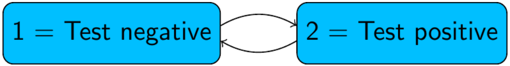
# Making semi-Markov multistate models for intermittent observations easily usable

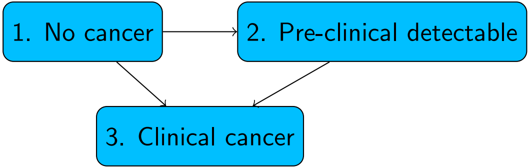
Christopher Jackson

Seminar: Center for Health Decision Science, Harvard School of Public Health

2026-04-23

# Multi-state models



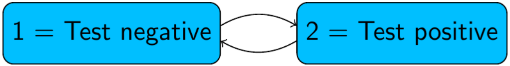

... or any other state and transition structure

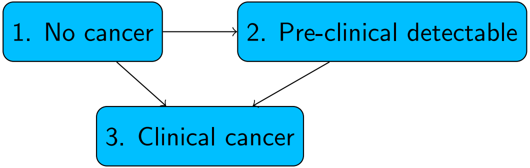
Parameters: continuous-time models with transition intensities / rates / hazards $q_{rs} = \exp(\beta_{rs}\mathbf{x})$, or time-dependent, $q_{rs}(t) = \exp(\beta_{rs}\mathbf{x}(t))$

Estimate:, e.g.,

- expected time spent in a state (e.g. duration of an infection)
- probabilities of transition between states, over periods of time ... (e.g. chance of developing clinical cancer within $t$ years — informing choice of screening interval $t$)

# Multi-state models




... or any other state and transition structure

Parameters: continuous-time models with transition intensities / rates / hazards
$q_{rs} = \exp(\beta_{rs}\mathbf{x})$, or time-dependent, $q_{rs}(t) = \exp(\beta_{rs}\mathbf{x}(t))$

Estimate:, e.g.,

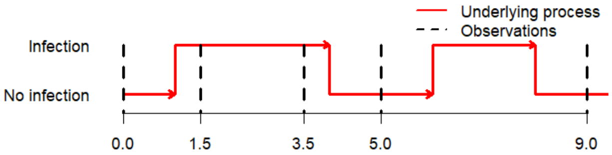
- expected time spent in a state (e.g. duration of an infection)
- probabilities of transition between states, over periods of time ... (e.g. chance of developing clinical cancer within $t$ years — informing choice of screening interval $t$)

# Data

Multi-state models get applied to a range of data structures.

(a) Sometimes we know the times of all transitions within follow-up. Extension of survival analysis (see e.g. tutorial by Putter et. al, Stat Med 2007) Easier: not covered here

(b) But here, we have intermittent observations:. We only know the state at a finite set of times — e.g. when person is tested for infection




Don't know transition times between states:

- e.g. when someone got the infection, when it cleared

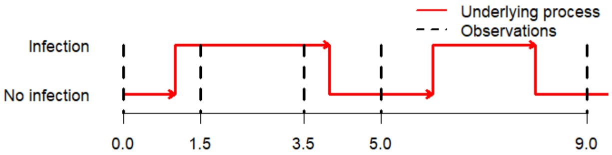
Some infections may be completely unobserved for people in the data

# Data

Multi-state models get applied to a range of data structures.

(a) Sometimes we know the times of all transitions within follow-up. Extension of survival analysis (see e.g. tutorial by Putter et. al, Stat Med 2007) Easier: not covered here

(b) But here, we have intermittent observations:. We only know the state at a finite set of times — e.g. when person is tested for infection




Don't know transition times between states:

- e.g. when someone got the infection, when it cleared

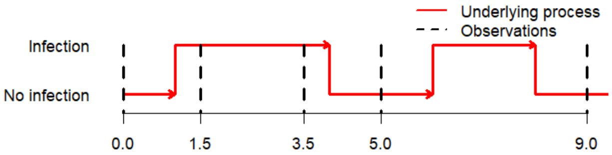
Some infections may be completely unobserved for people in the data

# Data

Multi-state models get applied to a range of data structures.

