risk of entering to 4 states (not 5)

- At 53.95th month, patient entered CKD4 which is at risk entering to 3 states (not four)

row	PID	trans	from	to	status	tstart	tstops	time
1	2	12	CKD4	CKD5A	1	53.95	57.47	7.92
2	2	13	CKD4	CKD5B	0	53.95	57.47	7.92
3	2	15	CKD4	DEAD	0	53.95	57.47	7.92



	CKD3A	CKD3B	CKD4	CKD5A	CKD5B	CVD	DEATH
CKD3A		t_1	t_2	t_3	t_4	t_5	t_6
CKD3B			t_7	t_8	t_9	t_{10}	t_{11}
CKD4				t_{12}	t_{13}	t_{14}	t_{15}
CKD5A					t_{16}	t_{17}	t_{18}
CKD5B						t_{19}	t_{20}
CVD		t_{21}	t_{22}	t_{23}	t_{24}		t_{25}

Data Preparation: CKD Multi-State Model

Suppose a patient has the following disease pathway: **CKD3A → CVD → CKD3B → CKD4 → CKD5A**

- When CVD patient entered CKD3B at 49.55th month,

row	PID	trans	from	to	status	tstart	tstops	time
1	2	7	CKD3B	CKD4	1	49.55	53.95	4.41
2	2	8	CKD3B	CKD5A	0	49.55	53.95	4.41
3	2	9	CKD3B	CKD5B	0	49.55	53.95	4.41
4	2	11	CKD3B	DEAD	0	49.55	53.95	4.41

patient is at risk of entering to 4 states (not 5)

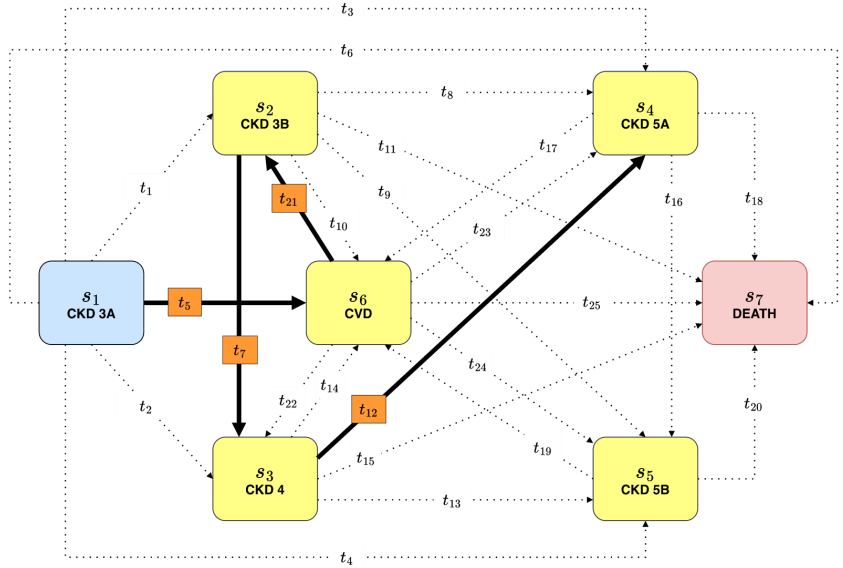
- At 53.95th month, patient entered CKD4 which is at risk entering to 3 states (not four)

row	PID	trans	from	to	status	tstart	tstops	time
1	2	12	CKD4	CKD5A	1	53.95	57.47	7.92
2	2	13	CKD4	CKD5B	0	53.95	57.47	7.92
3	2	15	CKD4	DEAD	0	53.95	57.47	7.92

- At 54.47th month, patient entered CKD5A state, which is at risk of entering to 2 states (not 3)

row	PID	trans	from	to	status	tstart	tstops	time
1	2	16	CKD5A	CKD5B	0	57.47	150.97	93.50
2	2	18	CKD5A	DEAD	0	57.47	150.97	93.50

- From CKD5A state, this patient should have 2 rows.



