Swiss TPH
Swiss Tropical and Public Health Institute
Schweizerisches Tropen- und Public Health-Institut
Institut Tropical et de Santé Publique Suisse

Department of Epidemiology & Public Health

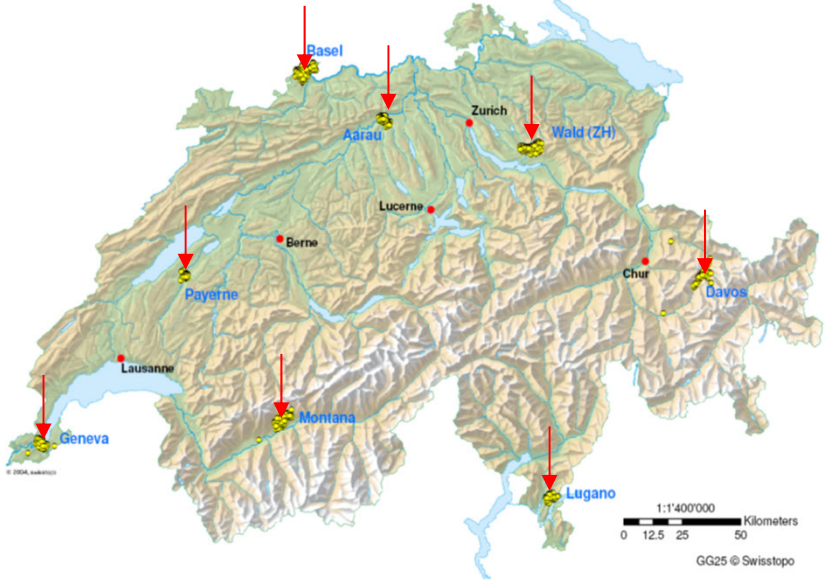
Overview of statistical methods used and envisaged to handle attrition bias in SAPALDIA

