THE UNIVERSITY OF QUEENSLAND
AUSTRALIA
CREATE CHANGE

Visualization and modelling changes in categorical variables in longitudinal studies

Professor Gita Mishra

NHMRC Principal Research Fellow

School of Public Health

The University of Queensland, Australia

@mishra_gita

Outline

Life course approach

Visualisation of longitudinal data

Life course models

SLCMA methods

Example – SES and Mortality

Latest developments and use cases

Conclusion & future directions

The seven ages of woman
– Hans Baldung Grien 1544

Why take a life course approach?

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

Focus on timing and duration of factors across life

Genetics

Gestation

Childhood

Adolescence

Young adulthood

Middle adult life

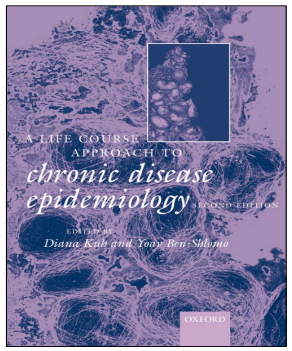
Later adult life

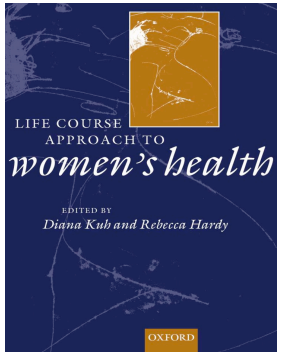
Across generations

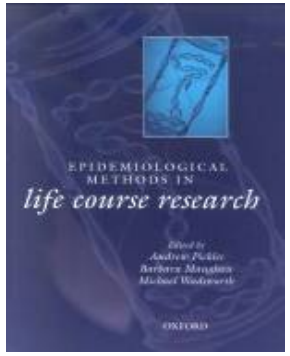
What is life course epidemiology?

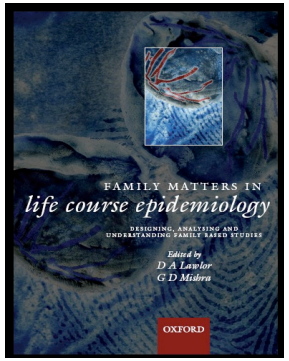
It studies the biological, behavioural and psychosocial pathways that operate across the life course and influence the development of chronic diseases.

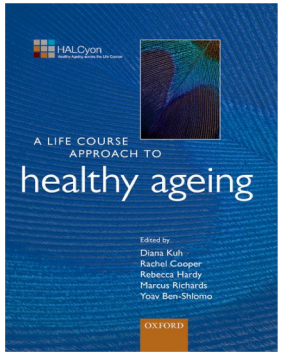
Kuh & Ben Shlomo 1997, 2004

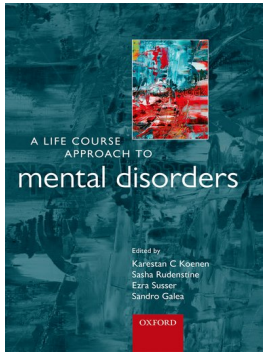












Overarching aim of life course epidemiology

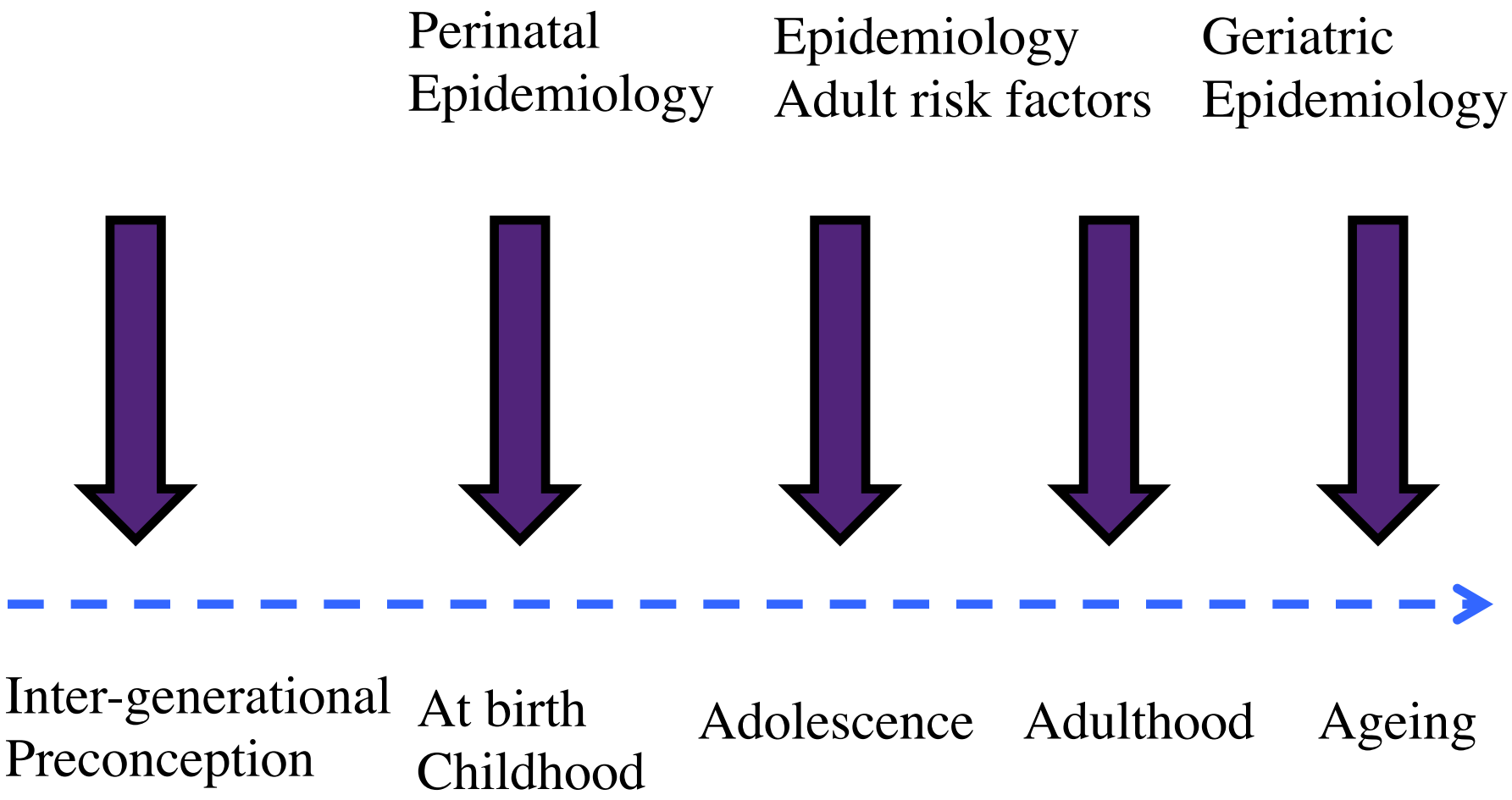
Life Course Epidemiology

Gita D Mishra, School of Population Health, The University of Queensland, Herston, QLD, Australia
Diana Kuh, MRC Unit for Lifelong Health and Ageing, London, UK
Yoav Ben-Shlomo, School of Social and Community Medicine, University of Bristol, Bristol, UK

