Survival analysis of longitudinal data: the case of English population aged 50+

Marjan Qazvini

Heriot-Watt University, Dubai

Longevity 16 Conference, August 2021

Outline

Introduction

- Analysis of survey data from the English Longitudinal Study of Ageing (ELSA)
- Discrete-time survival analysis
- The impact of demographic and self-rated health variables on the survival of the population aged 50+

Questions:

- What are longitudinal data?
- Why discrete-time survival analysis?
- How to deal with missing values?

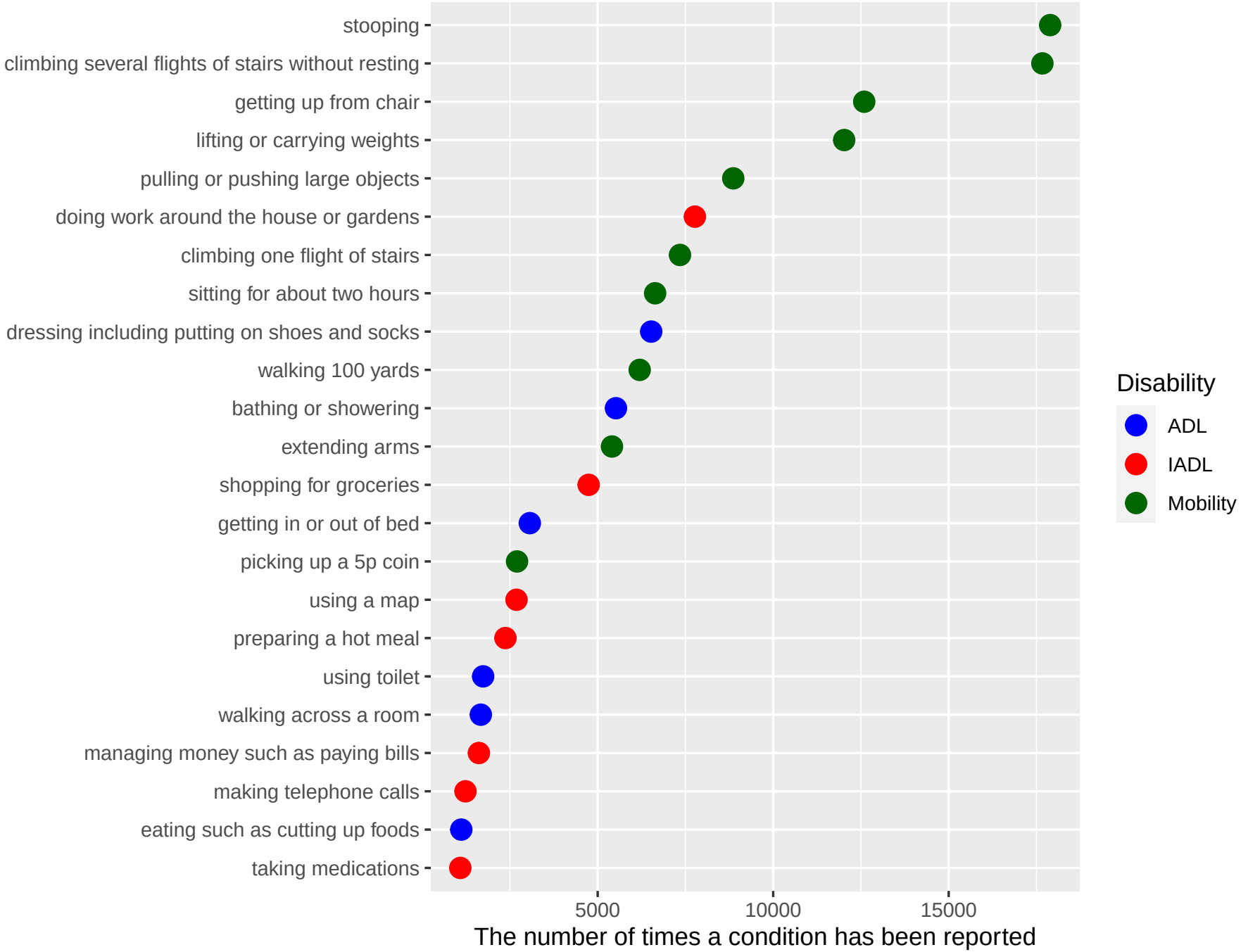
Data: ELSA (core members)

