**Assessing the uncertainty around age-mixing patterns inferred from generalized HIV epidemic using phylogenetic tree data**

Abstract

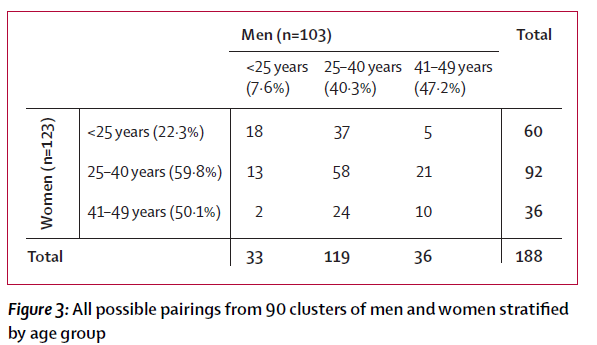
*Understanding age-mixing patterns in HIV transmission networks is key to design and implement more effective HIV prevention programmes. However, age-mixing patterns are difficult to investigate directly due to ethical constraints and stigma associated with HIV infection. A recent study in KwaZulu-Natal, South Africa, concluded that “Sexual partnering between young women and older men, who might have acquired HIV from women of similar age, is a key feature of the sexual networks driving transmission”, using an indirect phylogenetic approach. However, this age-mixing patterns was inferred from only a relatively small fraction of infected individuals. Only 4% of the census-defined population was included in the study, and among HIV-infected study participants, sequence coverage was rather low (37.3% for women and 48.0% for men). This means that there may be a high likelihood of a participant’s connections in the transmission network not being included in the analysis. As a result, it becomes very difficult to give a precise and valid meaning to “phylogenetically linked pairings”. We hypothesized that the range of compatible age-mixing patterns is much wider than the age-mixing patterns hypothesized in the study. To test this hypothesis, we conducted a simulation study with an agent-based simulation model and synthetic HIV sequence data. In this way, we are able to compare the actual age-mixing pattern in the simulated transmission network and the age-mixing pattern inferred by phylogenetic analysis with synthetic sequence data sampled from the model population.*

# **Study design**

1. Simulate an epidemic
2. Compute age mixing statistics in the entire epidemic
3. Considering missingness scenario and compute age mixing statistics among selected people
4. Construct phylogenetic tree
5. Compute transmission clusters
6. Construct pairings: by filtering the entire MRCA matrix from the tree by gender, transmission cluster IDs, tMRCA, CD4 count and viral load.
7. Compute true age structure pairings of people in transmission clusters (pairings between different age groups)
8. Compute proportion of pairings across different age groups in transmission clusters
9. Compute true age structure pairings of ALL selected individuals in the transmission network
10. Compute proportion of pairings across different age groups in true transmission network
11. Analysis and interpretation of descriptive table of pairings in different age groups: comparison of tables (true age structure transmission table and pairings table built from transmission clusters)
12. **Goodness of fit (GoF)**: Measure how the age structure table is far away or close from the true table (this will be done by comparing proportions of women/men paired with men/women across different age groups, and also the mean, median, and standard deviation of age difference in pairings): use root mean relative error.

**Age-mixing metrics of interest from Tulio et all. paper: *proportion of pairings* across age groups and *mean of age difference* between pairings**

A comparison of all possible pairings across age and sex categories (figure 3) identified two more frequent sets of linkages: between women aged 25–40 years and men aged 25–40 years (58 [30·9%] pairings) and between women younger than 25 years and men aged 25–40 years (37 [19·7%] pairings]), which constituted just over 50% of the linkages. In the 60 possible pairings between the 43 women younger than 25 years who were linked to 41 probable male partners, the mean age difference was 8·7 years (95% CI 6·8–10·6; p<0·0001). In these 60 possible pairings, 18 (30·0%) of the probable male partners were younger than 25 years, 37 (61·7%, 95% CI 49·7–74·3) were aged 25–40 years, and five (8·3%) were aged 41–49 years. Of the 41 probable male partners linked to a woman younger than 25 years, 16 (39·0%) were also linked to a 25–40-year-old woman (figure 4); the mean age of these male partners was 29·6 years (SD 7·1) and they were linked to both a younger woman (mean age 22·1 years) and an older woman (mean age 32·6 years). In the 92 possible pairings between the 56 women aged 25–40 years who were linked to 36 probable male partners, the mean age difference was 1·1 years (95% CI –0·6 to 2·8; p=0·111). Of 79 men (mean age 31·5 years) linked to women younger than 40 years, 62 (78·5%) were unaware of their HIV-positive status, 76 (96·2%) were not on ART, and 29 (36·7%) had viral loads of more than 50 000 copies per mL.