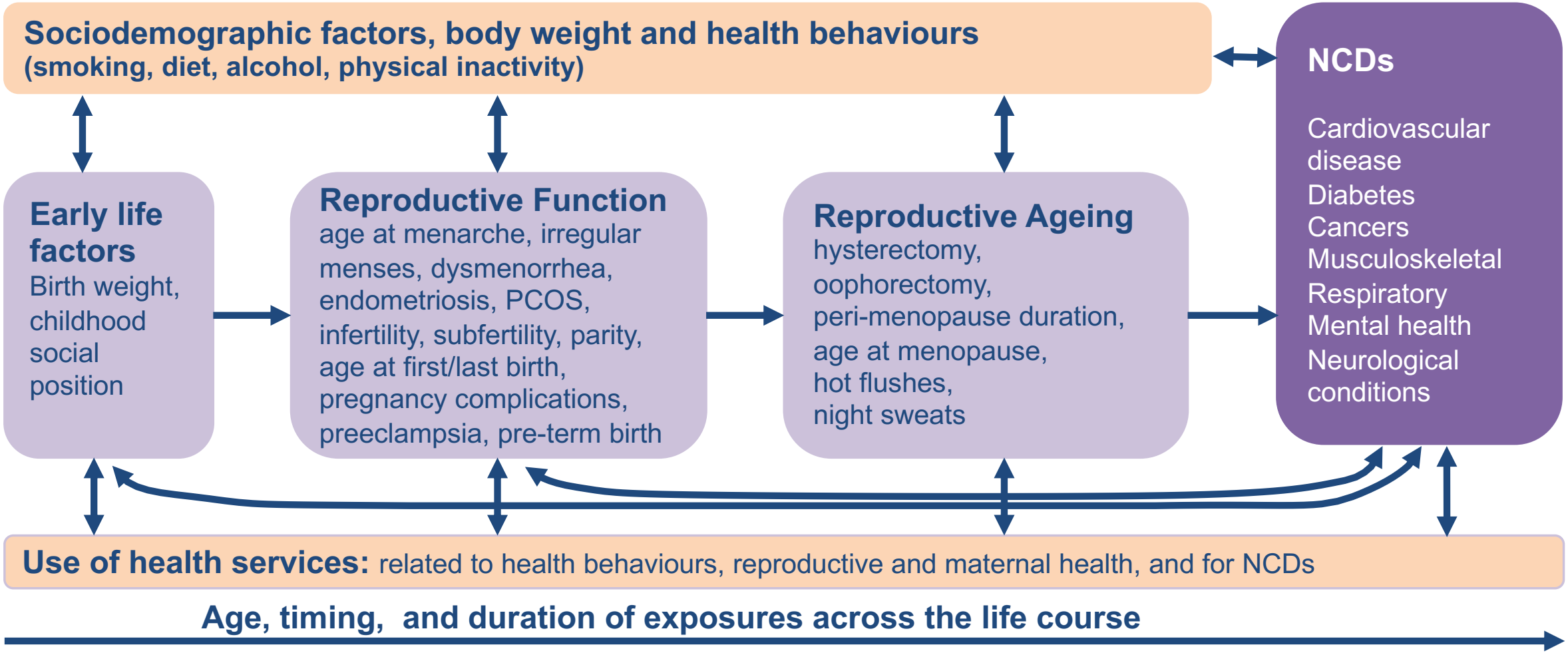
"…produce findings that not only can provide insights on the underlying biological processes or causal pathways operating from the earliest stages of life, but also have *implications for public health policy in terms of the optimal timing and targeting of preventive health strategies throughout life*."

Mishra, Kuh, Ben-Shlomo. The International Encyclopedia of Social and Behavioral Sciences, 2nd Edition (2015)

Life course epidemiology



Life course approach to non-communicable diseases



Adapted from Mishra, Anderson et al. Maturitas 2013; 74:235-40

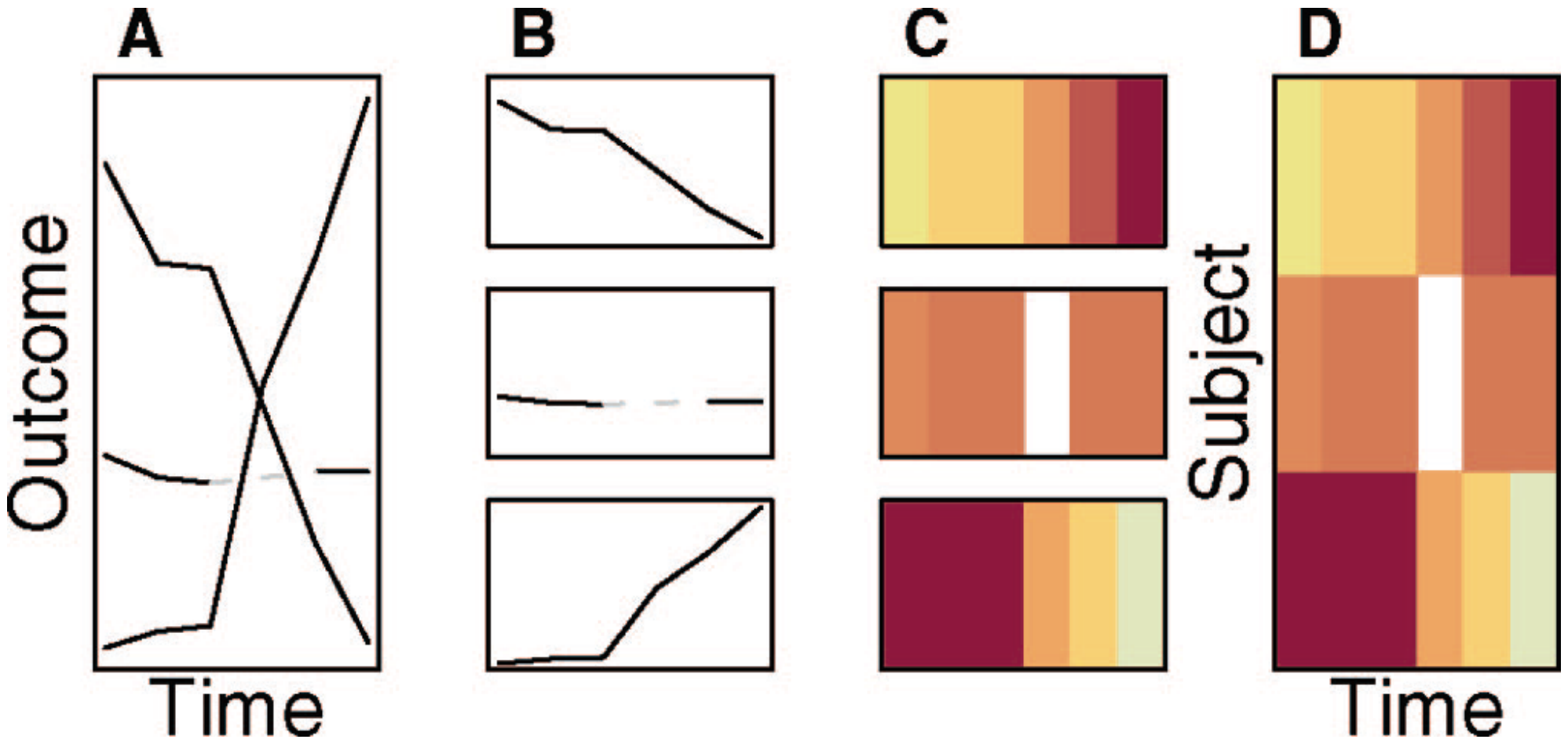
Longitudinal Data Visualisation

Step 1:

Need to have a clear sense of the longitudinal data to understand the relationship between data collection time points and exposures.