Waves	Cohort 1	Cohort 3	Cohort 4	Death
1) 2002-2003	11,391	-	-	-
2) 2004-2005	8,780	-	-	133
3) 2006-2007	7,535	1,275	-	369
4) 2008-2009	6,623	972	2,291	234
5) 2010-2011	6,242	936	1,912	-
6) 2012-2013	-	-	-	240

NatCen: Technical Report (Wave 6)

Variables

- Age: 50-90
- Gender: male, female
- Marital status: single, couple (including civil partnership)
- Employment status: retired, permanently sick or disabled, self-employed, other (looking after home or family, semi-retired, unemployed and other)
- Mobility: $0, 1, 2, \ldots, 10$
- Activities of daily living: $0, 1, 2, \ldots, 6$
- Instrumental activities of daily living: $0, 1, 2, \ldots, 7$
- Diseases: yes, no
- Status: dead, alive


stooping
climbing several flights of stairs without resting
getting up from chair
lifting or carrying weights
pulling or pushing large objects
doing work around the house or gardens
climbing one flight of stairs
sitting for about two hours
dressing including putting on shoes and socks
walking 100 yards
bathing or showering
extending arms
shopping for groceries
getting in or out of bed
picking up a 5p coin
using a map
preparing a hot meal
using toilet
walking across a room
managing money such as paying bills
making telephone calls
eating such as cutting up foods
taking medications
5000
10000
15000
The number of times a condition has been reported
Disability
ADL
IADL
Mobility

Longitudinal data and missing values (example)

Person 1 joins in the interval $[t_1, t_2)$ and dies in $[t_4, t_5)$

Person 2 joins in $[t_1, t_2)$ and permanently withdraws in $[t_2, t_3)$ (right-censored)

Person 3 joins in $[t_2, t_3)$ and dies in $[t_2, t_3)$ (left-censored)

Person 4 joins in $[t_1, t_2)$, temporary withdraws in $[t_3, t_4)$, rejoins in $[t_4, t_5)$ and dies in $[t_4, t_5)$

ID	Time	V_1	V_2	V_3	...	Status
1	$[t_1 - t_2)$	1	0	0	...	0
1	$[t_2 - t_3)$	*NA*	0	0	...	0
1	$[t_3 - t_4)$	1	*NA*	0	...	0
1	$[t_4 - t_5)$	0	1	0	...	1
2	$[t_1 - t_2)$	0	*NA*	0	...	0
2	$[t_2 - t_3)$	1	0	0	...	0
3	$[t_1 - t_2)$	*NA*	*NA*	*NA*	...	0
3	$[t_2 - t_3)$	1	0	0	...	1
4	$[t_1 - t_2)$	0	1	0	...	0
4	$[t_2 - t_3)$	1	0	0	...	0
4	$[t_3 - t_4)$	*NA*	*NA*	*NA*	...	0
4	$[t_4 - t_5)$	0	1	0	...	1

V_1, V_2, ... are variables such as age, ADL score, etc.

Missing values

- "Refusal", "don't know" and "schedule not applicable" are set as *NA*.
- "Not applicable" is set as 0. We consider both current symptoms and newly-diagnosed symptoms.
- Missing values include the waves interviewees who have not participated like $[t_1, t_2)$ for Person 3 and $[t_3, t_4)$ for Person 4.
- Participants above 60 are considered as "retired".
- Single imputation: Last Observation Carried Forward (LOCF) is used to impute *NAs* for "employment and marital status".
- Multiple imputation: MICE is used to impute other *NAs*.
- Predictive mean matching is used for "mobility", "ADL" and "IADL" and logistic regression for other variables.
- MICE can be carried out in R using the package "mice".

Random effects model

- T: a discrete-time random variable, where $T = t$ means the event has happened in the interval $[a_{t-1}, a_t)$.
- $\mathbf{x}_{it}$: a vector of covariates for individual i at time t.
- b_i: a random intercept specific to individual i.
- γ_{0t}: a fixed intercept which may depend on time.
- γ: a vector of parameters for covariates.
- $h(.)$: a suitable function to ensure the hazard function is in $[0, 1]$.

