Longitudinal data monitoring for Child Health Indicators

Vincent Were

Statistician, Senior Data Manager and Health Economist

Kenya Medical Research institute [KEMRI]

Presentation at Kenya Paediatric Association

25th April, 2017
Outline

• Definitions
• Designing Longitudinal Studies
• Advantages of Longitudinal Studies
• Data Management Process for monitoring longitudinal Studies
• Statistical Considerations
• Way forward
Definitions

• Longitudinal studies are defined as studies in which the outcome variable is repeatedly measured;
  • i.e. the outcome variable is measured in the same individual on several different occasions.

• In longitudinal studies, the observations of one individual over time are not independent of each other
  • It is necessary to apply special statistical techniques, which take into account the fact that the repeated observations of each individual are correlated

• Independent variables and outcome variables are measured repeatedly over time
Classification of study designs

Did investigator assign exposure?

Yes
- Experimental study
  - Random allocation
    - Randomised trial
    - Quasi-randomised trial

No
- Observational study
  - Comparison group
    - Analytical study
    - Descriptive study

Direction?

Forward
- Cohort study

Backward
- Case-control

At same time
- Cross sectional study

There are Both Observation and Experimental Longitudinal Studies
Four advantages of modern longitudinal methods

1. You have much more flexibility in research design
   ✓ Not everyone needs the same rigid data collection schedule—cadence can be person specific
   ✓ Not everyone needs the same number of waves—can use all cases, even those with just one wave!

2. You can identify temporal patterns in the data
   ✓ Does the outcome increase, decrease, or remain stable over time?
   ✓ Is the general pattern linear or non-linear?
   ✓ Are there abrupt shifts at substantively interesting moments?

3. You can include time varying predictors (those whose values vary over time)
   ✓ Participation in an intervention
   ✓ Family composition, employment
   ✓ Stress, self-esteem

4. You can include interactions with time (to test whether a predictor’s effect varies over time)
   ✓ Some effects dissipate—they wear off
   ✓ Some effects increase—they become more important
   ✓ Some effects are especially pronounced at particular times.
EXAMPLE: Integrated Data System provides longitudinal perspective
Observational longitudinal studies

• In observational longitudinal studies investigating individual development
  • Each measurement taken on a subject at a particular time-point is influenced by three factors [Confounders]:
  • (1) age (time from date of birth to date of measurement),
  • (2) period (time moment at which the measurements taken), and
  • (3) birth cohort (group of subjects born in the same year).
Data Structure: Long vs Broad

**'long' data structure**

<table>
<thead>
<tr>
<th>ID</th>
<th>Y</th>
<th>time</th>
<th>$X_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.5</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>3.7</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>3.9</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>3.0</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>3.2</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>3.2</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>4.1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>4.1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$N$</td>
<td>5.0</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>$N$</td>
<td>4.7</td>
<td>6</td>
<td>2</td>
</tr>
</tbody>
</table>
### 'broad' data structure

<table>
<thead>
<tr>
<th>ID</th>
<th>$Y_{t1}$</th>
<th>$Y_{t2}$</th>
<th>$Y_{t3}$</th>
<th>$Y_{t4}$</th>
<th>$Y_{t5}$</th>
<th>$Y_{t6}$</th>
<th>$X_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.5</td>
<td>3.7</td>
<td>3.9</td>
<td>3.0</td>
<td>3.2</td>
<td>3.2</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>4.1</td>
<td>4.1</td>
<td>4.2</td>
<td>4.6</td>
<td>3.9</td>
<td>3.9</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>3.8</td>
<td>3.5</td>
<td>3.5</td>
<td>3.4</td>
<td>2.9</td>
<td>2.9</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>3.8</td>
<td>3.9</td>
<td>3.8</td>
<td>3.8</td>
<td>3.7</td>
<td>3.7</td>
<td>1</td>
</tr>
<tr>
<td>$N$</td>
<td>4.0</td>
<td>4.6</td>
<td>4.7</td>
<td>4.3</td>
<td>4.7</td>
<td>5.0</td>
<td>2</td>
</tr>
</tbody>
</table>
Type of Effects / Confounders of effects

