# Guidelines for multiple imputations in repeated measurements with time-dependent covariates: a case study

Maria Magdalena Pradita Eka K.

Journal Club – Bangkok, 18 March 2022

Missing data are one of the central problem that one encounter during the analysis of longitudinal data. If we fill in missing values with wrong data, we are adding bias.

# Imputations

- What is imputations?

**The process of replacing missing data with substituted values.**

- What is multiple imputations?

**Replace missing value with more than one imputed value, randomly drawn from a distribution of possible value.**

# Time- Dependent covariates

- Time – dependent covariates or time – varying covariates.
- What is time – varying covariates?
  - Variables whose values can change across time
- Example of time – varying covariates
  - C-reactive protein (CRP) and smoking status

# Classification Of Missing Data

MISSING DATA

- Missing completely at random (MCAR)
- Missing at random (MAR)
- Missing not at random (MNAR)

# MCAR

The missing data mechanism depends neither on observed nor on unobserved values.

# MAR

The missing data mechanism depends only on the observed values (and not on the unobserved values).

# MNAR

The missing data mechanism depends on the unobserved values (and perhaps also on observed values).

ELSEVIER

Journal of Clinical Epidemiology 102 (2018) 107–114

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

# Guidelines for multiple imputations in repeated measurements with time-dependent covariates: a case study

Frans E.S. Tan[a,*,1], Shahab Jolani[a,1], Hilde Verbeek[b]

[a]*Department of Methodology and Statistics, CAPHRI Care and Public Health Research Institute, Maastricht University, Maastricht, The Netherlands*
[b]*Department of Health Service Research, CAPHRI Care and Public Health Research Institute, Maastricht University, Maastricht, The Netherlands*

Accepted 14 June 2018; Published online 28 June 2018

https://www.jclinepi.com/article/S0895-4356(17)31193-9/fulltext

# Introduction

# Introduction

- Analysing longitudinal data over cross – sectional data is the possibility to describe individual profiles overtime.
- Characteristics of subject may vary over time.
- Problem ➔ missing data during longitudinal data analysis.

# Introduction

- This article provide : Research for practical guidelines to handle the most common missing repeated measurements data problems in observational studies.
- Key :
  - How to analyse longitudinal data if there are missing observation in the outcome only and / or if missing observation are extended to independent variables too.
  - Practicalities in producing imputations when there are many time – varying variables and repeated measurements.
  - Some common statistical package SPSS, SAS and R that are ready to use.

# Introduction

How to handle missing data in longitudinal study ?

Simulations study

Maastricht study on long-term dementia care environments (MLTD)

# Case Study

# Case Study

- Maastricht study on long-term dementia care environments (MLTD).
- This study investigated the effect of innovative dementia care environments (i.e. small scale, homelike) compared with traditional nursing homes (i.e. large scale) on residents' daily lives.
- Case study : **To compare the mood**

| The elderly living traditional large – scale wards (LSW) | VS | The elderly living innovative small – scale wards |
|---|---|---|

- N on this study is115

# Case Study

- Randomized observation schedule. Every participants observed for 1 minutes during 20 minutes period within 4.5 hours observation.
- Get break half hour in the each block.
- Each participants was observed on 7 days:
  - Two weekday mornings (07.00 – 11.00)
  - Two weekday afternoons (11.30 – 16.00)
  - Two weekday evenings (16.00 – 20.30)
  - One Saturday afternoon (16.00 – 20.30)
- Total = 12 (observation minutes per block in a day) x 7 (observations days) = 84 moment per participants.

# Case Study

- Mood and engagement in activity (activity) assessed by the Maastricht Electronic Daily Life Observation tool.
- 7 range of mood:
  - 1 = great sign of negative mood
  - 7 = very high positive mood
- The variable of activity measures :
  - Household activity
  - Musical activity

# Case Study

- What is missing in this data set?
  - **Outcome mood**
  - **Activity : 5 – 25 %**
  - **Observation across time (one dayparts) : 1 – 18%**

# Method

# Method

- Naïve Method :
  - Complete case analysis (CCA)
    - All missing values are deleted.
  - Available case analysis (AC)
    - Calculate observed values of the relevant variables(s).
  - Mean substitution (MS)
    - Missing value replace by arithmetic mean of that variables.

# Method

- Naïve method :
  - Missing indicator method (MIM)
    - Fill missing observation with fixed number and then add a dummy variable to the analysis model to indicate whether value of that variable was missing.
  - Last observation carried forward (LOCF)
    - Use the last observation to fill the next missing value.

