Check what is missing with initial data analysis

Lara Lusa[1,2] , Katherine J. Lee[3], Carsten Oliver Schmidt[4] and Marianne Huebner[5]

[1]University of Primorska, Koper/Capodistria, Slovenia

[2]University of Ljubljana, Ljubljana, Slovenia

[3]Clinical Epidemiology and Biostatistics Unit, Murdoch Children’s Research Institute, Melbourne, Australia

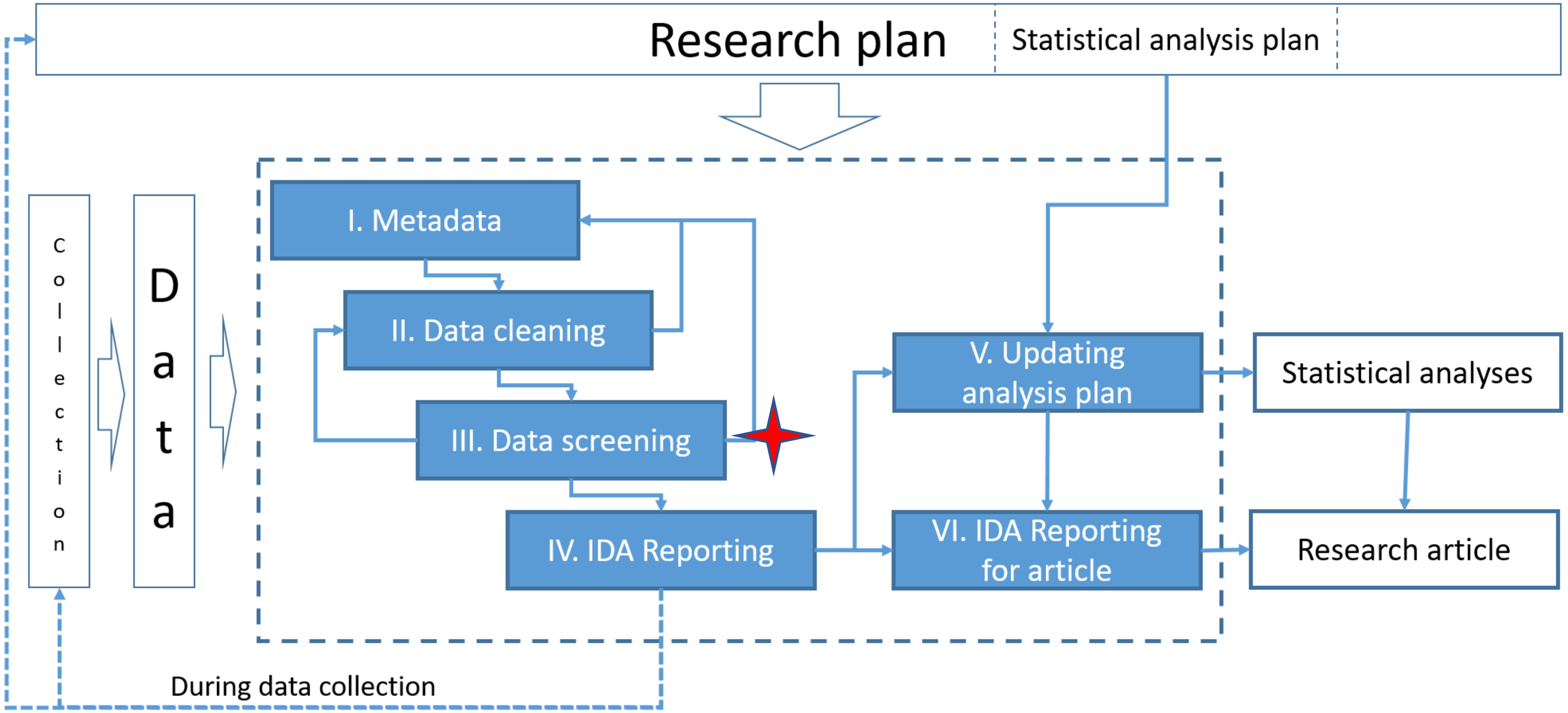
[4]Institute for Community Medicine, SHIP-KEF University Medicine of Greifswald, Greifswald, Germany

[5]Michigan State University, East Lansing, MI, USA

on behalf of the Topic Group (TG3) “Initial Data Analysis” of the STRATOS Initiative (STRengthening Analytical Thinking for Observational Studies).

Initial Data Analysis

- AIM: provides analysis-ready data set including reliable information about data properties to answer the research question



III. Data screening

Aim: understand data properties

Missing data ✦

Univariate distributions

Multivariate distributions

Possible consequence: changes in analyses and interpretations

Huebner M, le Cessie S, Schmidt CO, Vach W. A contemporary conceptual framework for initial data analysis. Observational Studies 2018; 4: 171-192. https://doi.org/10.1353/obs.2018.0014

Reporting on missingness is (still) incomplete

Missing values are present in reporting checklists. STROBE* example:

12. Statistical methods

    (c) explain how missing values were addressed

13. Participants – unit missingness

    (a) Report number of individuals at each stage (from eligible to analyzed)