Dr. C. Schindler

* Swiss cohort study on air pollution and lung and heart diseases in adults

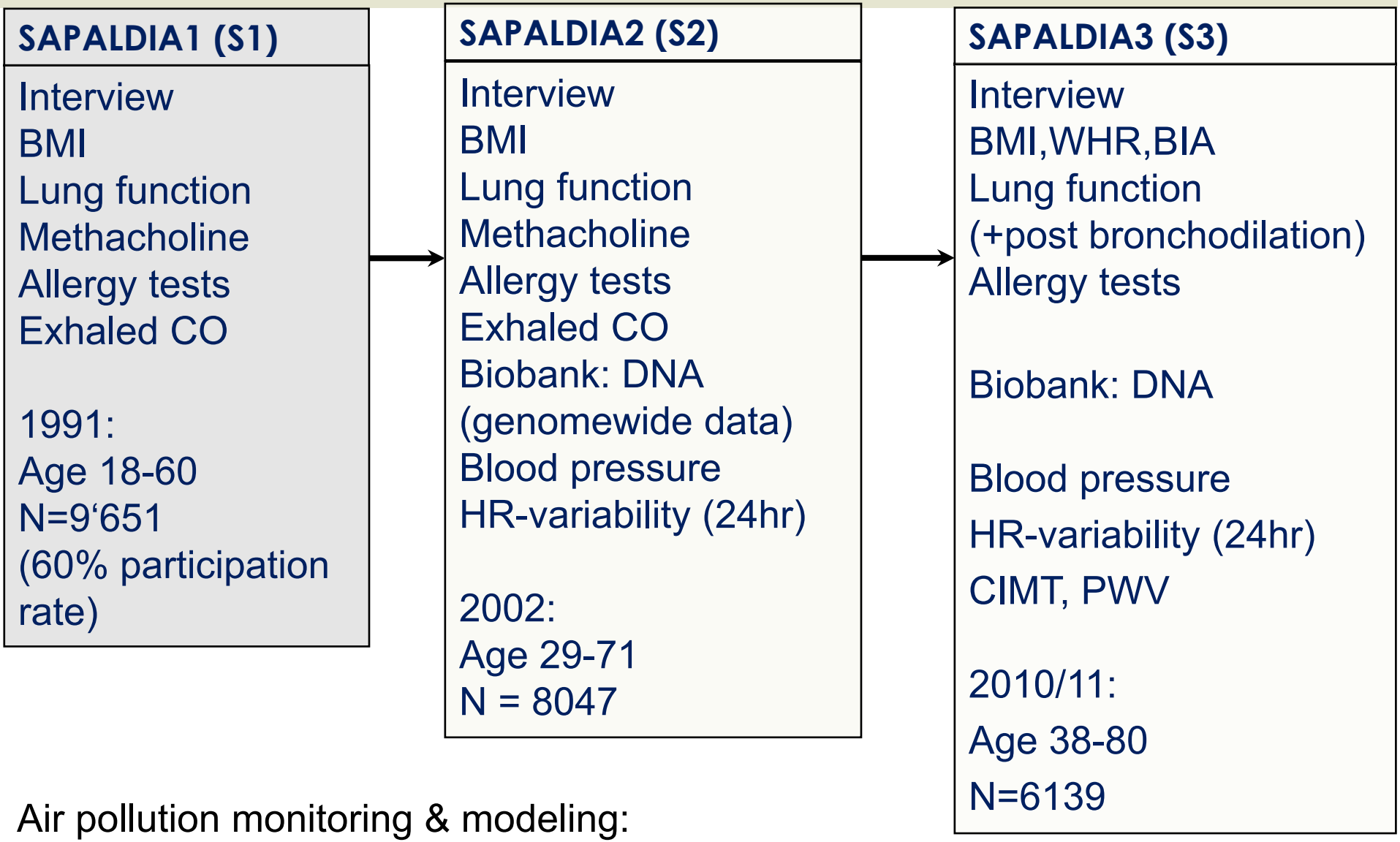
SAPALDIA study population

1991: age = 18 to 60 yrs



2002: age = 29 to 71 years

Swiss TPH



Air pollution monitoring & modeling:
PM10, NO_2, O_3 (since 1991) | PM2.5 (since 2002) | Ultrafine particles (2010/11)

Cohort management and the fight against attrition bias

Updates of address histories and vital status

Addresses are updated and validated

- during surveys
- after mailings of results / newsletters
- regularly by matching addresses with the data base of the Swiss postal services (covers about 60% of all moves in CH)

Sources of new addresses

- data base of postal services
- registries of inhabitants
- participants

Mortality data

Ascertainment of vital status and date of death

families of participants (during surveys and after mailings)

postal services

registries of inhabitants (upon query)

} match with record in the Swiss mortality data base (sex, date of birth, date of death, region of residence)

Provision of cause of death

Cause of death must be provided by the team of the Swiss National Cohort Study which is officially mandated for this. They use links between the mortality and the national census data base.

SAPALDIA cohort end of 2011

Member status at the end of 2011		%
Participated in S3	6088	63.1
Declined participation in S3	1524	15.8
Had moved abroad before S3	429	4.4
Had died before S3	689	7.1
Lost to follow-up before S3	921	9.5
	9651	

Analyses of disease incidence in SAPALDIA

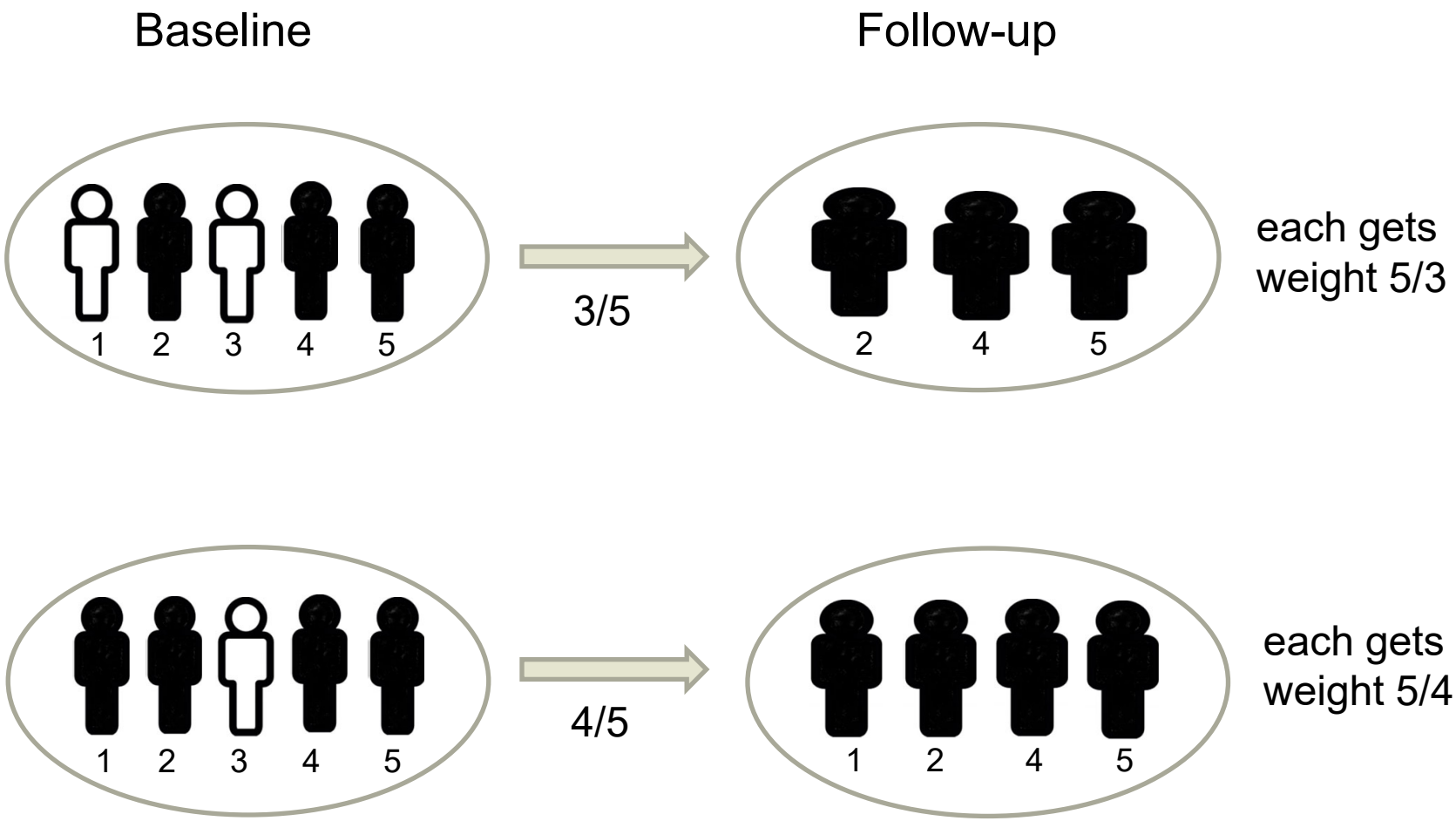
Chronic diseases other than respiratory or allergic ones were not yet assessed in the baseline survey (S1, 1991) but from the first follow-up survey (S2, 2002) on.