(a) Sometimes we know the times of all transitions within follow-up. Extension of survival analysis (see e.g. tutorial by Putter et. al, Stat Med 2007) Easier: not covered here

(b) But here, we have intermittent observations:. We only know the state at a finite set of times — e.g. when person is tested for infection




Don't know transition times between states:

- e.g. when someone got the infection, when it cleared

Some infections may be completely unobserved for people in the data

# Model estimation and challenges

Standard framework based on maximum likelihood estimation (Kalbfleisch and Lawless, JASA 1985)

- Likelihood function for intensities $\mathbf{q}$ built by multiplying transition probabilities $p_{rs}(t|\mathbf{q})$ between observed states $r, s$ over successive time intervals $t$.
- `msm` R package (CRAN, Jackson 2011 J. Stat. Soft.) widely used for this

Consequences of intermittent observation:

Strong model assumptions needed to easily calculate $p_{rs}$

Estimation challenges for intensities in continuous-time...

# Model estimation and challenges

Standard framework based on maximum likelihood estimation (Kalbfleisch and Lawless, JASA 1985)

- Likelihood function for intensities $\mathbf{q}$ built by multiplying transition probabilities $p_{rs}(t|\mathbf{q})$ between observed states $r, s$ over successive time intervals $t$.
- `msm` R package (CRAN, Jackson 2011 J. Stat. Soft.) widely used for this

Consequences of intermittent observation:

Strong model assumptions needed to easily calculate $p_{rs}$

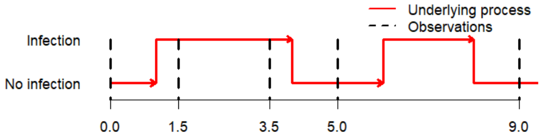
Estimation challenges for intensities in continuous-time...

# Strong model assumptions

Markov assumption

- Hazard is constant or piecewise-constant function of time since start of process
- ... doesn't depend on previous states / transition times
- exponentially-distributed *sojourn* time (time since entry to current state)

# Estimation challenges




The model parameters are the transition intensities in continuous time.

Since we can't easily tell from this data

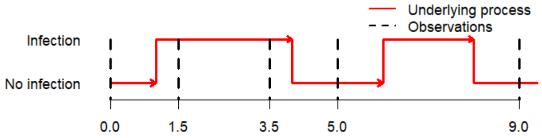
- how many times a state was visited, how long people spend there
- effects of covariates on these,

...hard to know which transition intensites are well informed by data.

Users tend to fit over-parameterised models $\rightarrow$ optimisation algorithm doesn't converge

"Is my model wrong for the data, or can the estimation be tweaked until it works??"

# Estimation challenges




The model parameters are the transition intensities in continuous time.

Since we can't easily tell from this data

- how many times a state was visited, how long people spend there
- effects of covariates on these,

...hard to know which transition intensites are well informed by data.

Users tend to fit over-parameterised models → optimisation algorithm doesn't converge

"Is my model wrong for the data, or can the estimation be tweaked until it works??"

# Previous approaches to relaxing the Markov assumption

Integrate over unknown times of state entry to get likelihood (e.g. Wei and Kryscio (2016), Aastveit et al (2023), Akwiwu et al BMC Med Res Meth (2022))

- computation scales badly with more states and more unknown times
- does not fully generalise to cyclic structures

Nonparametric approach (Gu et al, Biometrika 2024, JASA 2025)

- cyclic structures not supported, code not provided

Simulate pathways through states to approximate the likelihood (Aralis & Brookmeyer, SMMR (2019); Barone & Tancredi, Stat Med (2022)).

- MCMC or EM-type algorithm: needs custom software

# Previous approaches to relaxing the Markov assumption

Integrate over unknown times of state entry to get likelihood (e.g. Wei and Kryscio (2016), Aastveit et al (2023), Akwiwu et al BMC Med Res Meth (2022))

- computation scales badly with more states and more unknown times
- does not fully generalise to cyclic structures

Nonparametric approach (Gu et al, Biometrika 2024, JASA 2025)

- cyclic structures not supported, code not provided

Simulate pathways through states to approximate the likelihood (Aralis & Brookmeyer, SMMR (2019); Barone & Tancredi, Stat Med (2022)).