    (b) Give reasons for non-participation at each stage

    (c) Consider use of a flow diagram

14. Descriptive data – item missingness

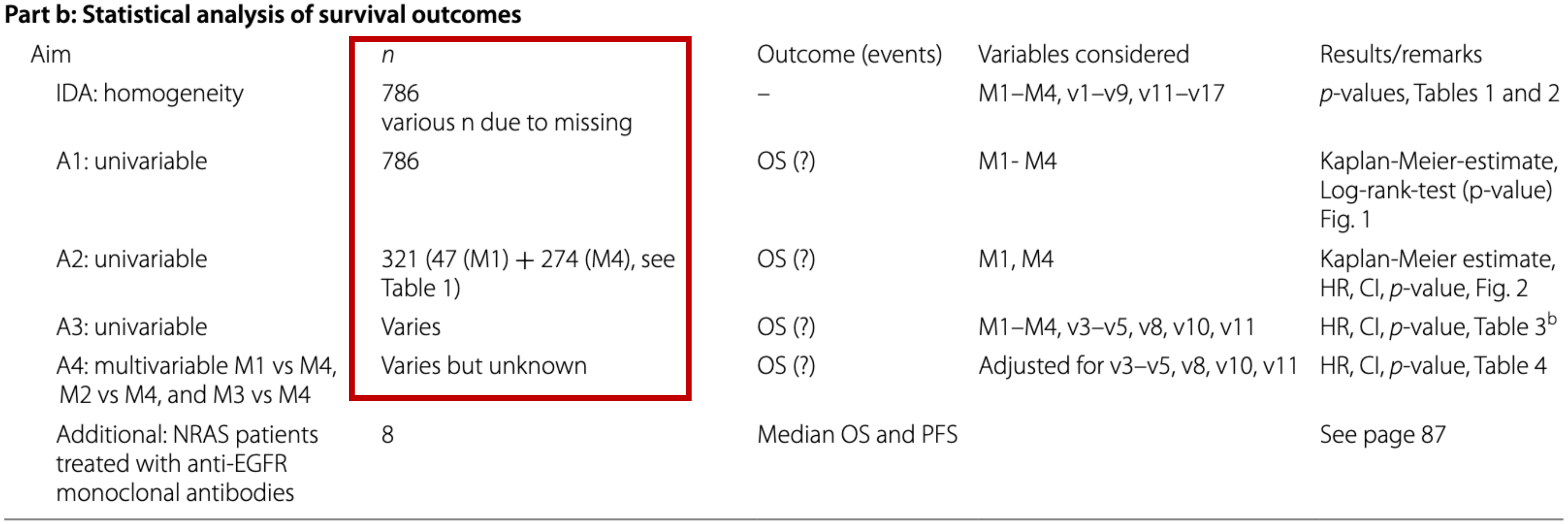
    (b) Indicate number of participants with missing data for each variable of interest

Missing data statements, review of 25 papers in top clinical journals, Huebner et al. 2020	Mentioned in papers, n (%)
Item missingness (in exploratory variables)	19/25 (76%)
Missing values for outcome variables	12/25 (48%)
Unit missingness (participants)	15/25 (60%)
Changes in analysis plan due to missing data	5/25 (20%)

*** Strengthening the reporting of observational studies in epidemiology**

Reporting on missingness is (still) incomplete

Sample sizes for models are insufficiently reported!



Part b: Statistical analysis of survival outcomes

Aim	*n*	Outcome (events)	Variables considered	Results/remarks
IDA: homogeneity	786 various n due to missing	–	M1–M4, v1–v9, v11–v17	*p*-values, Tables 1 and 2
A1: univariable	786	OS (?)	M1- M4	Kaplan-Meier-estimate, Log-rank-test (p-value) Fig. 1
A2: univariable	321 (47 (M1) + 274 (M4), see Table 1)	OS (?)	M1, M4	Kaplan-Meier estimate, HR, CI, *p*-value, Fig. 2
A3: univariable	Varies	OS (?)	M1–M4, v3–v5, v8, v10, v11	HR, CI, *p*-value, Table 3[b]
A4: multivariable M1 vs M4, M2 vs M4, and M3 vs M4	Varies but unknown	OS (?)	Adjusted for v3–v5, v8, v10, v11	HR, CI, *p*-value, Table 4
Additional: NRAS patients treated with anti-EGFR monoclonal antibodies	8	Median OS and PFS		See page 87

W Sauerbrei, T Haeussler, M Huebner. BMC Medicine (2022) 20:184; Structured reporting to improve transparency of analyses in prognostic marker studies

Literature review of biomarker studies published in top Cancer journals: Breast Cancer Research and Treatment, European Journal of Cancer, International Journal of Cancer, Journal of Clinical Oncology

Why explore the missing data with IDA?

- Missing data need to be described when reporting the analysis (STROBE, TARMOS framework, Lee et al. TG1,...)

- The exploration of missing data provides a deeper understanding of the data

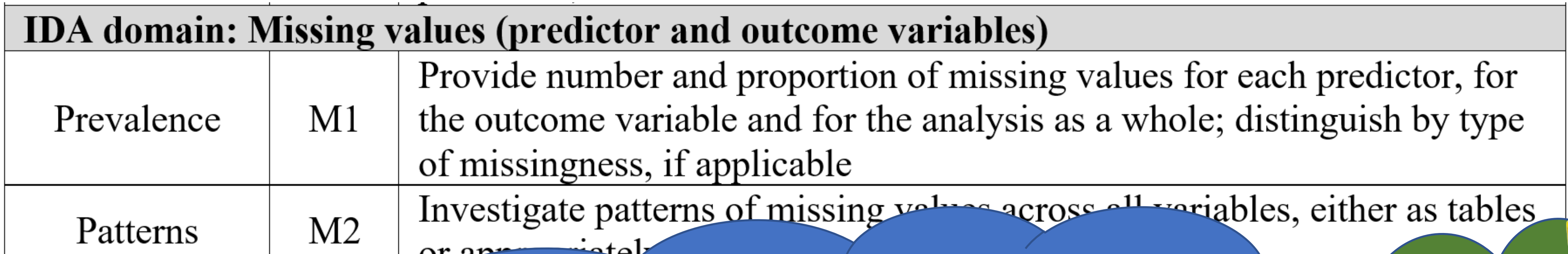
- The statistical analysis plan (SAP) should specify how to handle missing values in the analyses (TARMOS framework)
  - Some choices might depend on the missing data characteristics observed in the data (complete case vs multiple imputation? Sensitivity analyses?)

- The findings can be used when addressing the research question with statistical modeling, as its validity should be assessed.
  - Choice of modeling strategy conditional on missing data characteristics, if specified in the SAP
  - ⚠ Some models rely on assumptions about the missing value mechanism that might not be valid and can (sometimes) be explored
    - not always possible based on data exploration (MNAR vs MAR)
  - Auxillary variables that can be used in MI can be identified
  - Sensitivity analyses might be suggested

What aspects should be addressed in IDA?

- Which variables are observed and which are missing (missing data patterns)?
  - Number (%) of missing per variable (item missingness)
  - Co-occurrence of missing values in variables
  - Description of subjects that did not participate/respond (unit missingness) or interrupted participation (attrition in longitudinal studies)
  - If relevant, distinguish by type of missingness (by design, in longitudinal studies: lost to follow-up, intermittent, death, administrative censoring)
- Is there a (possible) relation between missing data and the values of the variables (missing data mechanisms)
  - Comparison of participants with complete/incomplete data
  - Assessment of which variables that can predict missingness

Missing values in IDA checklists: regression*



IDA domain: Missing values (predictor and outcome variables)		
Prevalence	M1	Provide number and proportion of missing values for each predictor, for the outcome variable and for the analysis as a whole; distinguish by type of missingness, if applicable
Patterns	M2	Investigate patterns of missing values across all variables, either as tables or appropriately
	ME1	

How much data would be available to fit models with:

Key predictors only?	93.9%
Key predictors + additional variables?	63.9%
All predictors?	27.1%