Therefore the present methodological considerations and associated simulations are for incidence of disease between the second (S2, 2002) and the third survey (S3, 2010/11)

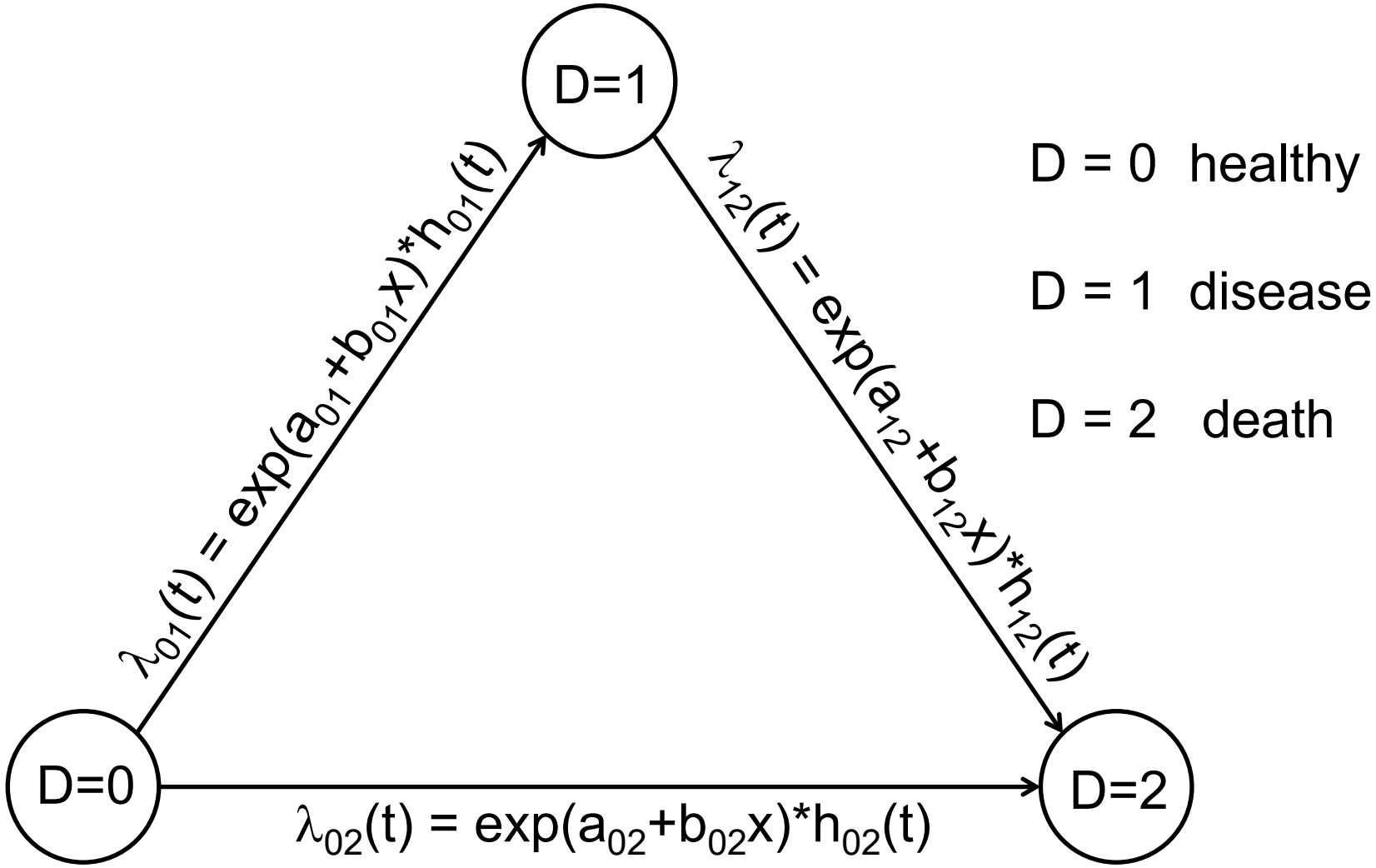
Terminology

Non-participation	Subject can be reached but declines participation Symbol: 
Loss to follow-up	Subject can no longer be reached (but not because it has knowingly died) Symbol: **?**