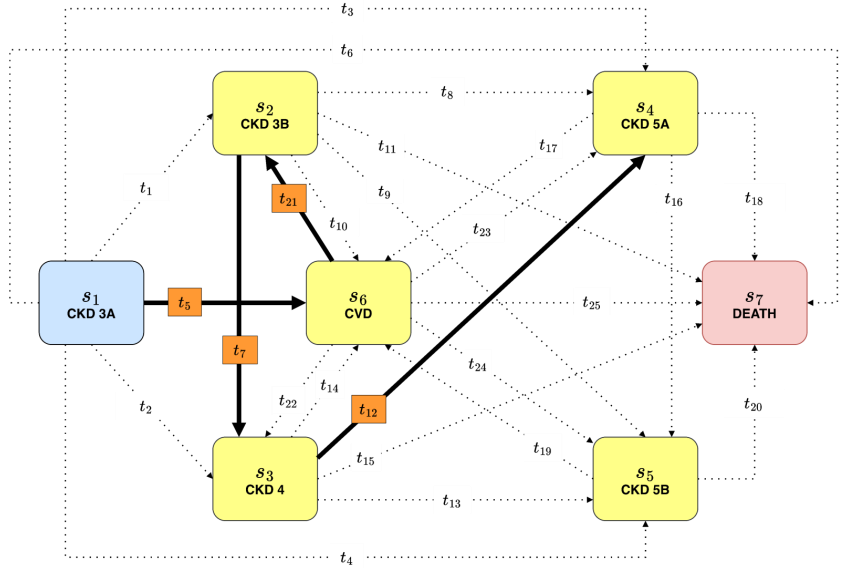
	CKD3A	CKD3B	CKD4	CKD5A	CKD5B	CVD	DEATH
CKD3A		t_1	t_2	t_3	t_4	t_5	t_6
CKD3B			t_7	t_8	t_9	t_{10}	t_{11}
CKD4				t_{12}	t_{13}	t_{14}	t_{15}
CKD5A					t_{16}	t_{17}	t_{18}
CKD5B						t_{19}	t_{20}
CVD		t_{21}	t_{22}	t_{23}	t_{24}		t_{25}

Data Preparation: CKD Multi-State Model

Suppose a patient has the following disease pathway: **CKD3A → CVD → CKD3B → CKD4 → CKD5A**

- Overall, patients who have this disease pathway should have 20 rows.