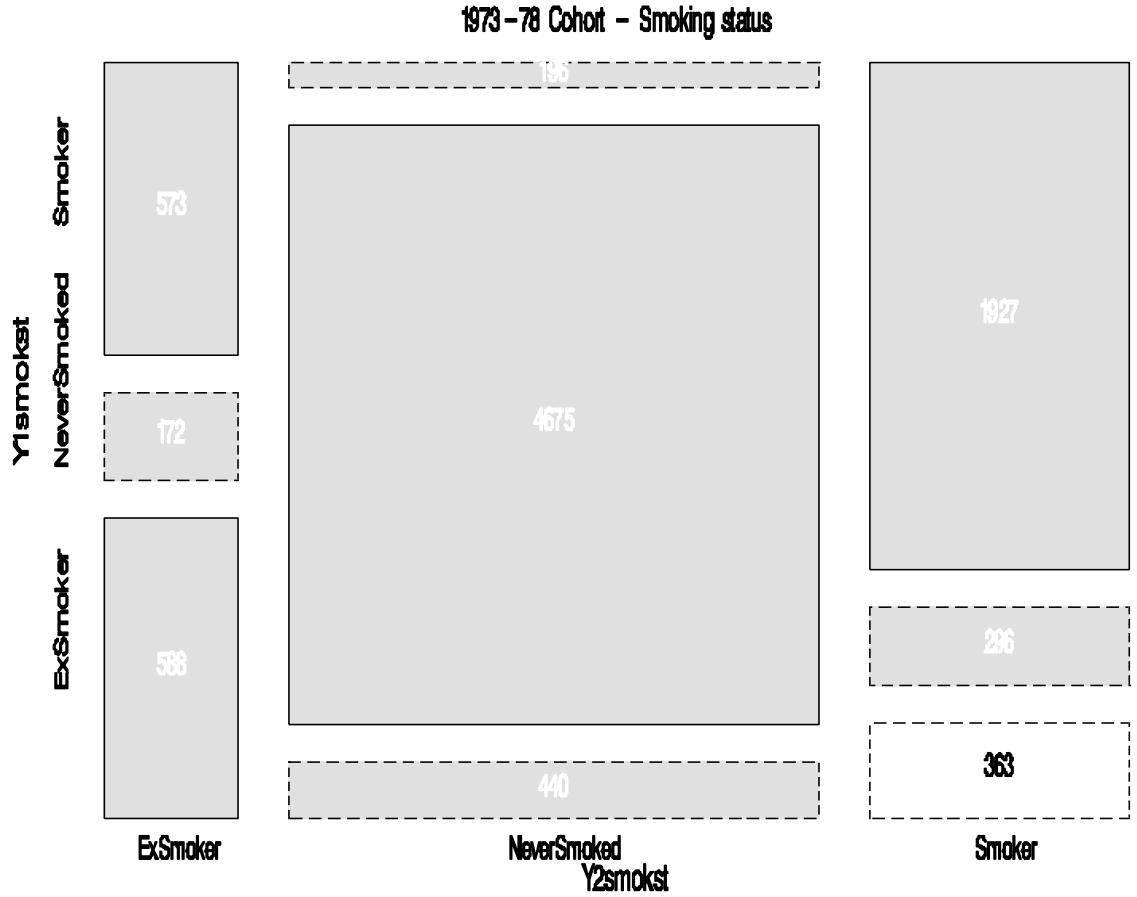
- current visualisation methods need improvement

Examples: Spaghetti and Lasagne plots



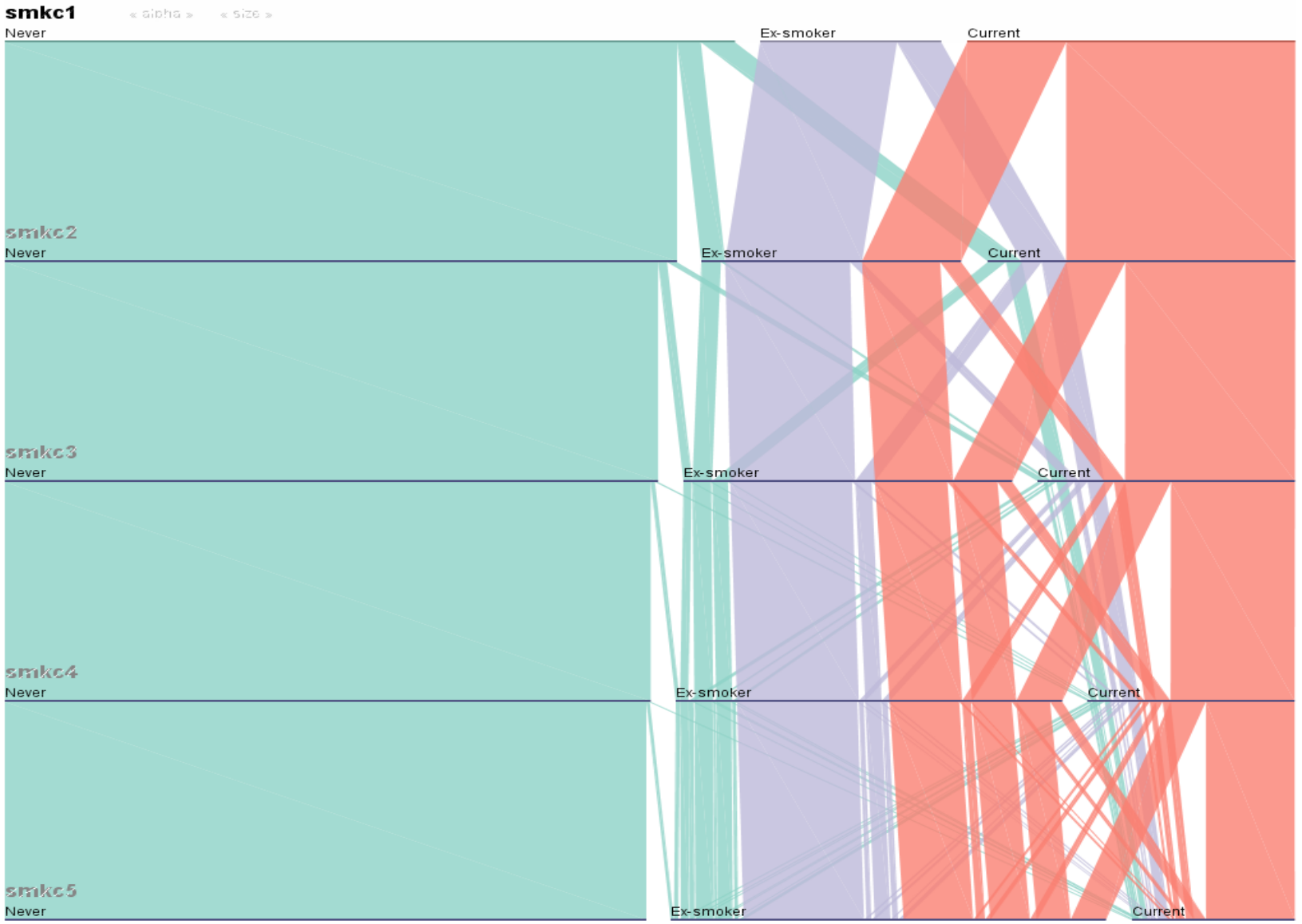
Swihart, B., et al. Lasagna Plots: A Saucy Alternative to Spaghetti Plots. Epidemiology, 2010, 21 (5), 621-625.

Examples: Mosaic plot



Friendly, M. Mosaic displays for multi-way contingency tables. Journal of the American Statistical Association, 1994, 89, 190-200.