*TG2/TG3 Project: IDA for regression led by Georg Heinze (TG2) and Marianne Huebner (TG3)
Extended for longitudinal studies – ongoing TG3 project in collaboration with Kate Lee (TG1)

Longitudinal data example

SHARE
SURVEY OF HEALTH, AGEING
AND RETIREMENT IN EUROPE
SLOVENIJA

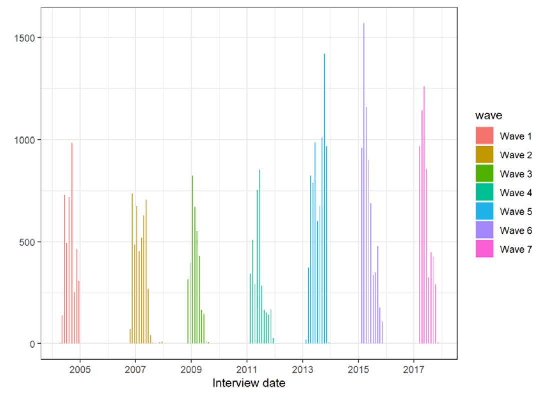
Data on health and socioeconomic variables of non-institutionalized individuals **aged 50 and older** across **27 European countries and Israel**. 140 000 participants, collected in years 2004 to 2018 in 7 waves. Thousands of questions about demographics, health and socio-economic status. Publicly available to researchers.

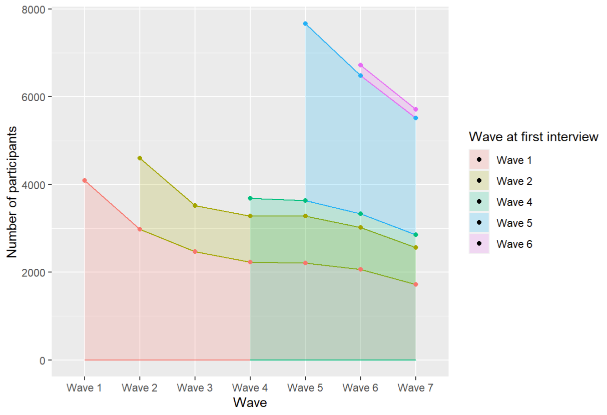
Subset: Denmark from 2004 to 2018 (7 waves)
Aim: Investigating age-associated change in max grip strength stratified by sex
Outcome: maximum grip strength
Covariates measured at first interview: sex, height
Time-varying covariates: age, weight, physical activity (vigorous or low intensity), smoking status
Population characteristics: education level, depression and other comorbidities (cancer, stroke, heart attack, lung disease, cancer)
Sampling design: simple random sampling with refreshment samples

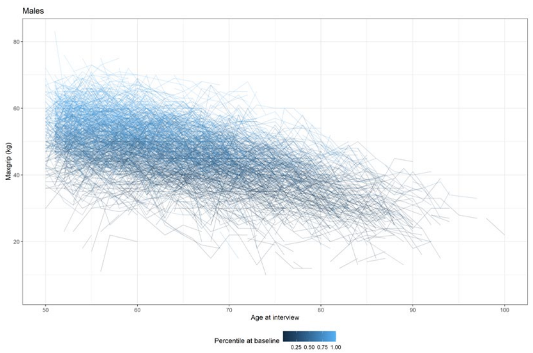
- The complexity of the data makes it a very interesting and difficult example, far from the “toy-data” often used in teaching or as examples in methodological papers.

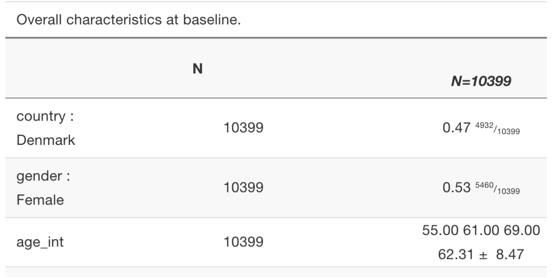
The complete IDA, based on reproducible R code, will be made fully publicly available, can be used by others as a template
A lot of attention is given to graphical displays and effective summaries

Distribution of the dates where the interviews were carried out, stratified by wave.



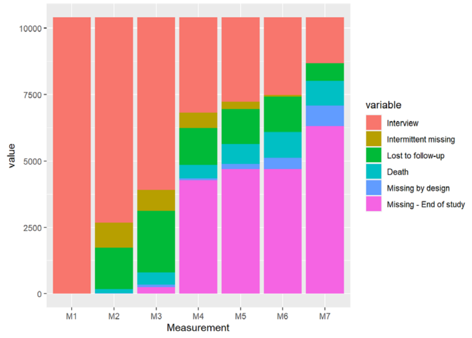




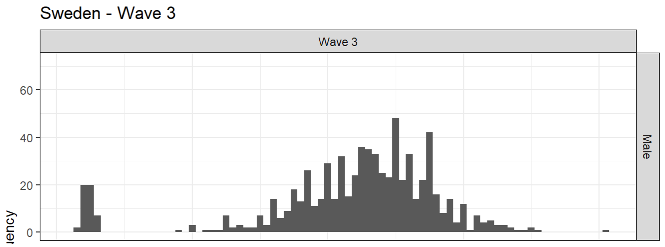


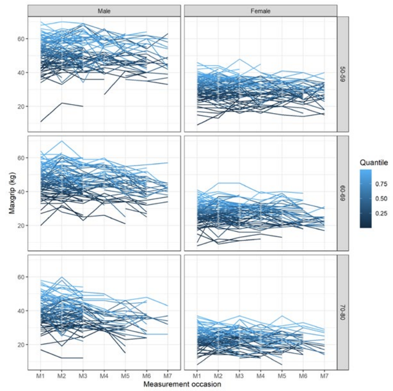
Overall characteristics at baseline.