Discrete-time hazard function is defined as

$$\lambda(t|\mathbf{x}_{it}, b_i) = \Pr(T = t | T \geq t, \mathbf{x}_{it}, b_i) = h(\eta_{it}),$$

where η is the linear predictor given by

$$\eta_{it} = b_i + \gamma_{0t} + \mathbf{x}_{it}^T \gamma.$$

Random effects model (continued)

Discrete-time survival probability is defined as

$$S(t|\mathbf{x}_{it}, b_i) = \Pr(T > t|\mathbf{x}_{it}, b_i).$$

We use complementary log-log (clog-log) function for h, given by

$$h(\eta) = 1 - \exp\left(-\exp(\eta)\right).$$

- The random effect models can be easily applied in R using the package "glmer".

- We have $59,265$ observations and $14,964$ unique individuals.
- We divide our dataset into 70% training set and 30% validation set.
- Fit a GLM model with a clog-log link function.
- Fit a random effects model to significant variables using package "glmer".
- We use Numerical Gauss-Hermite quadrature with 9 quadrature points, i.e. nAGQ $= 9$ for numerical optimisation.
- The number of function evaluations before termination is $100,000$.
- The variable age is scaled.
- AIC and BIC improve.
- The estimated parameters, AIC and BIC are reasonably close for all 5 created datasets after imputation.

Parameter estimation

Covariate	Estimate	Std. Error	Covariate	Estimate	Std. Error
Intercept	-5.4944^{***}	0.3492	Mobility	0.0623^{**}	0.0200
Time 2	-0.0488	0.1184	ADL	0.0880^{*}	0.0389
Time 3	-0.2683^{*}	0.1313	IADL	0.1591^{***}	0.0331
Time 4	-0.4687^{***}	0.1397	Heart attack(Y)	0.6959^{***}	0.1581
Time 5	-0.1371^{***}	0.1661	Heart failure(Y)	0.9739^{***}	0.2768
Gender(F)	-0.5309^{***}	0.0991	Diabetes(Y)	0.4037^{**}	0.1228
Age(scaled)	1.2049^{***}	0.0883	Stroke(Y)	0.3811^{*}	0.1540
Employment 2	-1.0803	1.0344	Lung disease(Y)	0.6829^{***}	0.1362
Employment 3	1.4866^{***}	0.3399	Arthritis(Y)	-0.2298^{*}	0.0941
Employment 4	0.0892	0.2951	Cancer(Y)	1.2938^{***}	0.1276
Employment 5	1.0352^{*}	0.4253	Alzheimer's(Y)	0.7538^{*}	0.3179
Marital status(S)	-0.1456	0.0985			
Random effect:					
Variance		1.346			
Number of observations		41,543 groups; idauniq: 10,475			