Examples: Parallel sets



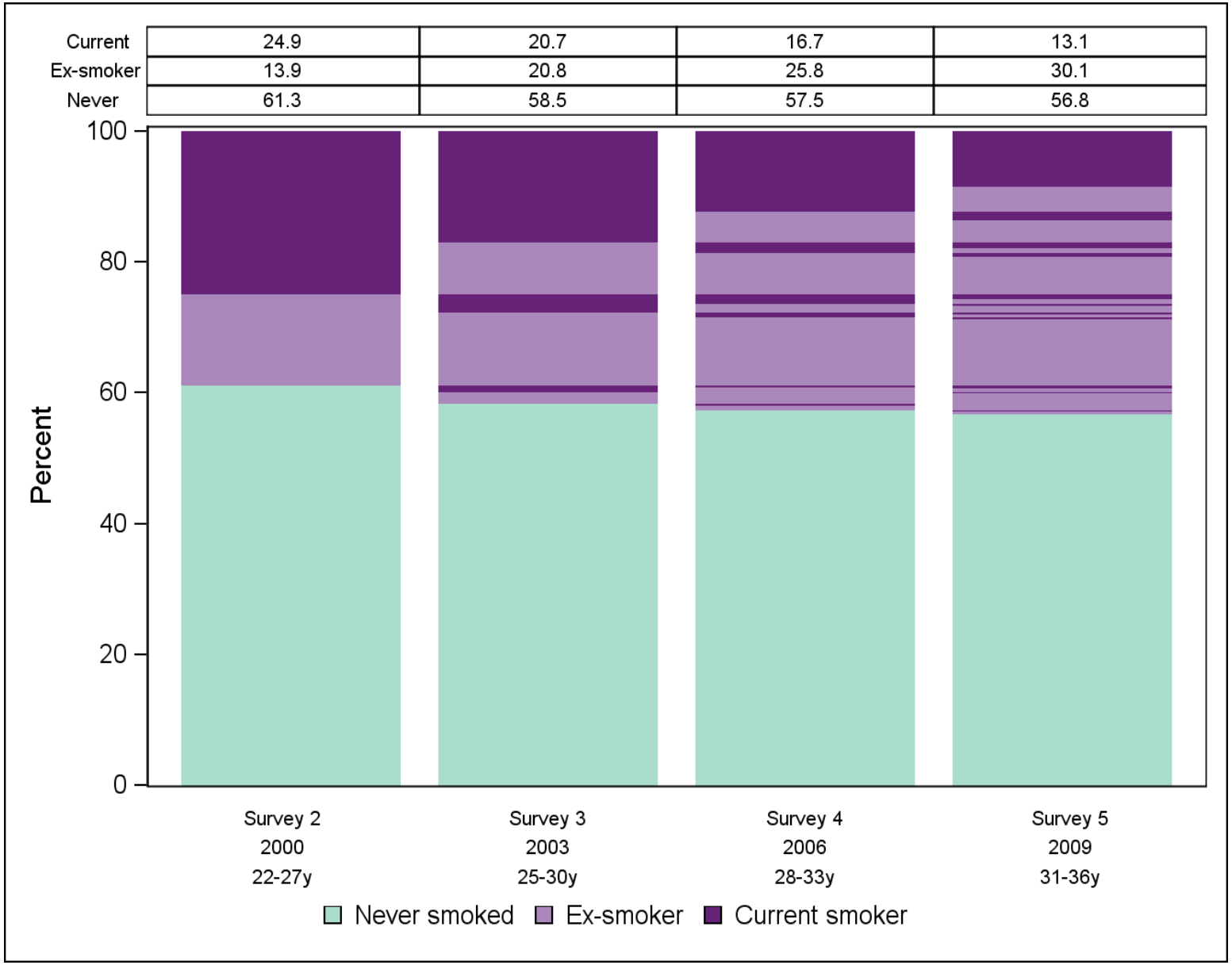
Alternative method of visualisation

A plot for visualising transitional distributions of categorical variables in longitudinal studies (similar to Sankey diagrams)

Jones M, Hockey R, Mishra GD, Dobson A. Visualising and modelling changes in categorical variables in longitudinal studies. BMC Med Res Methodol. 2014 Feb 27;14:32. doi: 10.1186/1471-2288-14-32.

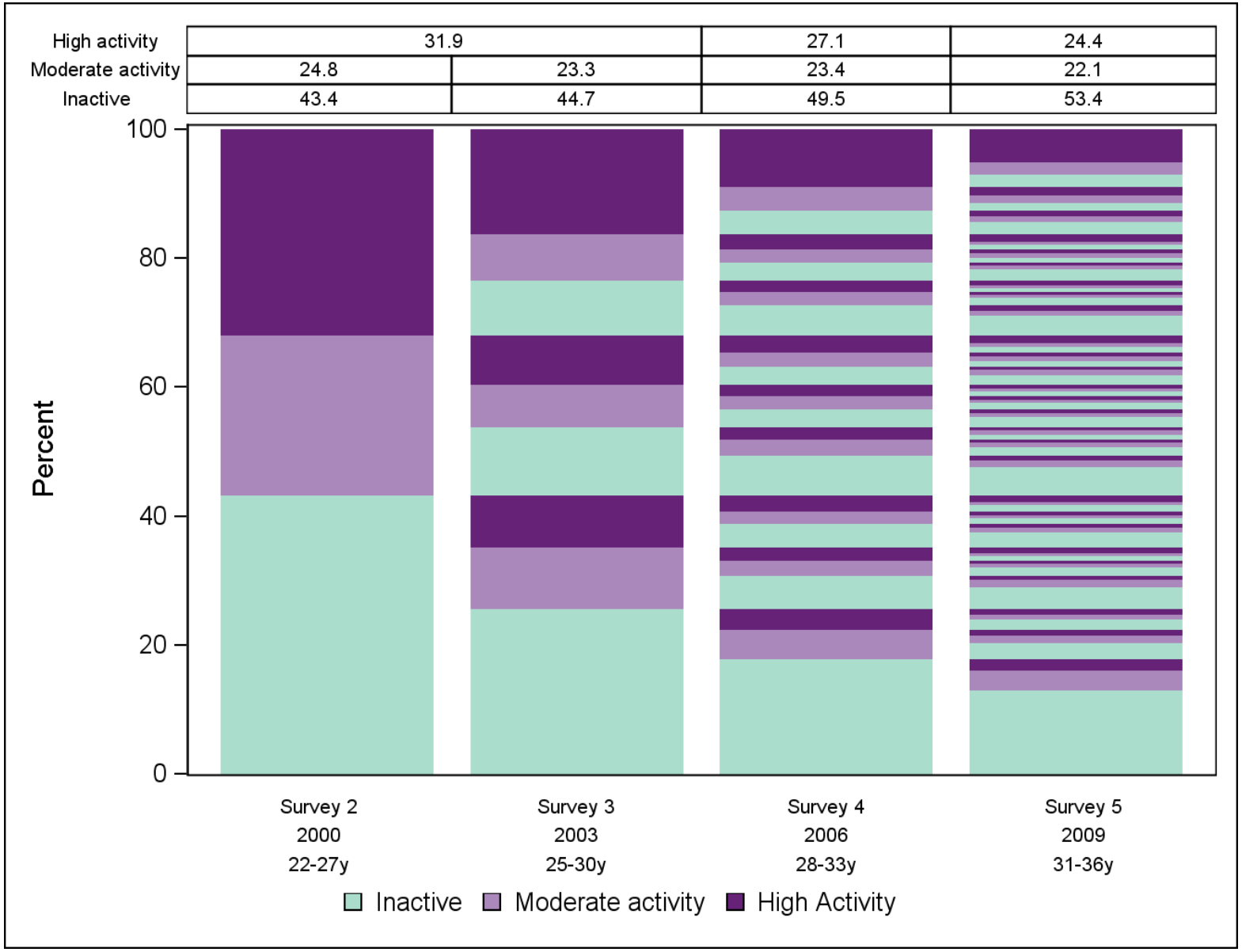
Smoking status

	Survey 2 2000 22-27y	Survey 3 2003 25-30y	Survey 4 2006 28-33y	Survey 5 2009 31-36y
Current	24.9	20.7	16.7	13.1
Ex-smoker	13.9	20.8	25.8	30.1
Never	61.3	58.5	57.5	56.8