- MCMC or EM-type algorithm: needs custom software

# Previous approaches to relaxing the Markov assumption

Integrate over unknown times of state entry to get likelihood (e.g. Wei and Kryscio (2016), Aastveit et al (2023), Akwiwu et al BMC Med Res Meth (2022))

- computation scales badly with more states and more unknown times
- does not fully generalise to cyclic structures

Nonparametric approach (Gu et al, Biometrika 2024, JASA 2025)

- cyclic structures not supported, code not provided

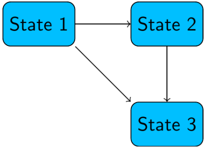
Simulate pathways through states to approximate the likelihood (Aralis & Brookmeyer, SMMR (2019); Barone & Tancredi, Stat Med (2022)).

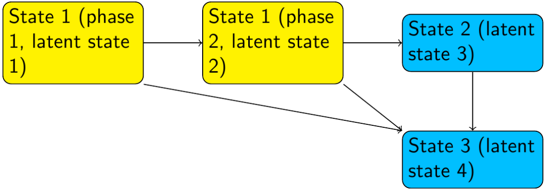
- MCMC or EM-type algorithm: needs custom software

# Semi-Markov "phase-type" model

(Titman and Sharples Biometrics 2012).

Allows the rate of transition out of some state to change with the time spent in that state. Example:



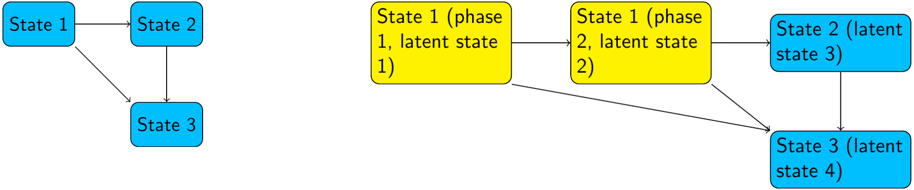

Replace an observable state (state 1 in this picture) with a set of latent states ("phases"). Latent states follow a Markov model

Hidden Markov model: hence likelihood calculation tractable. Sum over potential pathways among the $S$ states, equivalent to a product (over observations) of $S \times S$ matrices.

# Semi-Markov "phase-type" model

(Titman and Sharples Biometrics 2012).

Allows the rate of transition out of some state to change with the time spent in that state. Example:




Replace an observable state (state 1 in this picture) with a set of latent states ("phases"). Latent states follow a Markov model

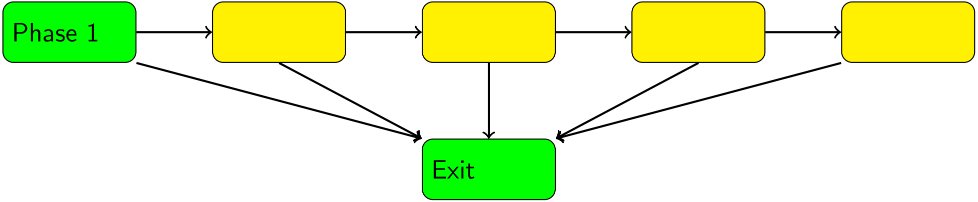
Hidden Markov model: hence likelihood calculation tractable. Sum over potential pathways among the $S$ states, equivalent to a product (over observations) of $S \times S$ matrices.

# Phase-type sojourn distributions

In a Markov model, the sojourn time in every state is exponentially distributed

In this semi-Markov model, the sojourn time in some state follows a "phase-type" sojourn distribution

Time from entering state 1 to reaching the "Exit" state in a continuous-time Markov model structure like this:



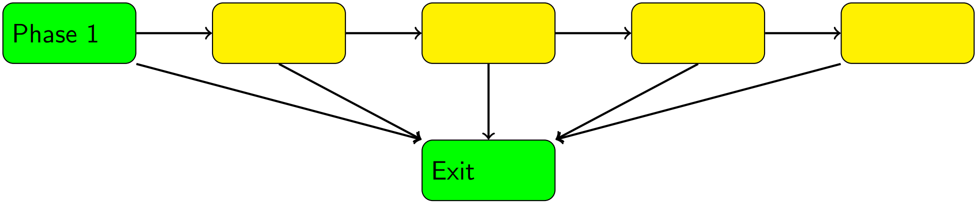

- More latent phases and transitions → more flexible sojourn distribution, more parameters ... estimation challenges, particularly with coarse data

# Phase-type sojourn distributions

In a Markov model, the sojourn time in every state is exponentially distributed

In this semi-Markov model, the sojourn time in some state follows a "phase-type" sojourn distribution

Time from entering state 1 to reaching the "Exit" state in a continuous-time Markov model structure like this:




- More latent phases and transitions → more flexible sojourn distribution, more parameters ... estimation challenges, particularly with coarse data

# Phase-type families inspired by common distributions

Motivation:

- Want a more stable / parsimonious sojourn distribution than the phase-type
- Gamma, Weibull typical but hard to calculate likelihood directly (need to integrate over unknown state entry times)
- Phase-type models infinitely flexible, can calculate their likelihood...