Inverse probability weighting



3-state model

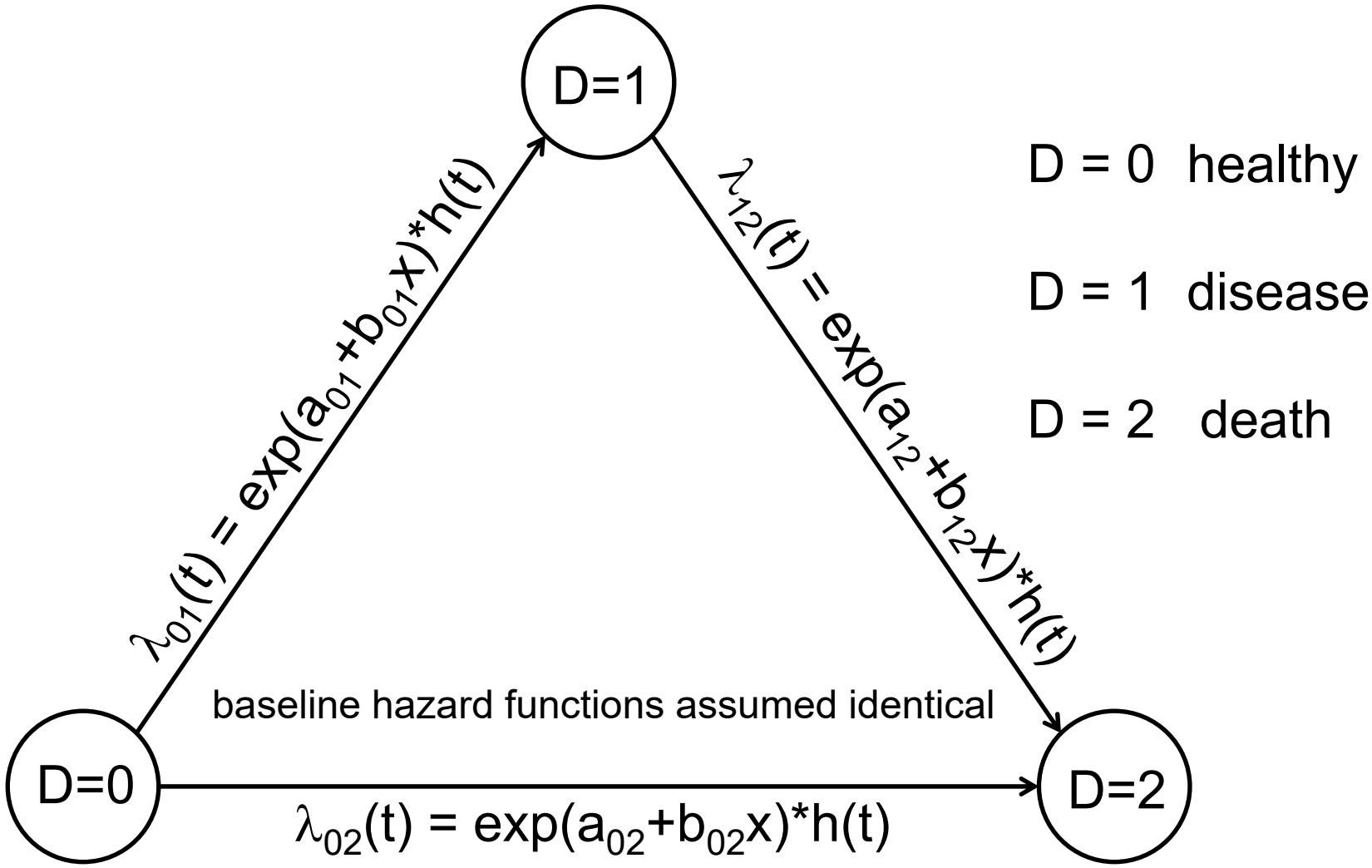


D = 0 healthy

D = 1 disease

D = 2 death

3-state model (simplified)



Modeling h(t)

a) Weibull model: $h(t;p) = p \cdot t^{p-1} \ (p > 0) \quad => \quad H(t;p) = t^p$

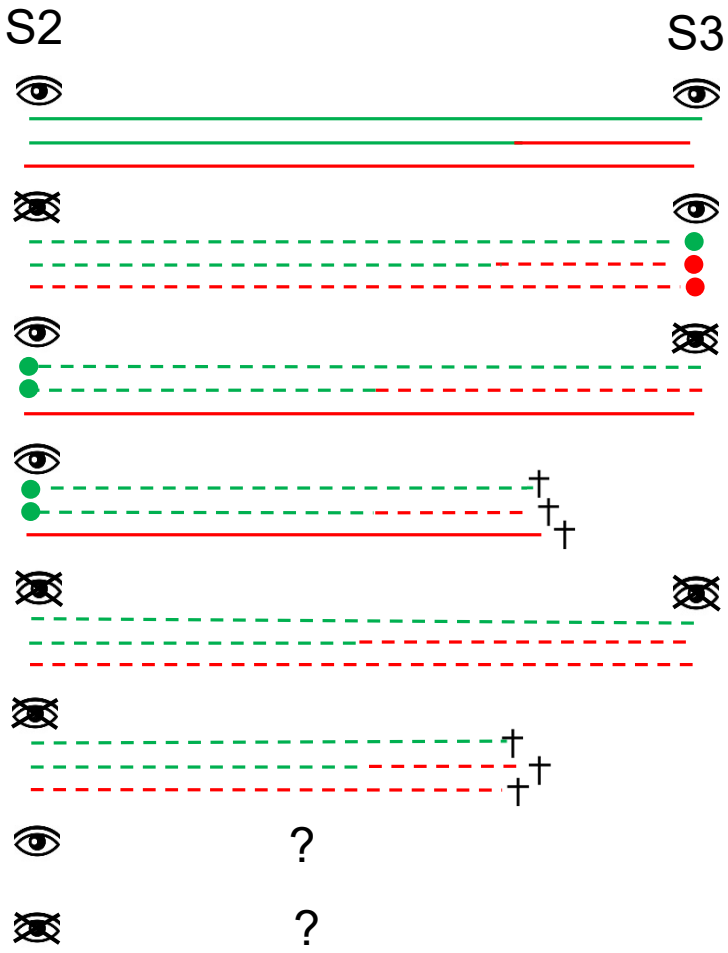
b) Exponential model: $h(t) = 1 \ = h(t;1) \quad => \quad H(t;1) = t$