**Data missingness / sequence coverage scenarios:**

When we consider sequence missing mechanisms, we consider: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing not at Random (MNAR). And in all these types of data missingness (MCAR, MAR, MNAR), different levels of missingness will be considered, we may consider scenarios between 5-95% of sequence coverage, but we cannot have 0 and 100%.

**Epidemic simulation and parameter combination:**

An HIV epidemic was simulated in an age and gender structured population. With an initial population of 15000 men and 15000 women, the simulation time was 45 years, and HIV was seeded in the population at the 10th year among randmly selected individuals between 20 and 50 of age. During simulation, demographic events like birth, and death were considered together with sexual partnerships. Refering to real world history of HIV, mainly ART treatment, we gradually allow interventions based on CD4 counts. Regarding mapping the epidemic on calendar time, we assume that the seeding time to be 1987, and hence the as time increase the eligibility to ART increase to reach diagnosis and treat guidelines.

Age-mixing parameters are:

|  |  |  |
| --- | --- | --- |
| ***Parameter*** | ***Value*** | ***Description*** |
| formation.hazard.agegapry.baseline | 2 | baseline in the expression for the hazard, allowing one to establish a baseline value |
| person.agegap.man.dist.normal.mu | 10 | mean of preferred age differences distribution for men |
| person.agegap.woman.dist.normal.mu | 10 | mean of preferred age differences distribution for women |
| person.agegap.man.dist.normal.sigma | 5 | standard deviation of preferred age differences distribution for men |
| person.agegap.woman.dist.normal.sigma | 5 | standard deviation of preferred age differences distribution for women |

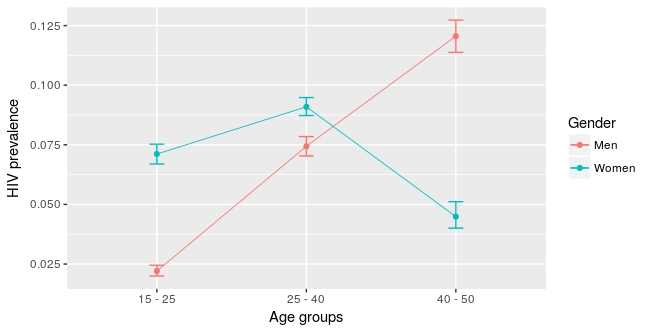
Age groups of interest were 15 – 24 years, 25 – 39 years, and 40 – 50 years, and sampling population was among these who were alive between 36 and 37 of simulation time.

**Simulation study work-flow**

1. Defining master model: it is the default model with well defined parameter combination - it will produce statistics of age mixing patterns across different age groups in different missingness scenarios (MCAR & MAR), and the truth ones as well.
2. With same parameter combination, run the master model ***1120*** times and produce statistics of age mixing patterns across different age groups, the truth one and these from different missingness scenarios (MCAR & MAR), and compute the RMSE and MAE between these statistics and the truth (*output data frame*: statistics of age mixing patterns in MCAR & MAR + true statistics of age mixing patterns + RMSE and MAE between the two sets of statistics + SEED)

**HIV Epidemic**

The epidemic was characterized by increasing prevalence for men across age groups whereas it reaches a maximum among 25 – 39 years among women, and drop thereafter within 36 and 37 time interval of the study.

***Figure****: Prevalence across men/women age groups, the values are median and error-bars are 1st quantile and 3rd quantile for the 1180 simulations.*

**Results for Missing Completely at Random (MCAR)**

Sampled individuals are drawn randomly, without any condition/preference.