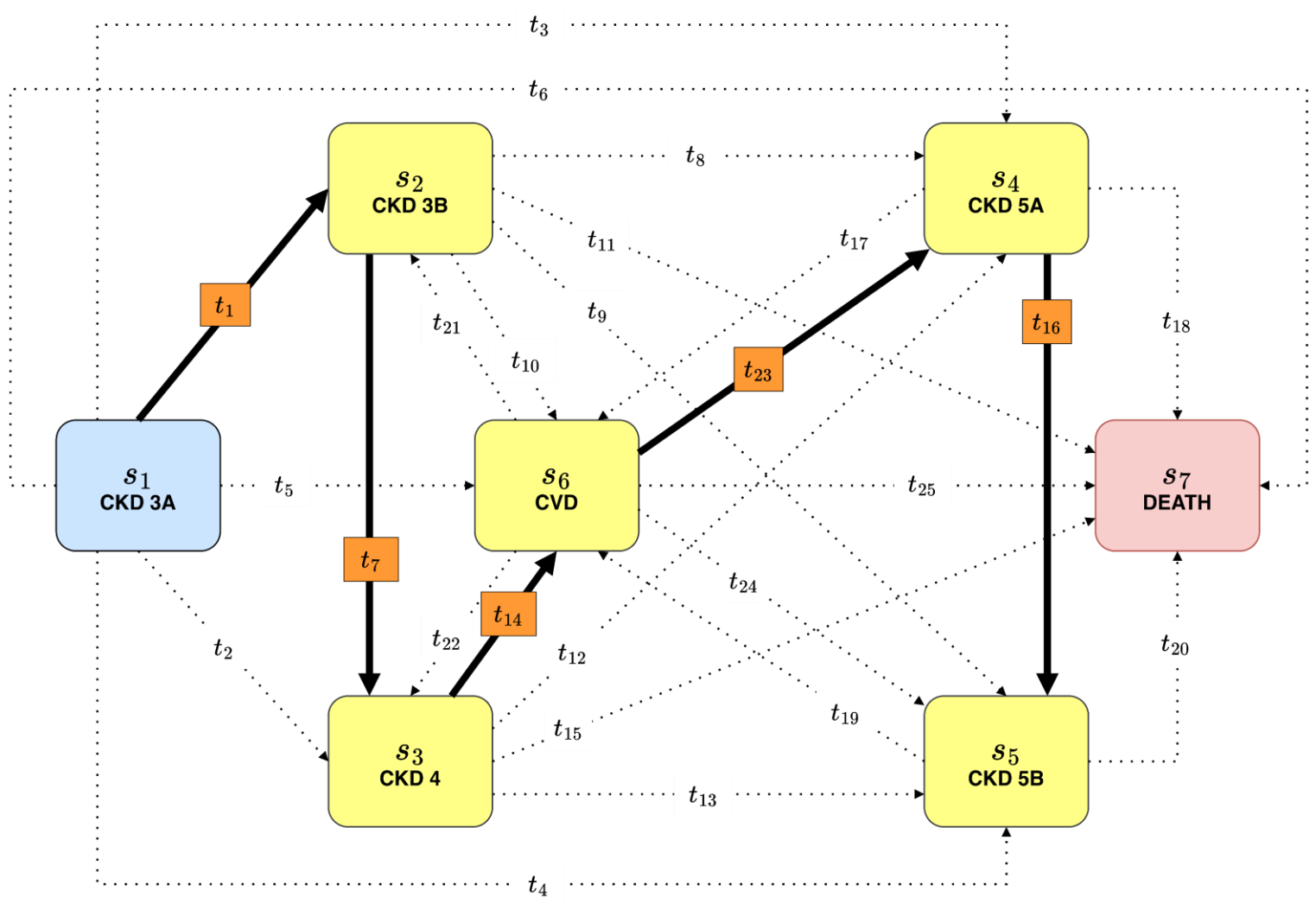
Row	PID	trans	from	to	status	tstart	tstops	time
1	2	1	CKD3A	CKD3B	0	0.00	47.05	47.05
2	2	2	CKD3A	CKD4	0	0.00	47.05	47.05
3	2	3	CKD3A	CKD5A	0	0.00	47.05	47.05
4	2	4	CKD3A	CKD5B	0	0.00	47.05	47.05
5	2	5	CKD3A	CVD	1	0.00	47.05	47.05
6	2	6	CKD3A	DEAD	0	0.00	47.05	47.05
7	2	7	CKD3B	CKD4	1	49.55	53.95	4.41
8	2	8	CKD3B	CKD5A	0	49.55	53.95	4.41
9	2	9	CKD3B	CKD5B	0	49.55	53.95	4.41
10	2	11	CKD3B	DEAD	0	49.55	53.95	4.41
11	2	12	CKD4	CKD5A	1	53.95	57.47	7.92
12	2	13	CKD4	CKD5B	0	53.95	57.47	7.92
13	2	15	CKD4	DEAD	0	53.95	57.47	7.92
14	2	16	CKD5A	CKD5B	0	57.47	150.97	93.50
15	2	18	CKD5A	DEAD	0	57.47	150.97	93.50
16	2	21	CVD	CKD3B	1	47.05	49.55	2.50
17	2	22	CVD	CKD4	0	47.05	49.55	2.50
18	2	23	CVD	CKD5A	0	47.05	49.55	2.50
19	2	24	CVD	CKD5B	0	47.05	49.55	2.50
20	2	25	CVD	DEAD	0	47.05	49.55	2.50



