Diabetes register research and multistate models

Bendix Carstensen Steno Diabetes Center Copenhagen
Herlev, Denmark
http://BendixCarstensen.com

SDCA, Aarhus, 8th October 2024

http://BendixCarstensen.com/PMM — Practical Multistate Modeling

Topics

- Registers
- Demography
- Scales
- Follow-up representation
- Multistate data
- Multistate likelihood
- Multistate modeling

What's in a register

One record per event (diagnosis):

- person-id
- time of event (a date, usually)
- type of event (T1 / T2 / other)

Some events can occur at most once (diabetes, cancer),
other any number of times (CVD, hypoglycemia)

Some registers contain multiple events of a type (NPR, e.g.)

It is **you** who define what an event is

Diabetes register use: Look-up

- Persons from some study cohort, such as a population survey or a clinical study—what is their:
  - diabetes **status** (noDM/T1/T2) at a given date
  - diabetes **date** (T1 / T2)
- by exclusion we also know if a person does **not** have diabetes (completeness assumption)
- $\Rightarrow$ data input to existing (cohort) studies where follow-up is already known
  - explanatory variable for known outcome
  - outcome event in an existing cohort

Diabetes register use: Demography

Demographic **analysis** of **population**

- incidence and
- mortality rates,
- prevalence
- —and derivatives of basic demographic measures:
  - state probabilities
  - lifetime risk
  - expected lifetime in noDM / T1 / T1
  - lifetime lost

  . . . but note that these measures need further assumptions
- register events are outcome **events**,
  FU-time in population is outcome **risk time**

Diabetes demography: Scales of inference

-1. Occurrence rates
—the scale of **observed** register data, (d, y) (empirical rate),
measured in time^{-1} (events per person-time)

0. State probabilities (survival function)
—the **integral** of rates w.r.t. time
—requires an origin (such as date of diagnosis)
measured in time^{0} (dimensionless)

1. Sojourn times (time spent in a state)
—the **integral** of state probabilities w.r.t. time
—requires an origin and endpoint
measured in time^{1}

Demographic quantities—functions of time

- occurrence rate:

$$\lambda(t) = \lim_{h \to 0} \mathrm{P}\big\{\text{event in } (t, t+h) \,|\, \text{alive at } t\big\}/h$$

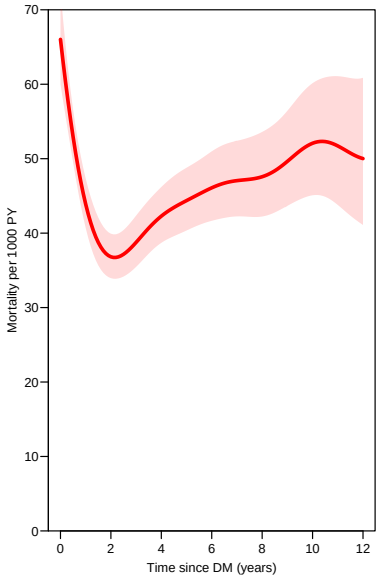
- survival probability (since time a):

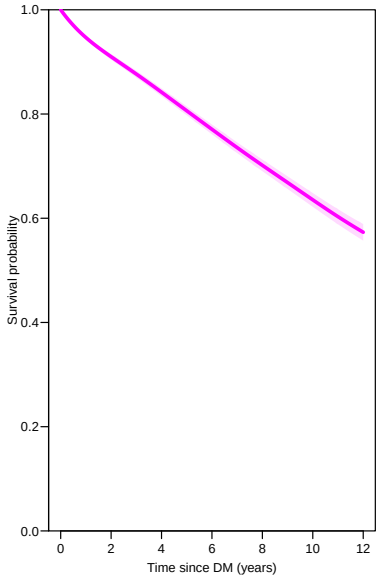
$$\mathrm{S}_a(t) = \exp\left(-\int_a^t \lambda(u)\,\mathrm{d}u\right)$$

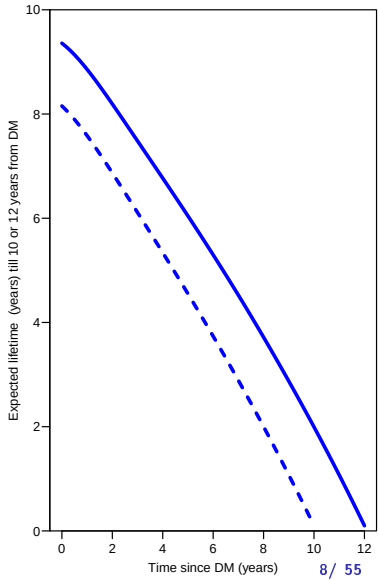
- sojourn time (between t and b)
  (restricted mean survival time to b, RMST):

$$\mathrm{L}(t) = \int_t^b \mathrm{S}_t(u)\,\mathrm{d}u$$

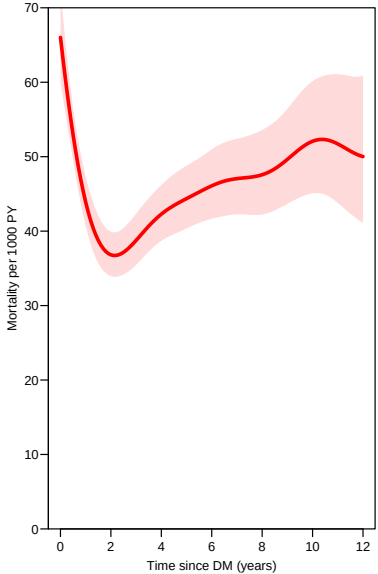
Mortality / survival / life time after DM

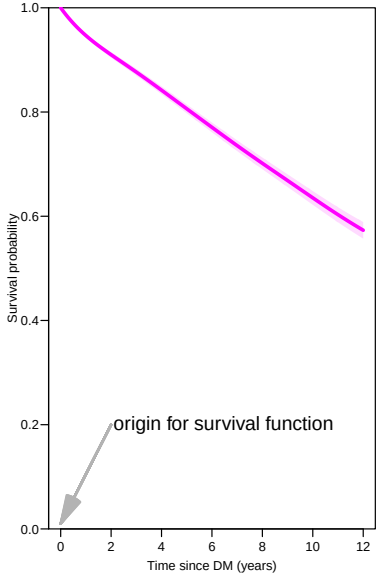


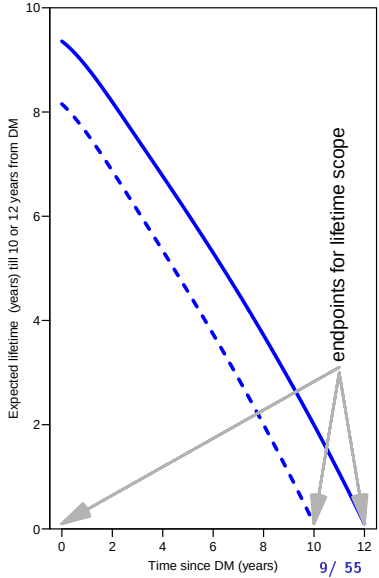




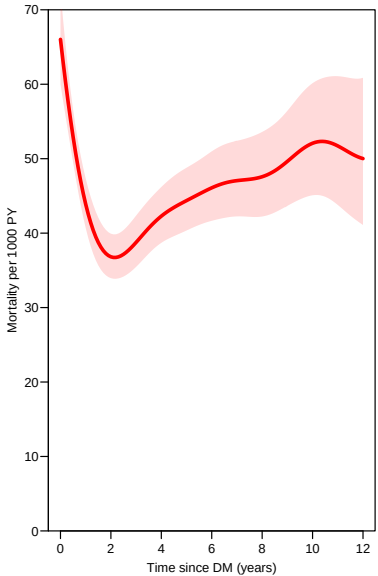
Mortality / survival / life time after DM