c) h(t) = step function — allows $h_{ij}(t)$'s to be different, since step heights can be absorbed into the parameter a_{ij} of u_{ij} .

d) h(t) = spline of t

In our simulation, we use an exponential model (time-invariant hazard).

Different data situations



assessed at survey

confirmed alive but not assessed

healthy

diseased

manifest trajectory

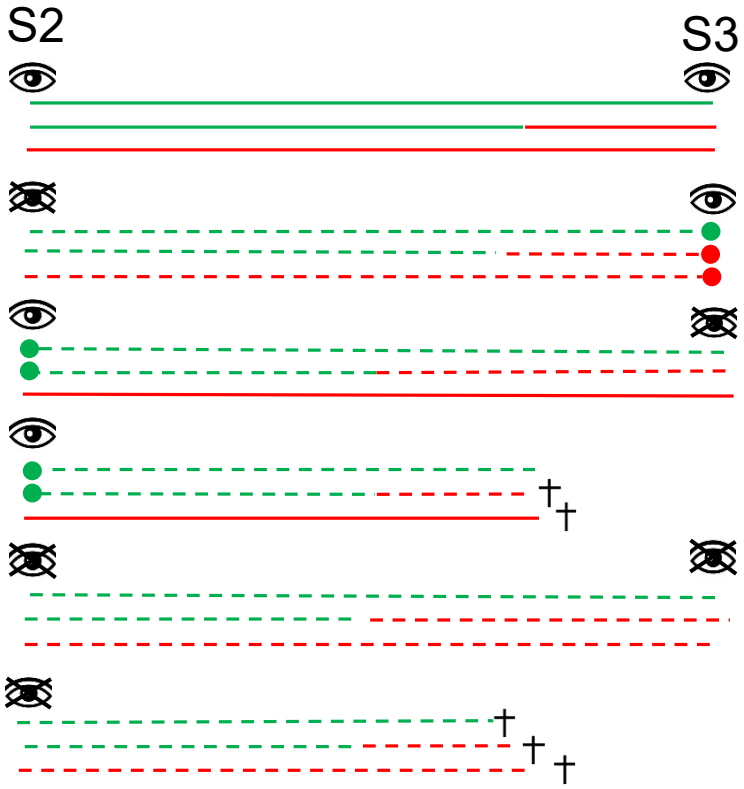
latent trajectory

S2 = SAPALDIA2 (2002)

S3 = SAPALDIA3 (2010/11)

Models and their domains of definition

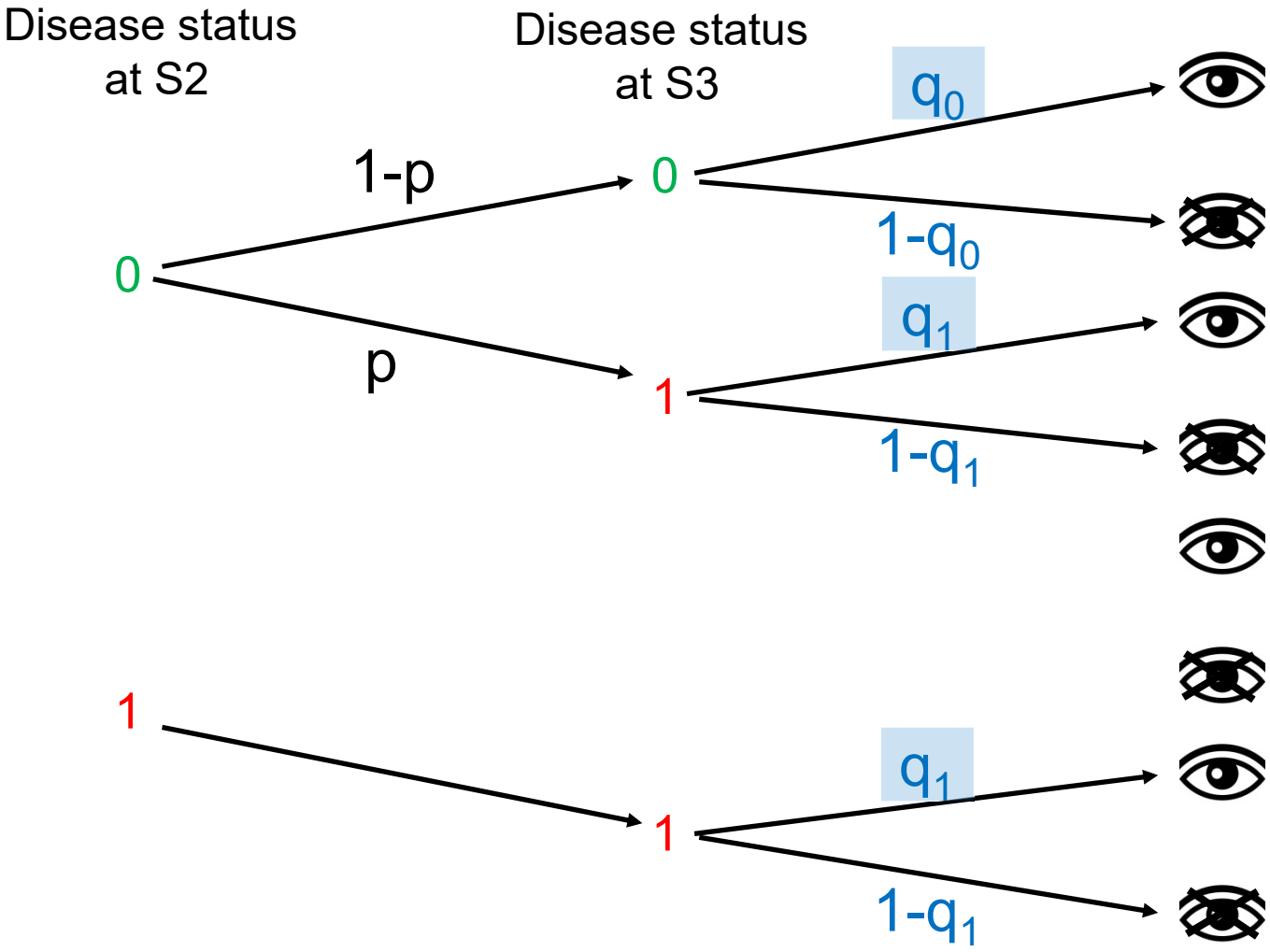
M1 = complete case analysis
M2 = inverse probability weighting
M3 = 3-state model
M4 = M3 + non-participation at S3
M5 = M4 + non-participation at S2