	CKD3A	CKD3B	CKD4	CKD5A	CKD5B	CVD	DEATH
CKD3A		t_1	t_2	t_3	t_4	t_5	t_6
CKD3B			t_7	t_8	t_9	t_{10}	t_{11}
CKD4				t_{12}	t_{13}	t_{14}	t_{15}
CKD5A					t_{16}	t_{17}	t_{18}
CKD5B						t_{19}	t_{20}
CVD		t_{21}	t_{22}	t_{23}	t_{24}		t_{25}

Data Preparation: CKD Multi-State Model

Example 3. CKD3A → CKD3B → CKD4 → CVD → CKD5A → CKD5B

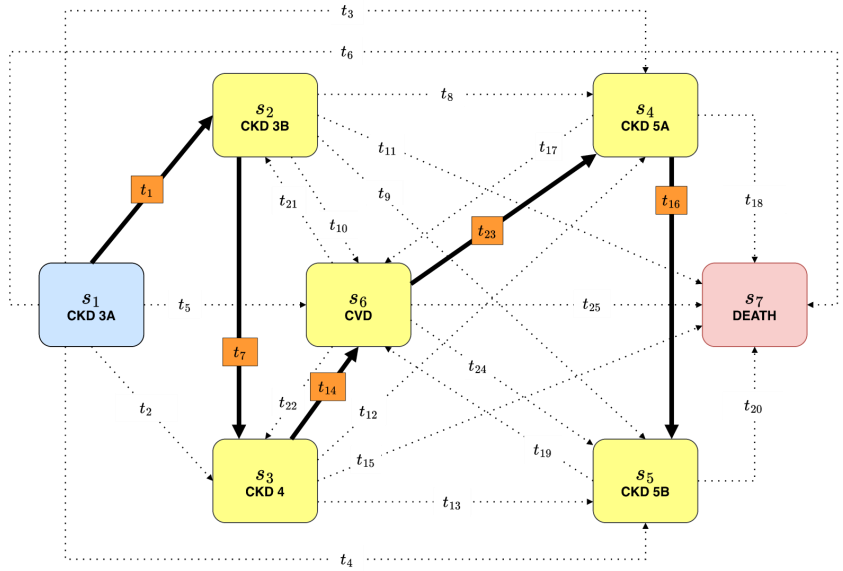


Data Preparation: CKD Multi-State Model

Suppose a patient has the following disease pathway: **CKD3A → CKD3B → CKD4 → CVD → CKD5A → CKD5B**

- For patients who were diagnosed with CKD3A to CKD3B to CKD4, initially, they should have 15 rows.
  - 6 (from CKD3A) + 5 (from CKD3B) + 4 (from CKD4)

row	PID	trans	from	to	status	tstart	tstops	time
1	3	1	CKD3A	CKD3B	1	0.00	8.68	8.68
2	3	2	CKD3A	CKD4	0	0.00	8.68	8.68
3	3	3	CKD3A	CKD5A	0	0.00	8.68	8.68
4	3	4	CKD3A	CKD5B	0	0.00	8.68	8.68
5	3	5	CKD3A	CVD	0	0.00	8.68	8.68
6	3	6	CKD3A	DEAD	0	0.00	8.68	8.68
7	3	7	CKD3B	CKD4	1	8.68	24.62	15.95
8	3	8	CKD3B	CKD5A	0	8.68	24.62	15.95
9	3	9	CKD3B	CKD5B	0	8.68	24.62	15.95
10	3	10	CKD3B	CVD	0	8.68	24.62	15.95
11	3	11	CKD3B	DEAD	0	8.68	24.62	15.95
12	3	12	CKD4	CKD5A	0	24.62	30.67	6.05
13	3	13	CKD4	CKD5B	0	24.62	30.67	6.05
14	3	14	CKD4	CVD	1	24.62	30.67	6.05
15	3	15	CKD4	DEAD	0	24.62	30.67	6.05