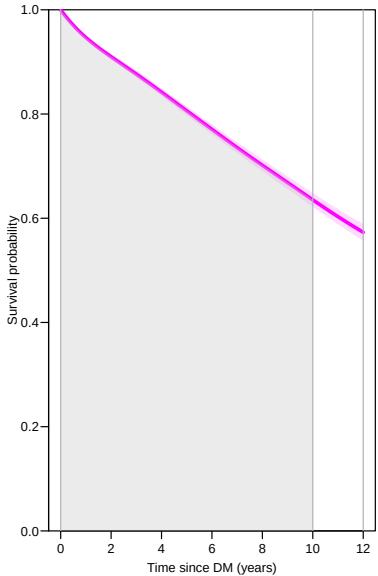


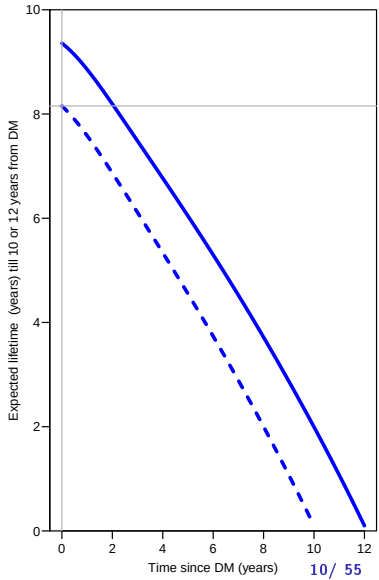




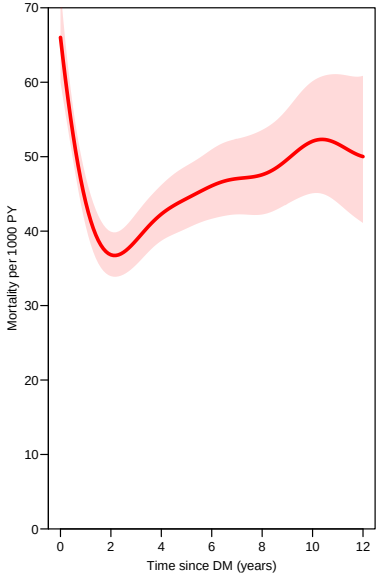
Mortality / survival / life time after DM

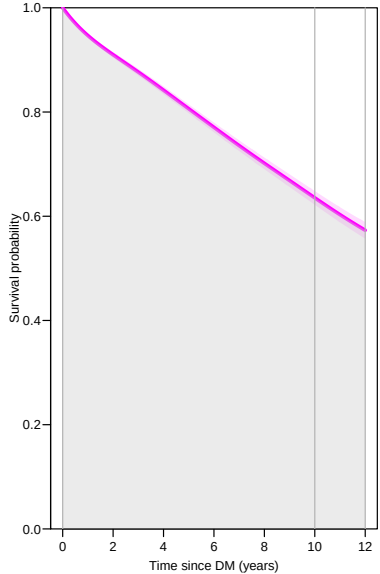


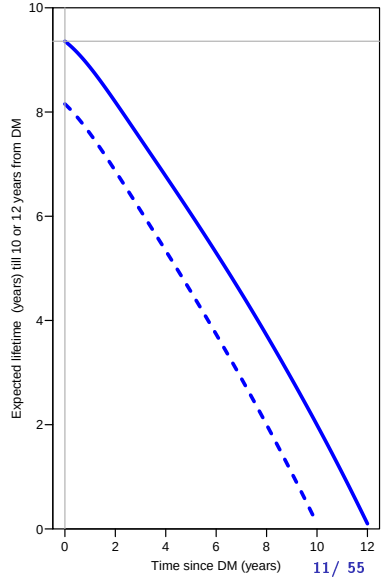




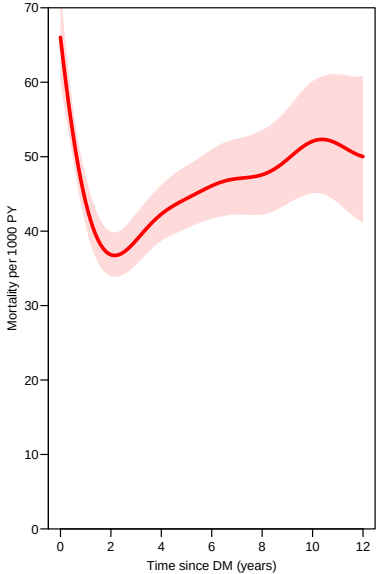
Mortality / survival / life time after DM

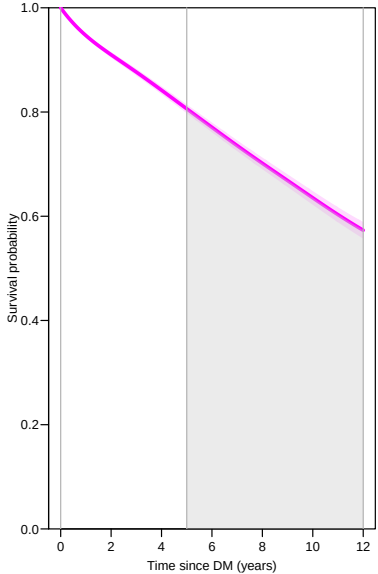


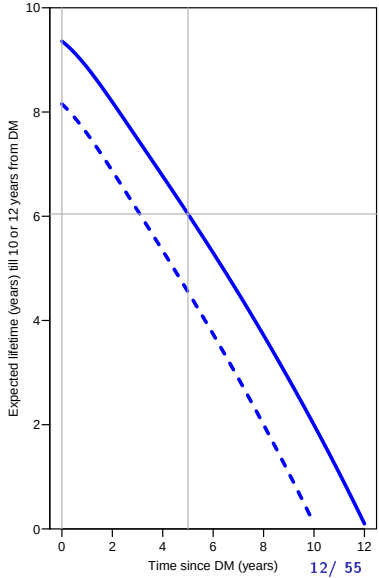




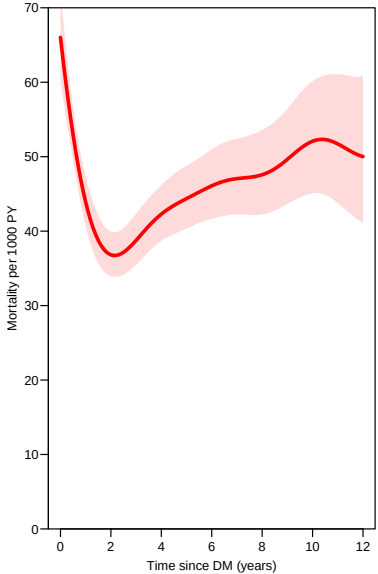
Mortality / survival / life time after DM