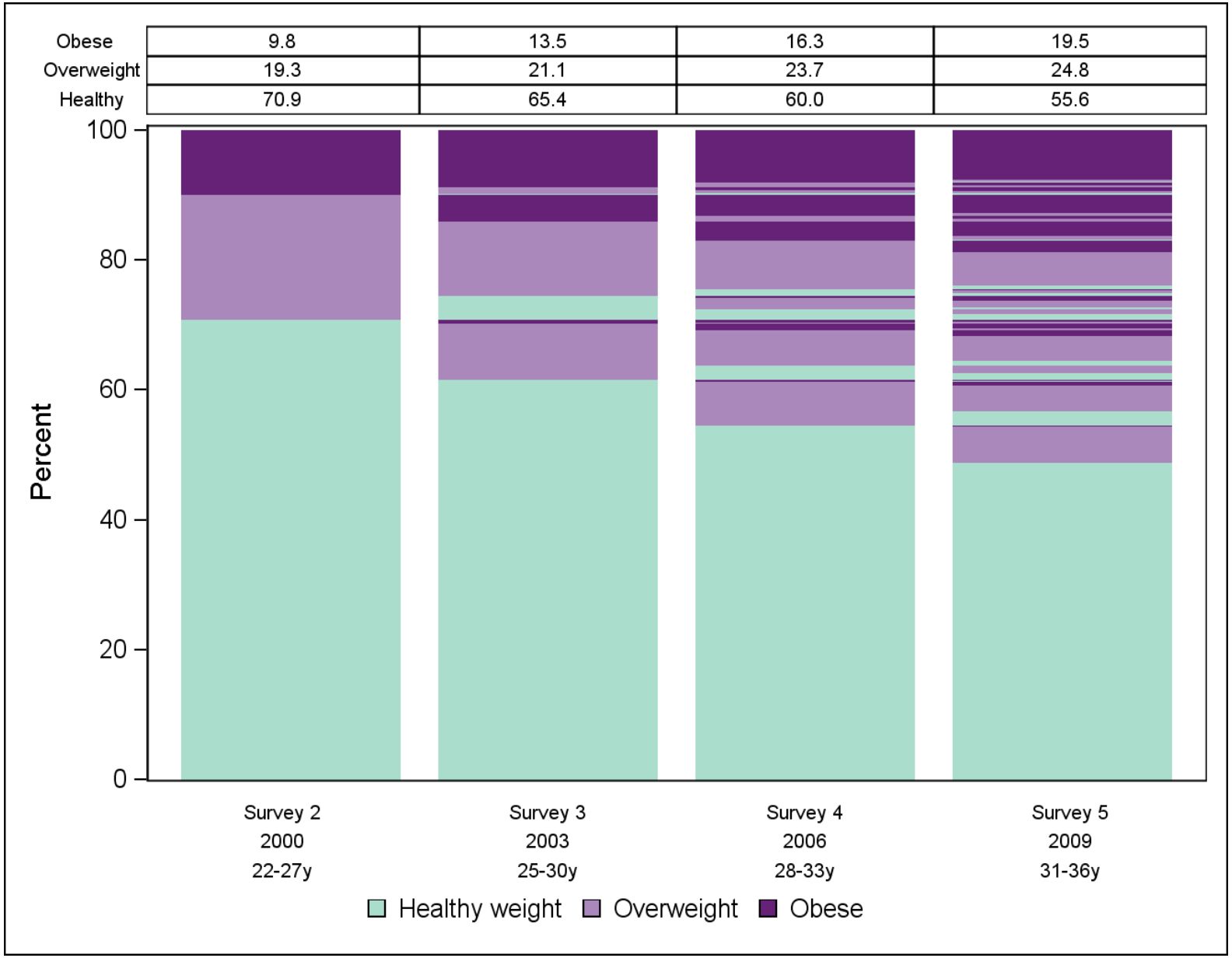
Physical activity level

High activity	31.9		27.1	24.4
Moderate activity	24.8	23.3	23.4	22.1
Inactive	43.4	44.7	49.5	53.4

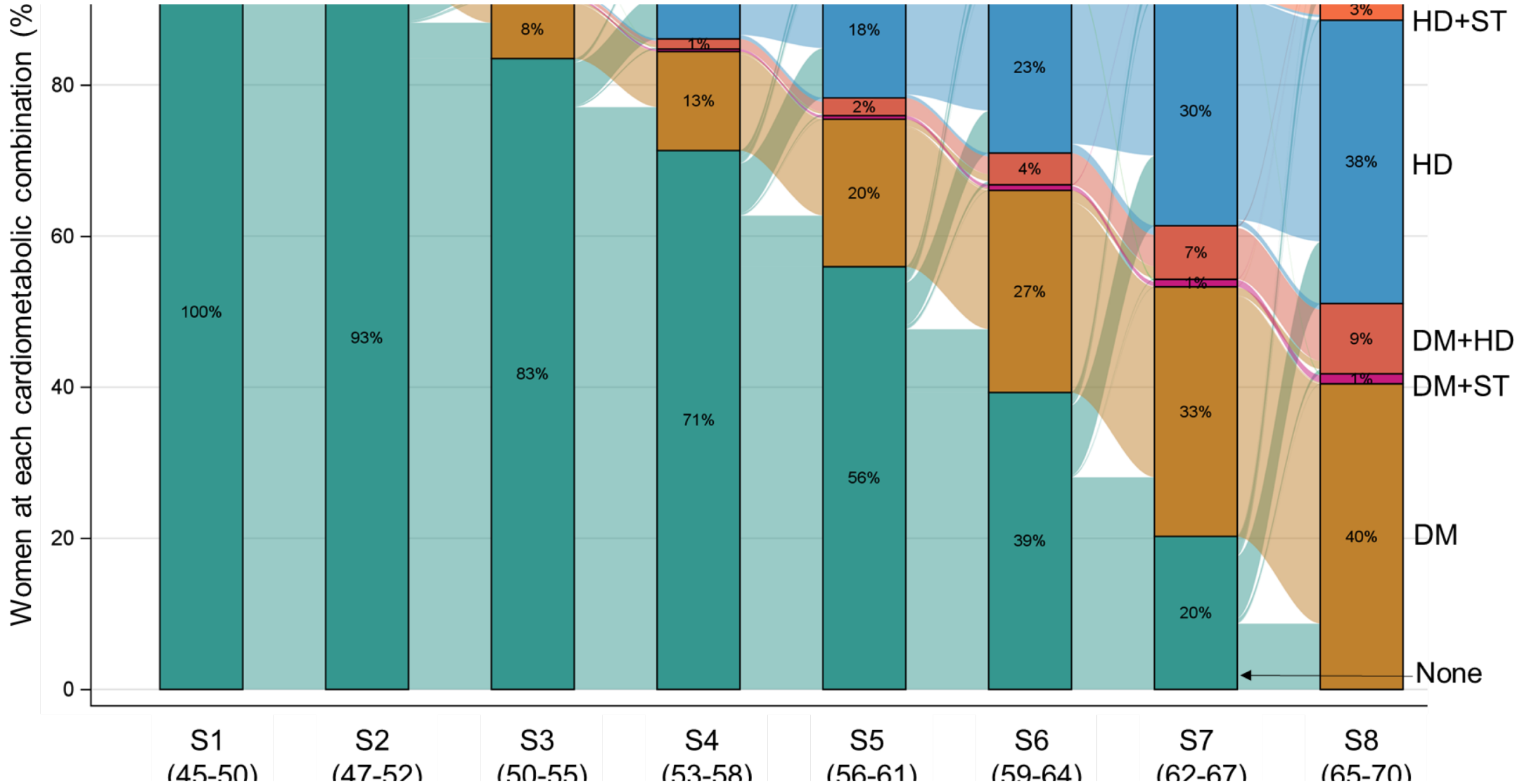


Body Mass Index group

	Survey 2 2000 22-27y	Survey 3 2003 25-30y	Survey 4 2006 28-33y	Survey 5 2009 31-36y
Obese	9.8	13.5	16.3	19.5
Overweight	19.3	21.1	23.7	24.8
Healthy	70.9	65.4	60.0	55.6



Longitudinal progression of multimorbidity over 20 years



Background

Plots are descriptive only

In addition may also want to:

- quantify marginal distribution trends over time
- estimate transitional probabilities
- determine how many previous survey measurements are needed to predict future levels
- estimate the overall level of predictability of future status based on previous status

Models to quantify marginal distribution trends over time

- Nominal logistic regression models
- Robust variances should be specified to take account of repeated measures on individuals across surveys
- Indicator variables for survey waves fitted to estimate changes in proportions of participants in each category of the outcome over time
- Goodness-of-fit of the models can be assessed by comparing estimated and observed marginal probabilities

Nominal logistic regression models

A type of regression model suitable for categorical outcomes that have no particular order, but can also be used for categorical outcomes that are ordered for j = 2, ..., J categories:

$$\text{logit}\,\pi_j = \log\left(\frac{\pi_j}{\pi_1}\right) = \mathbf{x}_j^{\mathrm{T}}\boldsymbol{\beta}_j$$

Marginal model (survey waves fitted as categorical variable)

Multinomial logistic regression — Number of obs = 35936
Wald chi2(6) = 935.60
Prob > chi2 = 0.0000
Log pseudolikelihood = -36742.974 — Pseudo R2 = 0.0100