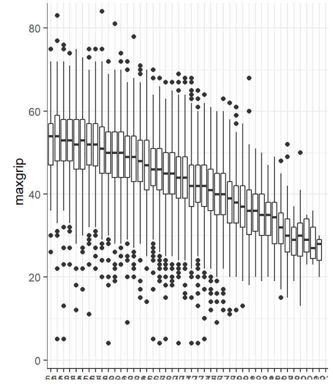
	N	N=10399
country : Denmark	10399	0.47 $^{4932}/_{10399}$
gender : Female	10399	0.53 $^{5460}/_{10399}$
age_int	10399	55.00 61.00 69.00 62.31 ± 8.47

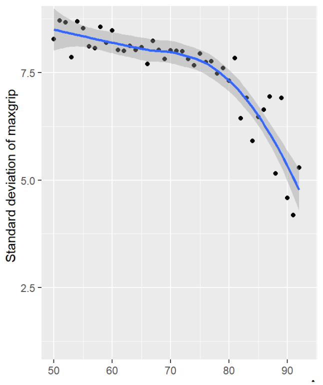


Sweden - Wave 3

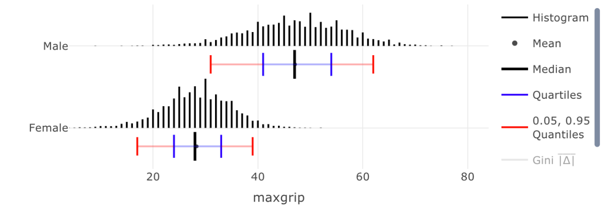




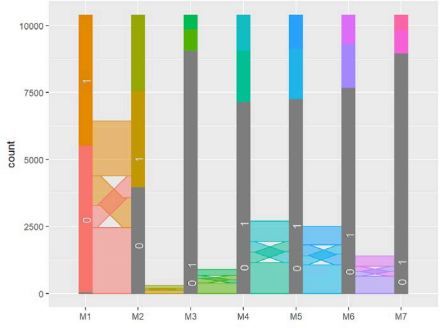


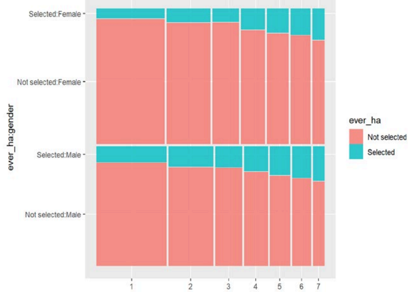


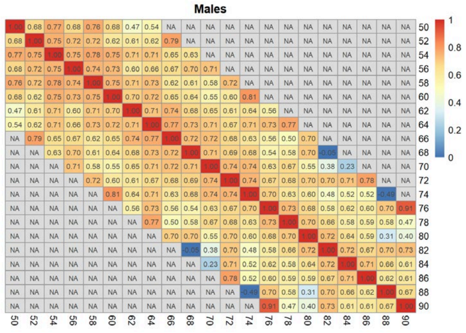
5.4.2.1.1 Grip strength



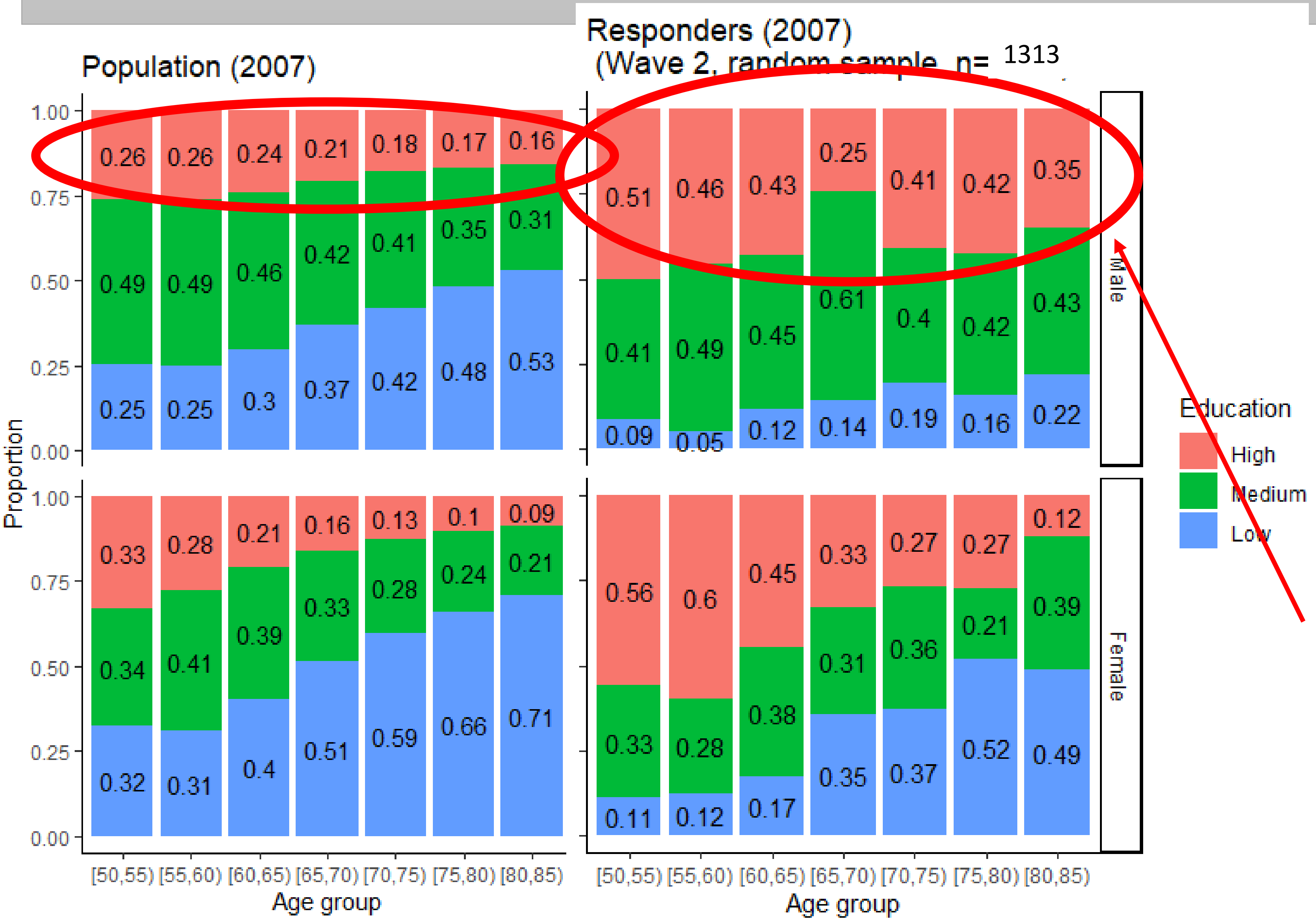
Vigorous physical activity across measurement occasions







