A synthesis of convenience survey and other data to estimate undiagnosed HIV infection among men who have sex with men in England and Wales

Kate Walker,1,2 Shaun R Seaman,3 Daniela De Angelis,3,4 Anne M Presanis,3 Julie P Dodds,1 Anne M Johnson,1 Danielle Mercey,1 O Noel Gill4 and Andrew J Copas1,5*

1Centre for Sexual Health and HIV Research, Research Department of Infection & Population Health, University College London, UK, 2Department of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, University of London, UK, 3MRC Biostatistics Unit, Cambridge, UK, 4Health Protection Agency, London, UK and 5MRC Clinical Trials Unit, London, UK

*Corresponding author. Research Department of Infection & Population Health, Centre for Sexual Health and HIV Research, University College London, The Mortimer Market Centre, Capper St, London, WC1E 6JB, UK. E-mail: a.copas@ucl.ac.uk

Accepted 14 July 2011

Background Hard-to-reach population subgroups are typically investigated using convenience sampling, which may give biased estimates. Combining information from such surveys, a probability survey and clinic surveillance, can potentially minimize the bias. We developed a methodology to estimate the prevalence of undiagnosed HIV infection among men who have sex with men (MSM) in England and Wales aged 16–44 years in 2003, making fuller use of the available data than earlier work.

Methods We performed a synthesis of three data sources: genitourinary medicine clinic surveillance (11380 tests), a venue-based convenience survey including anonymous HIV testing (3702 MSM) and a general population sexual behaviour survey (134 MSM). A logistic regression model to predict undiagnosed infection was fitted to the convenience survey data and then applied to the MSMs in the population survey to estimate the prevalence of undiagnosed infection in the general MSM population. This estimate was corrected for selection biases in the convenience survey using clinic surveillance data. A sensitivity analysis addressed uncertainty in our assumptions.

Results The estimated prevalence of undiagnosed HIV in MSM was 2.4% (95% confidence interval (95% CI 1.7–3.0%)), and between 1.6% (95% CI 1.1–2.0%) and 3.3% (95% CI 2.4–4.1%) depending on assumptions; corresponding to 5500 (3390–7180), 3610 (2180–4740) and 7570 (4790–9840) men, and undiagnosed fractions of 33, 24 and 40%, respectively.

Conclusions Our estimates are consistent with earlier work that did not make full use of data sources. Reconciling data from multiple sources, including probability-, clinic- and venue-based convenience samples...
can reduce bias in estimates. This methodology could be applied in other settings to take full advantage of multiple imperfect data sources.

**Keywords** Data synthesis, methods, bias, HIV, undiagnosed, men who have sex with men

---

**Introduction**

Reducing the prevalence of undiagnosed HIV infection and increasing the proportion of HIV-infected individuals who are aware of their status are important principles of HIV prevention. Diagnosing HIV infection provides an opportunity for risk-reduction counselling and to consider the initiation of anti-retroviral treatment. Sexual risk behaviours may decrease after HIV diagnosis and anti-retroviral treatment substantially reduces HIV infectivity. Individuals with undiagnosed infection may disproportionately contribute to new HIV infections. Accurately and efficiently estimating the prevalence of undiagnosed HIV infection among high-risk groups is therefore important for planning and evaluating HIV prevention and testing programmes.

In the UK the prevalence of HIV infection is much higher among men who have sex with men (MSM), those born in sub-Saharan Africa and injecting drug users, than in the general population. These groups are hard to reach in two respects. First, they constitute a small proportion of the population. Secondly, the census may not collect the information that would enable areas in which these groups are more prevalent to be identified and so allow efficient targeted sampling of them in representative household surveys. Consequently, convenience survey methods based on chain-recruitment methods, venue-based or time-location sampling may be employed to recruit many participants at low cost. In England and Wales (E&W), surveys of MSM have been conducted at gay venues (gay men sexual health survey, GMSHS), events and gyms and at genitourinary medicine (GUM) clinics. Convenience surveys may be susceptible to bias. For example, the MSM in the GMSHS are known to report behaviour suggestive of higher risk of HIV transmission than the MSM sampled in the British National Survey of Sexual Attitudes and Lifestyles (NATSAL), a probability sample survey. MSM in NATSAL are assumed to be representative of MSM in general.