(Std. Err. adjusted for 11818 clusters in idalias)

smokst	RRR	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
Never_smoked	(base outcome)					
Ex_smoker						
_Isurvey_3	1.24482	.0255677	10.66	0.000	1.195703	1.295954
_Isurvey_4	1.43581	.0318048	16.33	0.000	1.374808	1.499519
_Isurvey_5	1.618586	.0384846	20.25	0.000	1.544888	1.6958
Current_sm~r						
_Isurvey_3	.8815804	.0181198	-6.13	0.000	.8467722	.9178195
_Isurvey_4	.7069344	.0166591	-14.72	0.000	.6750257	.7403514
_Isurvey_5	.5158437	.0149064	-22.91	0.000	.4874396	.5459029

Marginal model (survey waves fitted as ordinal variable)

Multinomial logistic regression

Number of obs	= 35936
Wald chi2(2)	= 957.57
Prob > chi2	= 0.0000
Pseudo R2	= 0.0098

Log pseudolikelihood = -36749.32

(Std. Err. adjusted for 11818 clusters in idalias)

smokst	RRR	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
Never_smoked	(base outcome)					
Ex_smoker survey	1.172121	.0090314	20.61	0.000	1.154553	1.189957
Current_sm~r survey	.8082033	.0071725	-23.99	0.000	.7942671	.8223841

Marginal models

Physical activity (compared to being inactive):

- the relative risk of being in the moderate physical activity category decreased at each survey wave by 7% (RRR = 0.93; 95% CI: 0.90, 0.95)

- the relative risk of being in the high physical activity category decreased at each survey wave by 14% (RRR = 0.86; 95% CI: 0.85, 0.88)

BMI group (compared with being healthy or under-weight):

- the relative risk of being in a higher BMI category increased at each survey wave (RRR = 1.17; 95% CI: 1.15, 1.20 for overweight and RRR = 1.32; 95% CI: 1.29, 1.34 for obese).

Models to estimate transitional probabilities

- Used nominal regression models with independent variables that indicated the outcome status at previous surveys

- Used variance inflation factors and %change in log likelihood to guide model fitting

- Goodness-of-fit was assessed by comparing estimated and observed transition probabilities and McFadden's pseudo R^2

Variance inflation factor (VIF)

- Measures how much the variance of an estimated regression coefficient is increased because of collinearity
- Quantifies the severity of multicollinearity in a regression analysis
- Values >5 indicate multicollinearity.

Variable at survey wave 5	Previous survey	Two surveys previous	Three surveys previous
BMI group	1.4	2.2 to 2.6	2.0 to 2.9
Exercise group	1.1	1.1 to 1.2	1.1 to 1.3

Likelihood (and log likelihood)

- Used in regression modelling to obtain estimates of unknown parameters and assess how well a model fits the data

- The likelihood is equal to the probability of the observed outcomes given a particular set of parameter values

- The set of parameter values that maximise the likelihood are used to estimate the unknown parameter values

Changes in log likelihood

Variable at survey 5	Null model	Previous survey	Two surveys previous	Three surveys previous
BMI group	-8644.0	-4811.2 +44.3%	-4562.8 +2.9%	-4506.7 +0.6%
Exercise group	-8062.1	-7703.1 +4.5%	-7648.9 +0.6%	-7601.2 +0.6%

McFadden's pseudo R^2

- Based on the increase in (log) likelihood of the fitted model in comparison to the null model

$$R^2 = 1-[\ln L(M_{fitted})]/[\ln L(M_{null})]$$

  - If the likelihood of the fitted model is very high then R^2 is close to one
  - If the likelihood of the fitted model is very low then R^2 is close to zero

Pseudo R^2 values

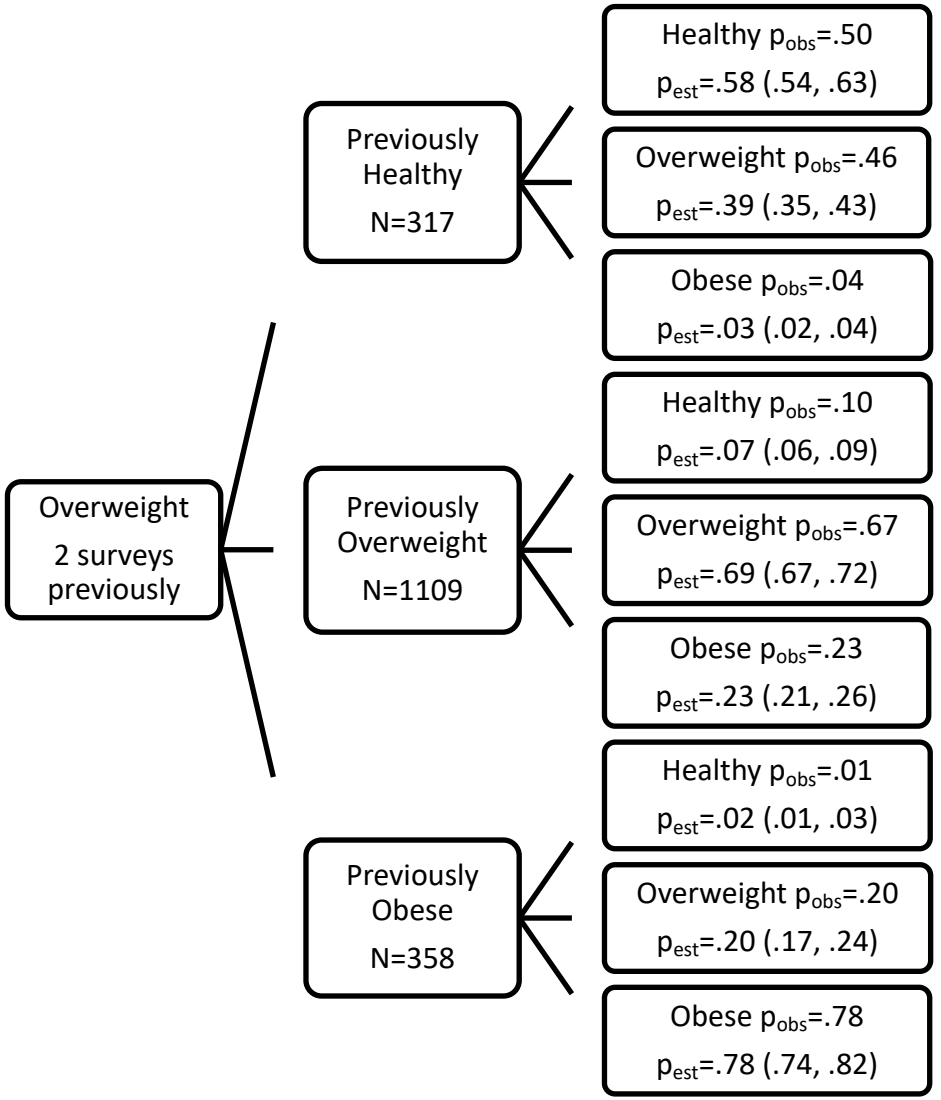
Variable	Previous surveys included in model	Pseudo R^2
BMI group	2	0.47
Physical activity level	1	0.045

Interpreting transitional probabilities

Based on ALSWH data and conditional on being overweight for two previous surveys, Australian women in their twenties:

- have 23% (95% CI: 21%, 26%) probability of being obese at the next survey

- but only 7% (95% CI: 6%, 9%) probability of being of healthy weight (or underweight)



Life Course Models

Step 2: take a more formal approach to identifying the potential relationships between exposures over time and outcomes.

Combine to construct one or more comprehensive life course models across the lifespan.

Illustrate the process here with three basic life course models.