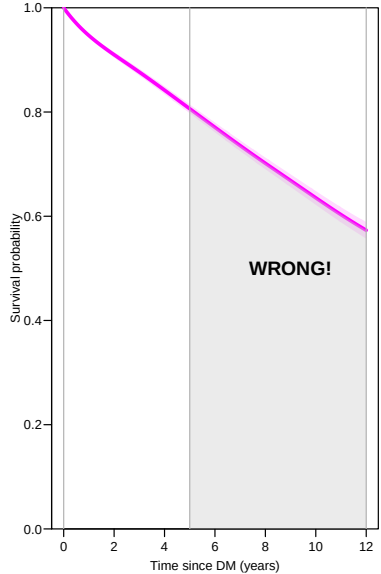


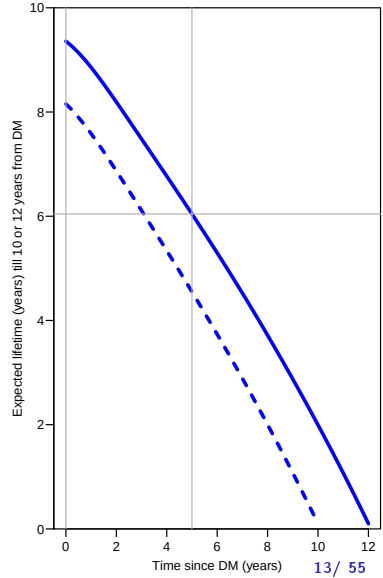




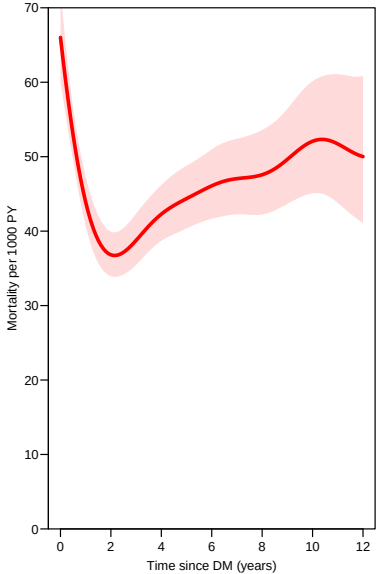
Mortality / survival / life time after DM

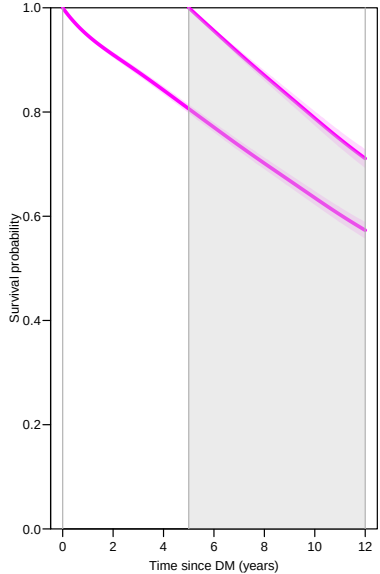


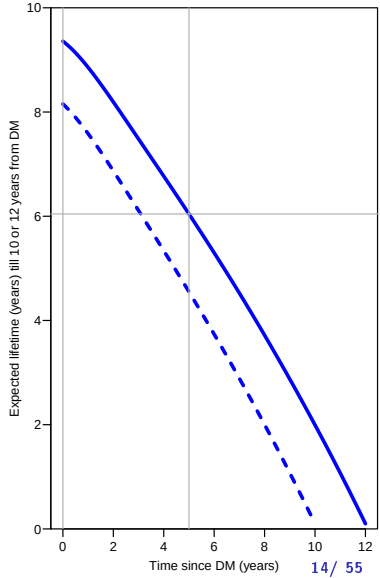




Mortality / survival / life time after DM







Diabetes demography

Demographic analyses of register event rates requires knowledge of **events** as well as **population time** covered by the register:

1. population size (number or risk time) by sex, age, date and other variables available both in the register and population. This will be **tabular** data, such as that available from Statistikbanken at DST.
2. **individual level** follow-up for **all** persons in the population — basically knowledge of entry (birth or immigration) and exit (death or emigration).
   Available as the LifeLines register at DST:
   individual follow-up of the entire DK population

How does follow-up look in a dataset

- One record per time **interval** (where nothing happens)
- Things happen at the **end** of the interval,
  the interval FU time belongs in a particular **state**, e.g.:
  - noDM / T1 / T2
  - noCKD / CKD
  - no comorb. / 1 comorb. / 2 comorb. / 3 comorb. / ...

How does follow-up look in a dataset

- Intervals may further be classified by **time-varying** variables:
  - quantitative deterministic variables (time scales):
    age, date of follow up, diabetes duration
  - quantitative random variables: HbA_{1c}, cholesterol, ...
  - categorical random variables: parity, marital status
- States are a special type of time varying covariates:
  targets of demographic measures (probability, sojourn time)

```
> library(Epi)
> data(DMlate)
> DMlate[13:19,]
```

	sex	dobth	dodm	dodth	dooad	doins	dox
119305	M	1938.107	1997.461	1998.35	NA	NA	1998.350
188248	F	1979.864	1999.684	NA	NA	NA	2009.997
38336	M	1944.420	2002.550	NA	NA	2005.354	2009.997
368534	F	1962.482	2000.355	NA	2001.559	NA	2009.997
139497	F	1956.439	1995.544	NA	NA	NA	2009.997
132331	M	1935.024	1996.746	NA	1997.915	2005.995	2009.997
228434	F	1949.622	2006.783	NA	2006.783	NA	2009.997

Each record: relevant dates for a person followed from date of diabetes till death or 2009-12-31 (end of study).

—combination of several registers

Total follow-up of diabetes ptt.

In terms of follow-up we must define:

- Entry time: `doDM`
- Exit time: `dox`
- Event death: `dodth` = `dox`

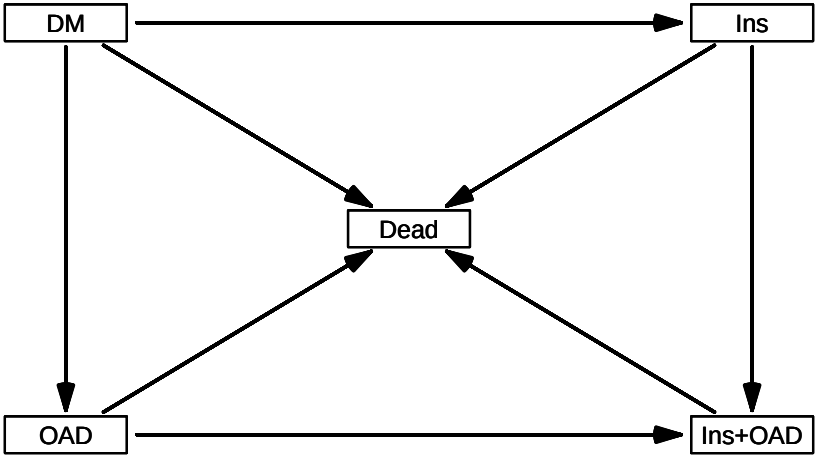
Intermediate register events

Other dates specify occurrence of intermediate events

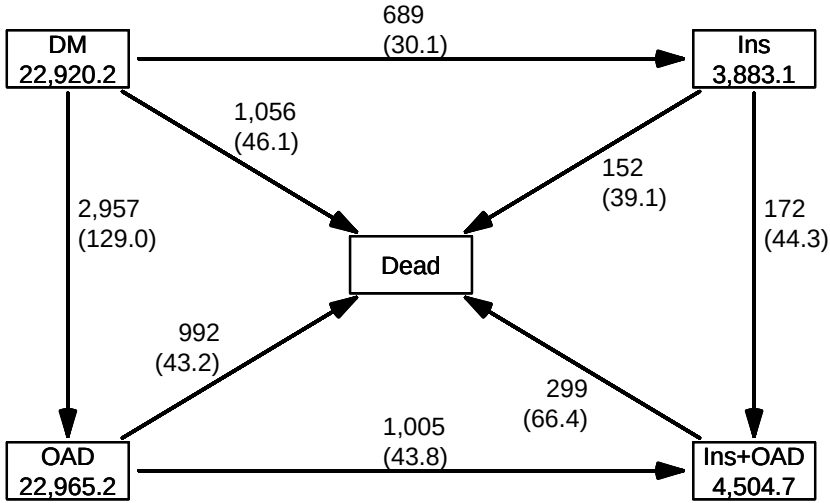
- start of OAD drugs at `doOAD`
- start of insulin at `doIns`
- possible states:
  - `DM`, no drug
  - `OAD` alone
  - `Ins` alone
  - both `OAD` & `Ins`
  - or:
    - `OAD` after `Ins`
    - `Ins` after `OAD`
  - `Dead`

States are not derived from data, they are defined by the investigator

Multi-state model — 5 states, 8 transitions



Multi-state data



Practical representation of follow-up

- provide an overview of the follow-up
- provide analytical possibility for **rate** models:
  modeling on the observation scale (observed rates (d, y))

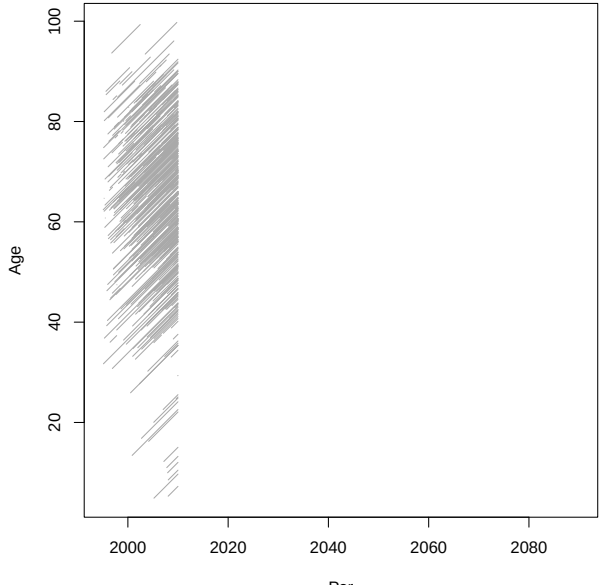
Multi-state data representation with Lexis

```
> dmL <- Lexis(entry = list(Per = dodm,
+                           Age = dodm - dobth,
+                         DMdur = 0 ),
+              exit = list(Per = dox),
+       exit.status = factor(!is.na(dodth),
+                            labels = c("DM", "Dead")),
+              data = DMlate)

NOTE: entry.status has been set to "DM" for all.
NOTE: Dropping  4  rows with duration of follow up < tol

> summary(dmL)

Transitions:
     To
From   DM Dead  Records:  Events: Risk time:  Persons:
  DM 7497 2499      9996     2499   54273.27      9996
```

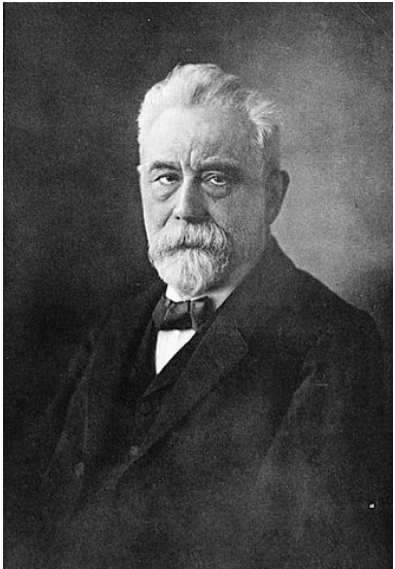
Multiple time scales: Per, Age, DMdur

A Lexis diagram

```
> plot(dmL)
```



Wilhelm Lexis



EINLEITUNG

IN DIE

THEORIE

DER

BEVÖLKERUNGSSTATISTIK

VON

W. LEXIS

DR. DER STAATSWISSENSCHAFTEN UND DER PHILOSOPHIE,
O. PROFESSOR DER STATISTIK IN DORPAT.

STRASSBURG

KARL J. TRÜBNER

1875.

Multi-state data representation with Lexis

```
> dmIO <- mcutLexis(dmL,
+                   wh = c("dooad", "doins"),
+              timescale = "Per",
+            new.states = c("OAD", "Ins"),
+            seq.states = FALSE,
+          ties.resolve = 1/365.25)

NOTE: Precursor states set to DM
NOTE: 15 records with tied events times resolved (adding 0.002737851 random unifo
      so results are only reproducible if the random number seed was set.

> summary(dmIO)

Transitions:
     To
From         DM Dead  OAD  Ins Ins+OAD  Records:  Events: Risk time:  Persons:
  DM       2830 1056 2957  689       0      7532     4702   22920.25      7532
  OAD         0  992 3327    0    1005      5324     1997   22965.24      5324
  Ins         0  152    0  462     172       786      324    3883.06       786
  Ins+OAD     0  299    0    0     878      1177      299    4504.72      1177
  Sum      2830 2499 6284 1151    2055     14819     7322   54273.27      9996
```

```
lex.id     Per   Age DMdur lex.dur lex.Cst lex.Xst
     2 2003.31 64.09     0    6.69      DM      DM
    15 2002.55 58.13     0    7.45      DM      DM
    18 1996.75 61.72     0   13.25      DM      DM
   770 1995.22 79.25     0    8.31      DM    Dead
```

```
lex.id     Per   Age DMdur lex.dur lex.Cst lex.Xst
     2 2003.31 64.09  0.00    4.14      DM     OAD
     2 2007.45 68.23  4.14    2.55     OAD     OAD
lex.id     Per   Age DMdur lex.dur lex.Cst lex.Xst
    15 2002.55 58.13   0.0    2.80      DM     Ins
    15 2005.35 60.93   2.8    4.64     Ins     Ins
lex.id     Per   Age DMdur lex.dur lex.Cst lex.Xst
    18 1996.75 61.72  0.00    1.17      DM     OAD
    18 1997.92 62.89  1.17    8.08     OAD Ins+OAD
    18 2005.99 70.97  9.25    4.00 Ins+OAD Ins+OAD
lex.id     Per   Age DMdur lex.dur lex.Cst lex.Xst
   770 1995.22 79.25  0.00    0.27      DM     Ins
   770 1995.49 79.52  0.27    0.15     Ins Ins+OAD
   770 1995.64 79.67  0.42    7.89 Ins+OAD    Dead
```