The number of undiagnosed HIV-infected people in Britain has been estimated over several years using the so-called direct method. The population is partitioned according to infection risk into groups, the sizes of which are estimated from the census together with NATSAL data. The prevalence of undiagnosed HIV infection in each group is estimated from surveillance data. MSM are defined as those men having sex with another man in the past 5 years. They are divided by whether resident in London and into higher- and lower-risk MSM according to whether or not they have attended a GUM clinic in the past 5 years. The prevalence of undiagnosed HIV infection in higher risk MSM is estimated from data on anonymized HIV tests conducted at GUM clinics among MSM (defined by orientation). The prevalence of undiagnosed infection in lower-risk MSM is assumed to be reduced by a factor (varying by region) compared with the higher-risk MSM; this factor is estimated using aggregate data from the GMSHS and NATSAL. Recent formal Bayesian data syntheses have extended the direct method approach using the same surveillance data sources, but accounting for uncertainty in the data and also including additional related data on prevalence of diagnosed HIV infection in MSM. However, these syntheses employed aggregate analysis of the GMSHS and NATSAL data, rather than a comprehensive individual-level analysis.

Other methods for estimating undiagnosed HIV prevalence have also been suggested, notably those based on an extended version of the back-calculation method (see, for example, refs). Back-calculation uses information about the incidence of an HIV-related endpoint (HIV or AIDS diagnosis) and data on HIV progression to re-construct the historic HIV incidence curve. This curve can then be used together with information on HIV mortality and prevalence of diagnosed infection to estimate the prevalence of undiagnosed infection. This approach, although interesting, requires much more temporal surveillance and natural history information than we consider here.

In this article, we develop and apply a method for synthesising data from the GMSHS, NATSAL and clinic surveillance to estimate the undiagnosed HIV prevalence in MSM in E&W. This method makes fuller use of these data than earlier work through an individual-level analysis. Our approach proceeds by first fitting a model for the probability that an MSM is undiagnosed HIV positive in the GMSHS, based on behaviour and demographic factors recorded also in the NATSAL survey. This model is then applied to the MSM in NATSAL, to predict the prevalence of undiagnosed HIV infection both among MSM who have attended a GUM clinic and among the general population of MSM. Finally the prevalence observed in the clinic surveillance data is compared with that predicted for MSM who have attended a GUM clinic. This comparison allows a
correction to the estimated prevalence of undiagnosed HIV infection in the general population of MSM to remove the selection bias affecting the GMSHS.

Methods

Data sources

The GMSHS is a repeated survey conducted in gay venues (bars, saunas, clubs) in London,6,21,22 involving a brief self-completion paper questionnaire, including a question on HIV diagnosis, and the invitation to provide an oral fluid sample for anonymous HIV testing. Further participants (questionnaire only) are recruited at GUM clinics, but are not included in this analysis as their HIV status is unknown. The fluid samples are sent to the Centre for Infections, London, to be tested for anti-HIV antibody. Details of the sampling device and testing procedures are published elsewhere.21 We use data from 3702 participants aged 16–44 years with an HIV test result in 2002–04, including 844, 1189 and 1097 who participated in London in each year, respectively. We selected this period because two other English cities each participated for only 1 year in the period: Brighton in 2003 (277 participants) and Manchester in 2004 (295 participants).22 We were, therefore, able to first investigate whether predictors of HIV infection and diagnosis differed across the cities. As there was little evidence of differences, these data are analysed together.