$*p < 0.05, **p < 0.01, ***p < 0.001$

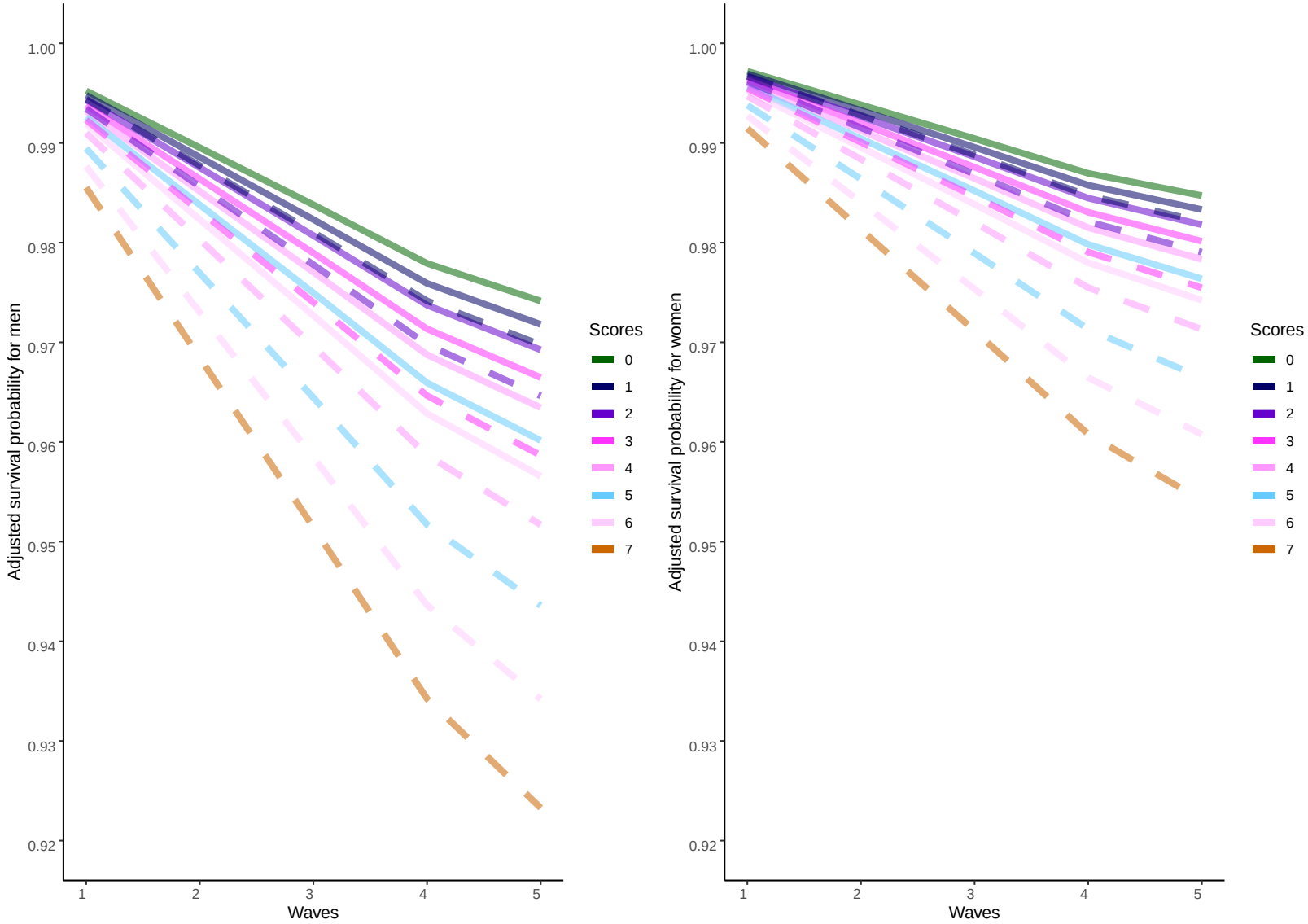


Figure: Adjusted survival probability for different ADL scores (solid line) and IADL scores (dashed line) for a male (left) and female (right) aged 65 in wave 1, retired, in relationship, with no disease (population)