**(i) Age mixing characteristic in partnerships**

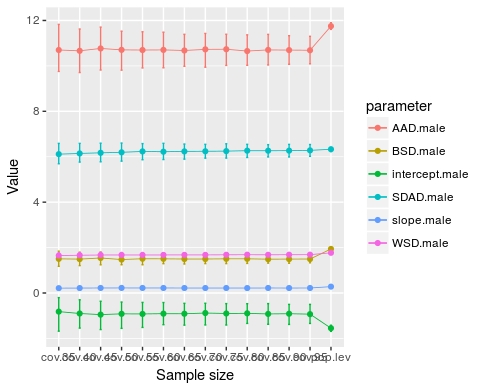
Age mixing characteristics among the population between 30 and 40 years time interval were average age gap, standard deviation of age gap, within variation of standard deviation, between variation of standard deviation, the slope and the intercept.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | AAD | SDAD | slope | WSD | BSD | intercept |
| Value | 11.75  [11.61;11.90] | 6.33 [6.23;6.41] | 0.28 [0.28;0.29] | 1.77 [1.74;1.80] | 1.94 [1.90;1.99] | -1.55  [-1.69 ;-1.42] |

When considering sampled individuals (who are infected between 36 and 37 time interval) at different sequence coverage scenarios we have, the age mixing characteristics



We can see that the average age difference from infected individuals in their partnership is not far from the mean of preferred age differences distribution for men/women which was 10; the difference between the two is less than 1. Same fact is observed from standard deviation of age difference in partnership of infected individuals and standard deviation of preferred age differences distribution for men, which was 5 ; the difference between the two is slightly greater than 1.



***Figure:*** *Partnership age mixing statistics in infected and general populations*

**(ii) Proportion of men/women in pairings**

The true pairings at 100% coverage were:



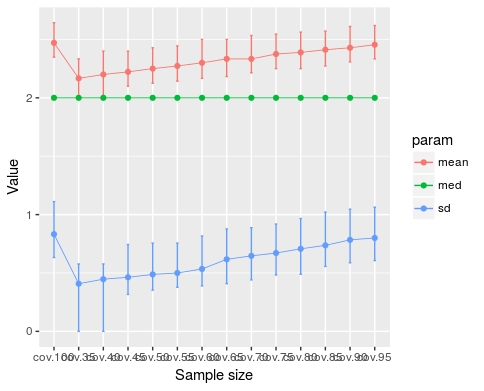
True aggregated pairings across men and women in different age groups were



In transmission clusters pairings across men and women in different age groups were



We computed transmission clusters’ statistics, mainly mean, median, and standard deviation of sizes of transmission clusters.



Proportions of pairing of men/women across different age groups from transmission clusters:

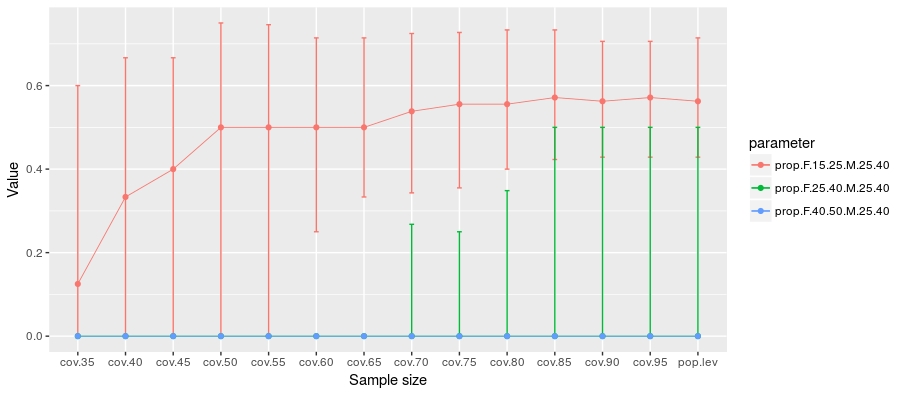


True proportions of pairing of men/women across different age groups from transmission clusters:



One of the highlight from the two tables above is that values of proportions of men and women in any given age group who are pairing with women and men in another age group obtained from transmission network built from transmission clusters are close to true values obtained from same individuals in the true transmission network, and this is true across all sequence coverage scenarios.