Design phase-type families with transition rates $\lambda = b\mathbf{h}(a)$ similar to the Weibull or Gamma (shape $a$, scale $b$) (Titman, Stat Comp 2014)

# Phase-type families inspired by common distributions

Motivation:

- Want a more stable / parsimonious sojourn distribution than the phase-type
- Gamma, Weibull typical but hard to calculate likelihood directly (need to integrate over unknown state entry times)
- Phase-type models infinitely flexible, can calculate their likelihood...

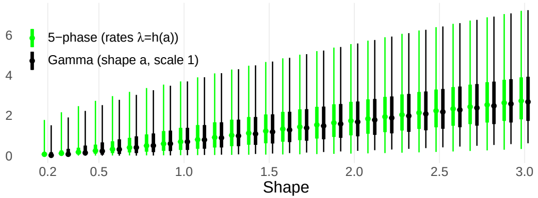
Design phase-type families with transition rates $\lambda = b\mathbf{h}(a)$ similar to the Weibull or Gamma (shape $a$, scale $b$) (Titman, Stat Comp 2014)

# Phase-type families inspired by common distributions

## Motivation:

- Want a more stable / parsimonious sojourn distribution than the phase-type
- Gamma, Weibull typical but hard to calculate likelihood directly (need to integrate over unknown state entry times)
- Phase-type models infinitely flexible, can calculate their likelihood...

Design phase-type families with transition rates $\lambda = b\mathbf{h}(a)$ similar to the Weibull or Gamma (shape $a$, scale $b$) (Titman, Stat Comp 2014)




Find a function $\mathbf{h}()$ that gives a family of time-to-event distributions that look like this picture in how they depend on shape.

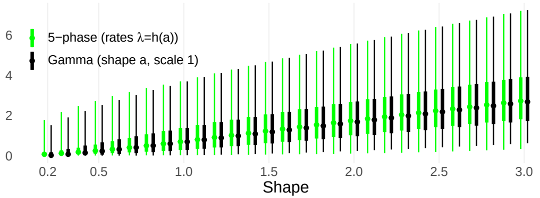
Lets us compute the likelihood as a function of $a$ (and $b$, via scaling time $t$)

# Phase-type families inspired by common distributions

## Motivation:

- Want a more stable / parsimonious sojourn distribution than the phase-type
- Gamma, Weibull typical but hard to calculate likelihood directly (need to integrate over unknown state entry times)
- Phase-type models infinitely flexible, can calculate their likelihood...

Design phase-type families with transition rates $\lambda = b\mathbf{h}(a)$ similar to the Weibull or Gamma (shape $a$, scale $b$) (Titman, Stat Comp 2014)



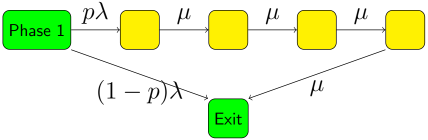

Find a function $\mathbf{h}()$ that gives a family of time-to-event distributions that look like this picture in how they depend on shape.

Lets us compute the likelihood as a function of $a$ (and $b$, via scaling time $t$)

# Defining phase-type families by moment-matching to Gamma/Weibull

Titman (2014) did a complicated spline fitting to find the mapping $h()$

Easier way: there is an analytic formula for the unique phase-type rates in this structure that give a particular mean, variance and skewness




(Bobbio et al. Stochastic Models 2005)

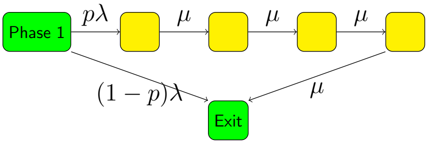
So to approximate a Gamma (or Weibull), given shape and scale,