# Method

- Multiple Imputations:
  - This method will replace missing value with more than one imputed value, randomly drawn from a distribution of possible value that determined using information from data.
  - Condition under MAR and MCAR.
  - Fully condition Specification (FCS) or chain equations is one popular method.
  - FCS will imputes missing data on a variable – by – variable.

# Statistical Analysis

Simulation Study

# Statistical Analysis

- Longitudinal design with three times point:
  - Correlated binary variables $X_1$, $X_2$, and $X_3$ were generated with equal marginal probabilities (i.e. $P(X_1 = 1) = P(X_2 = 1) = P(X_3 = 1) = 0.5$)
  - Also have equal correlation (i.e. $cor(X_1,X_2) = cor(X_1,X_3) = cor(X_2,X_3) = 0.5$)
  - Binary variables $X_1$, $X_2$ and $X_3$ generated using R package ‘bindata’.
  - Y generate using random intercept model.

$$Y_{it} = \beta_{0i} + \beta_1 X_{it} + \varepsilon_{it}$$

# Statistical Analysis

$$Y_{it} = \beta_{0i} + \beta_1 X_{it} + \varepsilon_{it}$$

- I = 1 … 115 (subject)
- t = 1, 2, 3 (time)
- $u_i$ = the random intercept (normal distribution with mean zero)
- $\varepsilon_{it}$ = the residual
- Note : the covariance structure implies that the outcome variable $Y_1$, $Y_2$, and $Y_3$ are correlated.

# Scenario 1

- Missing observation in the both outcome and independent variable under MCAR.
  - The outcome $Y_2$ or $Y_3$ (or both) were missing, constant probability 0.3.
  - Independent variable $X_2$ or $X_3$ (or both) were missing with same constant probability.
  - In total, 50% of the case was incomplete. The outcome and independent variables were never jointly.

# Analysis and Result

**Table 1.** Simulation results from five replications with a sample size of n = 115 and three repeated measurements

**Scenario 1: MCAR—x and y missing within total 50% incomplete.**

| | Statistics | | | | | |
|---|---|---|---|---|---|---|
| **Method** | $\hat{\beta}_0$ | $se(\hat{\beta}_0)$ | **95% CI coverage rate of** $\beta_0$ | $\hat{\beta}_1$ | $se(\hat{\beta}_1)$ | **95% CI coverage rate of** $\beta_1$ |
| REF | 2.01 | 0.095 | 0.95 | 0.50 | 0.101 | 0.95 |
| CCA | 2.01 | 0.136 | 0.95 | 0.50 | 0.145 | 0.95 |
| AC | 2.01 | 0.101 | 0.95 | 0.50 | 0.114 | 0.95 |
| MS | 2.05 | 0.091 | 0.90 | 0.42 | 0.104 | 0.88 |
| MIM | - | - | - | - | - | - |
| LOCF | 2.05 | 0.097 | 0.92 | 0.42 | 0.102 | 0.87 |
| MI | 2.03 | 0.101 | 0.95 | 0.45 | 0.117 | 0.93 |

Abbreviations: AC, available cases; CCA, complete case analysis; CI, confidence interval; LOCF, last observation carried forward; MI, multiple imputation; MIM, missing indicator method; MS, mean substitution; REF, reference.

# Scenario 2

- Missing observation in the outcome under MAR.
  - $Y_2$ or $Y_3$ (or both) were missing.
  - $Y_2$ was missing if $Y_1 \leq \bar{Y}_1$ ; $Y_3$ was missing if $Y_2 \leq \bar{Y}_2$.
  - The probability of missingness for $Y_2$ depends only on observed values of $Y_1$.
  - The probability of missingness for $Y_3$ depends only on observed values of $Y_2$.
  - Approximately 50% of the outcome variables was incomplete.