The NATSAL was last conducted in 2000.11 At each address sampled from the sampling frame, one person aged 16–44 years was randomly selected and invited to participate. Interviews were conducted in the participant’s home, with the more sensitive information collected through a computerized self-completion questionnaire. London addresses were over-sampled by design. A set of weights (denoted the ‘NATSAL weights’) was devised to reflect the variation in sampling probability, with post-stratification by sex, age and region. For our analysis, MSM are defined as men reporting sex with another man in the past year, rather than the previous 5 years as in earlier NATSAL work,23 in order to reflect the questions asked in the GMSHS. GUM clinic attendees are those who reported attending a sexually transmitted disease clinic (NATSAL) or a sexual health/GU clinic (GMSHS) in the past year. Of 4394 male participants in E&W, 134 were identified as MSM, of whom 132 provided information on GUM clinic attendance.

The ongoing Unlinked Anonymous survey (UA) is conducted by the Health Protection Agency (HPA) using data and samples collected from GUM clinic attendees.24 We use data collected in 2003–04 on the number of tests performed and number positive among MSM (defined by self-identified orientation) in clinics, stratified by location (London and outside London) and age. Prevalence of undiagnosed infection is defined as prevalence of previously undiagnosed infection, so that individuals diagnosed at the current clinic visit, through opting to have an HIV test, are included in the numerator, whereas previously diagnosed individuals are excluded from the numerator but included in the denominator. Data were available for 8219 MSM aged 16–44 years in London and 3161 outside London. The clinics outside London are all in urban areas, so may not be representative of the whole of E&W excluding London. Furthermore, if a participant attends a GUM clinic repeatedly in a year, he may contribute more than once to the data, though such ‘duplicates’ are eliminated within each quarter. These facts could result in the observed UA data giving an overestimate of prevalence (see sensitivity analyses later).

The ‘naïve’ approach to predict the prevalence of undiagnosed infection

We first build two logistic regression models, one for the probability that an individual is HIV infected ($\pi_1$) and the other for the probability that an infected individual is undiagnosed ($\pi_2$), using predictors (vector $X$) available in both GMSHS and NATSAL data sets (Table 2):

$$\logit(\pi_1) = X\theta_1$$
$$\logit(\pi_2) = X\theta_2$$

where $\theta_1$ and $\theta_2$ are vectors of parameters to be estimated. These models were fitted to the GMSHS data and backward step-wise selection used. The final models selected may be based on different subsets of $X$ (i.e. some components of $\theta_1$ and $\theta_2$ are estimated to be zero).

If we denote by $\hat{\pi}_1(x) = \pi_1(x,\hat{\theta}_1)$ and $\hat{\pi}_2(x) = \pi_2(x,\hat{\theta}_2)$, the estimates of $\pi_1$ and $\pi_2$ from our models for an individual with predictor values $x$, then the predicted probability of undiagnosed infection is $p_{\text{naive}}(x) = \hat{\pi}_1(x)\hat{\pi}_2(x)$.

We define four subgroups of MSM according to residence in London ($l = 1$) or not ($l = 0$) and GUM clinic attendance ($g = 1$) or not ($g = 0$). We evaluate $p_{\text{naive}}$ for each MSM in NATSAL, and then estimate the prevalence of undiagnosed infection in the general population of MSM, separately within each subgroup, as the average of $p_{\text{naive}}$ across the individuals in that subgroup, weighted by their NATSAL weights. For subgroup $(l, g)$ we denote this naïve estimate by $\hat{p}_{lg}^{\text{naive}}$.

Initial correction for bias

Since the GMSHS is a venue-based survey primarily in London, the models for $\pi_1$ and $\pi_2$ specified earlier may not apply to the general population of MSM, and so the naïve estimates $\hat{p}_{lg}^{\text{naive}}$ may be biased. Using the UA data, we can estimate these biases and correct the naïve estimates for them.

We assume that the bias in $\hat{p}_{lg}^{\text{naive}}$ consists of a basic bias that applies to all four subgroups of MSM and an additional London bias, which applies only to the two
London subgroups. Although similar associations between predictors and both HIV infection and the proportion diagnosed were seen in the GMSHS data for London, Brighton and Manchester, we allow the bias (i.e. the representativeness) of these data to be different in London and outside. For two of the four subgroups (MSM attending GUM clinics in London and outside), estimates of undiagnosed HIV prevalence are also available from the UA survey. Let \( \hat{q}_1^\text{clinic} \) and \( \hat{q}_0^\text{clinic} \) denote these estimates for inside and outside London respectively, and assume for now that they are unbiased.