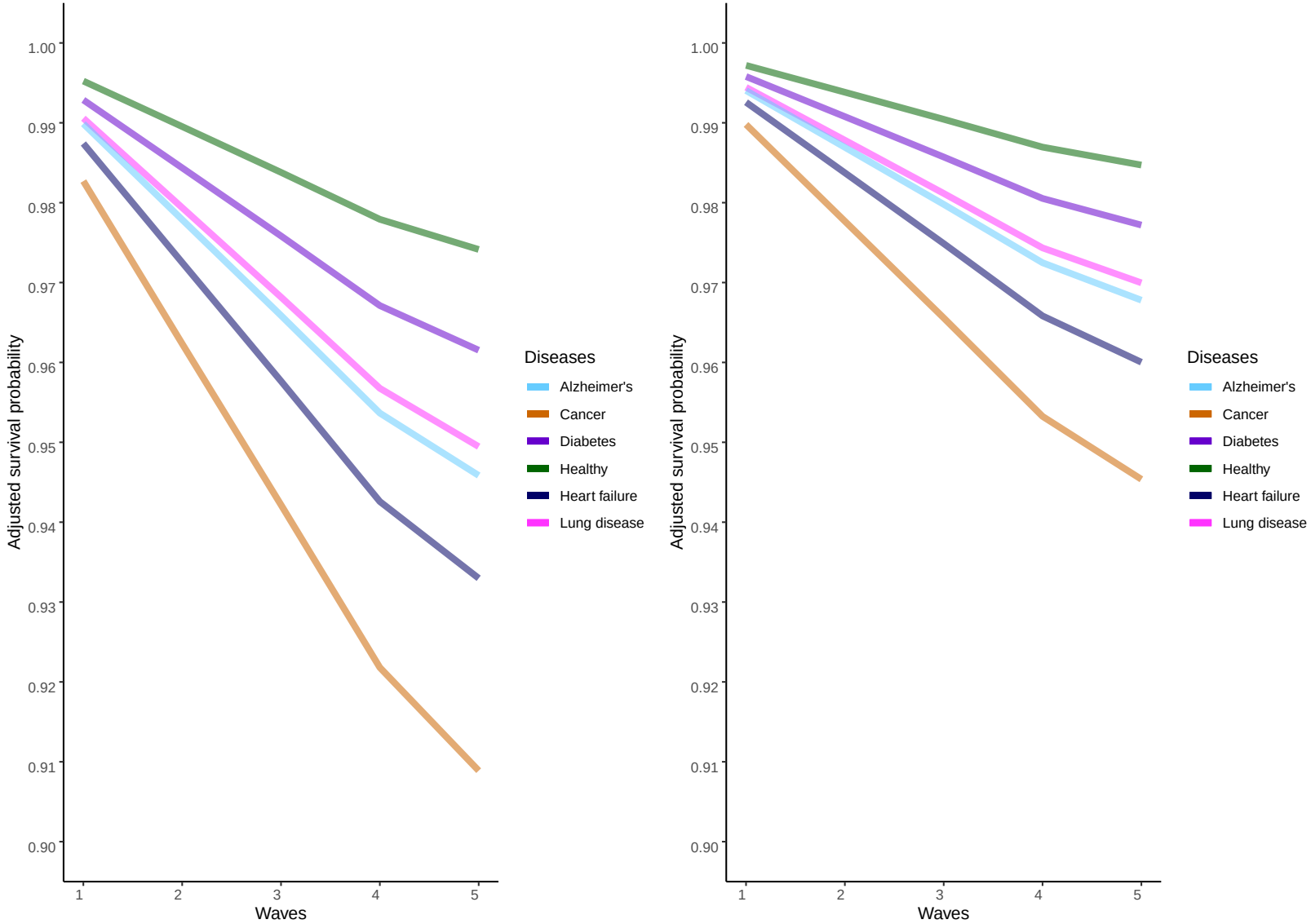


Figure: Adjusted survival probability for different diseases for a male (left) and female (right) aged 65 in wave 1, retired, in relationship (population)