# Result Scenario 2

| Scenario 2: MAR—y missing with approximately 50% of the outcome variable incomplete | | | | | | |
|---|---|---|---|---|---|---|
| | Statistics | | | | | |
| Method | $\hat{\beta}_0$ | $se(\hat{\beta}_0)$ | 95% CI coverage rate of $\beta_0$ | $\hat{\beta}_1$ | $se(\hat{\beta}_1)$ | 95% CI coverage rate of $\beta_1$ |
| REF | 2.00 | 0.095 | 0.95 | 0.51 | 0.101 | 0.94 |
| CCA | 2.66 | 0.116 | 0.00 | 0.41 | 0.135 | 0.88 |
| AC | 2.00 | 0.100 | 0.96 | 0.50 | 0.112 | 0.94 |
| MS | 2.16 | 0.084 | 0.55 | 0.41 | 0.102 | 0.86 |
| MIM | - | - | - | - | - | - |
| LOCF | 2.00 | 0.096 | 0.95 | 0.40 | 0.096 | 0.82 |
| MI | 2.03 | 0.100 | 0.94 | 0.46 | 0.115 | 0.95 |

Abbreviations: AC, available cases; CCA, complete case analysis; CI, confidence interval; LOCF, last observation carried forward; MI, multiple imputation; MIM, missing indicator method; MS, mean substitution; REF, reference.

# Scenario 3

- Missing observations in the independent variable under MAR.
  - Independent variables $X_2$ and $X_3$ (Or both) were missing.
  - $X_2$ was missing if $Y_2$ was smaller than or equal to its first quartile.
  - $X_3$ was missing if $Y_3$ was smaller than or equal to its first quartile.
  - Approximately 40% of independent variables was incomplete.
  - Comparable MAR mechanism as in scenario 2.

# Result Scenario 3

| Scenario 3: MAR—approximately 40% of the independent variables was incomplete | | | | | | |
|---|---|---|---|---|---|---|
| | Statistics | | | | | |
| Method | $\hat{\beta}_0$ | $se(\hat{\beta}_0)$ | 95% CI coverage rate of $\beta_0$ | $\hat{\beta}_1$ | $se(\hat{\beta}_1)$ | 95% CI coverage rate of $\beta_1$ |
| REF | 2.00 | 0.095 | 0.95 | 0.50 | 0.102 | 0.96 |
| CCA | 2.53 | 0.097 | 0.00 | 0.38 | 0.112 | 0.80 |
| AC | 2.19 | 0.092 | 0.45 | 0.43 | 0.105 | 0.90 |
| MS | 2.07 | 0.101 | 0.92 | 0.35 | 0.109 | 0.72 |
| MIM | 1.16 | 0.123 | 0.00 | 0.42 | 0.097 | 0.85 |
| LOCF | 2.03 | 0.099 | 0.95 | 0.42 | 0.110 | 0.89 |
| MI | 2.04 | 0.102 | 0.94 | 0.43 | 0.128 | 0.93 |
| Scenario 4: MAR—approximately 50% of the dependent and independent variables was incomplete | | | | | | |

Abbreviations: AC, available cases; CCA, complete case analysis; CI, confidence interval; LOCF, last observation carried forward; MI, multiple imputation; MIM, missing indicator method; MS, mean substitution; REF, reference.

# Scenario 4

- Missing observation in the both outcome and independent variable under MAR.
  - Missing values $Y_2$ or $Y_3$ created as in scenario 2, or missing value on $X_2$ or $X_3$ where created as in scenario 3.
  - But not on the both
  - Independent variables are incomplete
  - Approximately 50% of cases were incomplete
  - Comparable MAR mechanism as in scenario 2 or scenario 3.

# Result Scenario 4

| Scenario 4: MAR—approximately 50% of the dependent and independent variables was incomplete | | | | | | |
|---|---|---|---|---|---|---|
| | Statistics | | | | | |
| Method | $\widehat{\beta}_0$ | $se(\widehat{\beta}_0)$ | 95% CI coverage rate of $\beta_0$ | $\widehat{\beta}_1$ | $se(\widehat{\beta}_1)$ | 95% CI coverage rate of $\beta_1$ |
| REF | 2.01 | 0.095 | 0.95 | 0.50 | 0.102 | 0.95 |
| CCA | 2.61 | 0.106 | 0.00 | 0.39 | 0.123 | 0.83 |
| AC | 2.10 | 0.097 | 0.81 | 0.46 | 0.110 | 0.94 |
| MS | 2.10 | 0.093 | 0.79 | 0.39 | 0.106 | 0.84 |
| MIM | - | - | - | - | - | - |
| LOCF | 2.02 | 0.097 | 0.93 | 0.41 | 0.104 | 0.84 |
| MI | 2.04 | 0.101 | 0.93 | 0.44 | 0.122 | 0.94 |

Abbreviations: AC, available cases; CCA, complete case analysis; CI, confidence interval; LOCF, last observation carried forward; MI, multiple imputation; MIM, missing indicator method; MS, mean substitution; REF, reference.