The basic and London biases are then estimated as \( \hat{\beta}_\text{basic} \) and \( \hat{\beta}_\text{lon} \) where

\[
\hat{\beta}_\text{basic} + \hat{\beta}_\text{lon} = \logit(\hat{q}_{11}^\text{naive}) - \logit(\hat{q}_1^\text{clinic})
\]

\[
\hat{\beta}_\text{basic} = \logit(\hat{q}_{01}^\text{naive}) - \logit(\hat{q}_0^\text{clinic})
\]

and the corrected estimates of undiagnosed HIV prevalence are \( \hat{q}_l^\text{corrected} \) where

\[
\logit(\hat{q}_l^\text{corrected}) = \logit(\hat{q}_{01}^\text{naive}) - \hat{\beta}_\text{basic} - l\hat{\beta}_\text{lon}
\]

Note that this has ensured that \( \hat{q}_1^\text{corrected} = \hat{q}_1^\text{clinic} \) (\( l = 0,1 \)).

### Sensitivity analysis

We also consider a more complex bias model with additional effects for GUM clinic non-attendance, \( \beta_{gum} \), and for a GUM–London interaction \( \beta_{longum} \), in case our initial bias correction based on the clinic attendees is not wholly appropriate for clinic non-attendees. There is no information in our data for estimating \( \beta_{gum} \) or \( \beta_{longum} \), so we consider a range of plausible values. The corrected estimates are now given by

\[
\logit(\hat{q}_l^\text{corrected}) = \logit(\hat{q}_{01}^\text{naive}) - \hat{\beta}_\text{basic} - l\hat{\beta}_\text{lon} - (1 - g)\hat{\beta}_{gum} - (1 - g)\hat{\beta}_{longum}
\]

Up to now we have considered the estimates \( \hat{q}_1^\text{clinic} \) from the UA data to be unbiased, but for reasons outlined in our data sources section this may not be so. We assume that the bias in \( \hat{q}_1^\text{clinic} \) consists of a basic bias, \( \hat{\gamma}_\text{basic} \), and an additional London effect \( \hat{\gamma}_\text{lon} \). Again there is no information in our data for estimating these, so we consider a range of plausible values. \( \hat{\beta}_\text{basic} \) and \( \hat{\beta}_\text{lon} \) are now estimated not by Equations (1) and (2) but by

\[
\hat{\beta}_\text{basic} + \hat{\beta}_\text{lon} = \logit(\hat{q}_{11}^\text{naive}) - \logit(\hat{q}_1^\text{clinic}) - \hat{\gamma}_\text{basic} - \hat{\gamma}_\text{lon}
\]

\[
\hat{\beta}_\text{basic} = \logit(\hat{q}_{01}^\text{naive}) - \logit(\hat{q}_0^\text{clinic}) - \hat{\gamma}_\text{basic}
\]

That is, \( \hat{q}_1^\text{clinic} \) has been corrected for its assumed bias. Finally the corrected estimates of undiagnosed HIV prevalence are still given by Equation (4).

### Estimation of overall prevalence, number of undiagnosed infections and undiagnosed fraction

The overall prevalence of undiagnosed infection in MSM is estimated as the average of the prevalence estimates in each of the four subgroups of MSM, \( \hat{q}_l^\text{corrected} \), weighted by the proportion of MSM in each subgroup as estimated from NATSAL. The number of undiagnosed MSM is then estimated as the overall undiagnosed prevalence estimate multiplied by the total number of MSM in E&W, estimated from NATSAL together with the UK Census 2001 figure for the number of men resident in E&W aged 16–44 years. Finally, by comparison with the HPA figures for diagnosed cases in MSM in E&W, the undiagnosed fraction is estimated.

