Appendix: Estimating the number of people living with HIV and undiagnosed fraction in Spain in 2013

Table A1. Characteristics of National Registry of AIDS and new HIV diagnoses surveillance system in Spain

National Registry of AIDS	**HIV new diagnoses surveillance system**
Collects new AIDS cases following European case definition	Collects new HIV diagnosis following European case definition
Initiated in 1983 at national level. Collects data retrospectively since 1981	Initiated in 2007 at national level although some regional systems started earlier. Collects data retrospectively since 2003 in 9 out of 19 Autonomous Regions in Spain
Coverage: 100% from the beginning	Coverage increased from 34% in 2003 (9 Autonomous Regions reported data) to 100% in 2013.
Compulsory notification	Compulsory notification
Notifiers are clinicians but data are complemented from other data sources (mortality registry, clinical records, hospital discharge registries)	Notifiers are clinicians but data are complemented from other data sources (laboratory registries, clinical records, hospital discharge registries)
Main variables: age, sex, transmission mode, country of birth, date of diagnosis of all AIDS cases, date of the first positive HIV test	Main variables: age, sex, transmission mode, country of birth, date of diagnosis, CD4 count, date of first CD4 count after HIV diagnosis

Table A2. Estimated cumulative deaths, fraction of undiagnosed infections and number of persons living with HIV in Spain in 2013 by transmission category.

Transmission category	Figure	Estimates (95% CI)
MSM	Undiagnosed fraction (%)	18.8 (15.0 - 22.8)
	PLHIV	55079 (52025 - 58487)
	deaths	7533
IDU	Undiagnosed fraction (%)	3.5 (2.3 - 6.1)
	PLHIV	23753 (19255 - 28645)
	deaths	54265
Heterosexual	Undiagnosed fraction (%)	20.1 (15.2 - 26.3)
	PLHIV	28384 (26091 - 31240)
	deaths	12989
Total	Undiagnosed fraction (%)	18.0 (14.3 - 22.1)
	PLHIV	141000 (128000 - 155000)
	deaths	80390

Figure A1: Probabilities of diagnosis from CD4 stages over the studied period.

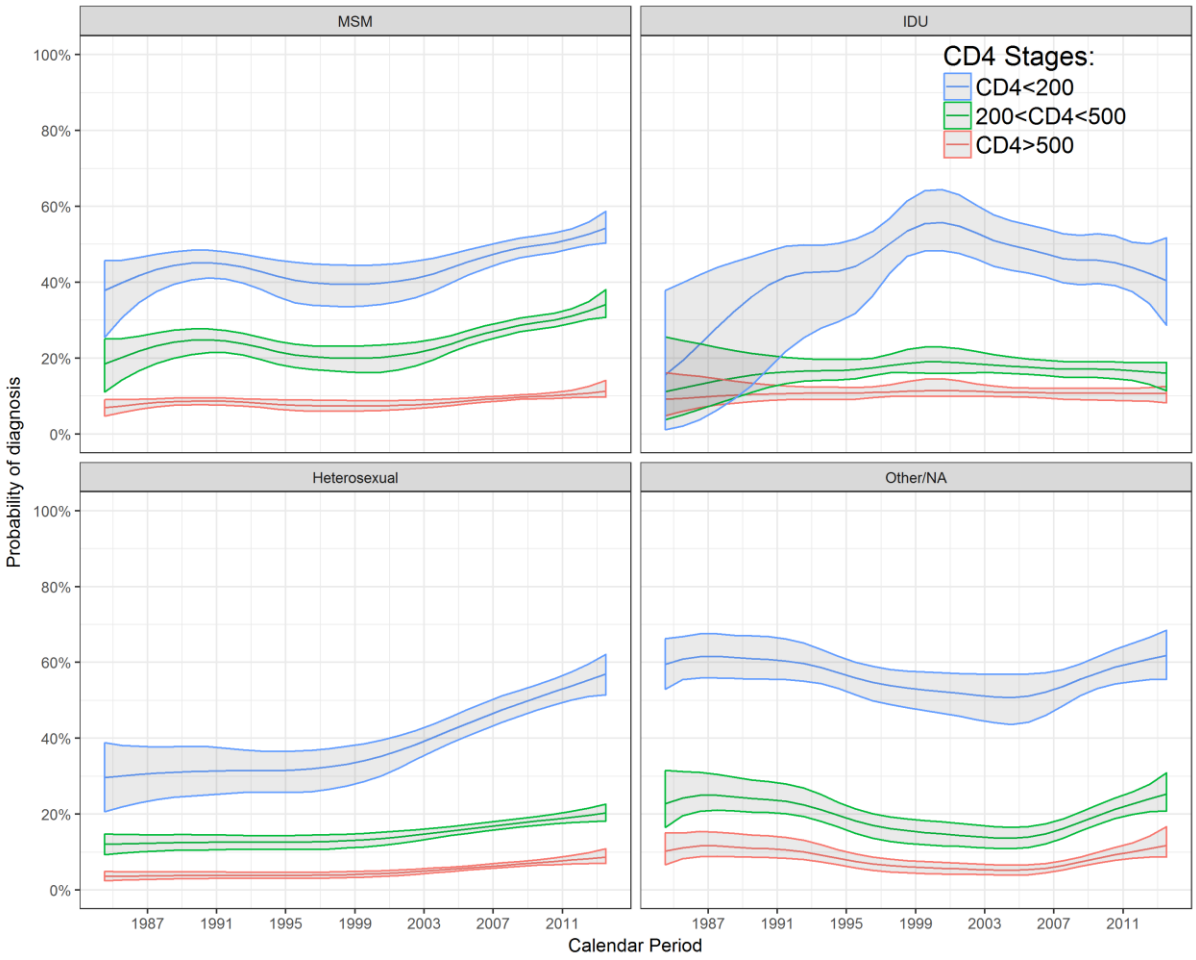


Figure A2: Estimated proportion of non-late HIV diagnoses coming from each CD4 stage over the period 2003-2013. Dots are the proportions observed in the HIV surveillance system.

