Web appendix

Search strategies for PubMed and Embase:

Search strategy for PubMed:

("HIV Infections/transmission"[Mesh] OR "HIV Infections/virology"[Mesh] OR ("HIV"[All Fields] AND ("Infection"[All Fields] OR "Transmission"[All Fields]))) AND("Disease Transmission, Infectious"[Mesh] OR "Transmission"[All Fields] OR "Transmit"[All Fields] OR "Infect"[All Fields]) AND ("Infectiousness"[All Fields] OR "Infectivity"[All Fields] OR "Transmission Rate"[All Fields] OR "Transmission Probability"[All Fields] OR ("Rate"[All Fields] AND "Transmission"[All Fields]) OR "Transmission Risk"[All Fields]) AND (("Viral Load"[All Fields] OR "Viral Load"[Mesh] OR "Viral Burden"[All Fields] OR ("RNA"[All Fields] AND "Plasma"[All Fields]) OR ("Concentration"[All Fields] AND "Blood"[All Fields]) ) OR ("Early Infection"[All Fields] OR "Acute Phase"[All Fields] OR "Early Stage"[All Fields] OR "Acute Disease"[Mesh] OR "Seroconversion Period"[All Fields] OR "Primary Infection" [All Fields] OR "Primary HIV Infection"[All Fields] OR "Concordant"[All Fields] )) NOT ("Infectious Disease Transmission, Vertical"[Mesh] OR "Vertical Transmission"[All Fields] OR "Perinatal"[All Fields] OR "Mother-To-Child"[All Fields] OR "PMTCT"[All Fields] OR "Mother To Child"[All Fields])

Search strategy for Embase:

 (('human immunodeficiency virus infection'/exp) OR ('HIV' NEAR/5 ('infection' OR 'transmission'))) AND (('disease transmission'/exp) OR ('sexual transmission'/exp) OR 'transmission' OR 'transmit' OR 'infect') AND ('infectiousness' OR 'infectivity' OR 'transmission rate' OR 'transmission probability' OR ('rate' NEAR/5 'transmission') OR 'transmission risk' OR 'infection rate') AND (('viral load' OR ('virus load'/exp) OR 'viral burden' OR ('RNA' NEAR/5 'plasma') OR ('concentration' NEAR/5 'blood')) OR (('acute disease'/exp) OR 'early infection' OR 'acute phase' OR 'early stage' OR 'acute disease' OR 'seroconversion period' OR 'primary infection' OR 'primary HIV infection' OR 'concordant')) NOT (('vertical transmission'/exp) OR 'vertical transmission' OR 'perinatal' OR 'mother-to-child' OR 'PMTCT' OR 'mother to child')

Transmission rates and probabilities

Let n be the number of sex acts in a year, p the transmission probability per sex act and r the transmission rate per year. The probability of getting infected in one year is either 1-(1-p)n (i.e. 1 minus the probability of not getting infected in any of the n acts) or 1-e-r (from exponential definition of a rate). By setting them equal we get

1-(1-p)n=1-e-r

=> (1-p)n=e-r

=> log((1-p)n)=log(e-r)

=> n\*log(1-p)=-r

=> r=-n\*log(1-p),

which is the formula used in the main text.

**Figure S1: Relationships between transmission risk and viral load using categorized viral load values**

4 studies reported overall step functions, 6 studies reported stratified step functions.

GUD: genital ulcer disease

**Figure S2: Rate ratio per log10 increase in viral load (all studies) **

**Re-analysis of Zambian data**

Fideli *et al* [1] was a serodiscordant couples study from Zambia, which included 1022 serodiscordant couples. This study was chosen because individual viral load values could be extracted. They observed a total of 162 transmissions during a mean follow-up time of 22 months. Viral loads were measured in 104 transmitters and 207 non-transmitters. We extracted the viral load values of Fideli *et al* with *PlotDigitizer 2.5.0*. We used the mean follow-up time as the time of transmission for all persons who transmitted HIV and as the censoring time for those who did not transmitters. Since the Zambian data only measured viral load values for a subset of all individuals (n=311), we used multiple imputation for the viral loads of the remaining HIV-positive partners (n=711). We separately imputed viral load values of transmitters and non-transmitters based on the observed viral load values using hot deck imputation.

We estimated the transmission rate based on three different models, namely the Cox proportional hazard model, the Hill function and a step function. We show the estimated transmission rates in Figure 2 (bottom panel) in the main text. The figure is qualitatively similar to Figure 2 (top panel), although the heterogeneity is reduced, especially for lower viral load values. This is not surprising, since Figure 2 (top panel) includes heterogeneity in both the data and the relationship between viral load and transmission rate, while Figure 2 (bottom panel) only includes the latter. We estimated the following parameter values for the functions r(x) or r(z) where x is the log10 viral load and z=10x is the viral load:

Cox function r(x) = c\*bx:

c=0.0041 (0.0021; 0.0075);

b=1.91 (1.70; 2.19)

Hill function $r\left(z\right)=\frac{β\_{max}z^{β\_{k}}}{z^{β\_{k}}+\left(β\_{50}\right)^{β\_{k}}}$:

βk=0.96 (0.59; 1.47);

βmax=0.15 (0.12; 0.22);

β50=17,617 (8,392; 83,650)

For the Cox function, this is in line with other studies. The estimate of Fideli *et al* was b=1.8 for Male-to-Female Transmission and b=2.5 for Female-to-Male transmission. The parameters for the Hill function are slightly different from the results in Fraser *et al* because we used multiple imputation techniques while they disregarded data from patients without viral load values. The main difference is that our βmax parameter is lower than in Fraser *et al*. (βk=1.02, βmax=0.317, β50=13,938). However, the functions are qualitatively similar.

In an additional analysis, we studied which underlying model most accurately represents the data. Since actual follow-up times were not available in the data by Fideli *et al*, we did not use standard model selection procedures. Instead, we simulated transmission times based on the three models and then compared the analysis of the simulated datasets with the analysis of the actual data. We considered three different simulated datasets. Each time we used the imputed viral load values from Fideli *et al* and simulated transmission times. We simulated times assuming the hazard function was a Cox function, a Hill function or a step function. For each of these simulated datasets we estimated the hazard functions based on all three models. Figure S3 shows the results of the analyses of these simulated datasets. In the simulations based on the Hill function and based on the step function we found similar results as in the real dataset (compare Figure S3, panels (b) and (c) with Figure 2). The data that was simulated based on the Cox model (Figure S3, panel (a)) led to different parameter estimates in the Hill function as compared with the real dataset (Figure 2). Figure S3 (a) also shows that even though the Hill function is mathematically saturating, this saturation may only occur above any meaningful viral load values. We conclude that the Cox model (which is most commonly used) does not reflect the data well. We therefore suggest investigating the Hill function further, since it leads to the best estimate with the fewest parameters.

*References*

1. Fideli, U.S., et al., *Virologic and immunologic determinants of heterosexual transmission of human immunodeficiency virus type 1 in Africa.* AIDS Res Hum Retroviruses, 2001. **17**(10): p. 901-10.

**Figure S3:** Transmission rates in simulated datasets, based on different assumptions about the underlying data. In panel (a), the simulated transmission times are based on the Cox model and the estimated transmission rates based on these data are shown. In panel (b) the data are simulated based on transmission rates from the Hill function and in panel (c) the data are simulated based on the step-function.