# Result from Scenario 1 - 4

- During 4 scenarios :
  - CCA can produce bias under MAR but can produce unbiased estimates under MCAR.
    - Scenario 1
  - AC analysis were unbiased when the outcome had missing observation. However will leaded biased estimated and lower coverage rare with missing data in the independent variables.
    - Scenario 2, 3 and 4 particularly $\beta_0$
  - Mean Substitutions produced biased estimates with lower coverage rates.
    - Bias in all scenarios

# Result from Scenario 1 - 4

- MIM only valid if missing data of the outcome conditional on the other independent variables. Also cannot handle missing observation in the outcome.
  - Bad performance in scenario 3
- LOCF leading biased estimates on all scenario.
  - Bad on the all situations
- MI provide best performance with negligible bias and acceptable coverage rates [~ 95%].
  - Work on the all scenario

# Statistical Analysis

Maastricht study on long-term dementia care environments (MLTD)

# Statistical Analysis for MLTD

- Compare the mood of participants in the large – scale wards and small – scale wards.
- Substantive model for analysis ➔ random intercept
  - Outcome variable : mood
  - Independent variables :
    - Large scale ward indicator (LSW = 1)
    - Activation indicator (activity = 1)
    - Part of the day (seven categories)
    - Repeated measurement of participants ( time tread as continuous)

# Statistical Analysis for MLTD

- Multiple Imputations:
  - Fully condition Specification (FCS) or chain equations.
  - Using R – Package MICE
  - Setting :
    - Number of imputation set to m = 20.
    - Data is formatted in wide format.
    - Applied "Just Another Variable"  and impute it separately.
    - The outcome is Mood10 ( mood multiple by a factor 10)

# Statistical Analysis for MLTD

- Tricks to work on the MLTD longitudinal study:
  - Data is formatted in wide format
  - Handle over parameterization
  - Apply “Just Another Variable” and impute it separately

# Statistical Analysis for MLTD

- Wide – format in MLTD :
  - Each person occupies only one record in the dataset, and observation made at different time points are coded as different column.
  - Because there 84 repeated measurements in MLTD:
    - Mood : 84
    - Activity : 84
    - Social Interaction : 84
    - Interaction where activity and social interaction are involved in the imputation model.
  - Total more than 300 time – varying covariates in wide formats.
  - N = 115 subjects

# Statistical Analysis for MLTD

- Handle over parameterization
  - This situation happen when the imputation include all variables as predictors for a particular variable cannot be fitted due to over parameterization.
  - In FCS ➔ for mood at time 1 need to imputed
    - Mood, activity, all interaction between activity and time
    - Use from time 2, 3 … at time 84

# Statistical Analysis for MLTD

- Applied "Just Another Variable" (JAV) and impute it separately:
  - The imputation of interaction term with missing value.
  - Example : Activity has missing observation and hence, its interaction with LSW has missing observations too.
  - In this study JAV → to imputed the interaction between Activity ( social interaction) and LSW

# Statistical Analysis for MLTD

- Why R – package MICE?
  - SPSS
  - SAS
  - R – package MICE

# Statistical Analysis for MLTD

SPSS

- Default FCS
- How interaction of categorical variables with missing values are handled?
- Not flexible enough to customize the variable’s role

# Statistical Analysis for MLTD

**SAS**

- FCS approach can optionally be used
- Control the role for each variable separately
- Interaction terms are passively imputed