- Calculate first three moments of the Gamma (or Weibull)
- $\rightarrow$ determine matching phase parameters $\lambda, p, \mu$

# Defining phase-type families by moment-matching to Gamma/Weibull

Titman (2014) did a complicated spline fitting to find the mapping $h()$

Easier way: there is an analytic formula for the unique phase-type rates in this structure that give a particular mean, variance and skewness



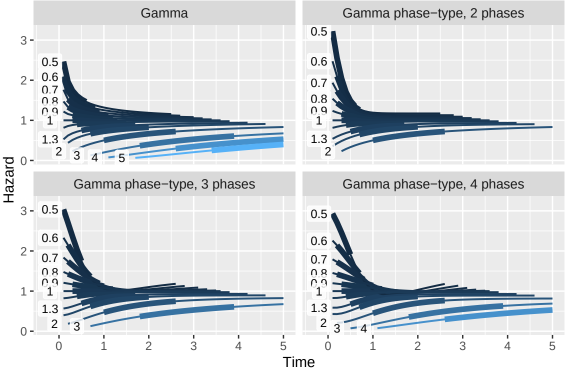

(Bobbio et al. Stochastic Models 2005)

So to approximate a Gamma (or Weibull), given shape and scale,

- Calculate first three moments of the Gamma (or Weibull)
- $\rightarrow$ determine matching phase parameters $\lambda, p, \mu$

# Hazard functions for phase-type Gamma

For scale $= 1$, compare different shape parameters




(thin/thick lines cover the 50%/95% probability intervals)

Phase-type family covers increasing / decreasing hazards, just like the classical Gamma.

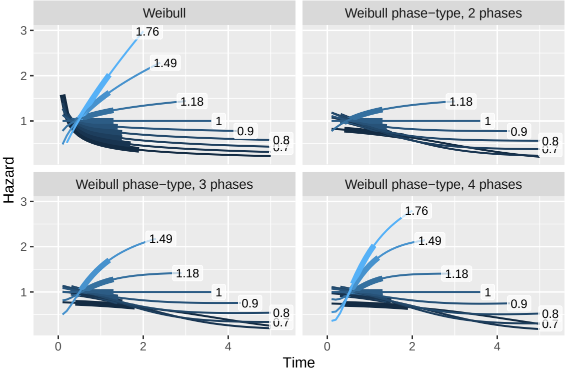
U-shaped for lowest shape parameters.

Don't need to match Gamma exactly — just needs to be useful in practice

More phases → Gammas with higher shape parameters can be moment-matched to a phase type

# Hazard functions for phase-type Weibull

For scale = 1, compare different shape parameters




(thin/thick lines cover the 50%/95% probability intervals)

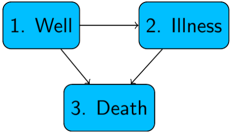
Phase-type family covers increasing / decreasing hazards just like the classical Weibull.

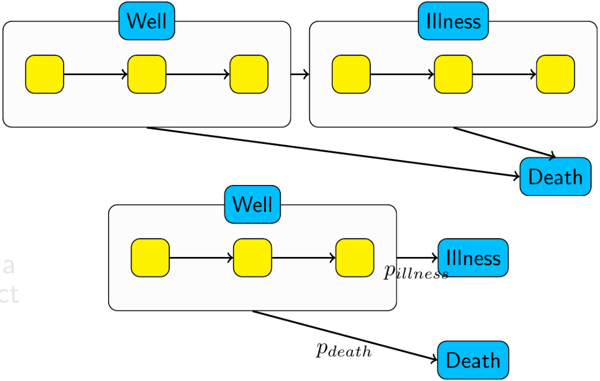
U-shaped for lowest shape parameters.

Note the Weibull is extremely skewed for lowest shape parameters (less common in practice?)

