Exploration of duplicates in HIV notifications

Richard T. Gray

01 November 2018

This Rmarkdown document describes an exploration of the number and proportion of duplicate notifications in the Australian HIV Registry. The primary reason for this exploration is to assess the validity of using the deduplication algorithm published in:

Law, M G, A M McDonald, and J M Kaldor. “Estimation of Cumulative HIV Incidence in Australia, Based on National Case Reporting.” Australian and New Zealand Journal of Public Health 20, no. 2 (April 1996): 215-7.

for use in the Australian HIV diagnosis and care cascades.

## Settings and assumptions

We applied the deduplication algorithm to the entire set of notifications in the HIV registry to the end of 2016 and separately for notifications from the states of New South Wales (NSW) and Victoria (Vic). For these settings we estimated annual and cumulative notifications (including duplicates), the cumulative number and proportion of notifications that are unique, and the annual number and proportion of notifications that are unique each year.

The algorithm is statistical in nature and is applied to all notifications since the start of Australia’s HIV epidemic in 1980 (year of first notification). Annual estimates for a given year are obtained by applying this algorithm to all the notifications between 1980 and each given year and then taking the difference. This can lead to invalid estimates where the annual proportion unique is > 1 and the number of duplicates is negative. However, adding the annual number of duplicates will produce the correct cumulative number of unique notifications and this value is the key estimate for the HIV cascades. Thus we pretend the annual estimates are valid for the purposes of our calculations and estimates.

In recent years there is likely to be very few duplicates in the registry due to validation processes as part of HIV surveillance especially following an audit of all notifications in 2002 (I believe). So far we have simply applied the algorithm to all notifications and assumed the estimates are valid, even if the number of duplicates seems a bit high in recent years, and assumed other (uncertain) components of the cascade calculations (such as emigration) balance out any errors in the estimates for duplicates. We explored the validity of this assumption by considering scenarios where all notifications since 1992 or 2002 overall and for each state are unique (arbitrarily chosen years). We then compared the estimates for each setting and scenario.

## Results

The following figures show selected results for each population. Figures 1 and 2 show the change in cumulative notifications over time for each state under various assumptions for when all annual notifications rare unique. Figure 1 shows the majority of duplicate notifications occurred prior to 1992 and the number of duplicates from Victoria is small. Zooming in to recent years (Figure 2) shows there is little difference between the algorithm output and the assumption all notifications are unique since 2002 (though the difference is a bit larger for NSW).

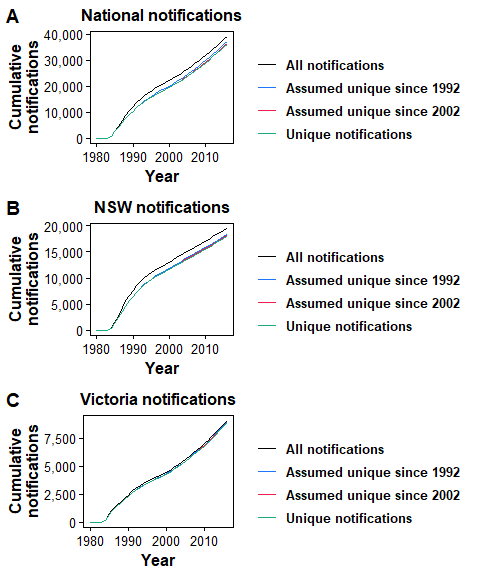
Figure 1: Cumulative total (black line) and unique notifications (green line) for A) Australia overall, B) NSW, and C) Victoria. The blue and red lines show the results when all notifications since 1992 and 2002 are assumed to be unique, respectively. 