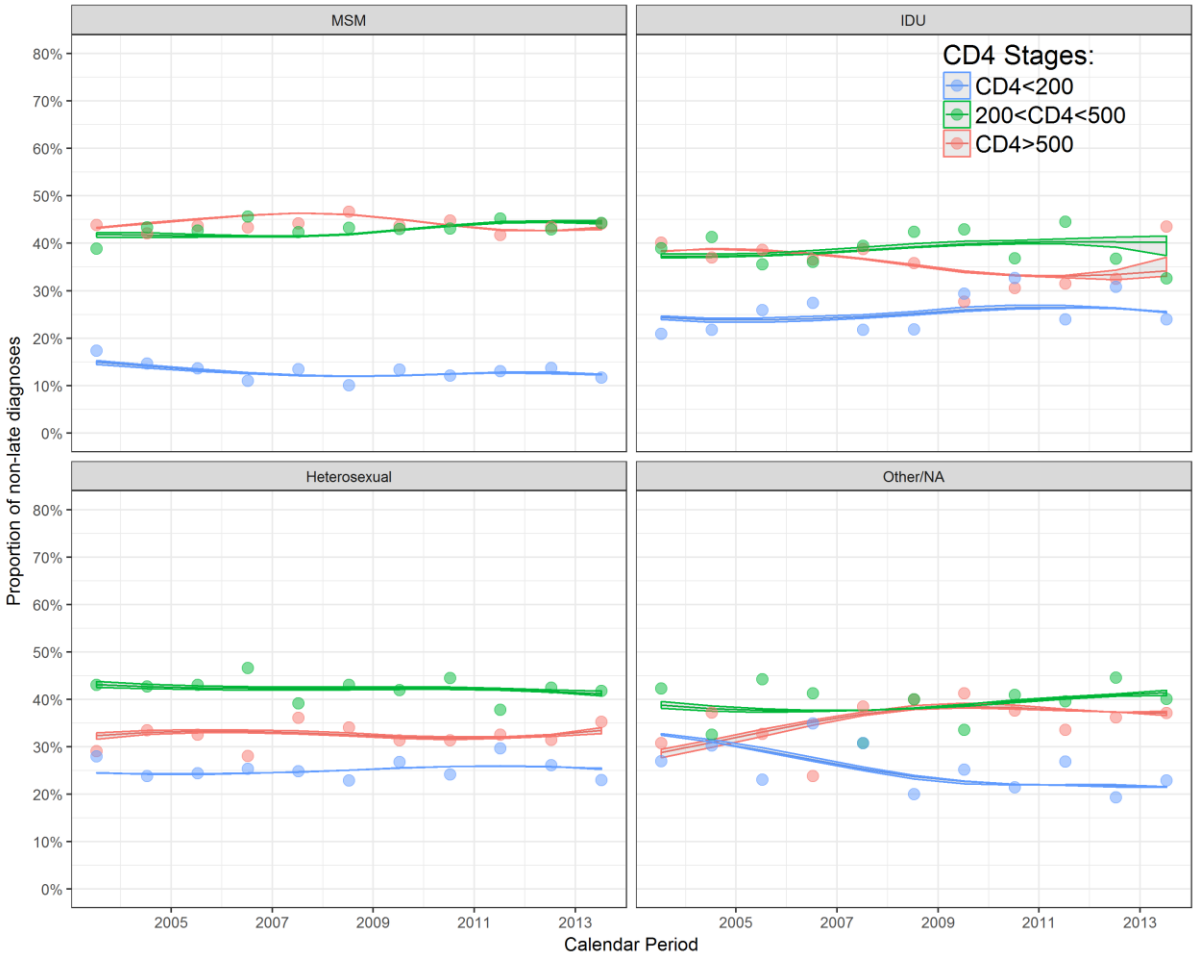


Figure A3: Estimated probability of HIV as a cause of death by transmission category.

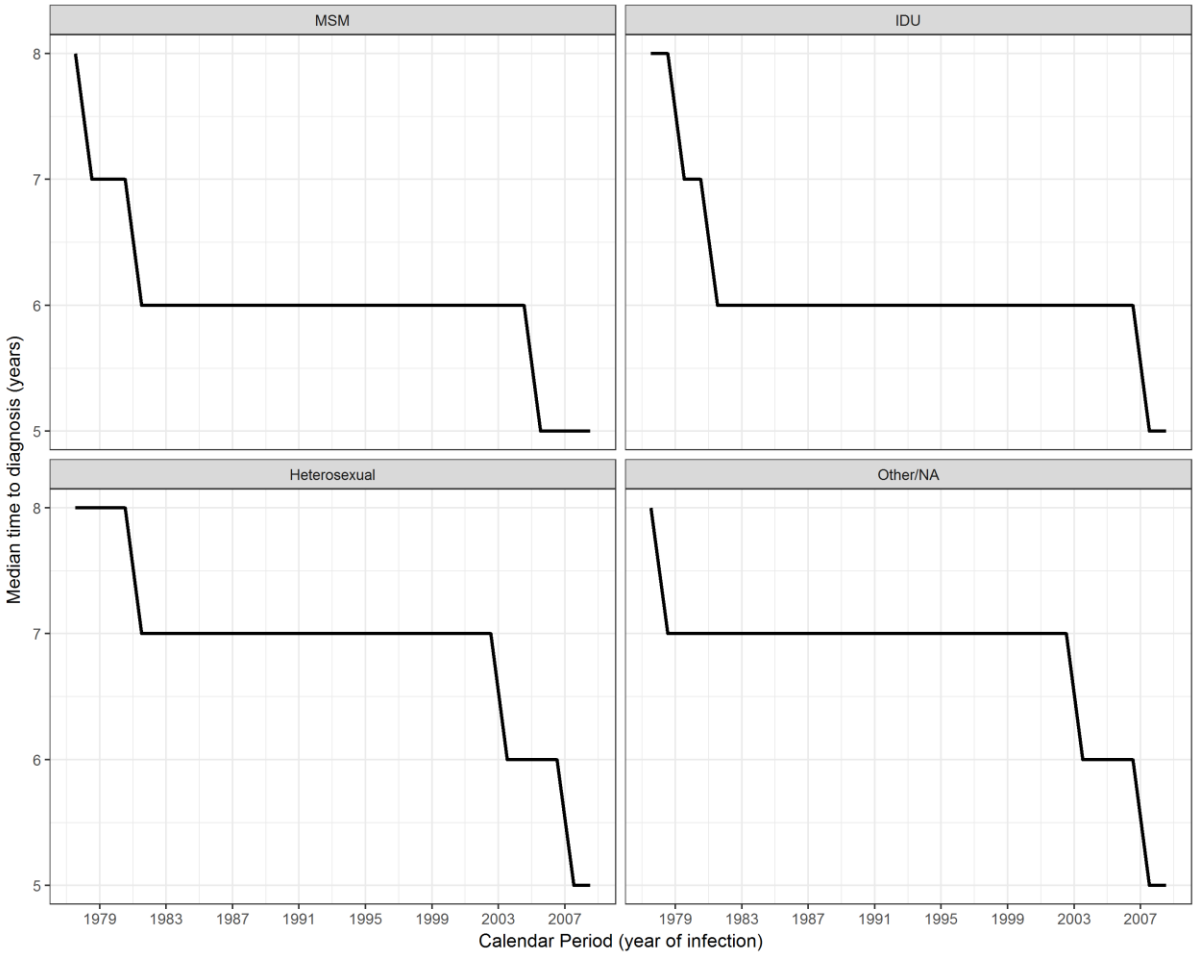


Figure A4: Estimation of the annual number of new diagnoses for different Spanish regions. Dots are the new diagnoses reported in the HIV surveillance system.