# Building multi-state models with "phase-type" states

One or more states $r$ can have a phase-type sojourn distribution defined by shapes $a$, scales $b$



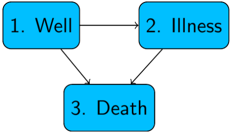



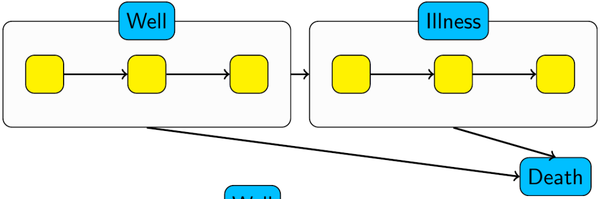

Covariates: can modify the scale $b$

Or if 2+ "competing risks" on leaving a "phase-type" state: covariates can affect probability of competing state (assumed independent of sojourn time)

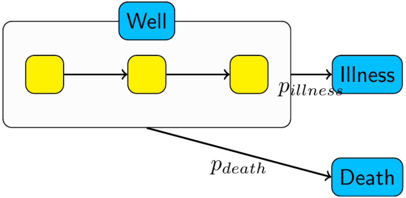
# Building multi-state models with “phase-type” states

One or more states $r$ can have a phase-type sojourn distribution defined by shapes $a$, scales $b$




Covariates: can modify the scale $b$




Or if 2+ “competing risks” on leaving a “phase-type” state: covariates can affect probability of competing state (assumed independent of sojourn time)

# Computation

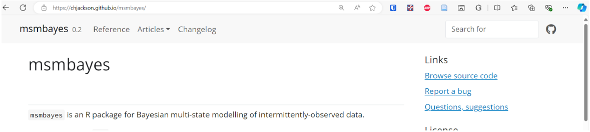
Bayesian inference, maximum likelihood or approximate Bayes (Laplace around posterior mode), using "off-the-shelf" algorithms in the **Stan** software

Advantages of a Bayesian approach:

- Background information can be supplied as an informative prior
- Weakly informative prior can stabilise estimation in situations where pure MLE fails due to infrequently-observed data
- Even if no information in data, we still get an answer: posterior which has not moved from the prior. More helpful than convergence failure.

# msmbayes R package

Extends the `msm` package for Markov models to do Bayesian inference and phase-type semi-Markov models

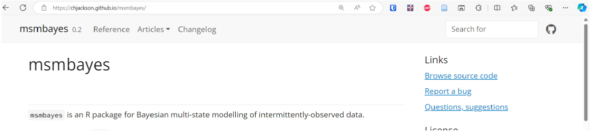



Familiar interface, like common R modelling packages

```
Q <- rbind(c(0, 1),
           c(1, 0)) # 2-state transition structure
priors <- list(
  msmprior("time(1,2)", median=10,  upper=30),
  msmprior("time(2,1)", median=0.5, upper=1)
)
mod <- msmbayes(data = infsim, state="state", time="months",
                subject="subject", qmatrix=q, priors=priors,...)
summary(mod)
```

# msmbayes R package

Extends the msm package for Markov models to do Bayesian inference and phase-type semi-Markov models




Familiar interface, like common R modelling packages

```
Q <- rbind(c(0, 1),
           c(1, 0)) # 2-state transition structure
priors <- list(
  msmprior("time(1,2)", median=10,  upper=30),
  msmprior("time(2,1)", median=0.5, upper=1)
)
mod <- msmbayes(data = infsim, state="state", time="months",
                subject="subject", qmatrix=q, priors=priors,...)
summary(mod)
```

# Simulation-based calibration study (as in Talts et al. 2018)

Assess correctness of Bayesian computation procedure:

- Simulate many datasets from prior predictive distribution
- Fit models to them: average of resulting posteriors should match the prior

Designed here around infection duration example (2 or 3-state).

Results:

- MCMC estimation accurate under a range of model structures, but slow
- Laplace approximation is mildly biased / underestimates uncertainty, but allows scalability

# Simulation-based calibration study (as in Talts et al. 2018)

Assess correctness of Bayesian computation procedure:

- Simulate many datasets from prior predictive distribution
- Fit models to them: average of resulting posteriors should match the prior

Designed here around infection duration example (2 or 3-state).