Critical and sensitive periods

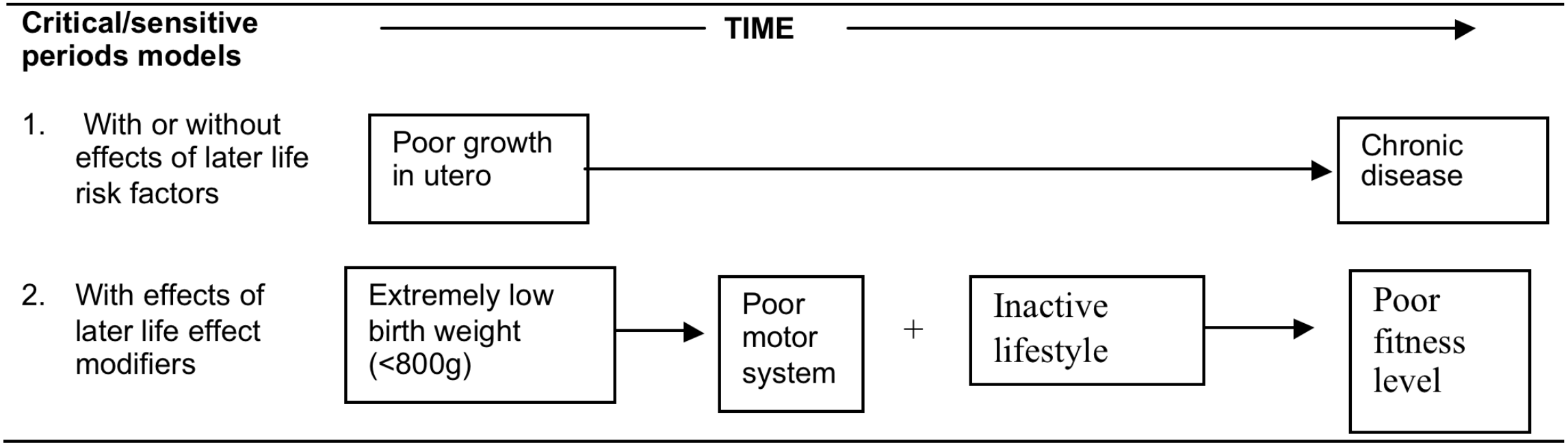
Critical period – when an exposure only has an effect during a specific time period

- Thalidomide exposure and limb abnormalities
- Specific viral infection during first trimester – schizophrenia

Sensitive period - when an exposure has greater effect during a specific time period than outside that period

- Learning a second language *in childhood*
- Exposure to air pollution during childhood (lung development) and incidence of asthma

Critical and sensitive periods



Accumulation of risk

Riley's concept of insult accumulation…

Life course exposures or insults gradually accumulate through episodes of illness and injury, adverse environmental conditions, and health damaging behaviour.

Kuh et al. JECH (2003)

Life course models: additional points

Life course models help to distinguish between variables acting through different pathways or mechanisms

- Other models include:
  - *recency (the most recent exposure is the most important)*
  - mobility (e.g. the effects of changes in social position on adult health)

The models force one to consider the timing (*critical, sensitive*), duration (*accumulation*), and temporal ordering of exposures

Basic models can be combined to form *a priori visual box diagrams* to illustrate potential aetiological pathways to be modelled and the need to distinguish confounders from intermediaries.

Structured Life Course Modelling Approach (SLCMA)

Step 3: Need a formal approach to select (determine) which life course model best fits the relationship between the longitudinal exposure data and the outcome.

METHODOLOGY

A structured approach to modelling the effects of binary exposure variables over the life course

Gita Mishra,[1*†] Dorothea Nitsch,[2†] Stephanie Black,[1] Bianca De Stavola,[2] Diana Kuh[1] and Rebecca Hardy[1]

IJE 2009 38(2):528-37

SLMCA: example on socio-economic position

Eur J Epidemiol
DOI 10.1007/s10654-013-9777-z

MORTALITY

Socio-economic position over the life course and all-cause, and circulatory diseases mortality at age 50–87 years: results from a Swedish birth cohort

Gita Devi Mishra · Flaminia Chiesa · Anna Goodman · Bianca De Stavola · Ilona Koupil

Received: 14 April 2012 / Accepted: 31 January 2013

Objectives

We test whether accumulation or critical period models of socio-economic position (SEP) provide the best fit to mortality data from a large Swedish birth cohort.

Uppsala Birth Cohort Study (UBCos)

Based on babies born in Uppsala University Hospital, Sweden during 1915-1929

Detailed information on course of pregnancy, conditions of newborns, basic socio-demographic information about the mothers, and occupation of fathers were obtained.

- *Koupil I, J Intern Med. 2007*

Exposure: SEP across the life course

Based on based on occupational social class (where A: advantage, D: disadvantage)

eg at birth –

11: unskilled manual, production----D

12: unskilled manual, service----D

21: skilled manual, production---D

22: skilled manual, services---D

55: house-daughters/sons-----D

….

56: higher non manual-----A

57: higher non manual, large company-----A

60: professionals----A

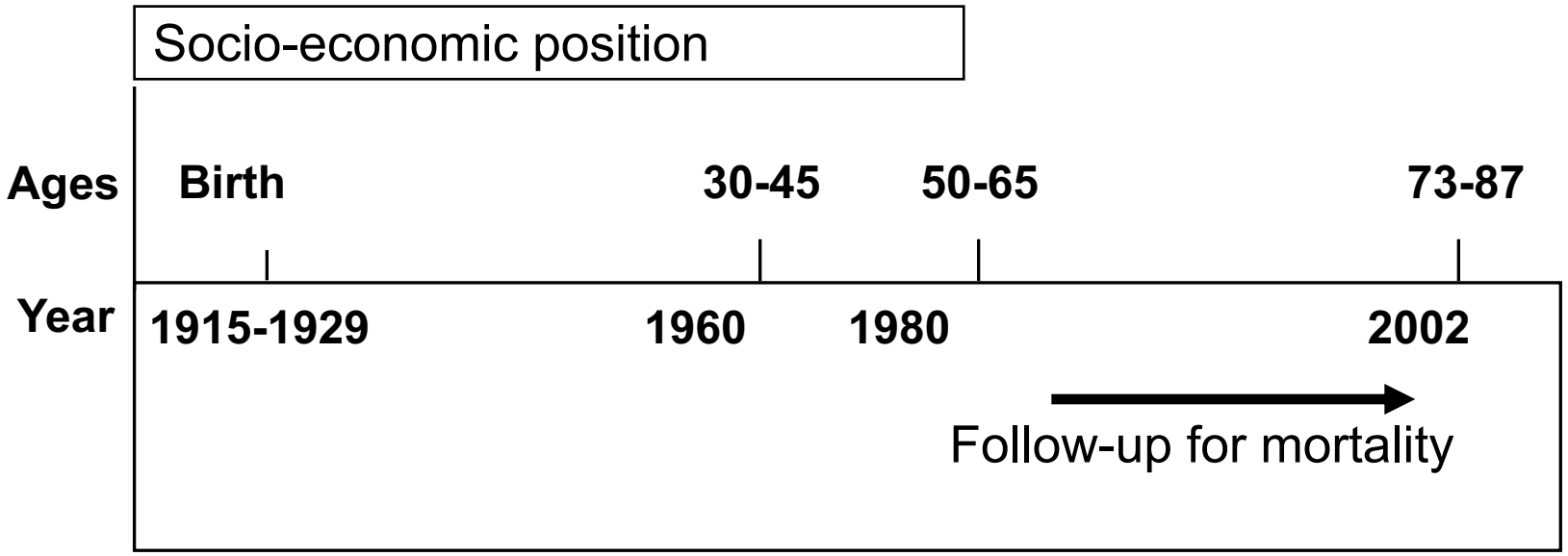
79: Self employed (not farmers)----A

89: farmer----A

98: not classified----missing

99: missing----missing

Relevant data (N= 10,207)



SEP indicators across life course

Binary SEP indicators at 3 time points:

S_1 (at birth)

S_2 (at age 30-45 y)

S_3 (at age 50-65 y)

$S_{1..3}$ = 0 (disadvantaged); $S_{1..3}$ =1 (advantaged)

Method

Step 1: We compared a series of nested models (representing either the accumulation or critical period models) with a fully saturated model.

Step 2: Model that is not significantly different to the fully saturated model is chosen to best described the data.