`lex.Cst` is the Current state `lex.Xst` is the eXit state

Multistate model: total (log-)likelihood

The log-likelihood contribution from a single person has:

- One contribution to the log-likelihood for each state visited
- . . . which is a sum of terms for each possible exit from the state
- If the model assumes **constant** rates, log-likelihood terms are $d\log(\lambda) - \lambda y$
  —a Poisson log-likelihood for variate d with mean λy
- $\Rightarrow$ total log-likelihood for a multistate model is a sum of terms, one per possible transition between states.
- a person only contributes terms from states actually visited

Multistate model data representation

- If all transition times are known (register data):
  - one record per follow-up interval (transient states)
    —representation of follow-up—`Epi` and `survival` package
    "Andersen-Gill" representation
  - one record per likelihood term (transitions)
    stacked data—`mstate` package
- state occupancy known at (some arbitrary) times
  (person `p` is in state `s` at time `t`)
  "prevalence", panel data—`msm` package

We stick to representation of follow-up time
—the most natural representation for register-based data

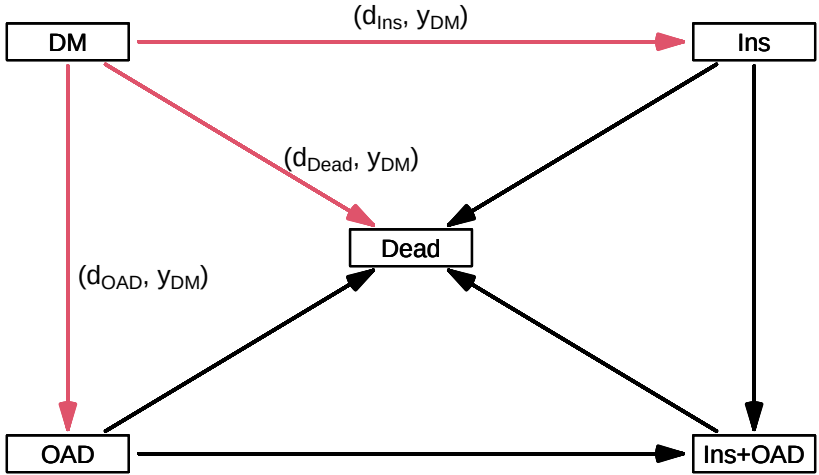
Likelihood for multistate transition rates

- assume all transitions and -times known exactly
- likelihood from one person is a **product** of terms with λ as argument
- $\Rightarrow$ log-likelihood a **sum** of terms like:

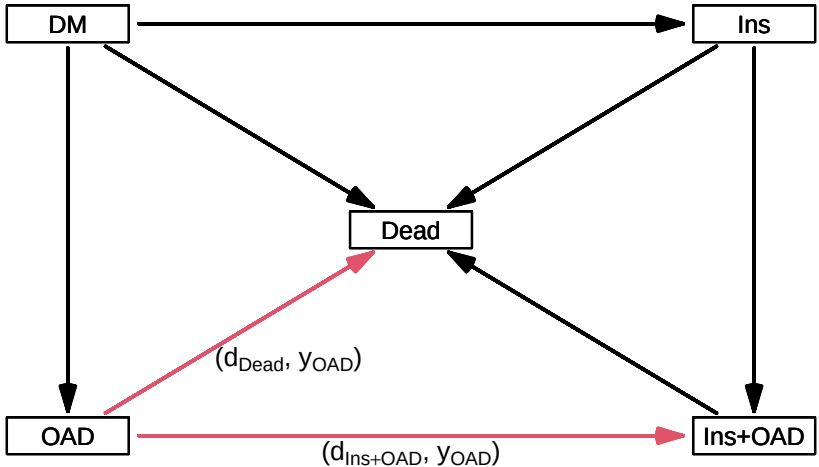
$$d \log(\lambda) - \lambda y$$

- —one term for each **possible** transition between states.
- for state DM **one** record but
  three likelihood terms, different ds, same y

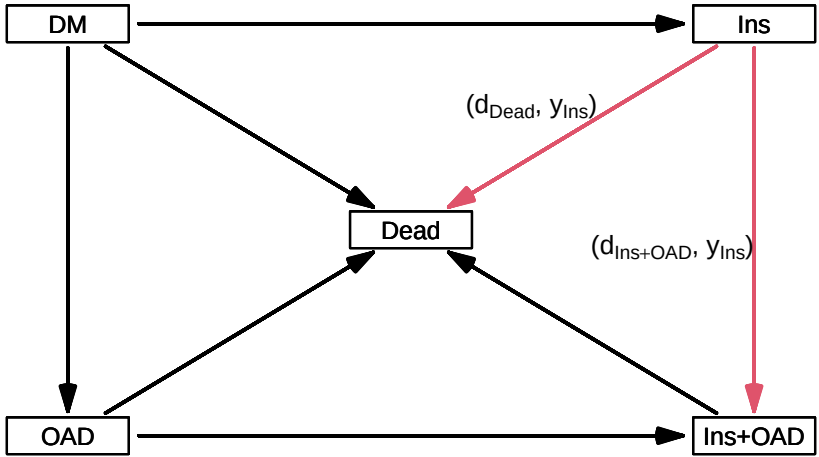
Total multi-state likelihood — 5 states, 8 transitions



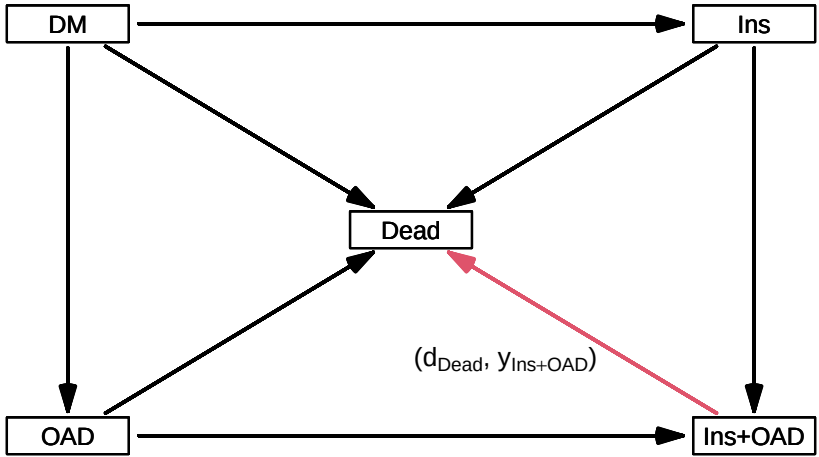
Total multi-state likelihood — 5 states, 8 transitions



