

Working Paper

John G. Gordon July 14, 2010

ESTIMATING HIV INCIDENCE FROM UNAIDS DATA

HIV incidence is a simple concept. It measures the rate of new HIV infections per year as a percentage of susceptible population. However, actually measuring incidence directly is extremely difficult, due in large part to the long period between HIV infection and the manifestation of AIDS symptoms – usually eight on nine years in adults. This paper will present a simple way of estimating total HIV incidence in a country or region from existing data.

The difficulties of calculating HIV incidence directly is discussed in a number of papers, including two recent publications: 1) UNAIDS has published a paper on “Methods for Estimating HIV Incidence” at http://data.unaids.org/pub/BaseDocument/2010/epi_alert_1stqtr2010_en.pdf ; 2) In a recent paper “Estimates of HIV incidence from household-based prevalence surveys” at http://journals.lww.com/aidsonline/Abstract/2010/01020/Estimates_of_HIV_incidence_from_household_based.19.aspx Hallett et al use prevalence data from DHS surveys (<http://www.measuredhs.com/>) to estimate HIV incidence for the Dominican Republic, Mali, Niger, Tanzania, and Zambia. The methods used for these studies required difficult-to-get data and advanced statistical skills.

In principle, HIV incidence can be estimated for a country or region if you have four pieces of data:

- The number of people living with HIV/AIDS (PLWHA) in a specific year (year t).
- The number of people living with HIV/AIDS in the year before year t (year t-1)
- The number of people who die of AIDS in year t
- The total population for the country or region in year t.

The formula is simple. First calculate total new infections TNI where $TNI = PLWHA(t) - PLWHA(t-1) + AIDS\ DEATHS(t)$. Then calculate the number of people susceptible to infection. There are various options on this. It would be possible to use the total population figure to estimate those susceptible to infection, but because I am using the estimates of HIV incidence to calculate new infections in a projection model I estimate the population susceptible to infection (Pop_adj) by subtracting PLWHA(t) from total population (POP). That is $Pop_adj = POP(t) - PLWHA(t)$. The formula for HIV incidence (HIV_I) is: TNI / Pop_adj as a percentage.

As an example, assume a country or a region has 1,500 people living with HIV/AIDS in 2007 (PLWHA(t)) and 1,250 people living with HIV/AIDS in 2006 (PLWHA(t-1)) and has had 100 aids deaths in 2007 year (t). The total population of the country in Year t is 100,000. In this case Total New Infections $TNI = 1,500 - 1,250 + 100 = 350$. Adjusted population $Pop_adj = 100,000 - 1,500 = 98,500$. HIV incidence = $350 / 98,500 = .0036$ or as a percentage 0.36%. Note that unless the proportion of PLWHA is quite high, the adjustment to the population figure has very little influence on the magnitude of incidence.

Fortunately, UNAIDS publishes time series data on PLWHA and AIDS deaths which can be combined with population data published by the World Bank, the United Nations Population Division or the US Census Bureau to make the calculations. Concrete examples and suggestions of using these resources is given below. In addition, Emily Oster in a paper “Roots of Infection: Exports and HIV Incidence in Sub-Saharan

Africa” (<http://faculty.chicagobooth.edu/emily.oster/papers/hivexports.pdf>) also uses UNAIDS data, but gives a more academic description of the calculations.

THE DATA.

This section will deal with accessing the data, preparing the data and calculating incidence.

Getting the Data-- UNAIDS data required for the calculations is in two Excel spreadsheets “**Estimated number of people living with HIV by country, 1990-2007**” and “**AIDS deaths in adults and children by country, 1990-2007**” both of which can be downloaded at

http://www.unaids.org/en/KnowledgeCentre/HIVData/GlobalReport/2008/2008_Global_report.asp .

The total population data is available from the World Bank World Development Indicators at

<http://databank.worldbank.org/ddp/home.do> , the United Nations Population Division at

<http://esa.un.org/unpp/> and the US Census Bureau at

<http://www.census.gov/ipc/www/idb/informationGateway.php> .

Preparing the Data – Setting up the data for a single country is relatively simple, but because of the formatting of the UNAIDS spreadsheets and the lack of consistency in country name spellings and country codes between the different organizations, setting up the calculations for multiple countries can be extremely time consuming. The data from UNAIDS provides three estimates for each year, a “best” estimate and high and a low estimates. Often there is data in only one field. If the best estimate data is available I use that, if only high and low estimates are available I use an average and if only a high or low estimate is available I use that. Further, numbers under 1,000 often include a “less than” sign in front of the number (<250)—the “<” must be removed before calculations and the number must be converted to numeric format (it may be necessary to convert all of the UNAIDS data into numeric format). For a single country this can be done by hand, for multiple countries it is best to use the Excel find and replace utility. An additional problem is that occasionally the data on deaths is not available. This is a particular problem with India which has a large HIV positive population. The India data was estimated from Nigerian death data. The Nigerian epidemic as roughly paralleled the Indian epidemic and the proportion of deaths to PLWHA in Nigeria was used to estimate the Indian data. When I am manipulating data for multiple countries, I use Microsoft Access database software, loading data from Excel after the data has been cleaned up. I find it is usually necessary to build a look-up table of the country name spellings to link the population data to the UNAIDS data.

Calculating Incidence—a sample spreadsheet with data is available [here](#). The spreadsheet contains all the necessary formulas for computing country HIV incidences. If you are not familiar with Excel, simply copy the data for the country you are dealing with into the data areas of the sample spreadsheet, remembering that PLWHA(t-1) is lagged one year. If you are calculating incidence for a region it is necessary to SUM all of the PLWHA(t), PLWHA(t-1), AIDS deaths and total population for that region for each year of the time series. Definitions of the World Bank regions are available at <http://data.worldbank.org/about/country-classifications/country-and-lending-groups> (there is a link for downloading the data at the bottom of the page).