M1	M2	M3	M4	M5
x	x	x	x	x
–	(x)	–	–	(x)
–	x	–	x	x
–	x	x	x	x
–	(x)	–	–	(x)
–	(x)	–	–	(x)

() if only S1-covariates and residential data are used

Inclusion of non-participation at S3



Assumption: participation only depends on the presence but not on the duration of the disease.

$$odds\left(D_3^{(obs)} = . \middle| D_2 = 0\right) = \frac{p \cdot (1 - q_1) + (1 - p) \cdot (1 - q_0)}{(1 - p) \cdot q_0 + p \cdot q_1}$$

$$P\left(D_3^{(obs)} = 1 \middle| D_2 = 0, D_3^{(obs)} \neq .\right) = \frac{p \cdot q_1}{(1 - p) \cdot q_0 + p \cdot q_1}$$

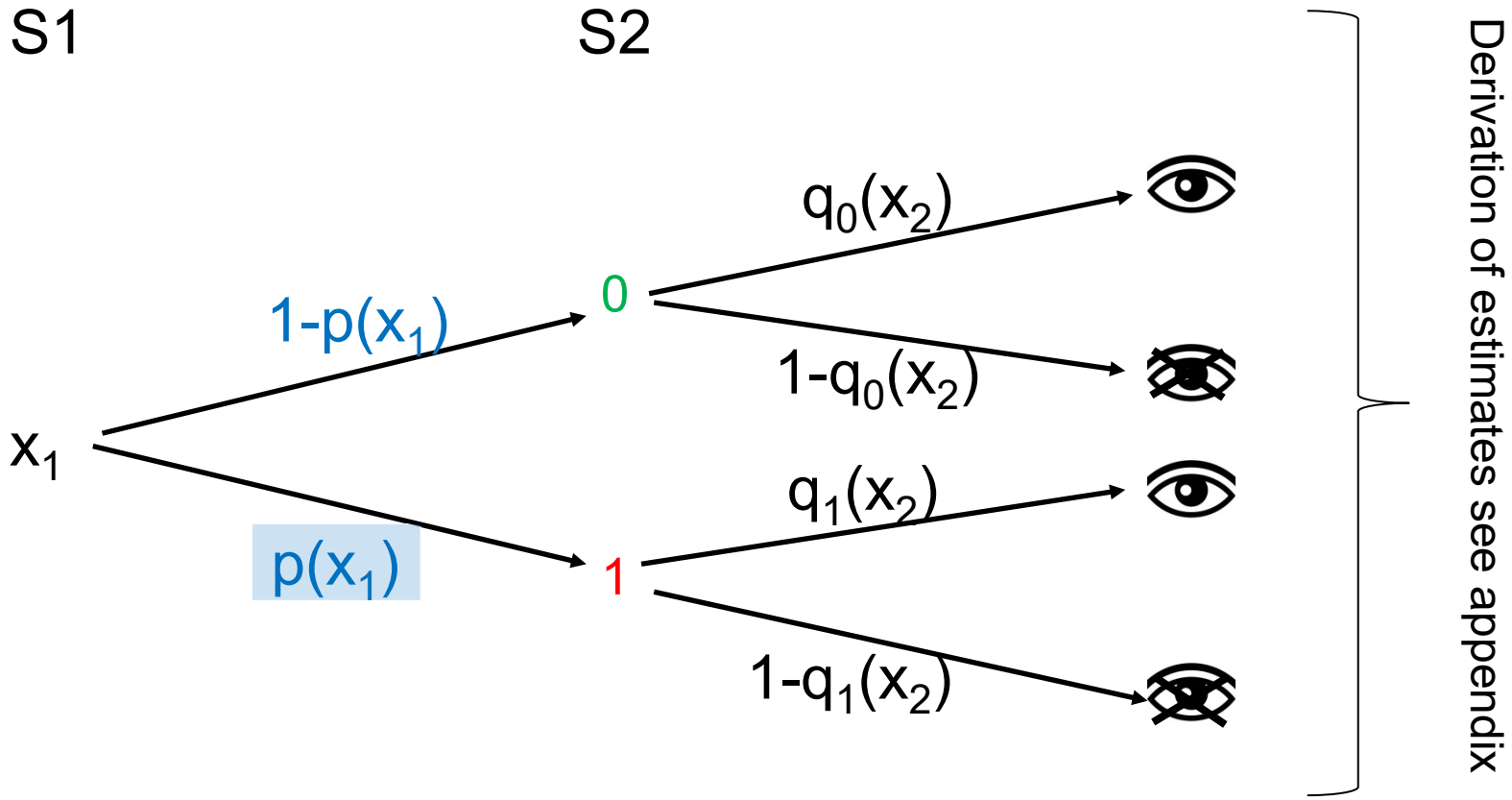
(A) $$odds\left(D_3^{(obs)} = . \middle| D_2 = 0\right) - \frac{1 - q_1}{q_1} \cdot P\left(D_3^{(obs)} = 1 \middle| D_2 = 0, D_3^{(obs)} \neq .\right)$$