Cox proportional hazard regression models adjusted for birth year (cohort/period effect).

Models

Full model (Saturated)

$$\text{Log } h(t) = \alpha + \beta_1 S_1 + \beta_2 S_2 + \beta_3 S_3 + \theta_{12} S_1 S_2 + \theta_{23} S_2 S_3 + \theta_{13} S_1 S_3 + \theta_{123} S_1 S_2 S_3$$

Critical period model:

$$\text{Log } h(t) = a + \beta_1 S_1$$

constraints: $\beta_2 = \beta_3 = 0;\ \theta_{12} = \theta_{23} = \theta_{13} = \theta_{123} = 0$

Accumulation model: summed score:

$$\text{Log } h(t) = \alpha + \beta \Sigma_j S_j$$

constraints: $\beta_1 = \beta_2 = \beta_3; \theta_{12} = \theta_{23} = \theta_{13} = \theta_{123} = 0$

Accumulation model: mutually adjusted:

$$\text{Log } h(t) = \alpha + \beta_1 S_1 + \beta_2 S_2 + \beta_3 S_3$$

constraints: $\beta_1 \neq \beta_2 \neq \beta_3; \theta_{12} = \theta_{23} = \theta_{13} = \theta_{123} = 0$

where S_i are binary indicator of socio economic circumstances as time *i*

Hazard ratios (95% CI) for all-cause mortality in men (N=5138)

Critical period model:		LRT test with saturated model	
At birth	N		<0.001
Disadvantaged	3365	Reference	
Advantaged	1773	0.84 (0.77, 0.91)	
At age 30-45			<0.001
Disadvantaged	2874	Reference	
Advantaged	2264	0.77 (0.71, 0.83)	
At age 50-65			<0.001
Disadvantaged	3118	Reference	
Advantaged	2020	0.69 (0.63, 0.76)	
Accumulation model			**0.08**
0 (always disadvantaged)	1694	reference	
1	1507	0.87 (0.78, 0.95)	
2	1261	0.71 (0.64, 0.80)	
3 (always advantaged)	676	0.57 (0.50, 0.66)	

Hazard ratios (95% CI) for all-cause mortality in men

Hazard ratios (95% CI) for all-cause mortality in men from **accumulation model** **(mutually adjusted):**

LRT test with saturated model	**0.63**
At birth (advantaged vs disadvantaged)	0.89 (0.81, 0.97)
Age 30-45	0.90 (0.81, 0.98)
Age 50-65	0.74 (0.67, 0.82)

Hazard ratios (95% CI) for all-cause mortality in women (N=5069)

Critical period model:		LRT test with saturated model	
At birth	N		<0.001
Disadvantaged	3434	Reference	
Advantaged	1635	0.85 (0.76, 0.95)	
At age 30-45			<0.001
Disadvantaged	2059	Reference	
Advantaged	3010	0.88 (0.79, 0.97)	
At age 50-65			**0.10**
Disadvantaged	3646	Reference	
Advantaged	1423	071 (0.62, 0.80)	
Accumulation model			**0.07**
0 (always disadvantaged)	1319	Reference	
1	1924	0.91 (0.80, 1.02)	
2	1334	0.71 (0.62, 0.82)	
3 (always advantaged)	492	0.68 (0.55, 0.84)	

Hazard ratios (95% CI) for all-cause mortality from in women

Hazard ratios (95% CI) for all-cause mortality in women from **accumulation model** **(mutually adjusted):**

LRT test with saturated model	**0.49**
At birth (advantaged vs disadvantaged)	0.87 (0.78, 0.97)
Age 30-45	0.95 (0.86, 1.06)
Age 50-65	0.73 (0.64, 0.83)

Conclusion

SEP across the life course were associated with all-cause mortality in a cumulative manner for men and sensitive periods for women

SLMCA recent developments: additional methods

Alternative/additional methods of model selection or of evaluating life course models:

- Bayesian approach - assessing the relative importance of a series of life course theories together compared with the estimates of posterior probability of each life course theory or hypothesis.
  - Madathil S, Joseph L, Hardy R, et al. A Bayesian approach to investigate life course hypotheses involving continuous exposures. Int J Epidemiol 2018;47(5):1623-1635

- Least angle regression
  - Smith AD, Hardy R, Heron J, et al. A structured approach to hypotheses involving continuous exposures over the life course. Int J Epidemiol. 2016;45(4):1271–1279.
  - Smith AD, Heron J, Mishra G, et al. *Model Selection of the Effect of Binary Exposures over the Life Course. Epidemiology. 2015;26(5):719–726.*

Model Selection of the Effect of Binary Exposures over the Life Course

Andrew D. A. C. Smith,[a,b] *Jon Heron,*[a] *Gita Mishra,*[c] *Mark S. Gilthorpe,*[d] *Yoav Ben-Shlomo,*[a] *and Kate Tilling*[a,b]

Abstract

Epidemiologists are often interested in examining the effect on a later-life outcome of an exposure measured repeatedly over the life course. When different hypotheses for this effect are proposed by competing theories, it is important to identify those most supported by observed data as a first step toward estimating causal associations. **One method is to compare goodness-of-fit of hypothesized models with a saturated model, but it is unclear how to judge the "best" out of two hypothesized models that both pass criteria for a good fit**. **We developed a new method using the least absolute shrinkage and selection operator to identify which of a small set of hypothesized models explains most of the observed outcome variation.** We analyzed a cohort study with repeated measures of socioeconomic position (exposure) through childhood, early- and mid-adulthood, and body mass index (outcome) measured in mid-adulthood. We confirmed previous findings regarding support or lack of support for the following hypotheses: accumulation (number of times exposed), three critical periods (only exposure in childhood, early- or mid-adulthood), and social mobility (transition from low to high socioeconomic position). Simulations showed that our least absolute shrinkage and selection operator approach identified the most suitable hypothesized model with high probability in moderately sized samples, but with lower probability for hypotheses involving change in exposure or highly correlated exposures. Identifying a single, simple hypothesis that represents the specified knowledge of the life course association allows more precise definition of the causal effect of interest.

Epidemiology. 2015;26(5):719–726

SLMCA recent developments: research areas

SLMCA methods increasingly applied across a range of areas to investigate time dependent exposures and provide insights of mechanisms, including:

- Childhood adversity, physical activity, BMI, socioeconomic position on psychological, metabolic, and disease outcomes:
  - Cooper R, Mishra GD, Kuh D. Physical Activity Across Adulthood and Physical Performance in Midlife: Findings from a British Birth Cohort. American Journal of Preventive Medicine. 2011;41(4):376–384.
  - Evans J, Melotti R, Heron J, et al. The timing of maternal depressive symptoms and child cognitive development: a longitudinal study. Journal of Child Psychology and Psychiatry. 2012;53(6):632–640.
  - Wills AK, Black S, Cooper R, et al. Life course body mass index and risk of knee osteoarthritis at the age of 53 years: evidence from the 1946 British birth cohort study. Annals of the Rheumatic Diseases. 2012;71(5):655–660.
  - Nicolau B, Madathil SA, Castonguay G, et al. Shared social mechanisms underlying the risk of nine cancers: A life course study. International Journal of Cancer. 2019;144(1):59–67.

- Also in relation to –omics, e.g. Childhood Adversity on DNA Methylation (DNAm)
  - Dunn EC, Soare TW, Zhu Y, et al. Sensitive Periods for the Effect of Childhood Adversity on DNA Methylation: Results From a Prospective, Longitudinal Study. Biological Psychiatry. 2019;85(10):838–849.

Mental health: how does it affect women and children?

Aim: use SLCMA to investigate the relationship between maternal depression around pregnancy (before, during, and post-natal) with subsequent child development outcomes.

Life course approach: conclusions

SLCMA is a useful additional tool for life course epidemiology

- should be used alongside other methods, including visualisation

By identifying the most appropriate model, SLCMA provides insights:

- On potential biological mechanisms at work
- For development and targeting of preventive health strategies

Considerable potential for further research and development:

- Especially with big data and –omics