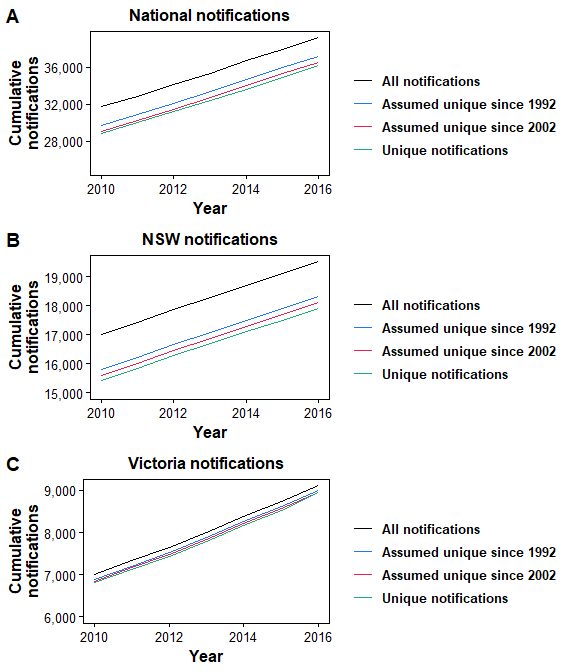
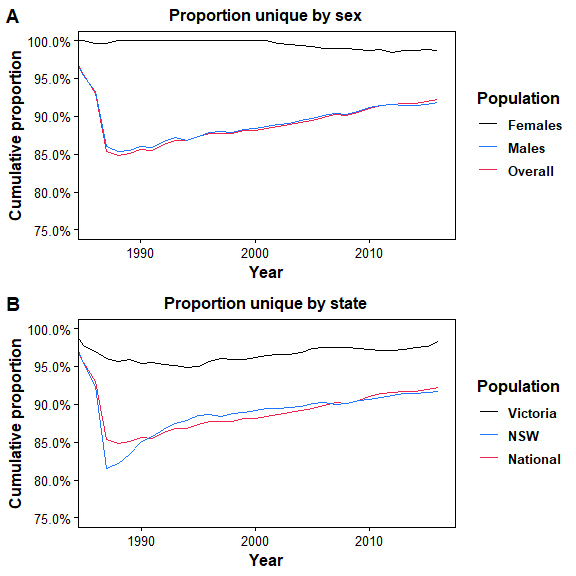
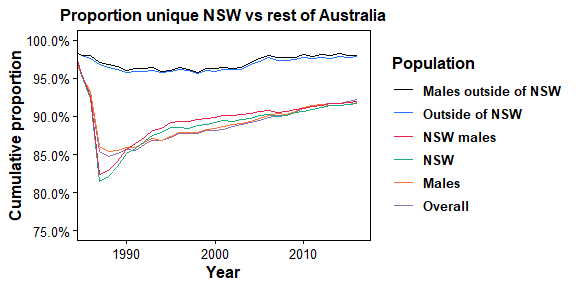
Figure 2: Zoom in on results from Figure 1. 

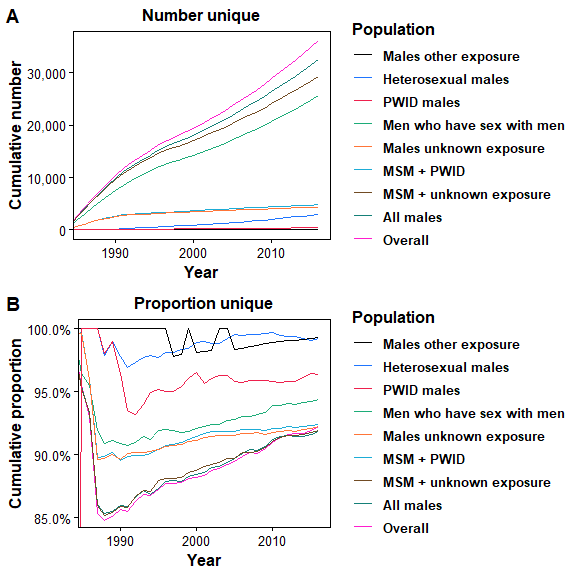
Figure 3 shows the cumulative proportion of notifications that are unique by sex and state. Over 95% of female and Overall Victorian notifications are unique. The proportion in males, overall NSW, and overall nationally are very similar according to the algorithm.