Another important fact is that as the sequence coverage increase the estimate proportion approaches the true value at 100% coverage, and also the defined confidence intervals[[1]](#footnote-2) are reduced as it can be seen in the figure below.



***Figure****: Example for proportion of men of 25-40 paired with women in different age groups inferred from transmission clusters*

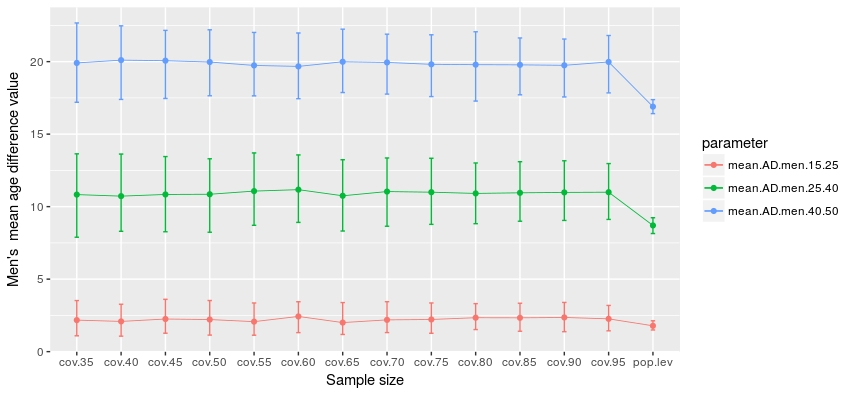
Another highlight is that the truth value at 100% coverage can never be greater than the maximum of the estimates values; hence estimates of proportions at any sequence coverage scenario can overestimate or underestimate the truth.

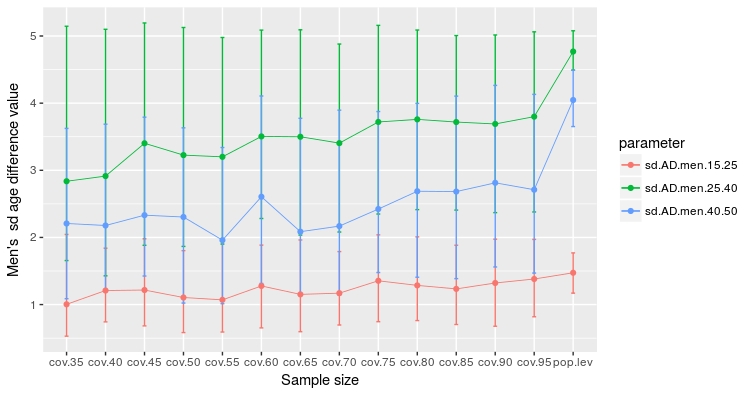
**(iii) Mean, median, and standard deviation of age gap in pairings**

From the transmission network constructed from transmission clusters, we compute mean, median, and standard deviation of age gap of females/males in a given age group paired with males/females regardless of age group.



The minimum age gap value for men in pairings is at 100% coverage (being the true value as well) as it can be seen in the figure below.

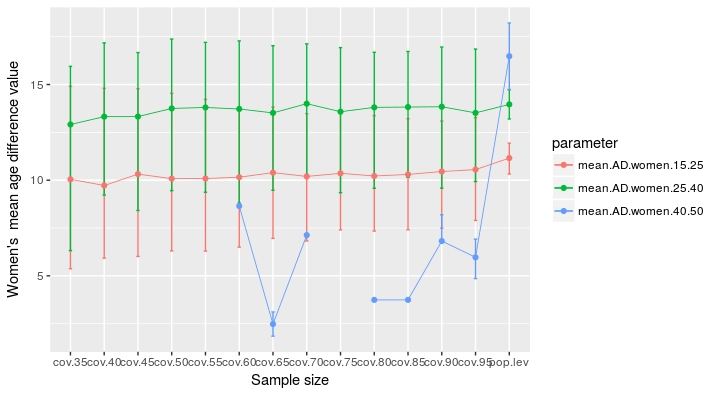
***Figure****: Men’s average age gap in pairings*