Results:

- MCMC estimation accurate under a range of model structures, but slow
- Laplace approximation is mildly biased / underestimates uncertainty, but allows scalability

# Computational stability and scalability

## Stability

- `msm` did not converge in majority of simulated datasets due to weak identifiability
- Bayesian estimation always produced a valid posterior — strongly influenced by prior, but more helpful than "it didn't work".
- Phase-type shape-scale approximation works where fitting the phase-type distribution directly doesn't

## Scalability

- expanding the state space makes likelihood harder, due to need to calculate matrix exponential.
- Stan "black-box" auto-differentiation particularly memory intensive here

# Computational stability and scalability

## Stability

- `msm` did not converge in majority of simulated datasets due to weak identifiability
- Bayesian estimation always produced a valid posterior — strongly influenced by prior, but more helpful than "it didn't work".
- Phase-type shape-scale approximation works where fitting the phase-type distribution directly doesn't

## Scalability

- expanding the state space makes likelihood harder, due to need to calculate matrix exponential.
- Stan "black-box" auto-differentiation particularly memory intensive here

# Computational stability and scalability

## Stability

- `msm` did not converge in majority of simulated datasets due to weak identifiability
- Bayesian estimation always produced a valid posterior — strongly influenced by prior, but more helpful than "it didn't work".

- Phase-type shape-scale approximation works where fitting the phase-type distribution directly doesn't

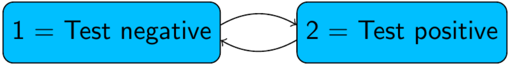
## Scalability

- expanding the state space makes likelihood harder, due to need to calculate matrix exponential.
- Stan "black-box" auto-differentiation particularly memory intensive here

# Ongoing / suitable applications

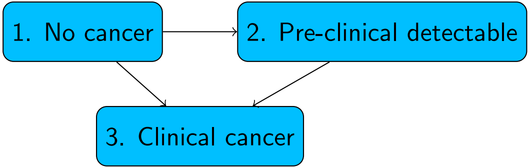
## (a) Estimating infection duration and incidence

Cohort of people tested intermittently.
Start / end times of person's infection unknown.
Work in progress with data from Coronavirus Infection Survey in the UK (developing Blake et al. https://arxiv.org/abs/2502.04824)




## (b) Cancer screening

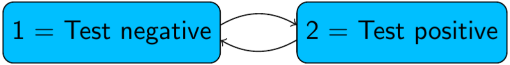
Incidence, time to progression not exponentially distributed.
Choose optimal screening interval.
(e.g. Akwiwu et al. BMC Med Res Meth 2022)



# Ongoing / suitable applications

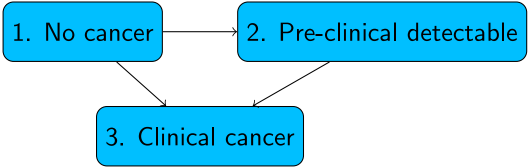
## (a) Estimating infection duration and incidence

Cohort of people tested intermittently.
Start / end times of person's infection unknown.
Work in progress with data from Coronavirus Infection Survey in the UK (developing Blake et al. https://arxiv.org/abs/2502.04824)

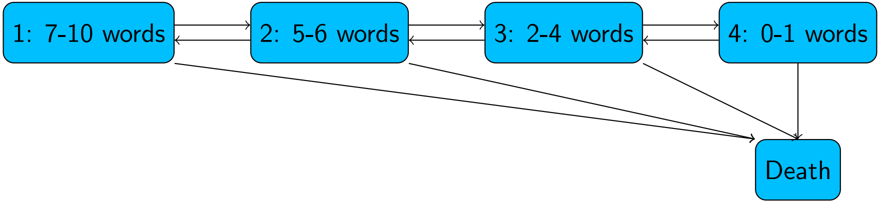



## (b) Cancer screening

Incidence, time to progression not exponentially distributed.
Choose optimal screening interval.
(e.g. Akwiwu et al. BMC Med Res Meth 2022)



# Realistically-complex illustrative application



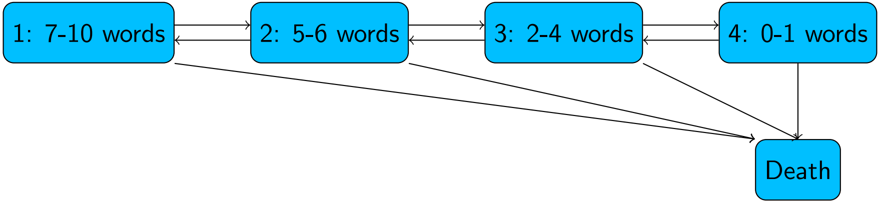

English Longitudinal Study of Ageing. Cognitive function test: how many words from a list of 10 recalled after a few minutes. $\approx$ 5000 observations from people aged 50+, every 2 years.

