Supplemental Appendix

December 23, 2013

1 Results By Race/Ethnicity

Table 1 shows the results broken down by race/ethnicity.

Ethnicity	TID Scenario	Incidence Model	Incidence Count (per quarter)	MSM Undiagnosed	Total HIV infected MSM*	Percentage Undiagnosed
White (n=1035)	Upper bound	Varying	33.1-40.4	402-441	4590-4629.2	8.8%-9.5%
		Constant	37	420	4608	9.10%
			33.3-40.5	203-229	4391-4417	4.6%-5.2%
Base case	Varying					
		Constant	37	214	4402	4.90%
African American (n=129)	Upper bound	Varying	3.4-7.2	61-81	519-539	11.8%-15.1%
		Constant	4.6	80	539	14.90%
	Base case	Varying	3.4-7.2	29-44	487-502	5.9%-8.7%
		Constant	4.6	423	501	8.60%
Hispanic (n=230)	Upper bound	Varying	3.6-10.2	96-122	668-694	14.4%-17.6%
		Constant	8.2	112	684	16.30%
	Base case	Varying	4.4-10.3	51-65	623-637	8.1%-10.2%
		Constant	8.2	58	631	9.30%

Table 1: Estimates of the number of undiagnosed HIV cases among MSM in King County stratified by ethnicity. (* Sum of cases thought to reside in King County based on HIV surveillance data (N= 4188, 458, and 572 respectively) and the estimated number of undiagnosed cases)

2 A Constant Incidence Example Calculation

It may be useful for some readers to see a simple illustrative example calculation using equation 1. Suppose that we observe 4 subjects, with time since last negative tests of $\{.25, .75, .25, 1\}$ years respectively and we have an incidence rate of 4 cases per year. Then if we use the upper bound estimate of the TID distribution then the probability that an infected individual remains undiagnosed for more that t is

$$P(T > t) = \begin{cases} 1, & \text{if } t < .25 \\ .5, & \text{if } .25 \leq t < .75 \\ .25, & \text{if } .75 \leq t < 1 \\ 0 & \text{if } t \geq 1 \end{cases},$$

and our estimated number of undiagnosed can then be calculated as

$$E(U) = E(Y) \int_0^\infty P(T > t)dt = 4 * (1 * .25 + .5 * .5 + .25 * .25) = 2.25.$$

3 Back-Calculation Background

Let Y_i be the number of individuals diagnosed with HIV at time $i \in \{1, ..., T\}$, and X_i be the (unobserved) number of infected at time i. It is assumed that the X_i are independently distributed Poisson with expectation λ_i, and thus the Y_i are also independently distributed Poisson with means $\sum_{j=0}^{i} \lambda_i f_{j,i-j}$, where $f_{j,d}$ is the probability that an individual infected at time j is diagnosed at time $j + d$.

Given the $f_{j,d}$ distribution, following the methodology of [Brookmeyer and Gail, 1988], which was adapted to the HIV/AIDS setting from work done in image cleaning for PET scans (See [Leahy and Qi, 2000] and references therein), we can express the log likelihood as

$$\ell(\lambda|Y = y) = \sum_i y_i log(\sum_{j=0}^{i} \lambda_i f_{j,i-j}) - \sum_{j=0}^{i} \lambda_i f_{j,i-j}.$$

Given the high dimensional nature of λ, maximizing this likelihood directly is impractical. Instead, we define a latent variable $N_{i,j}$ to be the number of infected at time i who are diagnosed at time j, which is distributed Poisson with mean $\lambda_i f_{j-i}$. The joint likelihood is then written as

$$\ell(\lambda|Y = y, N = n) = \sum_{i=1}^{T} \sum_{d=0}^{T-i} n_{i,i+d} log(\lambda_i f_d) - \lambda_i f_d$$

This likelihood can then be maximized via the EM algorithm, the E-step of which is

$$E(\ell(\lambda|Y = y, N = n)|Y = y, \lambda = \lambda') = \sum_{i=1}^{T} \sum_{d=0}^{T-i} \frac{\lambda'_i f_d}{\sum_{r=0}^{i+d} \lambda'_r f_{i+d-r}} log(\lambda_i f_d) - \lambda_i f_d,$$

which yields a fairly straightforward update in the M-step

$$\lambda_k^{(i+1)} = \frac{\lambda_k^{(i)}}{\sum_{d<T-k} f_d} \sum_{d+k<T-k} \frac{y_{k+d} f_d}{\sum_{r<k+d} \lambda_r^{(i)} f_{k+d-r}}. \quad (1)$$

3.1 Estimating the number of undiagnosed

Given a fit model, we may estimate the number of undiagnosed individuals at the mid-point of time interval j (U_j) as

$$E(U_j) = \sum_{i<j} \lambda_i (\frac{1}{2} f_{i,j-i} + \sum_{k>j-i} f_{i,k})$$

where $\sum_{k>j-i} f_{i,k}$ is the expected number of individuals infected at time i diagnosed after time j and $f_{i,j-i}$ is the expected number of individuals infected at time i and diagnosed during time period j.