• Period or Time of measurement effect
• Cohort Effect
• Missing data or drop-outs during follow-up
Experimental longitudinal studies

- Experimental (longitudinal) studies are by definition prospective cohort studies. Type of designs

- An illustration of a few experimental longitudinal designs: (1) ‘classic’ experimental design, (2) ‘classic’ experimental design with baseline measurement, (3) ‘Solomon four group’ design, (4) factorial design and (5) ‘cross-over’ design.
Data management Process
Missing data in longitudinal studies

• One of the main methodological problems in longitudinal studies is missing data or attrition, i.e. the (unpleasant) situation when not all $N$ subjects have data on all $T$ measurements

• When subjects have missing data at the end of a longitudinal study they are often referred to as drop-outs

• It is, however, also possible that subjects miss one particular measurement, and then return to

• the study at the next follow-up. This type of missing data is often referred to as intermittent missing data
Example of missing data dataset

<table>
<thead>
<tr>
<th>Subject</th>
<th>HRSD 1</th>
<th>HRSD 2</th>
<th>HRSD 3</th>
<th>HRSD 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject 1</td>
<td>20</td>
<td>13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subject 2</td>
<td>21</td>
<td>21</td>
<td>20</td>
<td>19</td>
</tr>
<tr>
<td>Subject 3</td>
<td>19</td>
<td>18</td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td>Subject 4</td>
<td>30</td>
<td></td>
<td>25</td>
<td>23</td>
</tr>
</tbody>
</table>
Types of Missing data

• (1) missing completely at random (MCAR: missing, independent of both unobserved and observed data)

• (2) missing at random (MAR: missing, dependent on observed data, but not on unobserved data, or, in other words, given the observed data, the unobserved data are random),

• and (3) missing not at random (MNAR: missing, dependent on unobserved data) (Little and Rubin, 1987).
Ignorable or informative missing data?

• Conduct analysis of the missing data for outcome or predictor variables
• Use test tests, chi-square tests or logistic regression
Handling missing data: continuous data

- Imputation:
- Calculation of the average value (mean or median) of the available data for a particular variable at a particular time-point. This average value is imputed for the missing values.

- The simplest longitudinal imputation method is called the ‘last value carried forward’ (LVCF) method.
  - Value of a variable at $t = 1$ for a particular subject is imputed for a missing value for that same subject at $t = 2$.

- Linear interpolation
  - With this method, for a missing value at $t = 2$ the average of the values at $t = 1$ and $t = 3$ is imputed, assuming a linear development over time of the variables with missing data.
Dichotomous and categorical outcome variables

• imputation of the category with the highest frequency for the subject(s) with missing data

• The most frequently used longitudinal imputation method available for dichotomous and categorical missing data is the ‘LVCF’ method.
Statistical Considerations: Continuous outcome variables

• **Research Question**
  ‘Does the outcome variable $Y$ change over time?’ Or, in other words: ‘Is there a difference in the outcome variable $Y$ between time periods $t_1$ and $t_2$?’

• The **paired t-test** is used to test the hypothesis that the mean difference between $Y_{t1}$ and $Y_{t2}$ equals zero.

• non-parametric equivalent of the paired t-test, the (Wilcoxon) signed rank sum test
Hypothetical dataset for a longitudinal study with two measurements

<table>
<thead>
<tr>
<th>( i )</th>
<th>( Y_{t1} )</th>
<th>( Y_{t2} )</th>
<th>Difference (( d ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.5</td>
<td>3.7</td>
<td>-0.2</td>
</tr>
<tr>
<td>2</td>
<td>4.1</td>
<td>4.1</td>
<td>0.0</td>
</tr>
<tr>
<td>3</td>
<td>3.8</td>
<td>3.5</td>
<td>0.3</td>
</tr>
<tr>
<td>4</td>
<td>3.8</td>
<td>3.9</td>
<td>-0.1</td>
</tr>
<tr>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
</tr>
<tr>
<td>( N )</td>
<td>4.0</td>
<td>4.6</td>
<td>-0.6</td>
</tr>
</tbody>
</table>

The test statistic of the paired t-test is the average of the differences divided by the standard deviation of the differences divided by the square root of the number of subjects.
More than two continuous measurements