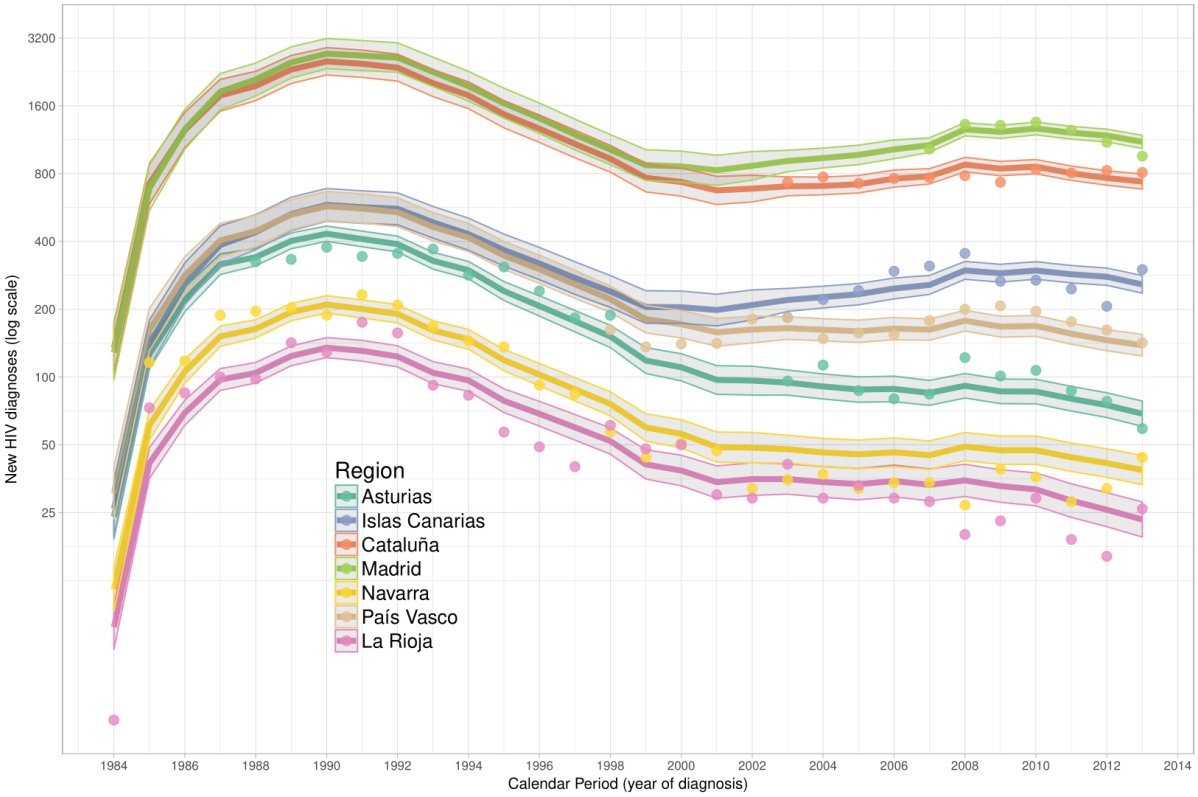
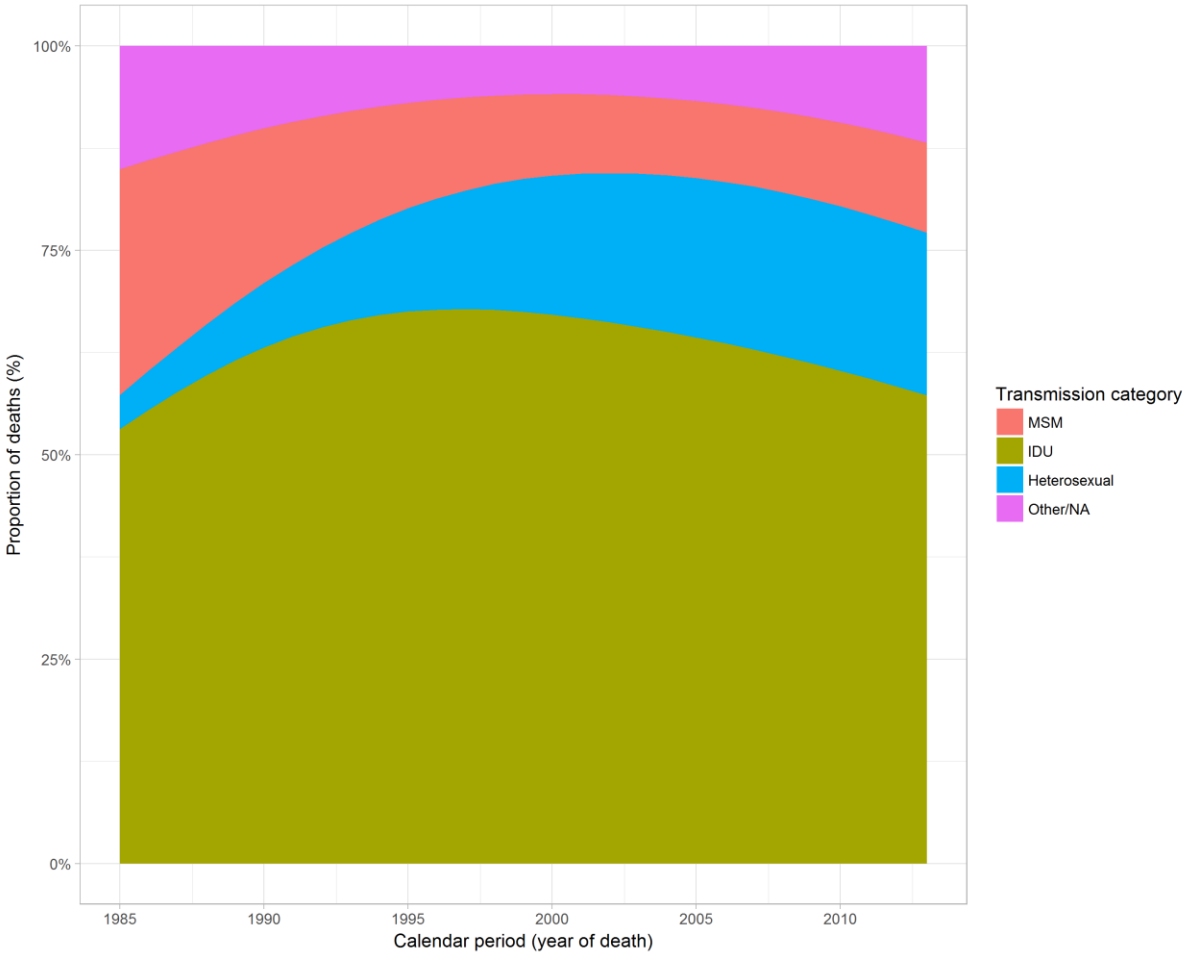