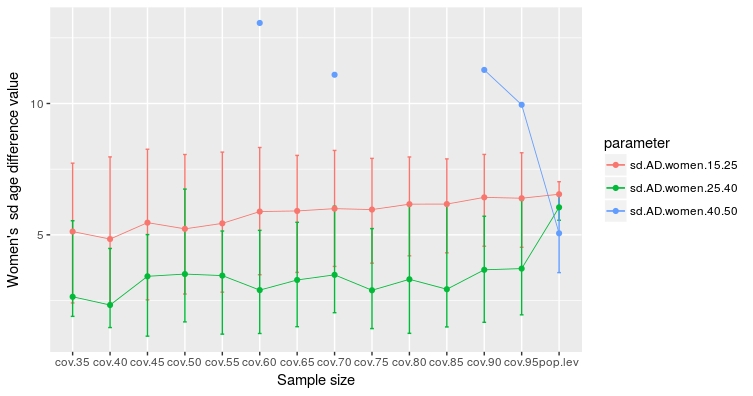
***Figure****: Men’s standard deviation age gap in pairings*

Men of 40 and 50 years old, 25 and 40 years, and 15 and 25 years had an average age gap of almost 20, 10, and 2 respectively. This means old men paired with younger women.

For women, we can see that women of 40 and 50 years old had much variable age gap with their pairs, spanning between 2 and 8. This means that sometimes they were paired with younger men and men of same age. For other women in 25 and 40 years, and 15 and 25 years, they had an average age gap of almost 10, and 13 respectively.



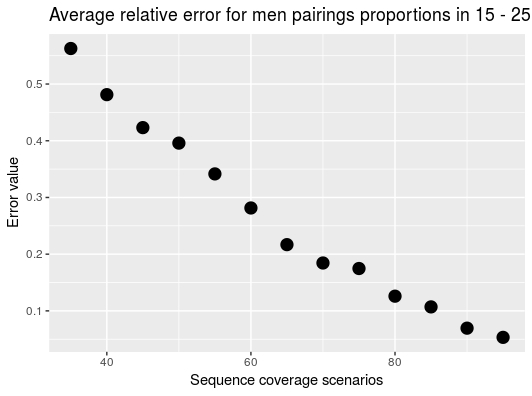
***Figure****: Women’s average age gap in pairings*

***Figure****: Women’s standard deviation age gap in pairings*

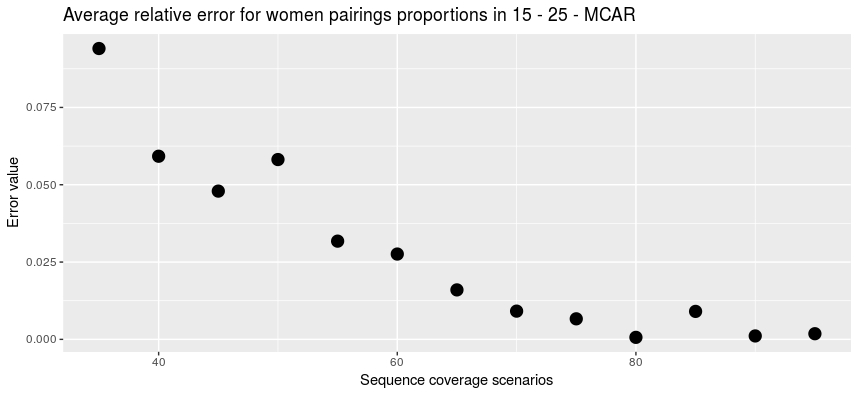
**(iv) Error: Goodness of fit**

The aim was to measure how far/close the inferences from transmission clusters are from the truth for all age mixing measurements (proportions of pairing and statistics of age gap) mentioned above. These values are computed by taking the true values from 100% of coverage and values from inference in transmission clusters. In some examples below, we considered mean relative error (MRE), which was computed by averaging the vector of difference between true value at 100% coverage and inferred values at different sequence coverage.