Total multi-state likelihood — 5 states, 8 transitions



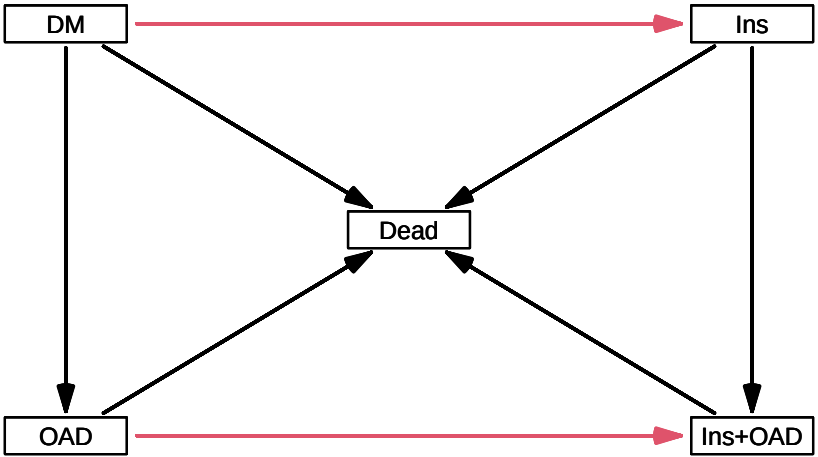
Total multi-state likelihood — 5 states, 8 transitions



Separate models for transition rates

- For rates in the same model: common parameters possible
  e.g. same age effect for different rates
- `Lexis` represents FU-time—**not** likelihood terms
- $\Rightarrow$ analysis of a model for different rates from **different** states can be done based on a `Lexis` object
- Analysis of a model for different rates from **the same** state requires a stacked data frame
- . . . but this is hardly ever relevant, e.g.:
  - do not expect age effect to be the same for rate of `OAD` and `Ins`
  - In practise only rates from **different** origin states are analysed together, such as `Ins` rates from `DM` resp. `OAD`

Partial multi-state likelihood — rates of `ins`

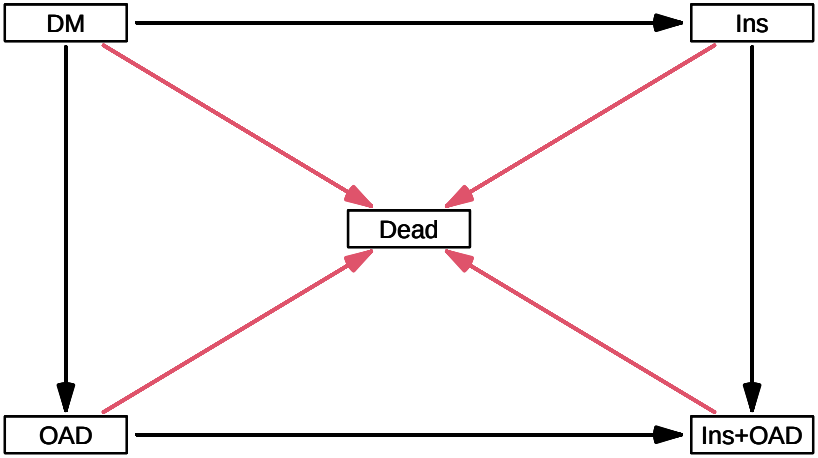


Modeling rates

- Poisson likelihood is for constant rates:
- $\Rightarrow$ model restricted to constant rate within each FU-record
- remedy: split records in many records with shorter length
  —so short that constant rates in intervals is reasonable
- `splitLexis` or `splitMulti` (from `popEpi` package)
- many records with `lex.Cst` = `lex.Xst`
- include timescales as quantitative variables

```
> summary(dmIO)

Transitions:
     To
From         DM Dead  OAD  Ins Ins+OAD  Records:  Events: Risk time:  Persons:
  DM       2830 1056 2957  689       0      7532     4702   22920.25      7532
  OAD         0  992 3327    0    1005      5324     1997   22965.24      5324
  Ins         0  152    0  462     172       786      324    3883.06       786
  Ins+OAD     0  299    0    0     878      1177      299    4504.72      1177
  Sum      2830 2499 6284 1151    2055     14819     7322   54273.27      9996

> sIO <- splitLexis(dmIO, seq(0,20,0.1), "DMdur")
> summary(sIO)

Transitions:
     To
From          DM Dead    OAD   Ins Ins+OAD  Records:  Events: Risk time:  Persons
  DM      228333 1056   2957   689       0    233035     4702   22920.25      753
  OAD          0  992 231721     0    1005    233718     1997   22965.24      532
  Ins          0  152      0 39203     172     39527      324    3883.06       78
  Ins+OAD      0  299      0     0   45923     46222      299    4504.72      117
  Sum     228333 2499 234678 39892   47100    552502     7322   54273.27      999
```

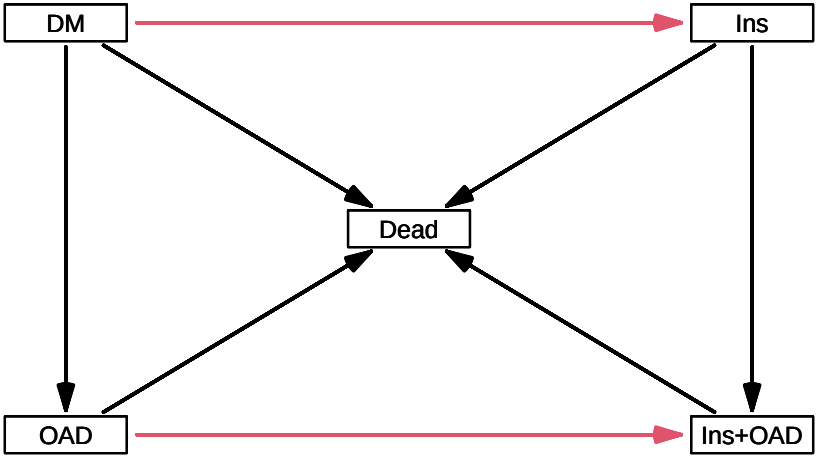
Multi-state likelihood — mortality rates



Mortality rates

```
> mdth <- glm.Lexis(sIO, ~ Ns(DMdur, knots=c(0,1,3,6,10)) + lex.Cst,
+                   to = "Dead")

stats::glm Poisson analysis of Lexis object sIO with log link:
Rates for transitions:
DM->Dead
OAD->Dead
Ins->Dead
Ins+OAD->Dead

> round(ci.exp(mdth), 3)

                                        exp(Est.)  2.5% 97.5%
(Intercept)                                 0.085 0.075 0.096
Ns(DMdur, knots = c(0, 1, 3, 6, 10))1       0.519 0.433 0.621
Ns(DMdur, knots = c(0, 1, 3, 6, 10))2       0.710 0.605 0.832
Ns(DMdur, knots = c(0, 1, 3, 6, 10))3       0.222 0.159 0.310
Ns(DMdur, knots = c(0, 1, 3, 6, 10))4       0.943 0.836 1.064
lex.CstOAD                                  0.973 0.891 1.063
lex.CstIns                                  0.880 0.742 1.045
lex.CstIns+OAD                              1.508 1.315 1.730
```

Multi-state likelihood — rates of Ins



Rates of insulin uptake

```
> mins <- glm.Lexis(sIO, ~ Ns(DMdur, knots=c(0,1,3,6,10)) + lex.Cst,
+                   from = c("DM","OAD"),
+                     to = c("Ins","Ins+OAD"))

stats::glm Poisson analysis of Lexis object sIO with log link:
Rates for transitions:
DM->Ins
OAD->Ins+OAD

> round(ci.exp(mins), 3)

                                      exp(Est.)  2.5% 97.5%
(Intercept)                               0.216 0.195 0.238
Ns(DMdur, knots = c(0, 1, 3, 6, 10))1     0.137 0.109 0.173
Ns(DMdur, knots = c(0, 1, 3, 6, 10))2     0.358 0.294 0.437
Ns(DMdur, knots = c(0, 1, 3, 6, 10))3     0.002 0.001 0.003
Ns(DMdur, knots = c(0, 1, 3, 6, 10))4     1.609 1.360 1.904
lex.CstOAD                                1.818 1.645 2.008
```