Figure 3: Cumulative proportion unique by sex and state. 

I was still a little concerned there could be a saturation effect when applying the algorithm to large numbers of notifications. For example the proportion unique is very similar for national, male, and NSW notifications. Figure 4 compares the cumulative proportion unique in notifications in NSW and outside of NSW (with similar numbers of duplicates). This figure suggests the majority of duplicates occurred in NSW males.

Figure 4: Cumulative proportion unique for NSW compared to rest of Australia. 

Looking at male notifications further, I ran the the deduplication algorithm on male notifications by exposure category. Figure 5 shows the cumulative proportion unique for each male exposure group. Considering the exposure groups separately results in an underestimate for the number of duplicates with an overall cumulative percentage of 94.6% compared to the overall male percentage of 91.9%. This suggests duplicates are being missed because duplicate notifications occur in distinct exposure categories (and hence are not classified as duplicates when the algorithm is run on the separate exposure categories). Figure 5 suggests combining the MSM and unknown exposure group notifications produces a cumulative proportion unique that aligns with the overall male and national populations.

Figure 5: Cumulative number (A) and proportion (B) unique for male exposure category groups. 

## Discussion

* Using the overall proportion unique by applying the deduplication to all notifications for males and the NSW population seems reasonable but separate calculations should be done for females and other states.
* For the 2017 ASR I did the deduplication calculations for males and females separately as required but the calculations should be done for each sub-population.
* Previously published estimates for the number unique in NSW obtained using the national estimates remain valid but NSW specific estimates will be generated in the future.
* Assuming all notifications since 2002 are unique will have little impact on the HIV cascade estimates.
* Seems like the majority of duplicates occurred in NSW males prior to 1992 particularly in MSM and males in the unknown exposure category (which makes sense I think).

## Appendix 1 - Annual duplicates and proportion unique

As stated previously when we estimated the number and proportion of duplicates annually we sometimes get a proportion > 1 and a negative number of duplicates. This is an artifact of the statistical method and calculation because we use the cumulative number to estimate annual duplicate numbers. Table 1 illustrates this using the national results from the last 10 years.

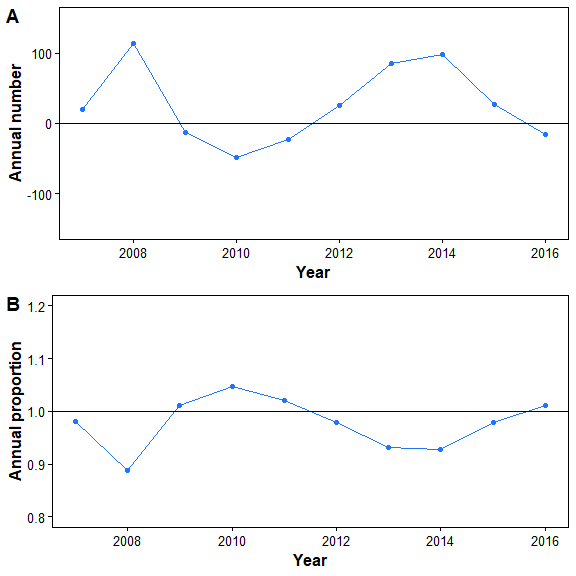
Even though proportions > 1 and negative number of duplicates don’t make sense, I pretend they do in the cascade calculations as the cumulative values do make sense and cumulative numbers are the important values in the cascade calculations (you may get a small of error going from year to year but that will be much smaller than the uncertainties in other aspects of the calculation).

Figure 6 shows the number and proportion of all annual notifications that are duplicates since 2007. The values tend to bounce around zero and 1 respectively. Note the number of annual duplicates can be a far from zero (> 100) because the numbers are obtained by taking the difference of cumulative notifications which are > 30,000.