# Statistical Analysis for MLTD

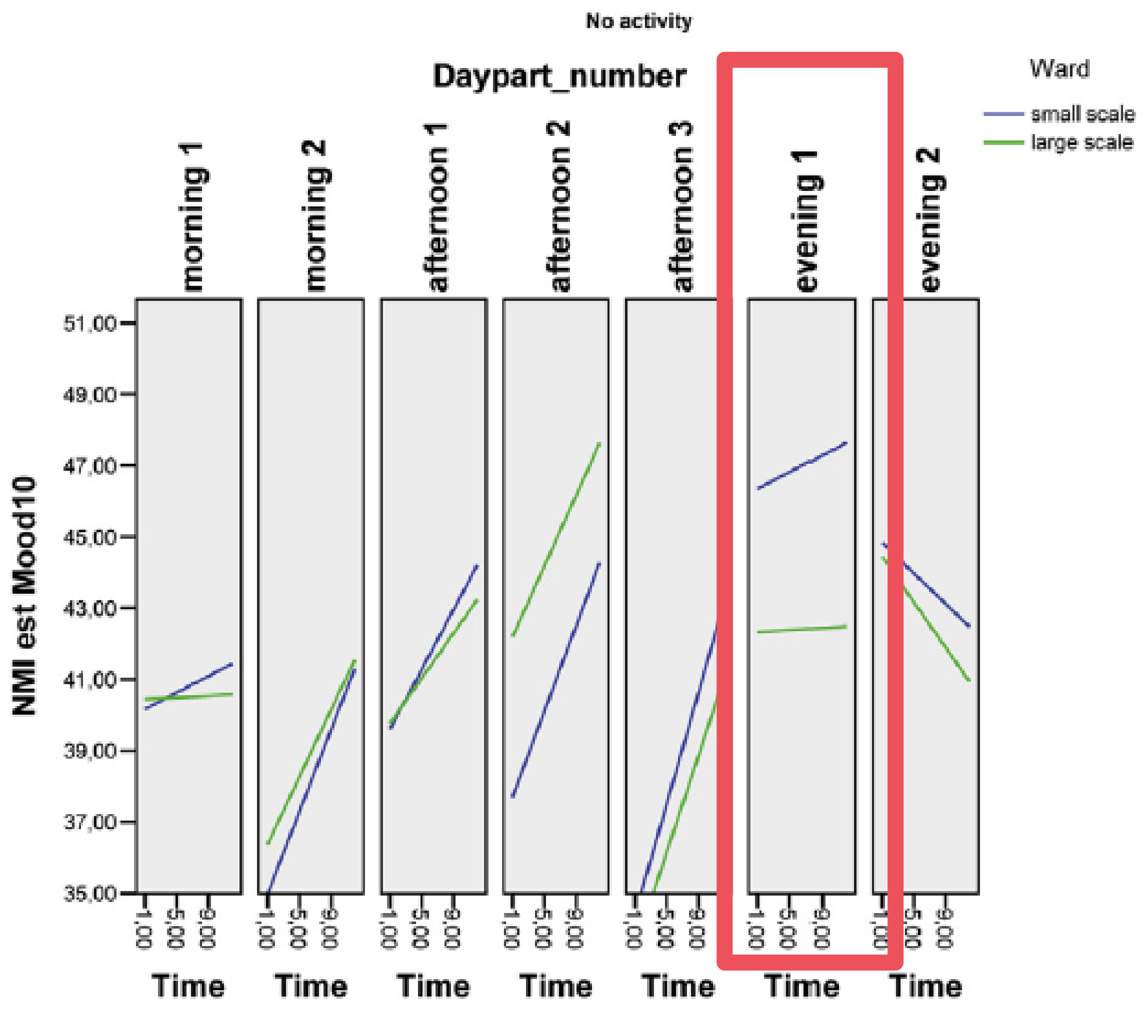
R - MICE

- FCS approach
- Customizable the role of variable
- JAV method can be use

**Table 2.** Relevant parameter estimates for Ward-effect with and without multiple imputation

| **Est LSW × effect** | **Substantive model** | |
|---|---|---|
| **Time, ward, activity, all first-order interactions as independent** | | |
| | **Complete case analysis (No multiple imputation)** | **MI, pooled estimates** |
| LSW | **−0.29 (0.46)** | **−0.11 (0.47)** |
| LSW × Daypart 1: Morning 1 | 0.67 (0.41) | 0.54 (0.43) |
| LSW × Daypart 2: Morning 2 | 0.89 (0.41) | 0.69 (0.45) |
| LSW × Daypart 3: Afternoon 1 | 0.19 (0.42) | 0.10 (0.46) |
| LSW × Daypart 4: Afternoon 2 | 1.12 (0.41) | 1.27 (0.44) |
| LSW × Daypart 5: Afternoon 3 | **−0.11 (0.41)** | **−0.02 (0.44)** |
| LSW × Daypart 6: Evening 1 | −0.60 (0.42) | −0.50 (0.44) |
| LSW × Daypart 7: Evening 2 | - | |
| LSW × time | −0.10 (0.03) | −0.09 (0.04) |
| LSW × Activity | **−0.09 (0.26)** | **−0.27 (0.32)** |

MI, multiple imputation.
Mood × 10 is the outcome.
Standard errors between brackets.