Semi-Markov model on all four states: 21 latent "phases"

Predictors of transitions: age, gender, education

MCMC not feasible for semi-Markov, use Laplace approximation to posterior

# Realistically-complex illustrative application



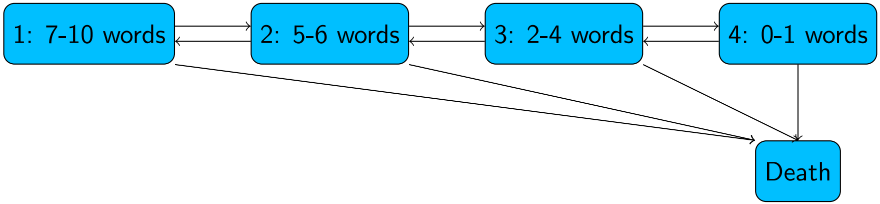

English Longitudinal Study of Ageing. Cognitive function test: how many words from a list of 10 recalled after a few minutes. $\approx$ 5000 observations from people aged 50+, every 2 years.

Semi-Markov model on all four states: 21 latent "phases"

Predictors of transitions: age, gender, education

MCMC not feasible for semi-Markov, use Laplace approximation to posterior

# Realistically-complex illustrative application




English Longitudinal Study of Ageing. Cognitive function test: how many words from a list of 10 recalled after a few minutes. $\approx$ 5000 observations from people aged 50+, every 2 years.

Semi-Markov model on all four states: 21 latent "phases"

Predictors of transitions: age, gender, education

MCMC not feasible for semi-Markov, use Laplace approximation to posterior

# Priors, model choice

Priors developed/calibrated to judgements about outcomes by simulating from them

- Strong priors on mortality rates from national statistics
- Markov model, specify judgements about mean sojourn time and hazard ratios: e.g. 95% credible interval (1/7, 7)
- Semi-Markov model: shape and scale of sojourn distribution, next-state probabilities?
  - Covariates affect both, so harder to give prior judgements
  - Ad hoc procedure of simulation/checking/modifying

Compare models via maximised log posterior. Formal cross-validation / absolute goodness-of-fit challenging with intermittent data

# Priors, model choice

Priors developed/calibrated to judgements about outcomes by simulating from them

- Strong priors on mortality rates from national statistics
- Markov model, specify judgements about mean sojourn time and hazard ratios: e.g. 95% credible interval (1/7, 7)
- Semi-Markov model: shape and scale of sojourn distribution, next-state probabilities?
  - Covariates affect both, so harder to give prior judgements
  - Ad hoc procedure of simulation/checking/modifying

Compare models via maximised log posterior. Formal cross-validation / absolute goodness-of-fit challenging with intermittent data

# Priors, model choice

Priors developed/calibrated to judgements about outcomes by simulating from them

- Strong priors on mortality rates from national statistics
- Markov model, specify judgements about mean sojourn time and hazard ratios: e.g. 95% credible interval (1/7, 7)
- Semi-Markov model: shape and scale of sojourn distribution, next-state probabilities?
  - Covariates affect both, so harder to give prior judgements
  - Ad hoc procedure of simulation/checking/modifying

Compare models via maximised log posterior. Formal cross-validation / absolute goodness-of-fit challenging with intermittent data

# Priors, model choice

Priors developed/calibrated to judgements about outcomes by simulating from them

- Strong priors on mortality rates from national statistics
- Markov model, specify judgements about mean sojourn time and hazard ratios: e.g. 95% credible interval (1/7, 7)
- Semi-Markov model: shape and scale of sojourn distribution, next-state probabilities?
  - Covariates affect both, so harder to give prior judgements
  - Ad hoc procedure of simulation/checking/modifying

Compare models via maximised log posterior. Formal cross-validation / absolute goodness-of-fit challenging with intermittent data

# Results

“Covariate effect parameters” hard to interpret. Instead:

Calculate expected total amount of time spent with no/mild cognitive impairment over next 10 years

Compare this between categories of one covariate (standardised over others)

# Discussion

Made semi-Markov models for intermittently-observed data practicable

Software to make Bayesian inference in general Markov and semi-Markov models accessible

Challenges: computational scalability, prior specification, model checking, more practical experience...

`https://chjackson.github.io/msmbayes`

Paper on ArXiV linked from there, with full details of the studies described here