Age and Racial/Ethnic Disparities in Per Contact Risk of HIV Seroconversion Among Men Who Have Sex with Men in the United States

Statistical Appendix

Our previous estimates of per-contact risks (PCRs)[1] were based on a so-called Bernoulli model. This model was also used by Jin *et al* [2] in their recent report, as well as other analyses of PCRs. The Bernoulli model assumes that seroconversion risk is exclusively determined by the numbers of contacts of each type reported, in combination with the corresponding PCRs. The basic Bernoulli model posits that $Pr(Y_{ij} = 1) = 1 - \prod_{k=1}^{K}(1 - \lambda_k)^{n_{ijk}}$, where Y_{ij} is an indicator for seroconversion for participant i at visit j, λ_k is the PCR for contact type $k = 1, \ldots, K$, and n_{ijk} is the number of contacts of type k reported by the participant at that visit. Infection risk is the complement of the probability of escaping infection on all contacts.

The basic Bernoulli model also assumes that the PCRs are constant for each type of contact, an assumption that almost surely does not hold. Recently, Hughes *et al* [3] used a Bernoulli model developed by Jewell and Shiboski[4] to assess PCRs for vaginal sex among serodiscordant couples in Africa. This model captures heterogeneity by allowing covariates to modify the number of contacts, using a complementary log-log link.

For MSM the problem is more difficult, because seroconversion risk is affected by both receptive and insertive anal sex contacts, with and without condom use, as well as contacts with multiple partners, including partners whose HIV serostatus is unknown or reported as negative. In our earlier work, we did attempt to let PCRs depend on covariates through a logit link, but this provided only very limited ability to model heterogeneity. To avoid these limitations, we have developed a new approach to more flexibly allow for heterogeneity in the PCRs.

Pooled logistic model

The new analysis uses a conventional pooled logistic model for seroconversion at each successive study visit. As in our earlier analysis, risk behaviors are treated as time-dependent covariates, updated at each visit, and used along with fixed demographic covariates to predict seroconversion at that same visit. In contrast to the Bernoulli model, this model allows seroconversion risk to depend directly on age, race/ethnicity, substance use, and STD, as well as the numbers of contacts; PCRs are estimated from the fitted model in a subsequent procedure, described below.

[1] Vittinghoff E, Douglas J, Judson F, McKirnan D, MacQueen K, Buchbinder SP. Per-contact risk of human immunodeficiency virus transmission between male sexual partners. Am J Epidemiol 1999;150:306-11.

[2] Jin F, Jansson J, Law M, et al. Per-contact probability of HIV transmission in homosexual men in Sydney in the era of HAART. AIDS 2010;24:907-13.

[3] Hughes JP, Baeten JM, Lingappa JR, et al. Determinants of per-coital-act HIV-1 infectivity among African HIV-1-serodiscordant couples. J Infect Dis 2012;205:358-65.

[4] Jewell NP, Shiboski SC. Statistical analysis of HIV infectivity based on partner studies. Biometrics 1990;46:1133-50.

Calculation of PCRs

After the pooled logistc model is estimated, using standard logistic regression software, we calculate the PCR for contacts of type $k = 1, \ldots, K$, in the following four steps:

1. The *actual* seroconversion R_{ij} risk for participant i at visit j is estimated with all contact counts and covariates at their observed levels.

2. The *potential* risk R_{ij}^k for participant i at visit j is estimated, assuming no contacts of type k, holding all other contact counts and covariates at their observed levels.

3. We then use the actual and potential risk to calculate the PCR for each participant-visit where at least one such contact is reported, using a Bernoulli formulation. Specifically, we assume $1 - R_{ij} = (1 - R_{ij}^k) \times (1 - \lambda_{ijk})^{n_{ijk}}$, where λ_{ijk} is the PCR for contact type k for that participant-visit, and R_{ij}, R_{ij}^k, and n_{ijk} are defined as before. This equation means that the probability of escaping infection given observed levels of exposure equals the potential probability of escaping infection from all *other* kinds of exposure, multiplied by the probabililty of escaping infection from all n_{ijk} contacts of type k. This gives $\lambda_{ijk} = 1 - [(1 - R_{ij})/(1 - R_{ij}^k)]^{1/n_{ijk}}$ for values of $n_{ijk} > 0$.

4. The marginal PCR λ_k is then calculated as average of the $\{\lambda_{ijk}\}$ across participant-visits with $n_{ijk} > 0$. Marginal PCRs for subgroups defined by age, race/ethnicity, numbers of partners, substance use, and STD are obtained by averaging over appropriate subsets of the $\{\lambda_{ijk}\}$.

We obtain confidence intervals for each PCR using the bias-corrected percentile bootstrap, with resampling of participants rather than participant-visits, to capture potential within-subject correlation. The resampling for the combined VPS, Explore, and Vaxgen data is also stratified on cohort, so that cohort sample sizes are preserved in each bootstrap sample. We also used bootstrapping for inference on pairwise differences in PCRs between subgroups, as well as between PCRs for different types of contact.

Capturing heterogeneity

Our new approach potentially captures heterogeneity two ways:

1. The pooled logistic model uses cubic spline transformations of the contact counts, allowing for PCRs that differ according to the numbers of contacts reported, possibly reflecting frailty selection, steady relationships, or differential reporting errors.

2. The model is multiplicative, so that for any given value of n_{ijk}, the difference between R_{ij} and R_{ij}^k, and hence λ_{ijk}, is larger if R_{ij} is increased by covariates including age, race/ethnicity, numbers of contacts, substance use, and STD.