Table 1: Annual numbers of duplicates for overall notifications. The letters and equations in brackets show how each column relates to the other.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Year | Cumulative notifications (A) | Cumulative proportion unique (B) | Cumulative number unqiue (C = A\*B) | Number unique each year (D = diff C) | Actual annual notifications (E) | Annual number of duplicates (F=E-D) | Annual proportion duplicates (G=F/D) |
| 2007 | 28,548 | 0.902 | 25,744.5 | 1,032 | 1,052 | 20 | 0.981 |
| 2008 | 29,563 | 0.901 | 26,645.5 | 901.1 | 1,015 | 113.9 | 0.888 |
| 2009 | 30,637 | 0.905 | 27,731.9 | 1,086.3 | 1,074 | -12.3 | 1.011 |
| 2010 | 31,700 | 0.91 | 28,844.1 | 1,112.2 | 1,063 | -49.2 | 1.046 |
| 2011 | 32,845 | 0.914 | 30,012 | 1,167.9 | 1,145 | -22.9 | 1.02 |
| 2012 | 34,102 | 0.916 | 31,243.4 | 1,231.4 | 1,257 | 25.6 | 0.98 |
| 2013 | 35,341 | 0.917 | 32,396.8 | 1,153.4 | 1,239 | 85.6 | 0.931 |
| 2014 | 36,680 | 0.917 | 33,638.2 | 1,241.3 | 1,339 | 97.7 | 0.927 |
| 2015 | 37,928 | 0.919 | 34,859.3 | 1,221.1 | 1,248 | 26.9 | 0.978 |
| 2016 | 39,195 | 0.922 | 36,141.4 | 1,282.1 | 1,267 | -15.1 | 1.012 |

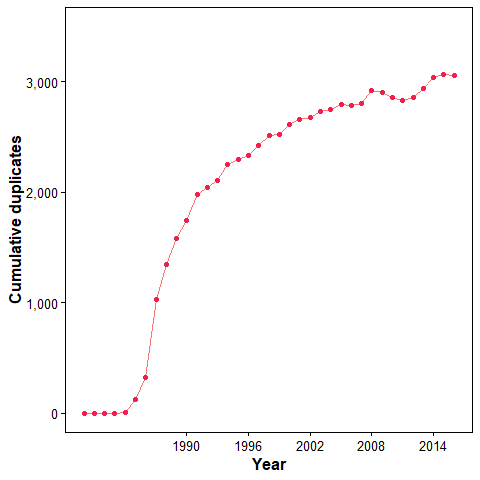
Figure 6: Number (A) and proportion (B) of annual  
notifications that are unique for all national notifications.



## Appendix 2 - Cumulative duplicates analysis

To see if there is a natural time point from where we could assume all notifications are unique I analysis the cumulative number of duplicates over 1980-2016 (shown in Figure 7. This figure suggests a rapid increase in duplicates until 1990 before a slow and steady increase since 1996. There has potentially been a stabilisation in duplicates since around 2005. In fact there is a fluctuation in cumulative diagnoses since 2007 which is unrealistic since cumulative duplicates should only increase.

Figure 7: Cumulative number of duplicate notifications over time.



To explore this further I performed a trend analysis including an exploration of change points. Figure 8 shows the results when the analysis is done on cumulative duplicates since 1980 and when cumulative duplicates are truncated from 1990. The trend in cumulative duplicates is well fitted by a segmented Poisson regression with a single change point in 1998. Since 1998 there is a steady increase in cumulative duplicates at around 1% per year. This suggests there is no clear time where we can assume all duplicates are unique. However, given the fluctuation in cumulative diagnoses since 2007, potentially the de-duplication algorithm could be stopped at this point.

Figure 8: (B) Trends in cumulative duplicates since 1980 with segmented analysis. (C) Trend and segmented trend analysis since 1990 (trends recalculated from the truncated data set).

