Using Multiple Imputation for Loss to Follow Up: Cohort of HIV-Positive Patients in Haiti

Deanna Jannat-Khah, DrPH, MSPH
Michelle Unterbrink

Margaret McNairy, Daniel Fitzgerald, Samuel Pierre, Jean W. Pape, Arthur Evans

2015 UK STATA Users Conference
Background

• In many public health program evaluations, cohorts of patients are followed for months-years
• A proportion of patients can not be found
• This patients are categorized as lost to follow-up (LTF)
• The challenge:
  • These patients do not have an outcome status (i.e., dead vs. alive)
  • How do programs estimate their outcomes status?
Objectives of talk

1. Presenting and comparing methods to estimate outcome status of patients who are LTF

2. Demonstrating the application of Multiple Imputation for estimating outcome status of LTF
LTF is not an outcome: a mixture of outcome statuses

1. Undocumented death
2. Alive and in care somewhere else
3. Alive and not engaged in care
Methods used to estimate LTF outcome status

1. Survival analysis (Kaplan Meier methods)
2. Tracing with Inverse Probability Weights (IPW)
3. Multiple Imputation with Chained Equations (MICE)
Study overview

**Study purpose:** Estimate 10 year survival among the first cohort of HIV patients receiving treatment in Haiti

**Study site:** Haitian Group for the Study of Kaposi’s Sarcoma and Opportunistic Infections (GHESKIO clinic)

**Study Population:** 910 adults aged ≥ 13 years enrolled in HIV care in 2003

**Study follow up period:** 10 years

**Primary Outcome:** Survival status at 10 years

**Secondary Outcome:** Predictors of survival
Study outcomes at 10 years: 20% LTF

Proportion of patients

Known dead 27%
Outcome unknown* 20%
Alive and in care 53%

Years after ART initiation

* 8% transferred, 12% lost
Applying 3 methods used to estimate survival to this cohort

1. **Survival Analysis** (Kaplan Meier methods): censor LTF

2. **Tracing with Inverse Probability Weights (IPW)**: probability weights generated from tracing

3. **Multiple Imputation with Chained Equations (MICE)**: impute LTF and baseline characteristics that are missing

---

1. Kaplan Meier: censor LTF

Estimated alive: 71%  95%CI (68%,74%)
2. Tracing with Inverse Probability Weights

- A field worker traced patients who were LTF to determine outcome status
- Assume the ones found are a random sample of all LTF
- 156 patients categorized as LTF
- 45 were found
- Estimated alive: 71%  95%CI (68%, 74%)

\[
\text{iweight} = \frac{1}{\frac{45}{156}} = 3.472
\]
3. Multiple Imputation with Chained Equations

- Imputes the outcome status by using baseline covariates
- Fill in missing values present in covariates\textsuperscript{3-5}
- Several equations are created to fill in missing values
- One must specify the number of datasets to generate, results will be averaged across datasets
- Assumptions:
  - Missing are only randomly different from patients with same set of covariates
  - LTF were assumed to have the same average survival as those not lost, conditional on covariates

Applying this to our cohort: missing covariates

Demographic characteristics:
• Sex, age, residence, income

Clinical characteristics:
○ CD4
  • Distribution 0-1400 cells/µL
  • Missing 12% of baseline CD4
  • “Missingness” associated with death: OR = 1.67 95% CI (1.09, 2.55)

○ Weight
  • Distribution 20-120 kg
  • Missing 3% of baseline weight
  • “Missingness” associated with death: OR = 4.39 95% CI (1.86, 10.35)
Using Multiple Imputation with Chained Equations to impute missing covariates and outcome status

Chained Equations:

1. **Weight**: regress weight CD4 status age sex stage TB income residence
2. **CD4**: regress CD4 weight status age sex stage TB income residence
3. **10 year survival**: logit status weight CD4 age sex stage TB income residence
4. Repeated **20 times** to fill in all missing
Covariates filled in by MICE are similar to non-imputed values

<table>
<thead>
<tr>
<th>Clinical Characteristic</th>
<th>Without Imputation</th>
<th>With Imputation</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD4+ count (cells/uL)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median (IQR) (range)</td>
<td>131 (51–212) [0–1400]</td>
<td>131 (51–212) [-330–1416]</td>
</tr>
<tr>
<td>Missing</td>
<td>12%</td>
<td>N/A</td>
</tr>
<tr>
<td>Body weight (kg)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Men median (IQR)</td>
<td>56 (50–63)</td>
<td>54 (46–61)</td>
</tr>
<tr>
<td>Women median (IQR)</td>
<td>49 (44–56)</td>
<td>47 (40–54)</td>
</tr>
<tr>
<td>Missing</td>
<td>3%</td>
<td>N/A</td>
</tr>
</tbody>
</table>

Estimated Survival: 67% (95% CI 64%-71%)
Primary outcome:
10 year survival estimated to be 67-71%
Secondary outcome: predictors of death