	CKD3A	CKD3B	CKD4	CKD5A	CKD5B	CVD	DEATH
CKD3A		t_1	t_2	t_3	t_4	t_5	t_6
CKD3B			t_7	t_8	t_9	t_{10}	t_{11}
CKD4				t_{12}	t_{13}	t_{14}	t_{15}
CKD5A					t_{16}	t_{17}	t_{18}
CKD5B						t_{19}	t_{20}
CVD		t_{21}	t_{22}	t_{23}	t_{24}		t_{25}

Data Preparation: CKD Multi-State Model

Suppose a patient has the following disease pathway: **CKD3A → CKD3B → CKD4 → CVD → CKD5A → CKD5B**

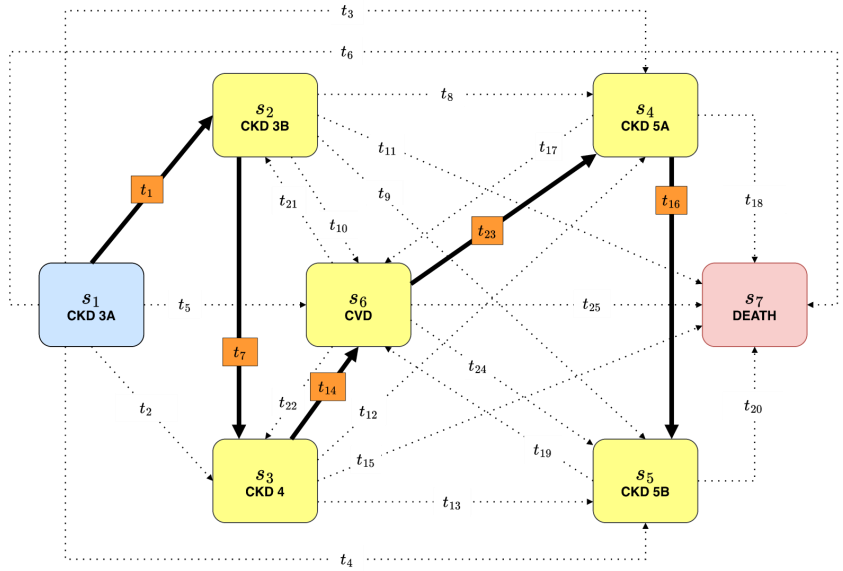
- However, when they entered CVD,

row	PID	trans	from	to	status	tstart	tstops	time
1	3	23	CVD	CKD5A	1	30.67	116.32	85.64
2	3	24	CVD	CKD5B	0	30.67	116.32	85.64
3	3	25	CVD	DEAD	0	30.67	116.32	85.64

only three rows will be added since patients were already diagnosed with CKD3B and CKD4 prior to their admission to CVD.

- When patients entered CKD5A and CKD5B, transitions 17 and 19 are also not included

row	PID	trans	from	to	status	tstart	tstops	time
1	3	16	CKD5A	CKD5B	1	116.32	125.36	9.04
2	3	18	CKD5A	DEAD	0	116.32	125.36	9.04
1	3	20	CKD5B	DEAD	0	125.36	134.23	8.87