$$= \frac{(1 - p) \cdot (1 - q_0)}{(1 - p) \cdot q_0 + p \cdot q_1}$$

(B) $$P\left(D_3^{(obs)} = 0 \middle| D_2 = 0, D_3^{(obs)} \neq .\right) = \frac{(1 - p) \cdot q_0}{(1 - p) \cdot q_0 + p \cdot q_1}$$

$$\frac{(B)}{(A)} = \frac{q_0}{1 + q_0} \quad \Longrightarrow \quad q_0 = \frac{(B)/(A)}{1 + (B)/(A)}$$

Inclusion of unobserved disease status at SAPALDIA 2



S1 = SAPALDIA1 (1991), S2 = SAPALDIA2 (2002)

Example of a pseudo-likelihood component

Situation: S2 and S3 =  (i.e., D2 ≠ 2 and D3 ≠ 2)

$$(1-p_0(x)) \cdot (1 - q_0(x)) \cdot l_{00}(x; a_{ij}, b_{ij}) \cdot (1-q_0(x))$$

$$+ (1-p_0(x)) \cdot (1 - q_0(x)) \cdot l_{01}(x; a_{ij}, b_{ij}) \cdot (1 - q_1(x))$$

$$+ \quad p_0(x) \cdot (1 - q_1(x)) \cdot l_{11}(x; a_{12}, b_{12}) \cdot (1 - q_1(x))$$

Simulation approach 1

Assumptions

1. Constant hazard for all three transitions
2. Only one covariate $x_1 = x_2 \sim U(0,1)$ $(\Rightarrow \lambda_{ij}(t) = \exp(a_{ij} + b_{ij} \cdot x),\ 0 \leq i < j \leq 2)$
3. $q_i = P(D_t^{(obs)} \neq . \mid D_t = i, x) = 1 / [1 + \exp(c_0 + c_1 \cdot i + c_2 \cdot x)]$ $(t=2,3,\ i=0,1)$
4. Exponential censoring independent of x and disease status for loss to follow-up.

5. Initial sample size = 10'000 (n = 9651 in S1)
6. Parameters for 3) and 4) chosen such as to produce participation rates similar to S2 and S3 ($c_0 = -1.7$, $c_1 = 0.6$, $c_2 = 0.6$).
7. $a_{01} = -5$ (=> 19-yr incidence (S1 to S3) ≈ 10%, if x = 0) (A)
8. $a_{02} = -5$ (=> 19-yr mortality among D=0 (S1 to S3) ≈ 10%, if x = 0) (B)
9. $a_{12} = -4$ (=> 19-yr mortality among D=1 (S1 to S3) ≈ 30%, if x = 0) (C)
10. $b_{01} = 0.3$, $b_{02} = 0.25$, $b_{12} = 0.55$

(10) implies that rates would be about 16% (A), 15% (B) and 45% (C), if x = 1.

Simulations were conducted in SAS using

PROC NLMIXED

with user-defined log-likelihood contributions for the different situations s and model statement

model s ~ general(ll);

R-package for 3-state models: SmoothHazard

Swiss TPH

Results of simulations around fixed parameter values (500 replications)