Unit missingness due to non-response



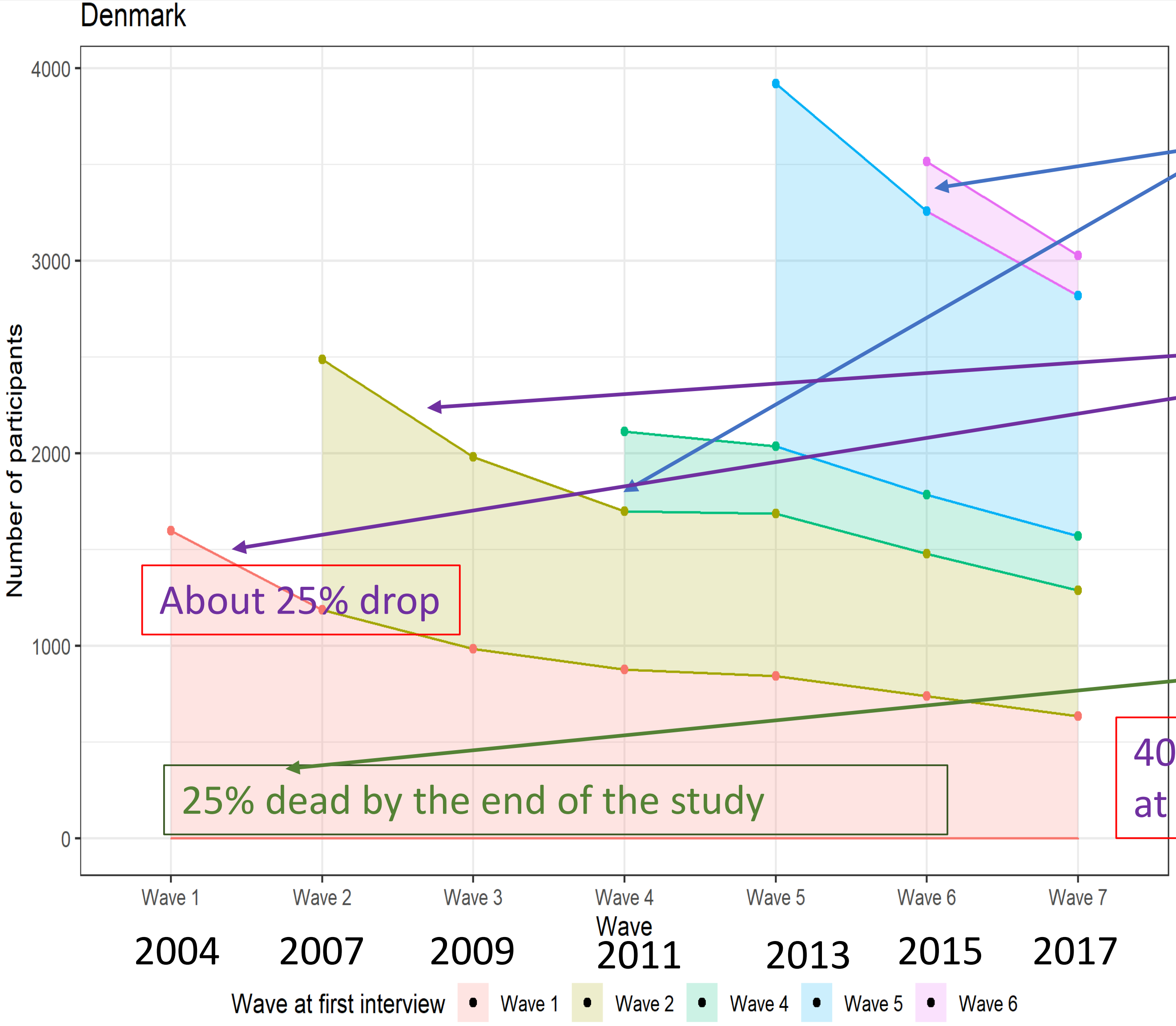
Cumbersome exploration due to non-availability of data about “complete” non-responders (some analyses can be based on EUROSTAT data)

Age/sex distribution is similar, younger men are somehow less likely to respond.

Responders have substantially higher education than expected (vs population data)

The available calibration weights for non-response are based only on age and sex -> not likely be helpful

Time frame and participation (conditional on entering the study)



2004 2007 2009 2011 2013 2015 2017

About 25% drop

25% dead by the end of the study

Refreshment samples are used

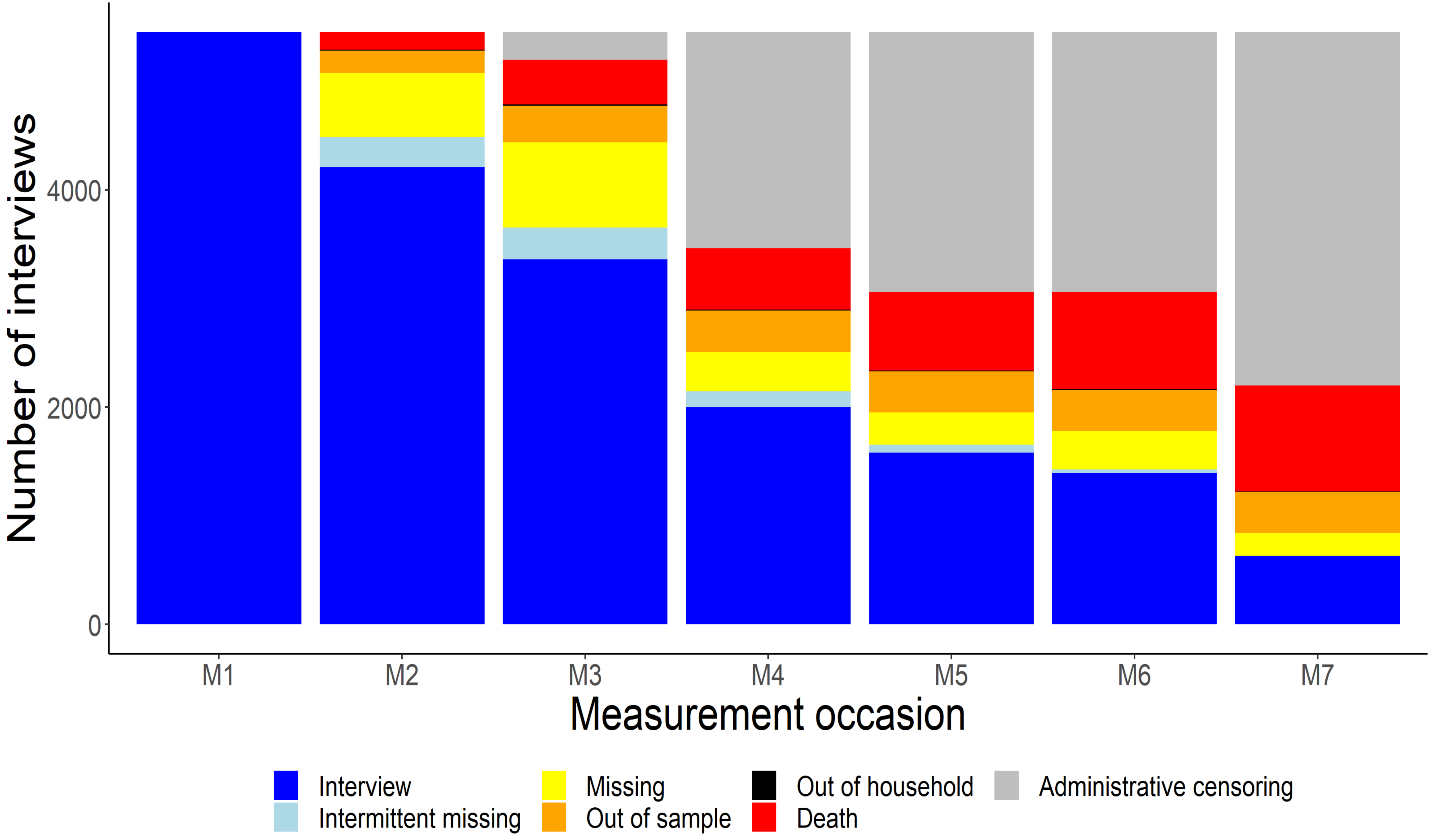
Small refreshment samples: only younger cohort