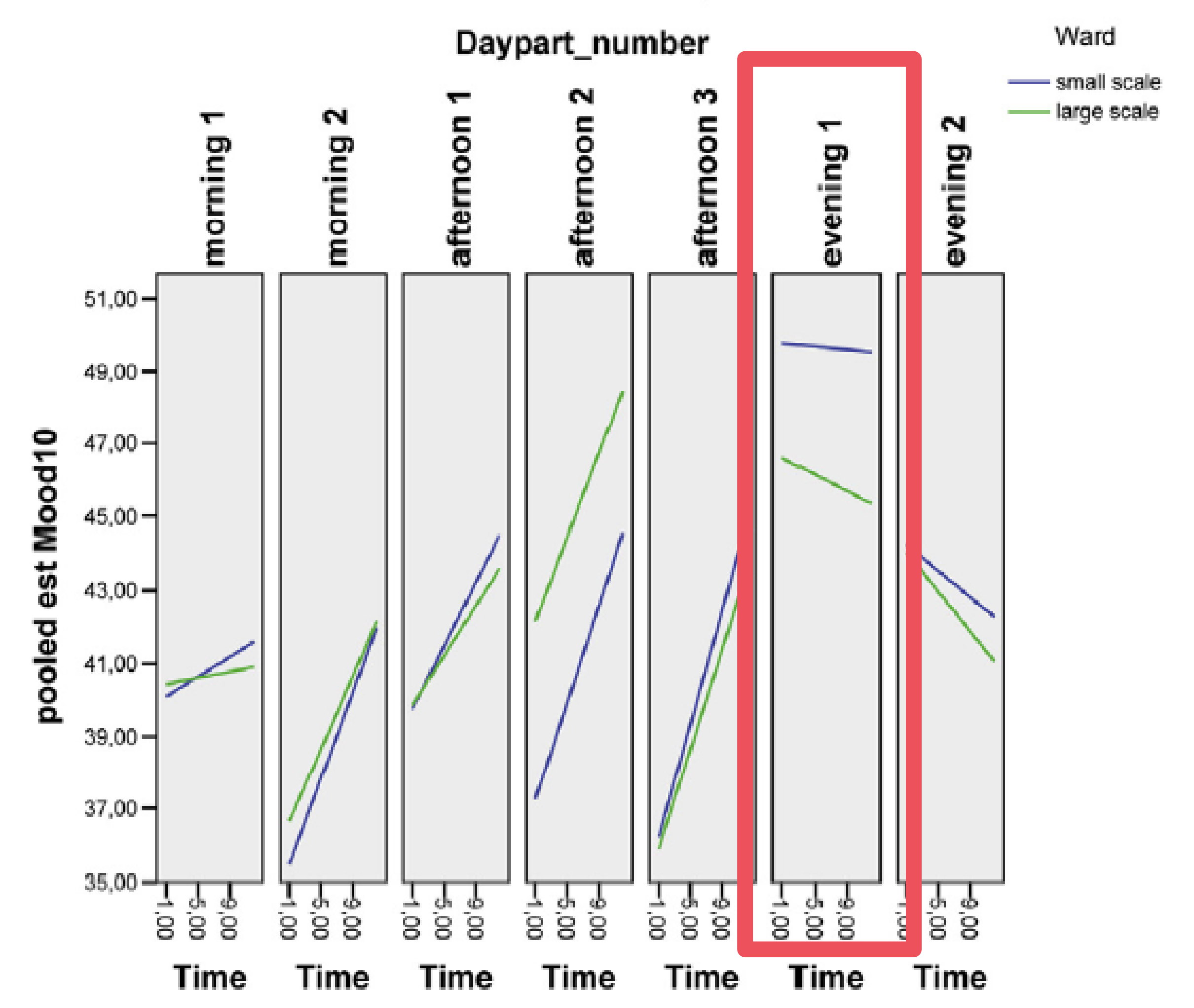




**Fig. 1.** Estimated difference between large- and small-scale wards and activity when not imputed and when imputed. NMI, no missing imputation.

# Discussion

# Key Finding

- Two situation require different approach :
  - When missing data on the outcome only (the independent variables are fully observed) ➔ use likelihood method and multiple imputation isn't important.
  - When missing data in the outcome and independent variables too ➔ multiple imputations
- Problem would be arise when there are more columns (variables per time points) than rows (subjects) ➔ wide format
- R – MICE is recommended application.

# Limitation

- Multilevel imputation has not been performed in this study. This study uses only standard FCS.
- Future study use multilevel imputations might be deal with the problem of higher level imputations.

# Thank you

Maria