What not to do

```
> mDM <- glm.Lexis(sIO, ~ Ns(DMdur, knots=c(0,1,3,6,10)), from = "DM")

NOTE:
Multiple transitions *from* state ' DM ' - are you sure?
The analysis requested is effectively merging outcome states.
You may want analyses using a *stacked* dataset - see ?stack.Lexis
stats::glm Poisson analysis of Lexis object sIO with log link:
Rates for transitions:
DM->Dead
DM->OAD
DM->Ins

> round(ci.exp(mDM), 3)

                                       exp(Est.)  2.5% 97.5%
(Intercept)                                1.170 1.115 1.229
Ns(DMdur, knots = c(0, 1, 3, 6, 10))1      0.217 0.188 0.250
Ns(DMdur, knots = c(0, 1, 3, 6, 10))2      0.178 0.151 0.211
Ns(DMdur, knots = c(0, 1, 3, 6, 10))3      0.004 0.003 0.005
Ns(DMdur, knots = c(0, 1, 3, 6, 10))4      0.513 0.447 0.588
```

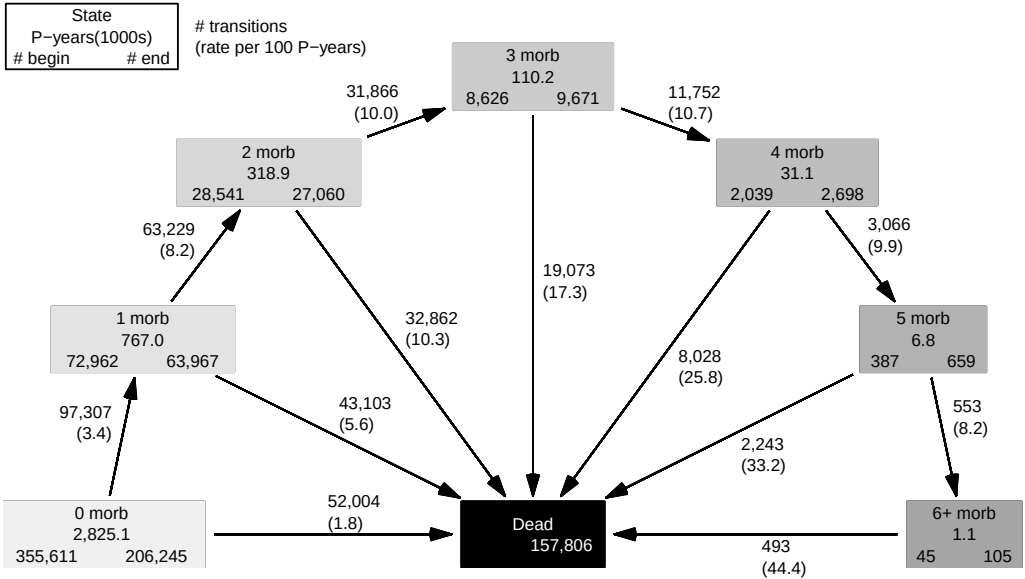
The model is meaningless, not statistically meaningless, but substantially meaningless

—not sensible to have same age effect for different event types

Multi-state model for no. vascular complications

- 9 types of complications (from NPR)
- two types of event rates:
  - Death
  - next complication
- determinants:
  - no. complications
  - age
  - calendar time

Multi-state model — 8 states, 13 transitions



```
> summary(sm)
```

Transitions:

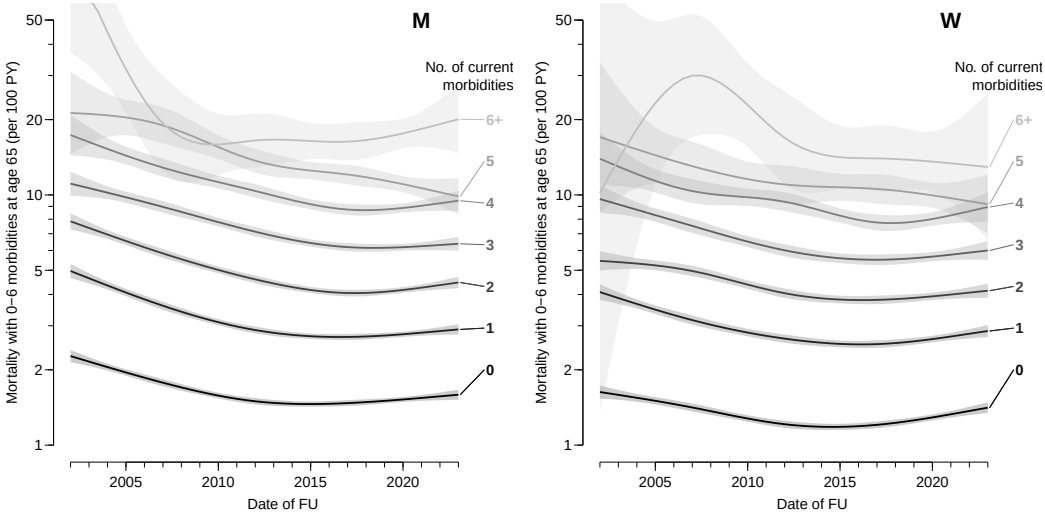
From \ To	0 morb	1 morb	2 morb	3 morb	4 morb	5 morb	6+ morb	Dead
0 morb	2,900,242	97,307	.	.	.	.	.	52,004
1 morb	.	793,775	63,229	.	.	.	.	43,103
2 morb	.	.	330,361	31,866	.	.	.	32,862
3 morb	.	.	.	114,715	11,752	.	.	19,073
4 morb	.	.	.	.	32,275	3,066	.	8,028
5 morb	.	.	.	.	.	7,075	553	2,243
6+ morb	.	.	.	.	.	.	1,236	493
Sum	2,900,242	891,082	393,590	146,581	44,027	10,141	1,789	157,806

From	Records:	Events:	Risk time:	Persons:
0 morb	3,049,553	149,311	2,825,104	355,611
1 morb	900,107	106,332	767,025	170,309
2 morb	395,089	64,728	318,920	91,793
3 morb	145,540	30,825	110,243	40,497
4 morb	43,369	11,094	31,057	13,793
5 morb	9,871	2,796	6,760	3,455
6+ morb	1,729	493	1,111	598
Sum	4,545,258	365,579	4,060,220	468,211

```
> mcmM <- glm.Lexis(subset(sm, sex == "M"), fcm, to = "Dead")

stats::glm Poisson analysis of Lexis object subset(sm, sex == "M") with log link:
Rates for transitions:
0 morb->Dead
1 morb->Dead
2 morb->Dead
3 morb->Dead
4 morb->Dead
5 morb->Dead
6+ morb->Dead

> round(cbind(ci.exp(mcmM), ci.exp(mcmW)), 3)
```

	exp(Est.)	2.5%	97.5%	exp(Est.)	2.5%	97.5%
(Intercept)	0.003	0.002	0.003	0.002	0.001	0.002
Ns(age - 40, knots = -1:4 * 10)1	2.611	2.187	3.117	3.206	2.541	4.044
...						
Ns(age - 40, knots = -1:4 * 10)5	15.319	13.844	16.951	18.405	16.078	21.069
lex.Cst1 morb	1.910	1.876	1.944	2.159	2.118	2.201
lex.Cst2 morb	2.958	2.902	3.015	3.237	3.168	3.307
lex.Cst3 morb	4.445	4.346	4.546	4.772	4.646	4.902
lex.Cst4 morb	6.370	6.181	6.564	6.859	6.589	7.141
lex.Cst5 morb	8.271	7.857	8.706	8.488	7.860	9.167
lex.Cst6+ morb	11.742	10.604	13.003	12.238	10.204	14.677
I(per - 2002)	0.978	0.977	0.980	0.986	0.985	0.988