Substantial attrition is observed during the study, especially between first and second interview

Ageing population (50+) -> missing values can be due to deaths

40% with available interview at last follow-up

Unit missingness by type



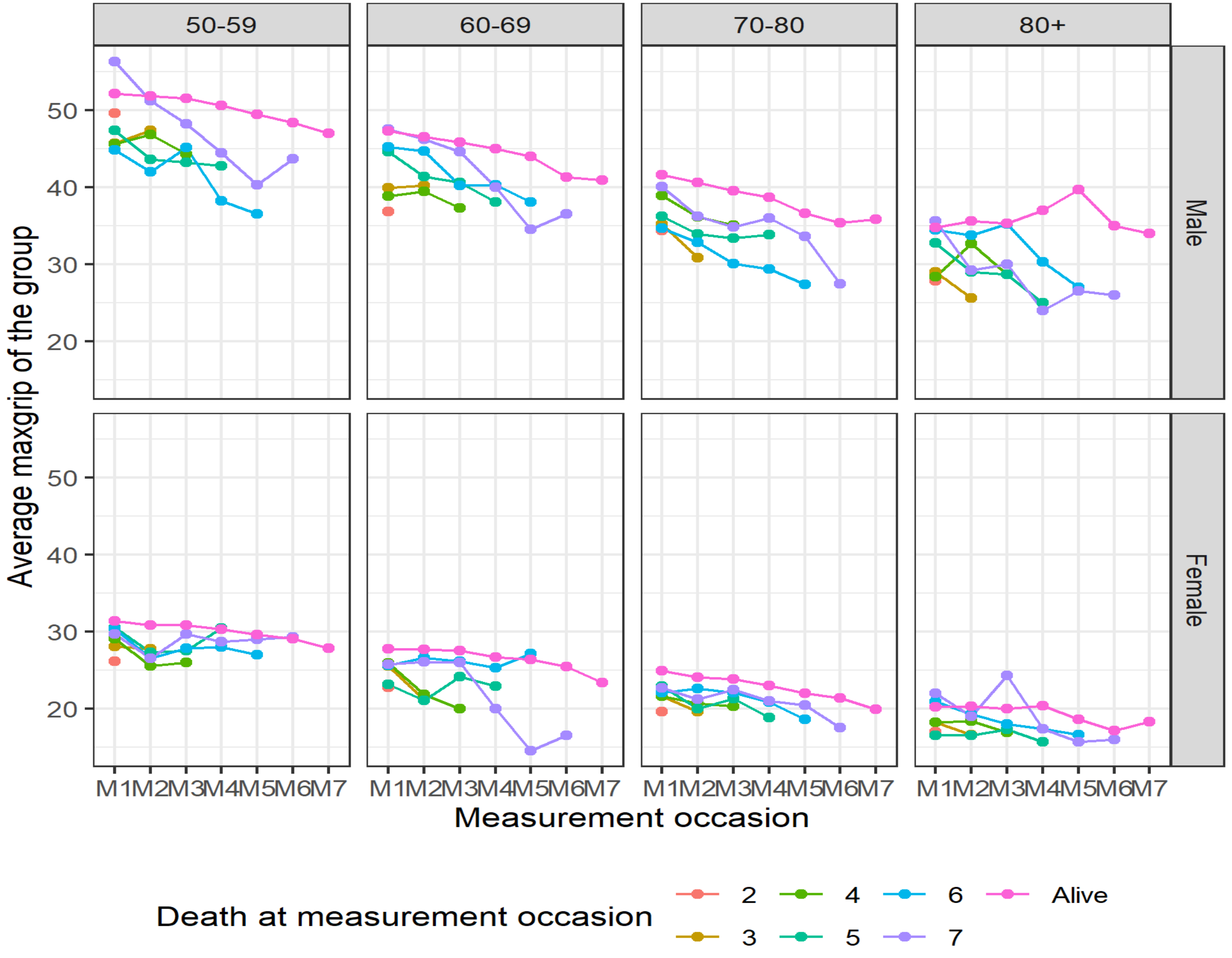
Median number of measurements: 3

Some participants have intermittent missingness,

Participants that die are on average: older, more frequently males, smokers, with less frequent vigorous physical activity -> as expected

Participants lost to follow-up have on average lower education and less healthy habits than complete responders (but are similar in terms of age and sex – investigated with descriptive statistics and multivariable logistic regression models)