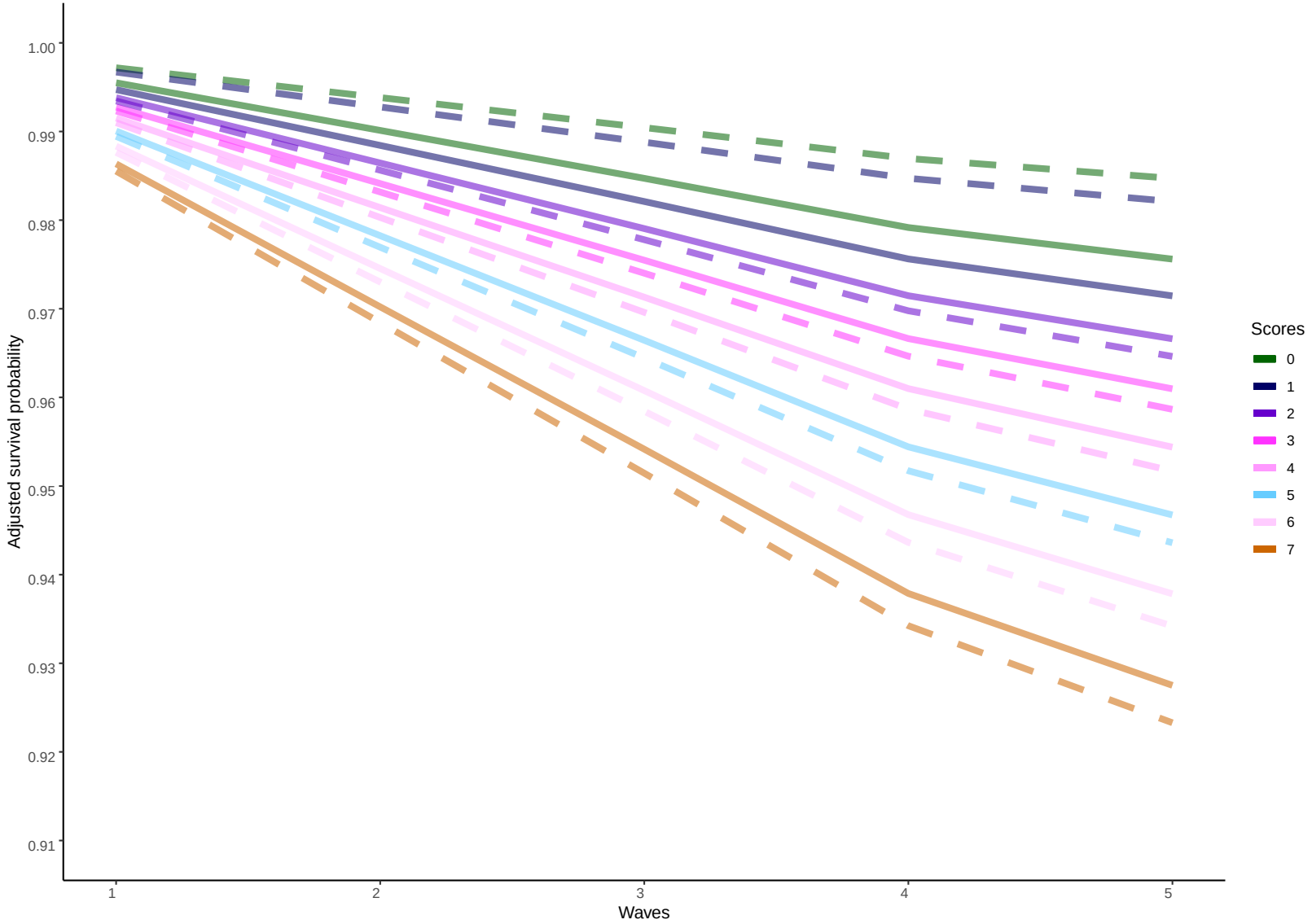


Figure: Adjusted survival probability for different IADL scores for a male, aged 65 in wave 1, retired, in relationship, individual (solid line) and average population (dashed line)

ROC and AUC:

- We consider two levels:
  - death at t
  - death after t
- We define True Positive Rate as $\Pr(\eta > c | T = t)$
- and False Positive Rate as $\Pr(\eta > c | T > t)$,where
  - c is a threshold
  - η is the linear predictor in our model
- The time-dependent ROC curve plots the True Positive Rate against the False Positive Rate for different thresholds.
- The AUC is the area under the time-dependent ROC curves for each time t (Tutz and Schmid, 2016).