Parameter	true value	M1*	M2*	M3*	M4*	M5*
b_{01}	0.3	0.190	**0.269**	**0.271**	**0.298**	**0.288**
		(-0.37, 0.73)	(-0.24 , 0.87)	(-0.23, 0.70)	(-0.38, 0.77)	(-0.27, 0.77)
b_{02}	0.25	---	---	0.365	**0.253**	**0.241**
		---	---	(-0.28, 1.07)	(-0.69, 1.16)	(-0.48, 0.93)
b_{12}	0.55	---	---	0.703	**0.539**	**0.564**
		---	---	(-0.28, 1.07)	(-0.17, 1.25)	(-0.03, 1.16)
a_{01}	-4.0	**-4.18**	**-4.08**	**-4.10**	**-4.00**	**-4.00**
		(-4.5, -3.9)	(-4.4, -3.8)	(-4.4, -3.8)	(-4.3, -3.7)	(-4.3, -3.7)
a_{02}	-5.0	---	---	**-4.83**	**-5.00**	**-5.00**
		---	---	(-5.3, -4.5)	(-5.5, -4.5)	(-5.4, -4.6)
a_{12}	-4.0	---	---	**-3.74**	**-3.98**	**-4.00**
		---	---	(-4.2, -3.4)	(-4.5, -3.6)	(-4.4, -3.6)

* Models 1 – 5, displayed are means and ranges of estimates

Simulation approach 2

Assumptions

Same as for approach 1, but with b_{01}, b_{02}, b_{12}, c_1, c_{20} and c_{21} randomly sampled from U(0,1), where c_{2i} = coefficient of x in q_i (i = 0,1)

After each simulation step, the estimates of b_{01}, b_{02} and b_{12} were compared with the respective ,,true'' parameter values and

a) | estimate – true value |

b) estimate – true value

were computed.

Both differences were averaged across all simulations

Swiss TPH

Results of simulations around varying parameter values (500 replications)

Parameter	Δ's	M1*	M2*	M3*	M4*	M5*
b_{01}	signed	-0.075	-0.022	-0.028	0.006	-0.006
	absolute	0.167	0.135	0.145	0.129	0.117
b_{02}	signed	---	---	0.094	0.001	-0.008
	absolute	---	---	0.190	0.174	0.169
b_{12}	signed	---	---	0.124	-0.006	-0.007
	absolute	---	---	0.195	0.171	0.146

* Models 1 – 5, displayed are means of absolute and signed estimation errors

Discussion

1. Models 4 and 5 performed well in the simulated scenarios with considerable differential non-participation. But model 5 might not perform well in realistic situations where the probability of non-participation at S2 also depends on unmeasured covariates and not only on disease status at S2. The same applies to model 4 if the probability of non-participation at S3 depends on unmeasured covariates at S3.

2. Both models are based on the assumption that the probability of non-participation only depends on disease status at the time of decision and not on the duration or severity of the disease.

3. Model 3 (ignoring non-participation) performed quite well in estimating the parameter b_{01} (effect of x on incidence of disease) but not with the parameters for the effect of x on mortality.

cont.

4. Model 2 (inverse probability weighting) performed quite well in estimating the parameter b_{01} (effect of x on disease incidence) and might be able to reduce bias from differential non-participation.

5. In our scenario, we assumed non-informative loss to follow-up. The present findings can therefore not be generalized to situations where loss to follow-up depends on disease status.

6. The algorithm estimating q_0 may produce values outside (0, 1) if separate logistic regression models are used to estimate the different probability components, but using a multinomial logistic regression regression model might help to avoid inconsistencies.

7. Truncation of inverse probability weights may be necessary to prevent some observations from becoming very influential.

8. Models will have to be extended to time-dependent hazards.

This preliminary work has been developed in collaboration with

Prof. Dr. Martin Schumacher and
Dr. Nadine Binder

Institut für Medizinische Biometrie und Statistik
University of Freiburg
Germany

Appendix

Example of a likelihood computation (3-state model)

$$L(D(t)=2 \mid D(0)=0) \quad = \exp\left(-(u_{01}+u_{02})\cdot H(t)\right)\cdot u_{02}\cdot h(t)$$

$$+\int_0^t \exp\left(-(u_{01}+u_{02})\cdot H(s)\right) * u_{01} * h(s) * \exp\left(-u_{12}\cdot\left(H(t)-H(s)\right)\right)\cdot u_{12}\cdot h(t)\cdot \mathrm{ds}$$

where $u_{ij} = \exp(a_{ij} + b_{ij}*x)$ and $H(t) = \int_0^t h(s)ds$.

After a few calculations one obtains

$$\mathrm{L}(D(t)=2 \mid D(0)=0) \quad = \exp\left(-(u_{01}+u_{02})\cdot H(t)\right)\cdot u_{02}\cdot h(t)$$

$$+\frac{u_{01}}{u_{01}+u_{02}-u_{12}}\left[\exp\left(-u_{12}\cdot H(t)\right)-\exp\left(-(u_{01}+u_{02})\cdot H(t)\right)\right]\cdot h(t)\cdot u_{02}$$

Inclusion of unobserved disease status at Sapaldia 2

$$odds\left(D_2{}^{(obs)} = 1 \,\middle|\, x_1, D_2{}^{(obs)} \neq .\right) = \frac{p(x_1) \cdot q_1(x_2)}{\left(1 - p(x_1)\right) \cdot q_0(x_2)}$$

$$\frac{p(x_1)}{1 - p(x_1)} = \frac{q_0(x_2)}{q_1(x_2)} \cdot odds\left(D_2{}^{(obs)} = 1 \,\middle|\, x_1, D_2{}^{(obs)} \neq .\right) \quad (C)$$

⇩

$$p(x_1) = \frac{(C)}{1 + (C)}$$ (probability of D_2=1 given x_1)

In my simulation, I have assumed that $x_2 = x_1$.