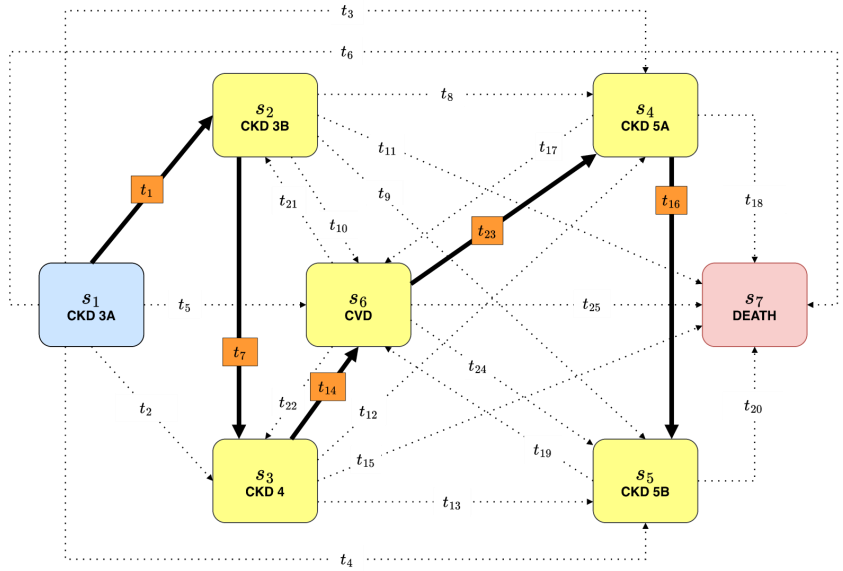
	CKD3A	CKD3B	CKD4	CKD5A	CKD5B	CVD	DEATH
CKD3A		t_1	t_2	t_3	t_4	t_5	t_6
CKD3B			t_7	t_8	t_9	t_{10}	t_{11}
CKD4				t_{12}	t_{13}	t_{14}	t_{15}
CKD5A					t_{16}	t_{17}	t_{18}
CKD5B						t_{19}	t_{20}
CVD		t_{21}	t_{22}	t_{23}	t_{24}		t_{25}

Data Preparation: CKD Multi-State Model

Suppose a patient has the following disease pathway: **CKD3A → CKD3B → CKD4 → CVD → CKD5A → CKD5B**

- Overall, patient with this pathway should have 21 rows.

row	PID	trans	from	to	status	tstart	tstops	time
1	3	1	CKD3A	CKD3B	1	0.00	8.68	8.68
2	3	2	CKD3A	CKD4	0	0.00	8.68	8.68
3	3	3	CKD3A	CKD5A	0	0.00	8.68	8.68
4	3	4	CKD3A	CKD5B	0	0.00	8.68	8.68
5	3	5	CKD3A	CVD	0	0.00	8.68	8.68
6	3	6	CKD3A	DEAD	0	0.00	8.68	8.68
7	3	7	CKD3B	CKD4	1	8.68	24.62	15.95
8	3	8	CKD3B	CKD5A	0	8.68	24.62	15.95
9	3	9	CKD3B	CKD5B	0	8.68	24.62	15.95
10	3	10	CKD3B	CVD	0	8.68	24.62	15.95
11	3	11	CKD3B	DEAD	0	8.68	24.62	15.95
12	3	12	CKD4	CKD5A	0	24.62	30.67	6.05
13	3	13	CKD4	CKD5B	0	24.62	30.67	6.05
14	3	14	CKD4	CVD	1	24.62	30.67	6.05
15	3	15	CKD4	DEAD	0	24.62	30.67	6.05
16	3	16	CKD5A	CKD5B	1	116.32	125.36	9.04
17	3	18	CKD5A	DEAD	0	116.32	125.36	9.04
18	3	20	CKD5B	DEAD	0	125.36	134.23	8.87
19	3	23	CVD	CKD5A	1	30.67	116.32	85.64
20	3	24	CVD	CKD5B	0	30.67	116.32	85.64
21	3	25	CVD	DEAD	0	30.67	116.32	85.64



	CKD3A	CKD3B	CKD4	CKD5A	CKD5B	CVD	DEATH
CKD3A		t_1	t_2	t_3	t_4	t_5	t_6
CKD3B			t_7	t_8	t_9	t_{10}	t_{11}
CKD4				t_{12}	t_{13}	t_{14}	t_{15}
CKD5A					t_{16}	t_{17}	t_{18}
CKD5B						t_{19}	t_{20}
CVD		t_{21}	t_{22}	t_{23}	t_{24}		t_{25}

How to check the overall multi-state model data?