Figure A5: Proportion of each transmission category among observed deaths in the AIDS surveillance system over the period 1985-2013.



Bayesian formulation of the model:

Step1: Estimation of the annual number of new HIV diagnoses

Let H_{rt} denote the number of new HIV diagnoses observed in a transmission category at year *t* and in region *r*, and $\widetilde{H}_{rt}$ the restriction of H_{rt} to people who developed AIDS. These numbers are jointly modeled in the following way:

$$H_{rt} = \text{Poisson}(N_{rt}\rho_{rt}),$$
$$\widetilde{H}_{rt} = \text{Poisson}(N_{rt}\tilde{\rho}_{rt}),$$

where the offset N_{rt} is the general population size, and the rates ρ_{rt} and $\tilde{\rho}_{rt}$ vary according to the following equations:

$$\log\tilde{\rho}_{rt} = \tilde{u}_r + \tilde{v}_t,$$
$$\log\rho_{rt} = \beta_1\tilde{u}_r + u_r + \beta_2\tilde{v}_t + v_t,$$

The spatial terms u_r and $\tilde{u}_r$ are random effects with normal prior $N(0,\tau_R)$ and $N(0,\tilde{\tau}_R)$, respectively. The smoothness on temporal terms v_t and $\tilde{v}_t$ is induced by assuming a second order random walk prior:

$$v_{t+2} = N(2v_{t+1} - v_t, \tau_T),$$
$$\tilde{v}_{t+2} = N(2\tilde{v}_{t+1} - \tilde{v}_t, \tilde{\tau}_T),$$

where v_1, v_2, $\tilde{v}_1$ and $\tilde{v}_2$ are treated as fixed unknown parameters with a normal prior $N(0,\tau_0)$.

Step2: Estimation of the HIV infection incidence

As previously, the smoothness over time of the annual HIV incidence h_t and the diagnosis probability d_{kt} from each CD4 stage (k=1,2,3) was obtained by assuming second order random walk priors on this terms:

$$h_{t+2} = N(2h_{t+1} - h_t, \tau_h),$$
$$\eta_{t+2} = N(2\eta_{t+1} - \eta_t, \tau_\eta),$$

where

$$d_{kt} = \text{logit}(a_k + b_k\,\eta_t),$$

and a_k, b_k are unknown parameters with a normal prior $N(0,\tau_0)$.

Let p_{tkl} denote the one-step transition probability from state k to state l at time t (year) associated with the back-calculation model formulated in Figure 1 of the manuscript. Using the notations used in this figure, we have that the transition probability matrix

$$P_t = (p_{tkl}) = \begin{pmatrix} p_{t11} & q_{12}(1-d_{1t}) & 0 & 0 & d_{1t} \\ 0 & p_{t22} & q_{23}(1-d_{2t}) & 0 & d_{2t} \\ 0 & 0 & p_{t33} & q_{34}(1-d_{3t}) & d_{3t} \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

where $p_{tkk} = 1 - \sum_{l>k}^{5} p_{tkl}$ for $k = 1,2,3$.

Given this matrix, the expected number (prevalence) E_{tk} of persons in state k at time t is obtained using the following recurrence equations [1]:

$$\begin{aligned} E_1 &= I_1, \\ E_t &= P_t' E_{t-1} + I_t, \qquad t \geq 2 \end{aligned}$$

Where P_t' is the transpose of P_t and I_t is a 5-dimensional column vector with the annual infection incidence h_t as the first component, and all other components equal to 0.

In the same way, the expected number of new arrivals (incidence) e_{tk} of persons in state k at time t is obtained using the recurrence relations:

$$\begin{aligned} e_1 &= I_1, \\ e_t &= \bar{P}_t' E_{t-1} + I_t, \qquad t \geq 2 \end{aligned}$$

where $\bar{P}_t$ is the same matrix as P_t but with zeros in its diagonal.

From this formulation, the expected proportion of non-late HIV diagnoses coming from each CD4 stage ($k = 1,2,3$) over time is

$$\theta_{tk} = \frac{E_{tk} d_{tk}}{e_{t5}}$$

Let denote r_{t1}, r_{t2} and r_{t3}, the observed numbers of new non-late HIV diagnoses at time t with an available CD4 count at diagnosis greater than or equal to 500, between 200 and 500, and less than 200, respectively. Then, it is assumed that

$$r_t = \text{Multinomial}(\theta_t, R_t),$$

where $R_t = r_{t1} + r_{t2} + r_{t3}$.

The expected number of new diagnoses for year t is

$$n_t = \sum_r N_{rt}\rho_{rt}.$$

Let α_t denote the corresponding proportion of late HIV diagnoses observed in the surveillance system. Expectations e_t^4 and e_t^5 can be approximated by the estimated number of late and non-late HIV diagnoses, respectively:

$$\hat{e}_t^4 = \alpha_t n_t,$$

$$\hat{e}_t^5 = (1-\alpha_t)n_t.$$

The two steps described in this Bayesian formulation are connected using an augmented model strategy by the creation of faked zero observations:

$$0 = (\log e_t^4 - \log \hat{e}_t^4) + \varepsilon_t^4$$

$$0 = \left(\log e_t^5 - \log \hat{e}_t^5\right) + \varepsilon_t^5 ,$$

where ε_t^4 and ε_t^5 are normal random deviations with arbitrary high precision:

$$\varepsilon_t^4 = N(0, \tau_\varepsilon),$$

$$\varepsilon_t^5 = N(0, \tau_\varepsilon).$$

Model implementation:

The following code is an implementation using JAGS [2] of the Bayesian model described in the paper to estimate the infection incidence. Flat priors were used for all parameters, except for temporal components whose prior distribution was slightly informative in order to obtain a smooth estimate of the HIV incidence curve. The estimation was perform using 40 MCMC chains with a burn-in period of 2,000. Algorithm convergence was monitored using the Gelman-Rubin diagnostic [3].

```
data
        Nt,# Number of years
        Nr,# Number of regions
        N, # Dimension of the data: number of years by number of regions (N=Nr*Nt)
        y[N], # Observed/missing annual HIV diagnoses from the HIV surveillance system
        y_s[N], # Observed annual HIV diagnoses from the AIDS surveillance system
        pop[N], # population sizes
        X[Nt,N], # design matrix
        alpha[Nt], # proportion of concurrent AIDS-HIV diagnoses
        zeros[Nt],zeros_s[Nt], # faked zeros observations
        CD4[Nt,3], # annual number of CD4 data by CD4 stage
        nCD4[Nt], # annual number of CD4 data
        prec, ## global precision for the parameters (prec=1e-4)
```

```
variables
        Mu[Nt] # Expected annual number of diagnoses
        r[N], # region index
        t[N], # time index
        u[Nr], # region effects
        v[Nt], # time effects (autoregressive prior)
        mu[N], # expected rates of HIV diagnosis in the HIV surveillance system
        mu_s[N], # expected rates of HIV diagnosis in the AIDS surveillance system
        P[Nt-1,5,5], # Matrix of transition probabilities
        P0[Nt-1,5,5], # Working matrix
        E[Nt,5], # Expected annual prevalence in the different states
        e[Nt,5], # Expected annual incidence in the different states
        d[Nt-1,3], # Annual diagnosis probabilities
        prop[Nt-1,3], # Expected proportion of non-late HIV diagnoses coming from each CD4
        h[Nt], # Expected annual incidence of the HIV infection
        w[Nt], # autoregressive prior for the time variations of the HIV infection
        g[Nt], # autoregressive prior for the time variations of the diagnosis probabilities

model{

  ###############################################################################
  ################## Step 1: Estimation of the annual number of diagnoses incidence ##
  ###############################################################################

  for(i in 1:N) { ## Response
       y[i] ~ dpois(mu[i]) #HIV diagnosis counts in the SINIHIV
       y_s[i] ~ dpois(mu_s[i]) #HIV diagnosis counts in the RNS
    }
  for(i in 1:N) { ## Expected response
      log(mu[i]) <- log(pop[i]) + u[r[i]] + v[t[i]] + beta.r*u_s[r[i]] + beta.t*v_s[t[i]]
      log(mu_s[i]) <- log(pop[i]) + u_s[r[i]] + v_s[t[i]] #
    }

  for(k in 1:Nr){ ##  prior for region effects
    u[k] ~ dnorm(0,tau.r)
    u_s[k] ~ dnorm(0,tau_s.r)
  }
  for(l in 3:Nt) { ## prior for period effects (2nd order random walk)
    v[l]  ~ dnorm(v[l-1]*2-v[l-2], tau.t)
    v_s[l]  ~ dnorm(v_s[l-1]*2-v_s[l-2],tau_s.t)
  }
  ## Priors for the parameters
  v[1] ~ dnorm(0,prec)
  v[2] ~ dnorm(0,prec)
  v_s[1] ~ dnorm(0,prec)
  v_s[2] ~ dnorm(0,prec)
  beta.r ~ dnorm(0,prec)
  tau.r ~ dgamma(1,prec)
  tau_s.r ~ dgamma(1,prec)
  beta.t ~ dnorm(0,prec)
  tau.t ~ dgamma(1,prec)
  tau_s.t ~ dgamma(1,prec)



  ###############################################################################
  ################## Step 2: Estimation of infection incidence ######################
  ###############################################################################

  Mu <- X.t %*% mu #annual number of HIV diagnoses across the country

  # Faked zero observations to link the two steps
  for (i in 1:Nt){
   zeros[i] ~ dnorm(log(e[i,5]+eps)-log(Mu[i]*(1-alpha[i])+eps),1e3)#non late HIV diagnosis
   zeros_s[i] ~ dnorm(log(e[i,4]+eps)-log(Mu[i]*alpha[i]+eps),1e3)#concurrent HIV-AIDS diag.
   }
  for(i in 1:(Nt-1)) {
     CD4[i+1,] ~ dmulti(prop[i,],nCD4[i+1])
     }
```

```
eps <- 1e-3  #to avoid numerical error with the log

# Expected proportion of non late HIV
prop <- E[1:(Nt-1),1:3] * d + eps

# Expected prevalence and incidence
E[1,] <- rep(0,5)
e[1,] <- rep(0,5)
h[1] <- exp(w[1])

for(i in 2:N){
   E[i,] <- t(P[i-1,,])%*%E[i-1,] + f*h[i]
   e[i,] <- t(P0[i-1,,])%*%E[i-1,] + f*h[i]
   h[i] <- exp(w[i])
}

# Fraction of recent infection to CD4 stages
#f <- c(.75,.25,0,0,0) ## Cori (2015)
f <- c(1,0,0,0,0) ## Aalen (1997)


# output: PLHIV and undiagnosed fraction (F)
PLHIVD <- (sum(e[,4])+sum(e[,5])-deaths)
PLHIV <- (sum(e[,1])-deaths)
F<-(1-PLHIVD/PLHIV)*100

# Transition Probability Matrix
for(i in 1:(Nt-1)){
   for(col in 1:4){ for(row in (col+1):5){ P[i,row,col] <- 0 }}
   for(col in 1:5){ for(row in col:5){ P0[i,row,col] <- 0 }}
   P[i,1,2] <- q[1]*(1-d[i,1])
   P[i,1,3] <- 0
   P[i,1,4] <- 0
   P[i,1,5] <- d[i,1]
   P[i,2,3] <- q[2]*(1-d[i,2])
   P[i,2,4] <- 0
   P[i,2,5] <- d[i,2]
   P[i,3,4] <- q[3]*(1-d[i,3])
   P[i,3,5] <- d[i,3]
   P[i,4,5] <- 0
   for(k in 1:5){ P[i,k,k] <- 1 - sum(P[i,k,index[k,]]) }
   for(row in 1:4){ for(col in (row+1):5){ P0[i,row,col] <- P[i,row,col] }}
}


#Diagnosis probabilities equation
for(i in 1:(n0-1)){ d[i,]= rep(0,3) } #before 1984
for(i in n0:(Nt-1)) {
   for(k in 1:3) {
         logit(d[i,k]) <- a[k]+b[k]*(g[i-n0+1]-mg)
         }}


#Probabilities of natural progression
q=c(1/5.5,1/4,1) ## Aalen (1997)
#q=c(1/3.32,1/8.20,1) ## Cori (2015)


# Prior on HIV incidence curve (2nd order random walk)
for (i in 3:Nt){
   w[i] ~ dnorm(diff.w[i],tau.w)
   diff.w[i] <- w[i-1]*2-w[i-2]
}

# Prior on diagnosis probabilities variations  (2nd order random walk)
mg=mean(g[1:(Nt-n0)])
for (i in 3:(Nt-n0)){
    g[i] ~ dnorm(diff.g[i],tau.g)
    diff.g[i] <- g[i-1]*2-g[i-2] #RW2
```

```
    }

    # initial values
    w[1] ~ dnorm(0,prec)
    w[2] ~ dnorm(0,prec)
    g[1] ~ dnorm(0,prec)
    g[2] ~ dnorm(0,prec)

    for(k in 1:3){
        b[k] ~ dnorm(0,prec)
        a[k] ~ dnorm(0,prec)
        }

    # smooth precision
    tau.w ~ dgamma(10,prec)
    tau.g ~ dgamma(10,prec)
}
```