4 Accounting for limited surveillance windows

If the historical data Y goes back to the beginning of the HIV epidemic, such as in [Cui and Becker, 2000], then Equation 1 is the correct update to use. However, if the diagnosis data is only available after a certain time t_0 after the start of the epidemic, then some of the Y_i are actually missing. This changes the E-step to

$$\begin{aligned} Q(\lambda|Y = y, \lambda') &= E(\ell(\lambda|Y = y, N = n)|Y_{t_0} = y_{t_0}, ..., Y_T = y_t, , \lambda = \lambda') \\ &= \sum_{i=0}^{T} \sum_{d+i<t_0} \lambda'_i f_d log(\lambda_i f_d) + \sum_{d=t_0}^{T-i} \frac{\lambda'_i f_d}{\sum_{r=0}^{i+d} \lambda'_r f_{k+d-r}} log(\lambda_i f_d) - \sum_{d=0}^{T-i} \lambda_i f_d \end{aligned}$$

and the M-step update equations become

$$\lambda_k^{(i+1)} = \lambda_k^{(i)} (a_k + \frac{c_k}{b_k})$$

where $a_k = \frac{\sum_{d-k<t_0} f_d}{b_k}$, $b_k = \sum_{d<T-k} f_d$ and $c_k = \sum_{d=t_0-k}^{T} \frac{y_{k+d} f_d}{\sum_{r<k+d} \lambda_r^{(i)} f_{k+d-r}}$.

5 Smoothing via quadratic penalties

It is well known that the maximum likelihood estimate yields noisy solutions [Leahy and Qi, 2000], whereas we expect a priori that the mean infection rates, year to year display smooth trends. [Silverman et al., 1990] proposed incorporating a smoothing step in the EM algorithm. This method was applied to HIV/AIDS data by [Becker et al., 1991]. More modern work from the PET literature focused on adding a penalty (or equivalently a prior distribution) to the log likelihood enforcing smoothness (see [Leahy and Qi, 2000] and references therein). The likelihood with a quadratic smoothing penalty is defined as

$$\ell_p(\lambda|Y = y, N = n) = \ell(\lambda|Y = y, N = n) - \gamma \sum_{i=2}^{T} (\lambda_i - \lambda_{i-1})^2$$

where γ is a positive smoothing parameter. The M-step of the EM algorithm then becomes

$$Q_p(\lambda|Y=y,\lambda') = Q(\lambda|Y=y,\lambda') - \gamma \sum_{i=2}^{T} (\lambda_i - \lambda_{i-1})^2.$$

If $\gamma > 0$, this penalty represents an a priori belief that the epidemic is in a stable state with constant incidence, which is constant with the current state of the epidemic in the United States. [Hall et al., 2008], for example, report stable infection rates from the early 1990s through 2007. [Bacchetti et al., 1993] have also applied smoothing penalties to HIV/AIDS data in order to remove noise, however they utilized a penalty that was quadratic in the log scale and implied an a priori belief that the trend in HIV infection rate is either growing or declining at an exponential rate.

5.1 Update algorithm using numeric root finding

Having the quadratic term couples the λ parameters, making the simple update equations inapplicable. However, it is possible to find a solution by maximizing the likelihood numerically. The gradient of the penalized likelihood is

$$\begin{aligned}\frac{\delta Q_p}{\delta \lambda_k} &= \frac{1}{\lambda_k}\lambda'_k(a_k b_k + c_k) - b_k - 2\gamma(\lambda_k - \lambda_{k-1}) + 2\gamma(\lambda_{k+1} - \lambda_k) \\ &= \frac{1}{\lambda_k}\lambda'_k(a_k b_k + c_k) - b_k - 4\gamma\lambda_k - 2\gamma(\lambda_{k+1} - \lambda_{k-1})\end{aligned}$$

and its hessian is a banded matrix with

$$\begin{aligned}\frac{\delta^2 Q_p}{\delta^2 \lambda_k} &= \frac{1}{\lambda_k^2}\lambda'_k(a_k b_k + c_k) - 4\gamma \\ \frac{\delta^2 Q_p}{\delta \lambda_k \delta \lambda_{k+1}} &= -2\gamma \\ \frac{\delta^2 Q_p}{\delta \lambda_k \delta \lambda_{k-1}} &= 2\gamma.\end{aligned}$$

Each step of the EM algorithm is therefore defined as the lambda such that the root of $\frac{\delta Q_p}{\delta \lambda_k}$ is attained. This can be computed via the Newton-Raphson algorithm using the banded hessian matrix. An efficient implementation of this is present in the R package rootSolve [Soetaert, 2009].

References

[Bacchetti et al., 1993] Bacchetti, P., Segal, M. R., and Jewell, N. P. (1993). Backcalculation of hiv rates. *Statistical Science*, 8(2):82–101.

[Becker et al., 1991] Becker, N. G., Watson, L. F., and Carlin, J. B. (1991). A method of non-parametric back-projection and its application to aids data. *Statistics in Medicine*, 10(10):1527–1542.

[Brookmeyer and Gail, 1988] Brookmeyer, R. and Gail, M. H. (1988). A method for obtaining short-term projections and lower bounds on the size of the aids epidemic. *Journal of the American Statistical Association*, 83(402):pp. 301–308.

[Cui and Becker, 2000] Cui, J. and Becker, N. G. (2000). Estimating hiv incidence using dates of both hiv and aids diagnoses. *Statistics in medicine*, 19(9):1165–1177.

[Hall et al., 2008] Hall, H., Song, R., Rhodes, P., and et al (2008). Estimation of hiv incidence in the united states. *JAMA*, 300(5):520–529.

[Leahy and Qi, 2000] Leahy, R. and Qi, J. (2000). Statistical approaches in quantitative positron emission tomography. *Statistics and Computing*, 10(2):147–165.

[Silverman et al., 1990] Silverman, B., Jones, M., Wilson, J., and Nychka, D. (1990). A smoothed em approach to indirect estimation problems, with particular, reference to stereology and emission tomography. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 271–324.

[Soetaert, 2009] Soetaert, K. (2009). rootsolve: Nonlinear root finding, equilibrium and steady-state analysis of ordinary differential equations.