Let's use this guide to represent the number of rows for each state

$$R = \sum_{i \in S}^{n} \left(|T(s_i)| - \left(\mathbb{1}_{\{\text{CVD} \in T(s_i)\}} \right) \left| T(s_i)_{\text{adj}} \right| \right)$$

- $|T(s_i)|$ be the number of possible transitions from state s_i before adjustments
- $\mathbb{1}_{\{\text{CVD} \in T(s_i)\}}$ indicator if 1 = CVD has occurred or 0 = hasn't occurred yet
- $\left| T(s_i)_{\text{adj}} \right|$ number of transitions removed from $|T(s_i)|$ at state s_i

Previous example: **CKD3A → CKD3B → CKD4 → CVD → CKD5A → CKD5B**

- CKD3A = 6 − (0)0 = **6**
- CKD3B = 5 − (0)0 = **5**
- CKD4 = 4 − (0)0 = **4**
- CVD = 5 − (1)2 = **3**
- CKD5A = 3 − (1)1 = **2**
- CKD5B = 2 − (1)1 = **1**

- R =**6 + 5 + 4 + 3 + 2 + 1 = 21 rows**

	CKD3A	CKD3B	CKD4	CKD5A	CKD5B	CVD	DEATH
CKD3A		t_1	t_2	t_3	t_4	t_5	t_6
CKD3B			t_7	t_8	t_9	t_{10}	t_{11}
CKD4				t_{12}	t_{13}	t_{14}	t_{15}
CKD5A					t_{16}	t_{17}	t_{18}
CKD5B						t_{19}	t_{20}
CVD		t_{21}	t_{22}	t_{23}	t_{24}		t_{25}

How to check your data?

1. Identify the unique disease progression (pathway) and calculate the number of patients who have the same disease progression

   1,840 patients: ('CKD3A', 'CKD3B', 'CKD4', 'CKD5A', 'CKD5B')

2. Calculate the number of rows for each pathway

   ('CKD3A', 'CKD3B', 'CKD4', 'CKD5A', 'CKD5B') = 20 rows

3. Determine the total rows by multiplying the number of patients with rows in that specific pathway

Transition Pathways	Patients	Rows	Total Rows
('CKD3A', 'CKD3B', 'CKD4', 'CKD5A', 'CKD5B')	1,840	20	36,800

4. Manual check 36800 of 313876 records found

How to check the overall multi-state model data?

Transition Pathways	No. of Patients	Rows	Total Rows
('CKD3A', 'CVD')	5,184	11	57,024
('CKD3A', 'CKD3B', 'CVD')	4,038	15	60,570
('CKD3A',)	3,477	6	20,862
('CKD3A', 'CKD3B')	2,679	11	29,469
('CKD3A', 'CKD3B', 'CKD4', 'CKD5A', 'CKD5B')	1,840	20	36,800
('CKD3A', 'CKD3B', 'CKD4', 'CKD5A', 'CKD5B', 'CVD')	1,118	21	23,478
('CKD3A', 'CKD3B', 'CKD4', 'CVD')	1,025	18	18,450
('CKD3A', 'CKD3B', 'CKD4')	880	15	13,200
('CKD3A', 'CVD', 'CKD3B')	824	15	12,360
('CKD3A', 'CKD3B', 'CVD', 'CKD4')	431	18	7,758
('CKD3A', 'CKD3B', 'CKD4', 'CKD5A')	269	18	4,842
('CKD3A', 'CKD3B', 'CKD4', 'CKD5A', 'CVD')	236	20	4,720
('CKD3A', 'CKD3B', 'CKD4', 'DEATH')	182	15	2,730
('CKD3A', 'CKD3B', 'CKD4', 'CKD5A', 'DEATH')	160	18	2,880
('CKD3A', 'CVD', 'CKD3B', 'CKD4')	152	18	2,736
('CKD3A', 'CKD3B', 'DEATH')	135	11	1,485
('CKD3A', 'CKD3B', 'CKD4', 'CKD5A', 'CKD5B', 'DEATH')	122	20	2,440
('CKD3A', 'CKD3B', 'CKD4', 'CVD', 'CKD5A')	112	20	2,240
('CKD3A', 'CKD3B', 'CKD4', 'CKD5A', 'CKD5B', 'CVD', 'DEATH')	77	21	1,617
('CKD3A', 'CKD3B', 'CVD', 'CKD4', 'CKD5A')	75	20	1,500