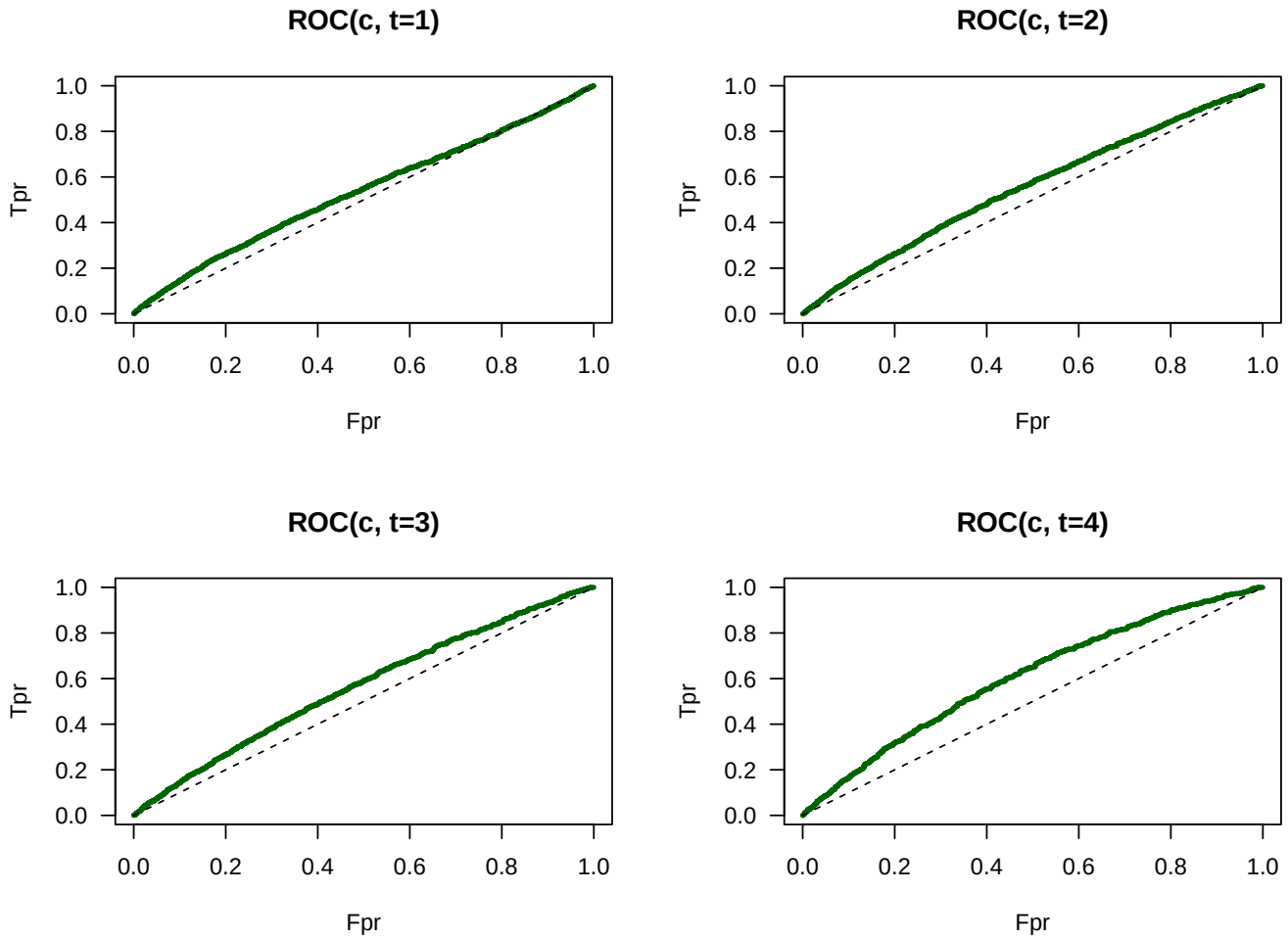


Figure: Random effects model for the population average based on dataset I: AUC(t=1) = 0.534, AUC(t=2) = 0.554, AUC(t=3) = 0.562, AUC(t=4) = 0.605.

Conclusion

- When we don't know the exact time of death, we can use discrete-time survival analysis.
- Random effects model can distinguish between the individual-level and population-level hazards.
- The adjusted survival probability for people with difficulty in performing IADL is less than the adjusted survival probability for people with difficulty in performing ADL.
- The AUC for the time interval $t = 4$ is greater than the AUC for the intervals $1, 2$ and 3.

References

- Steptoe, A., Breeze, E., Banks, J. and Nazroo, J. (2013) Cohort profile: The English Longitudinal Study of Ageing. *International Journal of Epidemiology*, 42: 1640–1648.
- Tutz, G. and Schmid, M. (2016). Modelling discrete time-to-event data. Springer, Switzerland.
- van Buuren, S and Groothuis-Oudshoorn, K. "(2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45 (3).
- van Burren, S. (2018). Flexible imputation of missing data, Second Edition. Chapman & Hall/CRC.