References

1. Sweeting MJ, De Angelis D, and Aalen OO. Bayesian back-calculation using a multi-state model with application to HIV. *Statistics in medicine* 2005; 24:3991–4007.
2. Plummer M. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. [JAGS] *Hornik K, Leisch F, Zeileis A (eds) Proceedings of the 3rd international workshop on distributed statistical computing (DSC 2003)* 2003, Vienna.
3. Gelman, A., and D. B. Rubin. 1992. Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science* 7: 457–511.

Fig1

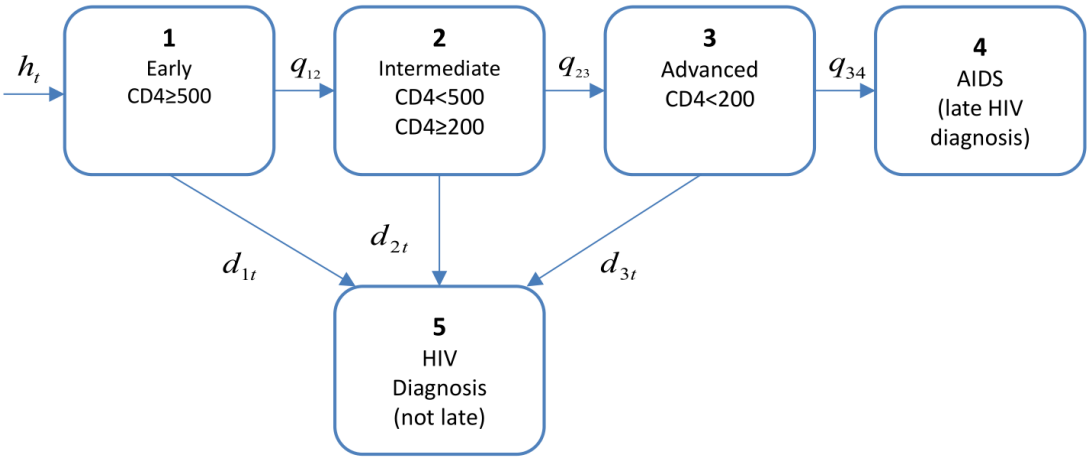


Fig. 1: Multi-state model for the disease progression from HIV infection to diagnosis from Sweeting *et al.* [6].

Fig2

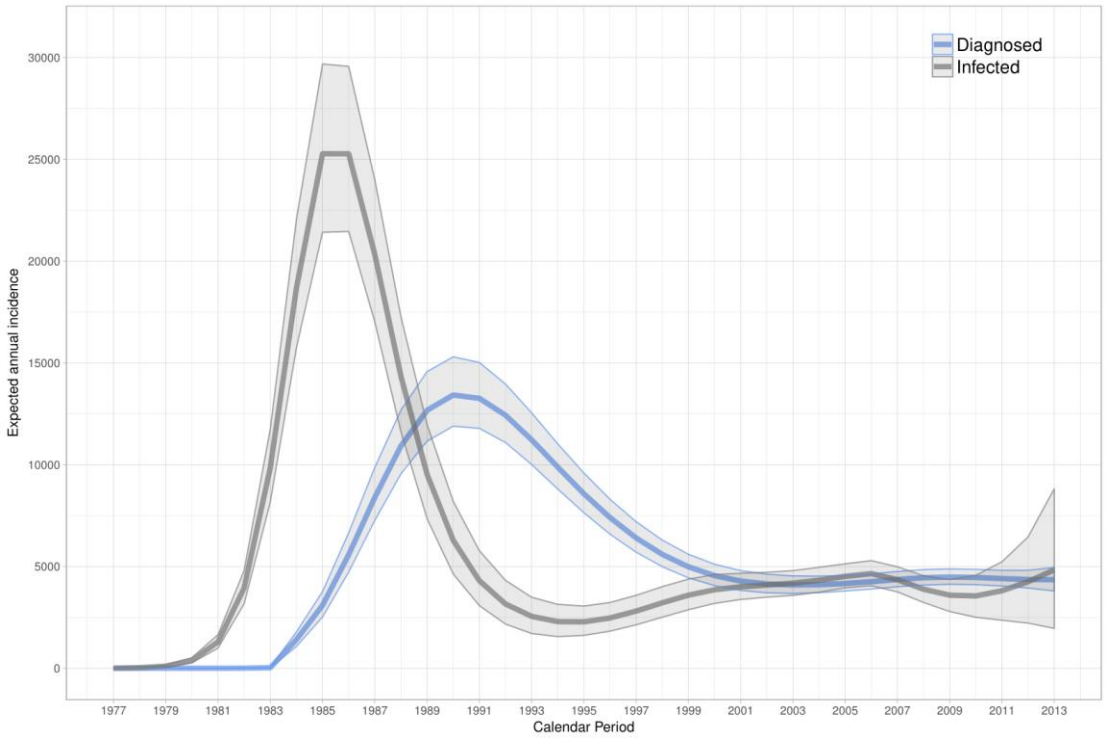


Fig. 2: Overall annual new cases of HIV infections and HIV diagnoses (Spain, 1977-2013).