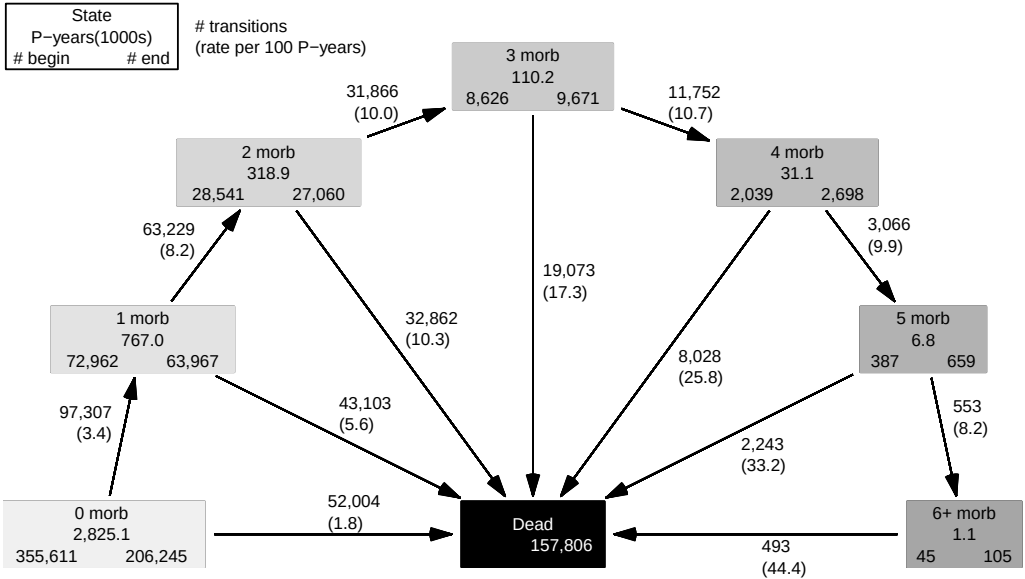
Multi-state model — state×non-linear date of FU



—gradual increase by no. comorbidities

```
> levels(sm)
[1] "0 morb"  "1 morb"  "2 morb"  "3 morb"  "4 morb"  "5 morb"  "6+ morb" "Dead"
> fcm
~Ns(age - 40, knots = -1:4 * 10) + lex.Cst + I(per - 2002)
> ccmM <- glm.Lexis(subset(sm, sex == "M"), fcm, to = levels(sm)[2:7])
stats::glm Poisson analysis of Lexis object subset(sm, sex == "M") with log link:
Rates for transitions:
0 morb->1 morb
1 morb->2 morb
2 morb->3 morb
3 morb->4 morb
4 morb->5 morb
5 morb->6+ morb
```

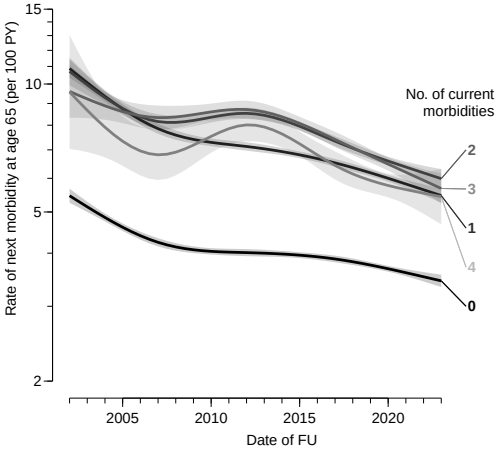
Multi-state model — 8 states, 13 transitions

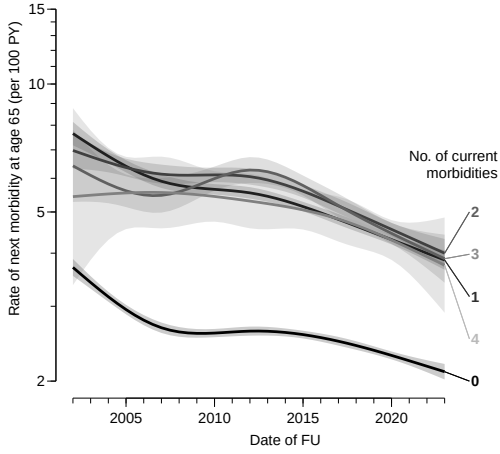


```
> round(cbind(ci.exp(ccmM), ci.exp(ccmW)), 3)
```

	exp(Est.)	2.5%	97.5%	exp(Est.)	2.5%	97.5%
(Intercept)	0.009	0.009	0.010	0.006	0.006	0.007
Ns(age - 40, knots = -1:4 * 10)1	3.590	3.252	3.962	3.238	2.847	3.682
Ns(age - 40, knots = -1:4 * 10)2	4.233	3.853	4.650	3.673	3.281	4.112
Ns(age - 40, knots = -1:4 * 10)3	5.652	5.278	6.052	6.631	6.079	7.234
Ns(age - 40, knots = -1:4 * 10)4	15.114	12.596	18.135	14.900	12.065	18.402
Ns(age - 40, knots = -1:4 * 10)5	6.513	6.165	6.880	8.002	7.405	8.646
lex.Cst1 morb	1.737	1.715	1.760	2.021	1.987	2.054
lex.Cst2 morb	1.933	1.902	1.965	2.137	2.091	2.185
lex.Cst3 morb	1.934	1.888	1.980	2.103	2.032	2.178
lex.Cst4 morb	1.723	1.650	1.798	1.948	1.819	2.085
lex.Cst5 morb	1.472	1.338	1.619	1.426	1.196	1.700
I(per - 2002)	0.980	0.979	0.981	0.979	0.977	0.980

Multi-state model — state×non-linear date of FU





—increase only from 0 to 1

Conclusion

- Registers provide **dates** of **events**
- defines **transition times** between **states**
- ...or time-dependent variables
- data representation in `Lexis` object
- `cut` to introduce intermediate states
- `split` to make intervals short to assume constant rate
- (parametric) models for rates:
  `glm.Lexis`, `gam.Lexis`, `coxph.Lexis`
- predicted rates used to predict survival and expected life time

Material

- Book on line: Practical Multistate Modeling
  https://bendixcarstensen.com/PMM/
- Book: Bendix Carstensen:
  Epidemiology with R, Oxford University Press, 2022
- Vignette in the Epi package:
  Analysis of follow-up data using the Lexis functions in Epi