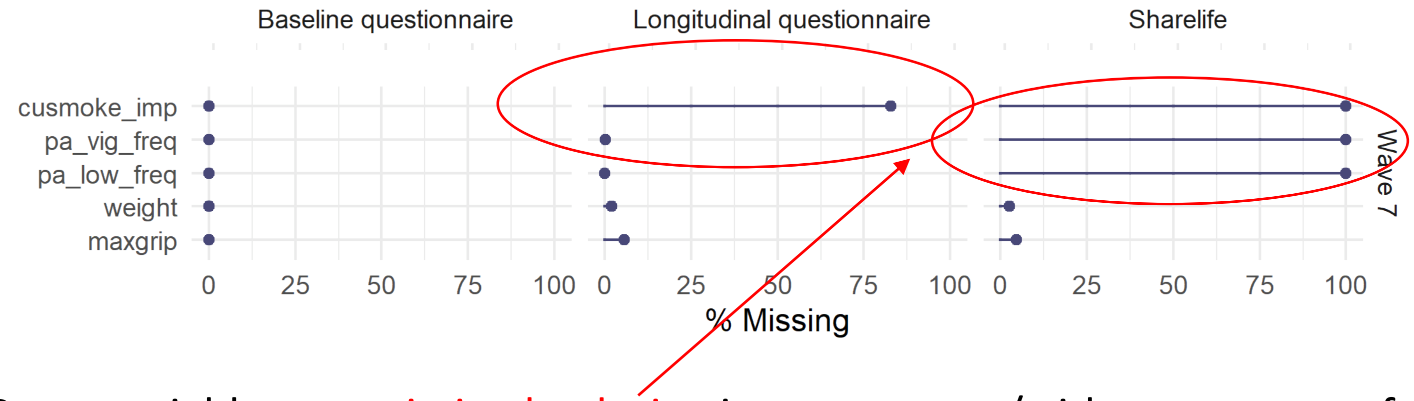
Drop-out effect on the outcome



Participants that die during the study tend to have lower values of grip strength than those that survive, especially among men.

The drop-out effect is not so strong when loss to follow-up occasion is analyzed.

Item missingness (and missing by design)



Some variables are missing by design in some waves/with some types of questionnaire ->

- summaries useful to "uncover" data properties difficult/absent in metadata
- Consequences on SAP
  - complete case analysis not sensible
  - use smoking as time-fixed variable (at entry)?
  - (not shown) not feasible to use the multiple imputations provided by the study

Small percentages of participants with item missingness, if not missing by design

- additional summaries provided for item missingness | not by design
- tables are used for understanding better the "small numbers"

Missingness in the grip strength outcome

All measurements	Percentage with missing outcome
Males	2.8% (244/8728)
Females	4.6% (459/9904)

Overall not many missing values

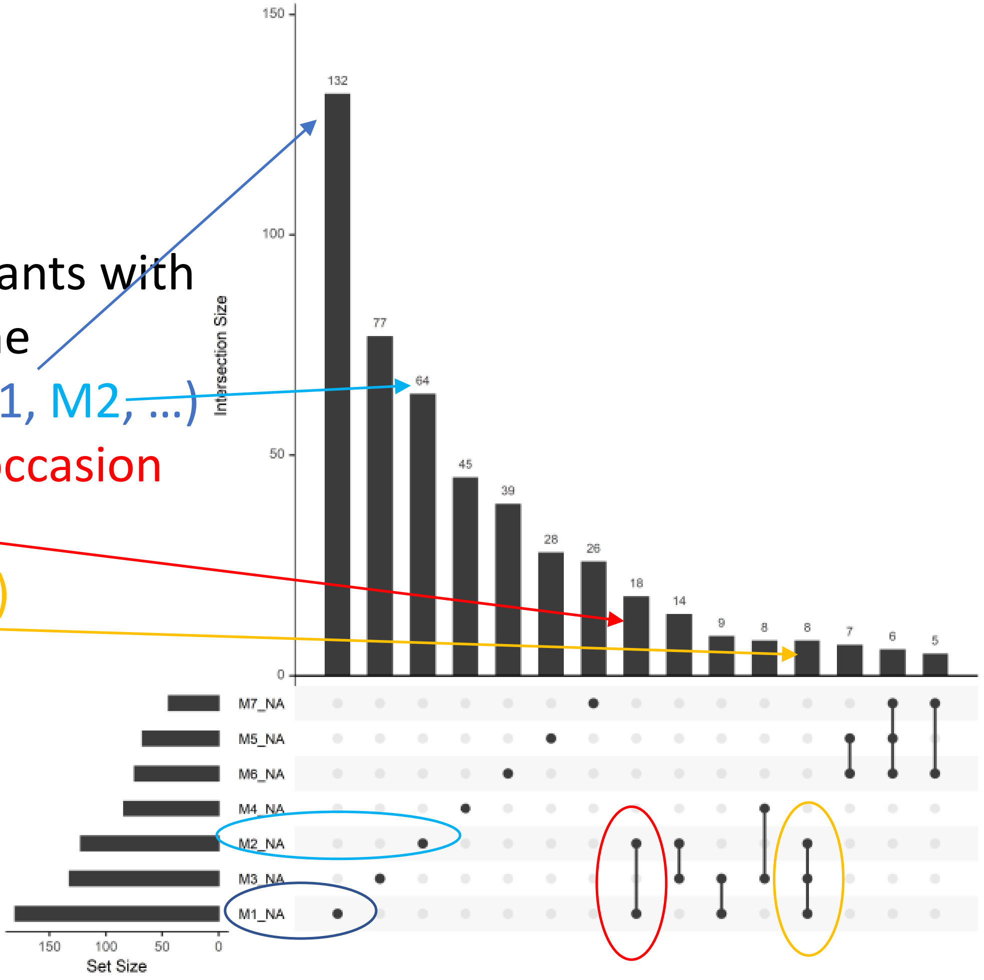
But there is a clear association between outcome missingness and age/sex

By measurement occasion	M1	M2	M3	M4	M5	M6	M7
Males	2.7% (71/2583)	2.0% (40/1983)	2.5 % (39/1562)	2.9% (27/940)	3.2% (23/720)	4.5% (29/646)	5.1% (15/294)
Females	3.8% (109/2869)	3.7% (82/2228)	5.2% (93/1801)	5.4% (57/1059)	5.1% (44/861)	6.0% (45/748)	8.6% (29/338)

By age	50-59	60-69	70-79	80+
Males	1.5% (42/2890)	1.9% (45/2890)	3.1% (57/2989)	11.4% (63/1994)
Females	2.4% (77/3159)	2.7% (89/3226)	6.2% (140/2104)	13.8% (153/956)

Co-occurrence of outcome missingness

Number of participants with missing values in the outcome in one (M1, M2, …) or more than one occasion (M1 and M2, M1, M2 and M3, …)



Outcome (item) missingness <u>does not co-occur frequently</u> for participants with valid interviews -> somehow surprising

Can be used to investigate co-occurrence for different variables

IDA in summary

IDA is the foundation for statistical modeling:
presentation, checking expectations, interpretation, model decisions

IDA takes time and planning

BUT: finding problems after modeling takes MORE time and
may miss issues (not systematic)
Help: code and workflow

IDA can detect features of a data set that could affect

the analysis
the interpretation
the presentation of results

It should also be reduced to only necessary steps, as in too lengthy default reports important findings could be overlooked.