Excel (manual check)

57024 of 313876 records found
60570 of 313876 records found
20862 of 313876 records found
29469 of 313876 records found
36800 of 313876 records found
23478 of 313876 records found
18450 of 313876 records found
13200 of 313876 records found
12360 of 313876 records found
7758 of 313876 records found
4842 of 313876 records found
4720 of 313876 records found
2730 of 313876 records found
2880 of 313876 records found
2736 of 313876 records found
1485 of 313876 records found
2440 of 313876 records found
2240 of 313876 records found
1617 of 313876 records found
1500 of 313876 records found

Libraries and Functions

- msprep, mstate packages in R
  - Data preparation, model estimation, and prediction in multi-state models
  - Only allow data preparation for irreversible multi-state models
  - See more: https://github.com/hputter/mstate

- pymsm package in Python
  - Model estimation and prediction in multi-state models
  - See more: https://hrossman.github.io/pymsm

- multistate package in Stata
  - Data preparation, model estimation, and prediction in multi-state models
  - Sample doc file:
    https://github.com/enochytchen/NordicStata2022/blob/main/manual_msset.do
  - Crowther MJ (2018); Chen YET (2022)

Details matter – small inaccuracies can cause model misfits

What are the implications of improper data preparation in multi-state models?

1. Suppose you have three states s_1, s_2, and s_3 where the transitions are as follows:
   - $s_1 \rightarrow s_2$
   - $s_2 \rightarrow s_3$
   - $s_2 \rightarrow s_1$

If you mistakenly include extra transitions, it affects transition probability p

$$p_{ij}(t) \approx \frac{\text{number of transitions from state } i \text{ to state } j}{\text{time spent in state } i}$$

- Adding extra rows, the numerator becomes inflated, leading to a higher estimated intensity
- This overestimation distorts the model's reflection of real-world transition behaviour.

Details matter – small inaccuracies can cause model misfits

What are the implications of improper data preparation in multi-state models?

2 Assume you're using a CoxPH to estimate hazard ratios for transitions
A CoxPH model's hazard for an individual i in state j transitioning to k at time t is given by

$$h_{jk}(t|X_i) = h_{jk,0}(t)\exp(\beta X_i)$$

where $h_{jk,0}(t)$ is the baseline hazard and β is the coefficient vector for covariates X_i

- For each transition, the risk set $R_{jk}(t)$ includes individuals who are at risk of transitioning from j to k at time t. If your data includes individuals in $R_{jk}(t)$ who shouldn't be there (e.g., due to irrelevant transitions), the CoxPH incorrectly considers these individuals at risk for this transition.

$$L(\beta) = \prod_{events\ \mathrm{e}} \frac{\exp(\beta X_e)}{\sum_{i \in R_{jk}(t_e)} \exp(\beta X_i)}$$

- Including individuals in $R_{jk}(t)$ who shouldn't be there biases the likelihood function, which results in incorrect β for covariates.

Summary

- Always put the multi-state model diagram or transition matrix aside
- Think about competing risk
  - What may happen next for a specific state? What are the risk sets?

- Thorough data preparation is the foundation of accurate multi-state modeling
  - Keep risk sets in mind
  - Anticipate transitions
  - Double-check your data for accuracy

- Be meticulous! Details matter – small inaccuracies can cause model misfits