<table>
<thead>
<tr>
<th>$i$</th>
<th>$Y_{t1}$</th>
<th>$Y_{t2}$</th>
<th>$d_1$</th>
<th>$Y_{t3}$</th>
<th>$d_2$</th>
<th>⋮</th>
<th>$Y_{t6}$</th>
<th>$d_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.5</td>
<td>3.7</td>
<td>−0.2</td>
<td>3.9</td>
<td>−0.2</td>
<td></td>
<td>3.0</td>
<td>0.2</td>
</tr>
<tr>
<td>2</td>
<td>4.1</td>
<td>4.1</td>
<td>0.0</td>
<td>4.2</td>
<td>−0.1</td>
<td></td>
<td>4.6</td>
<td>0.0</td>
</tr>
<tr>
<td>3</td>
<td>3.8</td>
<td>3.5</td>
<td>0.3</td>
<td>3.5</td>
<td>0.0</td>
<td></td>
<td>3.4</td>
<td>−0.4</td>
</tr>
<tr>
<td>4</td>
<td>3.8</td>
<td>3.9</td>
<td>−0.1</td>
<td>3.8</td>
<td>0.1</td>
<td></td>
<td>3.8</td>
<td>0.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$N$</td>
<td>4.0</td>
<td>4.6</td>
<td>−0.6</td>
<td>4.7</td>
<td>−0.1</td>
<td></td>
<td>4.3</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Parametric test: MANOVA; non-parametric Friedman test
Generalized Estimating Equations (GEE)

\[ Y_{it} = \beta_0 + \sum_{j=1}^{J} \beta_{1j} X_{ijt} + \beta_2 t + \sum_{k=1}^{K} \beta_{3k} Z_{ikt} + \sum_{m=1}^{M} \beta_{4m} G_{im} + \varepsilon_{it} \]

where \( Y_{it} \) are observations for subject \( i \) at time \( t \),
\( \beta_0 \) is the intercept,
\( X_{ijt} \) is the independent variable \( j \) for subject \( i \) at time \( t \),
\( \beta_{1j} \) is the regression coefficient for independent variable \( j \),
\( J \) is the number of independent variables
\( t \) is time,
\( \beta_2 \) is the regression coefficient for time, \( Z_{ikt} \) is the time-dependent covariate \( k \) for subject \( i \) at time \( t \),
\( \beta_{3k} \) is the regression coefficient for time dependent covariate \( k \), \( K \) is the number of time-dependent covariates,
\( G_{im} \) is the time-independent covariate \( m \) for subject \( i \)
\( \beta_{4m} \) is the regression coefficient for time-independent covariate \( m \), \( M \) is the number of time independent covariates, and \( \varepsilon_{it} \) is the ‘error’ for subject \( i \) at time \( t \).
Dichotomous and categorical outcome variables

• More complex analysis than continuous variable

• In general, for categorical outcome variables with $C$ categories, the change between subsequent measurements is another categorical variable with $C^2$ categories.
Analysis of experimental studies

- Experimental (longitudinal) studies differ from observational longitudinal studies in that experimental studies (in epidemiology often described as trials) include one or more interventions.
  - Manova analysis can be used correcting for baseline differences
- Example of GEE used analyse the effects of the therapy on systolic blood pressure, the following statistical model can used used

\[ Y_{it} = \beta_0 + \beta_1 \times \text{therapy} + \beta_2 \times \text{time} + \varepsilon_{it} \]

- where \( Y_{it} \) are observations for subject \( i \) at time \( t \), \( \beta_0 \) is the intercept, \( \beta_1 \) is the regression coefficient for therapy versus placebo, \( \beta_2 \) is the regression coefficient for time, and \( \varepsilon_{it} \) is the ‘error’ for subject \( i \) at time \( t \).
Considerations for Data Management System

Key matters to consider in designing a system to manage data are

- What is required by regulation?
- How would a person find relevant data?
- What makes the data fit for purpose?
- How will the data be shared and with whom?
- How will it be secured for future use?
• Thank you: Any Questions?

• Contact:
  vwere@kemricdc.org
  vincentwere@gmail.com

+254 721 876 573