Research studies need both: **Statistical analysis plan + IDA plan**

Initial Data Analysis Research Group

- Marianne Huebner, chair, (Michigan, USA)
- Carsten Oliver Schmidt, co-chair, (Greifswald, Germany)
- Saskia le Cessie (Leiden, Netherlands)
- Mark Baillie (Basel, Switzerland)
- Lara Lusa (Slovenia)

- Ackowledgements: Andrej Srakar (SHARE data) and Frank Lawrence (longitudinal modelling)

STRATOS
INITIATIVE

https://www.stratos-initiative.org/

https://www.stratosida.org/

References

Comprehensive IDA Framework

Huebner M, le Cessie S, Schmidt CO, Vach W on behalf of STRATOS-TG3. A contemporary conceptual framework for initial data analysis. Observational Studies 2018; 4: 171-192. Link

Ten Simple Rules for IDA

Baillie et al on behalf of STRATOS TG3. Ten simple rules for Initial Data Annalysis. PLoS Comp Biol https://doi.org/10.1371/journal.pcbi.1009819

IDA Reporting

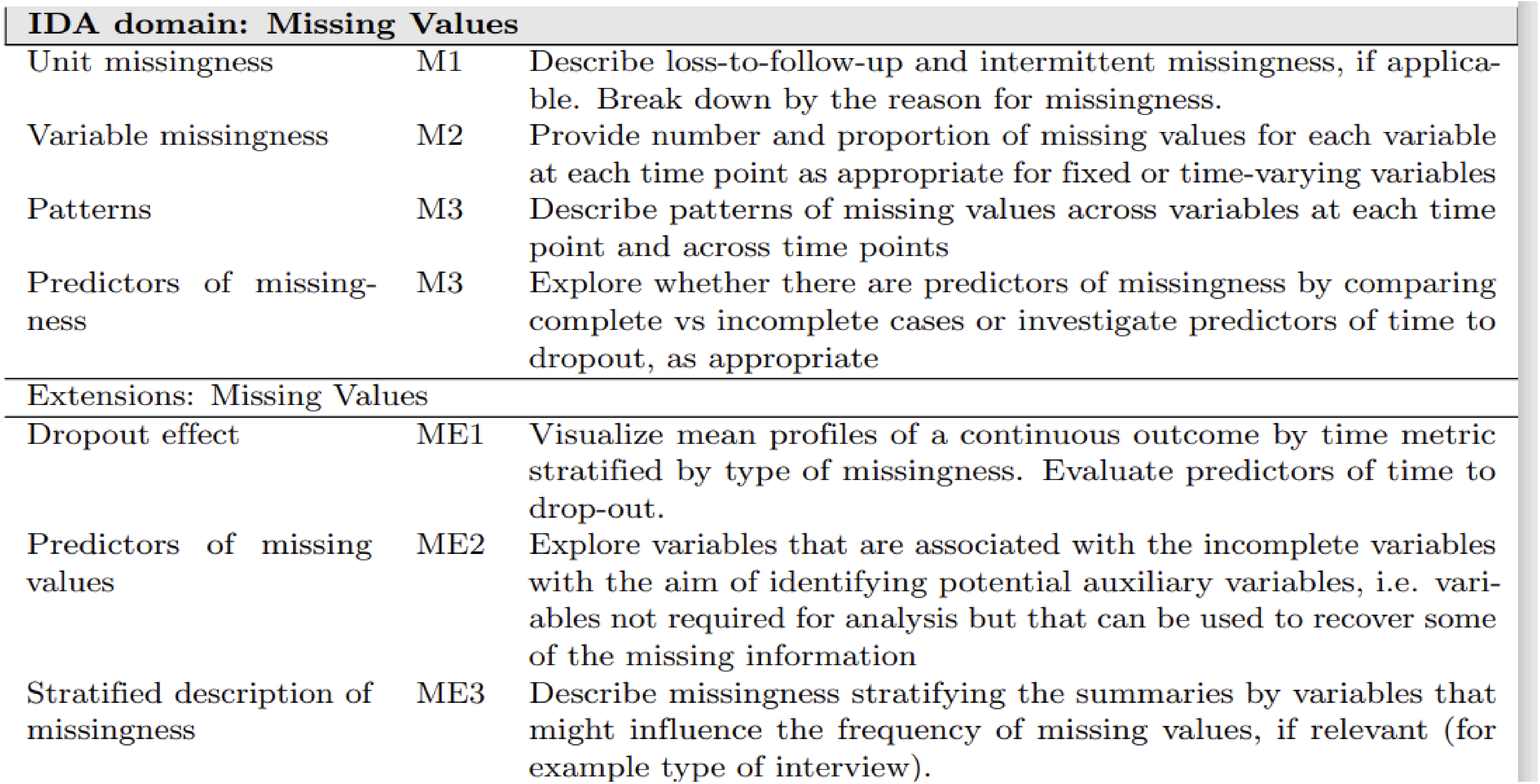
Huebner M, Vach W, le Cessie S, Schmidt C, Lusa L on behalf of STRATOS-TG3. Hidden Analyses: a review of reporting practice and recommendations for more transparent reporting of initial data analyses. BMC Med Res Meth 2020; 20:61. Link

An example

Lusa L, Huebner M. Organizing and Analyzing Data from the SHARE Study with an Application to Age and Sex Differences in Depressive Symptoms. IJERPH 2021;18(18):9684. doi: 10.3390/ijerph18189684 with open source R code: https://doi.org/10.17605/OSF.IO/KGTX6

Description of the TG3 projects and links to papers: https://www.stratosida.org/

Missing values in IDA checklists: longitudinal data



IDA domain: Missing Values		
Unit missingness	M1	Describe loss-to-follow-up and intermittent missingness, if applicable. Break down by the reason for missingness.
Variable missingness	M2	Provide number and proportion of missing values for each variable at each time point as appropriate for fixed or time-varying variables
Patterns	M3	Describe patterns of missing values across variables at each time point and across time points
Predictors of missingness	M3	Explore whether there are predictors of missingness by comparing complete vs incomplete cases or investigate predictors of time to dropout, as appropriate
Extensions: Missing Values		
Dropout effect	ME1	Visualize mean profiles of a continuous outcome by time metric stratified by type of missingness. Evaluate predictors of time to drop-out.
Predictors of missing values	ME2	Explore variables that are associated with the incomplete variables with the aim of identifying potential auxiliary variables, i.e. variables not required for analysis but that can be used to recover some of the missing information
Stratified description of missingness	ME3	Describe missingness stratifying the summaries by variables that might influence the frequency of missing values, if relevant (for example type of interview).