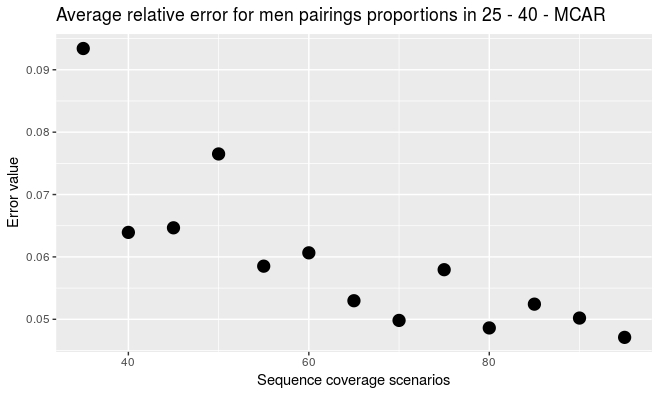
**For proportions: some decrease systematically other are sporadic**



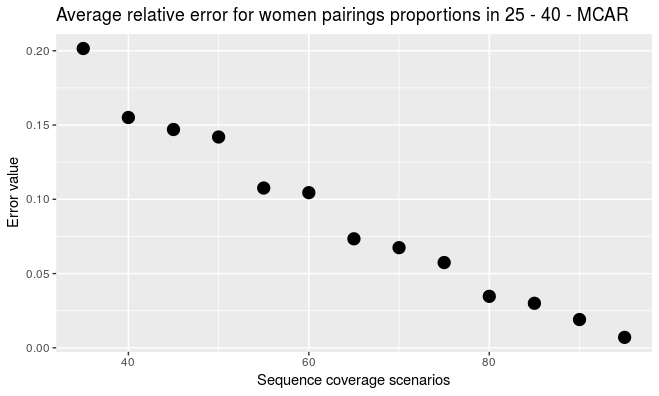
For example, for proportion of men aged between 15 – 25, we had a systematic decrease of the error as the sequence coverage increase as it can be seen in the figure above, whereas the mean relative error decrease sporadically for women in 15 and 25 years.



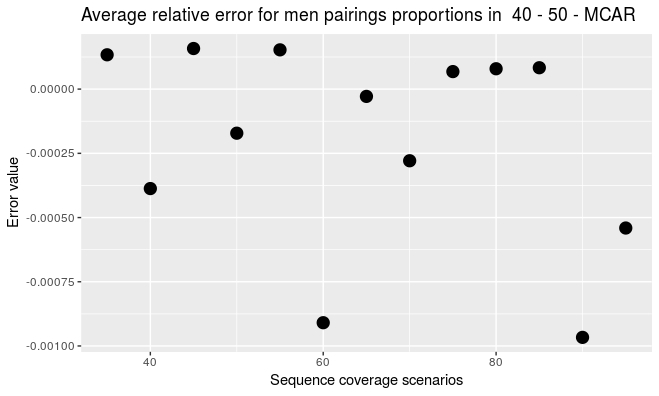
Same sporadic decrease of error is observed in pairings of men in 25 and 40 years in the figure below:

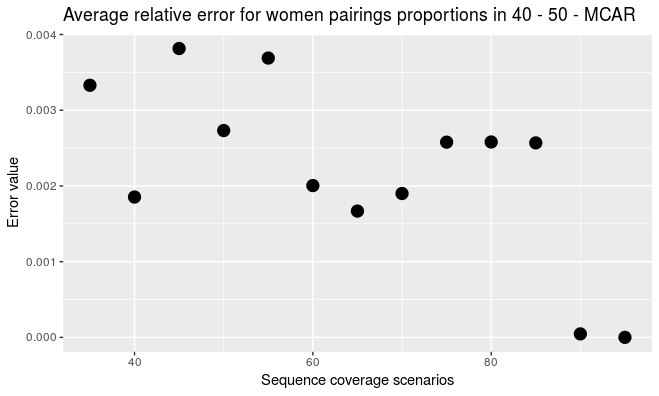


Surprisingly the error decrease systematically for women in that same age group as it can be seen in the figure below



For men and women in 40 and 50 years, the error does not seems to decrease or increase, it varies across age groups as it can be seen in the following two figures:





One explanation of a weird behavior of MRE for proportion of men/women pairings in 40 and 50 years, it may be the lack of clients in these age groups.

**Results for Missing at Random (MAR)**

Sampled individuals are drawn randomly but based on their age groups and gender. We first consider individuals in a given age group and thereafter we look on their gender. This says that if we want to sample …………..

1. Which is the 1st and 3rd quantiles [↑](#footnote-ref-2)