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Without Imputation</th>
<th></th>
<th>With Imputation</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Odds Ratio</td>
<td>95% CI</td>
<td>Odds Ratio</td>
<td>95% CI</td>
</tr>
<tr>
<td>Female</td>
<td>0.79</td>
<td>(0.55, 1.12)</td>
<td>0.61</td>
<td>(0.44, 0.87)</td>
</tr>
<tr>
<td>Age</td>
<td>1.03</td>
<td>(1.01, 1.04)</td>
<td>1.03</td>
<td>(1.01, 1.05)</td>
</tr>
<tr>
<td>Residence</td>
<td>1.16</td>
<td>(0.82, 1.64)</td>
<td>1.14</td>
<td>(0.81, 1.59)</td>
</tr>
<tr>
<td><strong>Income</strong></td>
<td><strong>1.56</strong></td>
<td><strong>(1.09, 2.23)</strong></td>
<td><strong>1.81</strong></td>
<td><strong>(1.27, 2.58)</strong></td>
</tr>
<tr>
<td>CD4</td>
<td>1.00</td>
<td>(0.99, 1.00)</td>
<td>1.00</td>
<td>(1.00, 1.00)</td>
</tr>
<tr>
<td>Base weight</td>
<td>0.97</td>
<td>(0.95, 0.99)</td>
<td>0.96</td>
<td>(0.94, 0.98)</td>
</tr>
<tr>
<td><strong>WHO stage</strong></td>
<td><strong>1.51</strong></td>
<td><strong>(1.06, 2.14)</strong></td>
<td><strong>1.83</strong></td>
<td><strong>(1.31, 2.55)</strong></td>
</tr>
<tr>
<td>Baseline TB</td>
<td>2.12</td>
<td>(1.24, 3.62)</td>
<td>1.59</td>
<td>(0.92, 2.73)</td>
</tr>
</tbody>
</table>
### Comparing methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Assumptions for LTF</th>
<th>How LTF is treated in the analysis</th>
<th>Missing Covariate Data</th>
</tr>
</thead>
</table>
| **Survival Analysis**| • LTF is unrelated to mortality  
• That is, they are a random sample of those who continue to be followed | • Censored                         | • Censored                                  |
| **Tracing w/ IPW**   | • Those unsuccessfully traced have the same mortality as those successfully traced | • Weighted                         | • Case-wise deletion                         |
| **Multiple Imputation** | • Missing are only randomly different from patients with same set of covariates | • Imputed                          | • Imputed  
• All observations used |
## Application of methods in our study

<table>
<thead>
<tr>
<th>Method</th>
<th>Limitations</th>
<th>Strengths</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Survival Analysis</strong></td>
<td>• Most studies found assumption to be incorrect</td>
<td>• Most common method</td>
</tr>
<tr>
<td></td>
<td>• Survival is usually overestimated</td>
<td>• Easy to perform</td>
</tr>
<tr>
<td><strong>Tracing w/ IPW</strong></td>
<td>• Tracing was done at the end of the 10 year follow up period on everyone</td>
<td>• Common method in HIV studies</td>
</tr>
<tr>
<td></td>
<td>• Case-wise deletion if covariates are missing</td>
<td>• Conceptually easy to understand</td>
</tr>
<tr>
<td></td>
<td>• Tracing can be difficult and expensive</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Only as successful as your tracing success</td>
<td></td>
</tr>
<tr>
<td><strong>Multiple Imputation</strong></td>
<td>• Relies on a good prediction model</td>
<td>• Use all observations</td>
</tr>
<tr>
<td></td>
<td>• Biologically impossible values</td>
<td>• Robust standard error</td>
</tr>
</tbody>
</table>
Summary

1. LTF is a common category of patients in cohort studies (public health studies)
2. LTF is a mixture of patients (dead, alive)
3. Three commonly used methods estimate survival among LTF
4. Multiple Imputation with Chained Equations is a valid method that is infrequently used in public health
5. MICE estimated survival was different than the traditionally used methods
6. Potentially we could use MICE to impute survival time
Acknowledgements

GHESKIO staff

Weill Cornell Medical College
Division of Hospital Medicine

Weill Cornell Medical College Center
for Global Health
Any thoughts on how to impute survival time or how to deal with violations of PH assumptions?

- Imputing survival time
- Augmenting/limiting imputations
- Recommendations for how to deal with violations of PH assumptions: Aalen models or time varying or both
References & Resources

**Multiple Imputation**


**GHESKIO**


We chose age, sex, WHO stage, baseline and incident TB, income, residence, being self referred, weight, CD4 and outcome status at 6m and 10 years based on clinical, programmatic and research experience.
MICE diagnostics

*check to see if the imputed values are close enough for all imputed covariates
midiagplots base_wt, m(1/5) combine

*trace plots
use impstats21915
reshape wide *mean *sd, i(iter) j(m)
tsset iter
tsline base_wt_kg_mean*, name(graph1b) nodraw legend(off)

graph combine graph1b graph2b graph3b graph4b graph5b graph6b graph7b graph8b graph9b graph10b, title(trace plots of summaries of imputed values from 20 chains) rows(5)

* check for proportions and confidence intervals:
mi estimate: proportion Itdead_10

→ Marchenko STATA presentation great reference!