Fig3

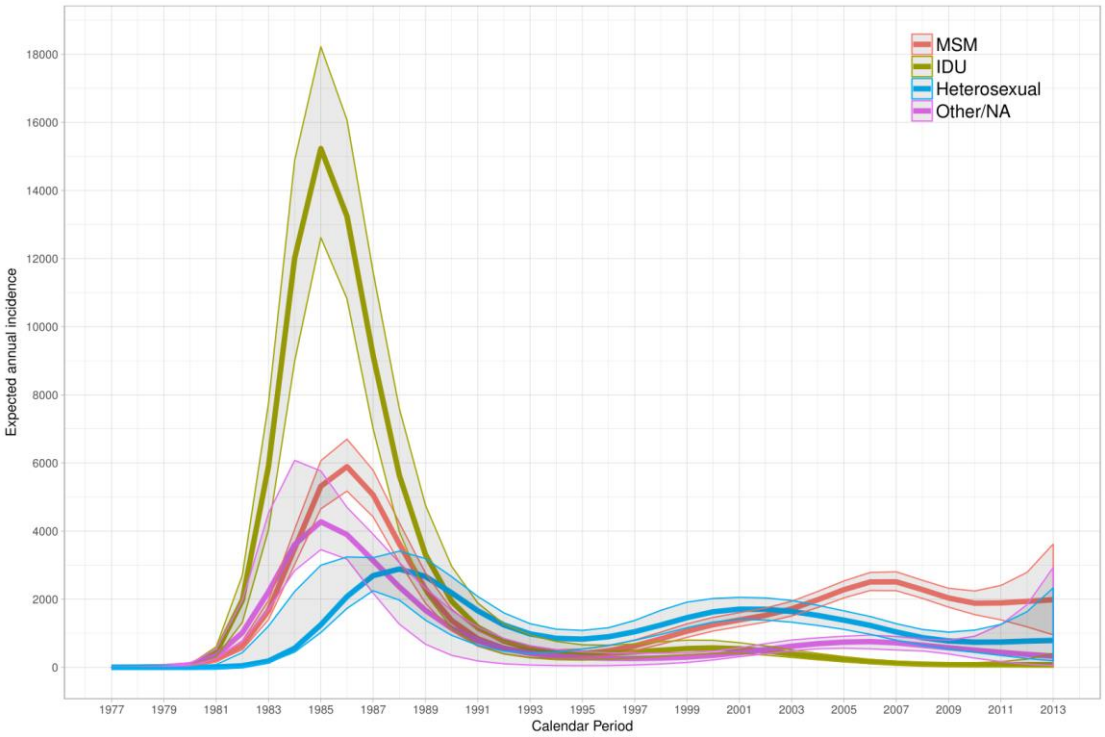


Fig. 3: Annual new cases of HIV infections by transmission category (Spain, 1977-2013).

Fig4

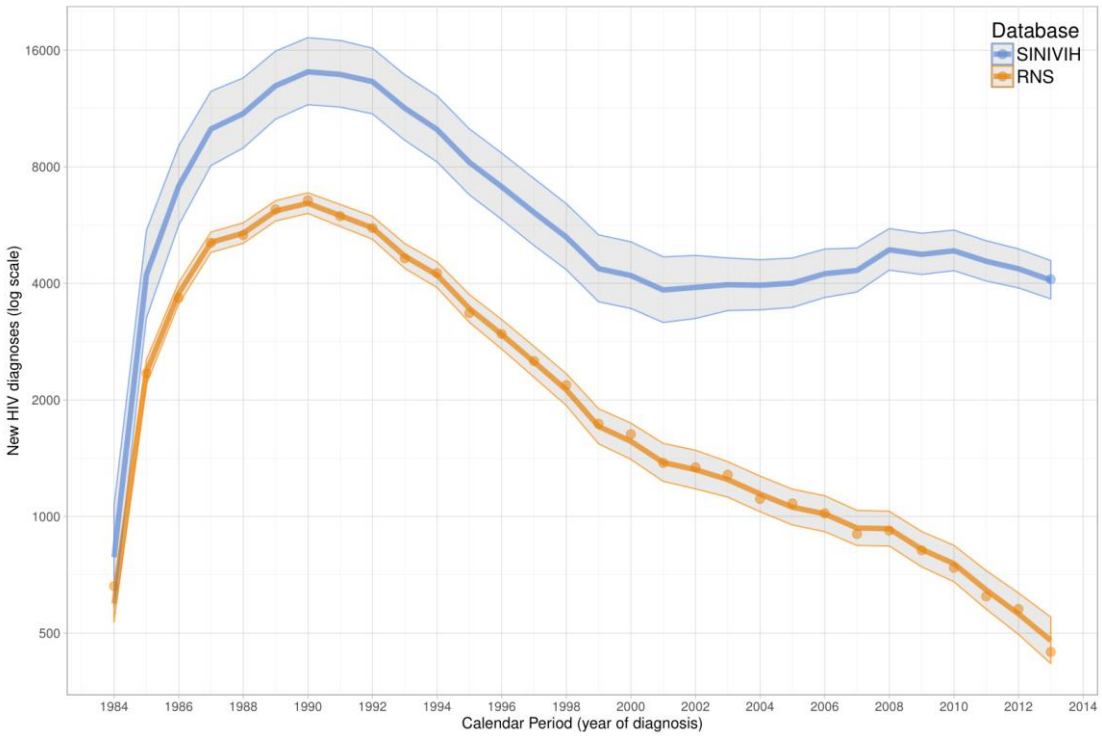


Fig. 4: Annual new HIV diagnoses estimates (in log scales) over the 1984-2013 period, in the HIV surveillance system (SINIVIH) and in the AIDS surveillance system (RNS).

Fig5

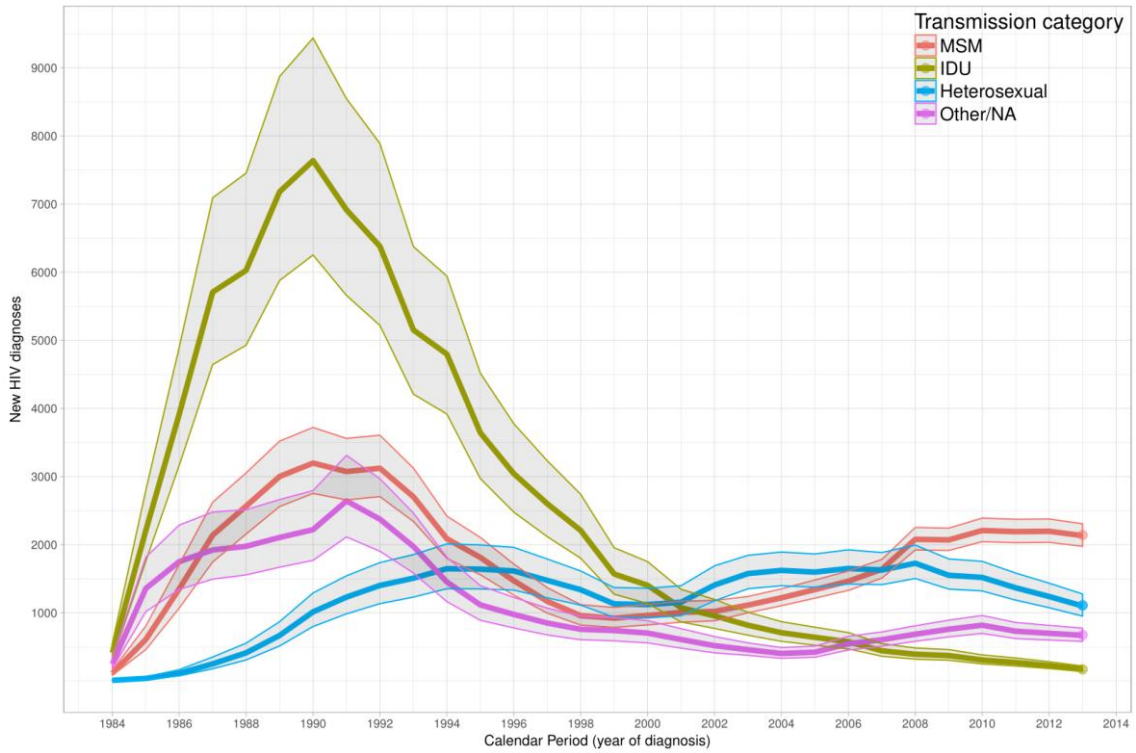


Fig. 5: Annual new HIV diagnoses estimates by transmission category (Spain, 1984-2013). MSM indicates men who have sex with men; IDU, injection drug users; Other/NA, haemophiliac/transfusion recipient, mother-to-child or unknown categories.

AIDS: Author's paper submission checklist

Title of paper:	► Estimating the number of people living with HIV and the undiagnosed fraction in Spain
Names of authors:	►Nuñez Olivier, Hernando Victoria, Díaz Asunción

AUTHORS SHOULD PLEASE ENSURE THAT ALL APPROPRIATE INFORMATION (EG. CONFLICT OF INTEREST STATEMENTS) ARE ALSO INCLUDED IN THE TEXT OF THE ARTICLE.

1. DUPLICATE PUBLICATION is not acceptable and includes papers, or letters to the Editor reporting the same data previously published in any journal. Abstracts of papers presented at meetings and published in the proceedings of such meetings do not constitute duplicate publication, but should be disclosed by including a note at the beginning of the paper, i.e.. "Data presented previously at (state meeting) and published as abstract in (give reference)". Have you published these data previously?

►A very brief summary has been included in "Gourlay A et al. The Human Immunodeficiency Virus Continuum of Care in European Union Countries in 2013: Data and Challenges. Clin Infect Dis. 2017 Jun 15;64(12):1644-1656. doi: 10.1093/cid/cix212"

2. CONFLICT OF INTEREST include financial support from the biomedical industry or other commercial sources in the form of research grants, bench fees, consultancy or lecture fees, travelling expenses, payment of registration fees, consultancy appointments, posts held in the biomedical industry or equipment manufacturers, stock holdings in the company, free supply of drugs and the like. These should be stated in relation to each author. Has any of the authors any conflict of interest? Please state details.

► No

3. CONSENT Please note that patient's, or normal control's, written consent is needed not only for full papers, but also for case reports. The written consent needs to include not only agreement to undergo treatment, or participate in an experiment or an randomised control trial, but also agreement for anonymised data to be published in a scientific journal. Was patient's consent obtained and in what form?

►According to Spanish legislation (Ley 33/2011, de 4 de octubre, General de Salud Pública), surveillance data not require patient's content. Patients included in the CoRIS cohort, have to sign a consent for being included in the cohort.

4. ETHICS All studies need to be approved by the local Ethical Committees. Was your study? Please provide the approval from your local Ethical Committees for any animal experimentation or human subject studies.

►

5. AUTHOR'S CONTRIBUTIONS AND APPROVAL OF TEXT Please state briefly how each of the authors contributed to the study, to data analysis and to the writing of your paper. Subject to your agreement, we will print this information, if the paper is accepted for publication. In addition, please confirm that all the authors have read and approved the text as submitted to AIDS. Justify individual's contributions when the author list exceeds 10.

►
AD was the main study researcher. She supervised all phases of the work, reviewed surveillance data and quality, and critically revised the article.
ON did the statistical analysis, interpretation of results, and drafted the manuscript.
VH provided mortality data and made important contributions to successive versions of the manuscript.
All authors have seen and approved the final manuscript.

6. STATISTICAL ANALYSIS Kindly please let me know who performed the statistical analysis of you data.

►Olivier Nuñez

7. ARTICLE LENGTH Please acknowledge that you have kept within the word and insert limits for your submission by ticking the relevant boxes below, and that you have indicated the word count on the title page of your article

Article length

- ✓ Original papers, 3500 words of text excluding references with no more than five inserts (figures/tables)
- ❑ Concise communications, 1800 words of text excluding references with maximum of two inserts (figures/tables)
- ❑ Research Letters, 1000 words excluding summary with no more than one insert (figure/table)
- ❑ Correspondence, 750 words excluding references with no more than one insert (figure/table)

Title page

- ✓ Title no more than 120 characters
- ✓ Running head, no more than 40 characters
- ✓ Justification of the number of contributors greater than 10 included in this document

Abstract page

- ✓ Abstract no more than 250 words
- ✓ 5–7 keywords listed

Word count

- ✓ Word count of text (excluding references) included on title page

8. CLINICAL TRIALS AND BEHAVIOURAL EVALUATIONS: Authors reporting results of randomized controlled trials should include with their submission a complete checklist from the CONSORT statement, see *JAMA* 1996; 227:637-639 or http://www.consort-statement.org. For behavioural and public health evaluations involving non-randomized designs, authors should include with their submission a complete checklist from the TREND statement, see Am J Public Health 2004; 94:361-366 or www.trend-statement.org .

REGISTRATION OF CLINICAL TRIALS: As a condition for publication of a clinical trial in *AIDS*, registration of the trial in a public registry is required. Registration of a trial must be at or before the enrollment of participants. The editors of *AIDS* also do not advocate one particular registry but require that the registry utilized meet the criteria set out in the statement of policy of the ICMJE (www.ijcme.org).

Please state that your article includes a clinical trial and that the conditions of submission above have been met.

► No

Other information for the Editor:

►	
Name of person completing this form:	►*Olivier Nuñez*
Date:	► *Friday 2 March 2018*