### Missing data

Approximately 20% of MSM in NATSAL, but only 2% in GMSHS, have at least one of the predictors missing. Multiple imputation was performed on the NATSAL data using the ‘mice’ package in R, with five imputed data sets, specifying predictive mean matching; a complete case analysis was used for the GMSHS data.

### Confidence intervals and sensitivity analysis

To estimate confidence intervals, we applied the parametric bootstrap to the GMSHS prediction model, but the non-parametric bootstrap to the NATSAL and UA data, as the UA data are simple proportions and we do not specify a model for the distribution of the NATSAL data. In the parametric bootstrap, the infection status and diagnosis status were simulated from the initial fitted model and then the whole estimation procedure, including the step-wise model selection, was repeated. In the non-parametric bootstrap procedure, the weights were scaled so that their sum within each sampling stratum matched that in NATSAL, to reflect the post-stratification in NATSAL. The multiple imputation for NATSAL data was performed within each bootstrap resample.

The bias effects \( \beta_{gum} \) and \( \beta_{longum} \) were set to either \(-0.22\) or \(+0.22\) corresponding to an odds ratio (OR) of either 0.8 or 1.25, considered to be moderate effects. Similarly we set \( \hat{\gamma}_\text{basic} \) to 0 or 0.22 and \( \hat{\gamma}_\text{lon} \) to 0 or \(-0.22\) so that the outside London UA data (from urban areas only—see data sources section) cannot underestimate the undiagnosed HIV prevalence in clinic attendees. We only consider non-zero values for \( \beta_{longum} \) when \( \beta_{gum} \) is non-zero. We consider 10 scenarios for our sensitivity analysis to reflect the impact of individual effects and the maximum impact of the effects together.

### Results

In Table 1, we see the demographic and behavioural characteristics of MSM in the GMSHS and NATSAL,
stratified by GUM clinic attendance and residence in London. Compared with NATSAL MSM overall, individuals participating in the GMSHS are more likely to be GUM clinic attendees, better educated, employed, have had an HIV test, have had a sexually transmitted infection (STI) diagnosed in the past year and have had 10 or more male partners in the past year. Among the GMSHS participants, GUM clinic attendees are more likely than non-attendees to have had an HIV test and an STI diagnosis.

The prevalence of HIV infection in the GMSHS is 11.9%. In Table 2, we see the strongest predictors of HIV infection are older age, unemployment, Black ethnicity, attending a GUM clinic, unprotected anal intercourse, STI diagnosis and reporting 10 or more sexual partners. These factors are selected by the backwards step-wise selection procedure (threshold $P = 0.05$) as predictors in our model for infection. The proportion undiagnosed among HIV-infected men in the GMSHS is 41.9%. The strongest predictors of being undiagnosed are Black and ‘other’ ethnicity, being employed and not attending a GUM clinic. These
### Table 2 Prediction models for HIV infection and being undiagnosed, GMSHS data

<table>
<thead>
<tr>
<th>Overall</th>
<th>Model for probability of being HIV positive</th>
<th>Model for probability of being undiagnosed, conditional on being HIV positive</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HIV+ / total (%)</td>
<td>Adjusted OR (95% CI)</td>
</tr>
<tr>
<td>Age</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age (1 year)</td>
<td>–</td>
<td>1.10 (1.08–1.12)</td>
</tr>
<tr>
<td>Age²</td>
<td>–</td>
<td>0.996 (0.993–0.998)</td>
</tr>
<tr>
<td>Ethnicity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>378/3260 (11.6)</td>
<td>1</td>
</tr>
<tr>
<td>Black</td>
<td>28/98 (28.6)</td>
<td>3.43 (2.07–5.67)</td>
</tr>
<tr>
<td>Other</td>
<td>33/331 (10.0)</td>
<td>–</td>
</tr>
<tr>
<td>Continued in education for ≥3 years post 16 years</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>176/1171 (15.0)</td>
<td>1</td>
</tr>
<tr>
<td>Yes</td>
<td>263/2531 (10.4)</td>
<td>0.65 (0.52–0.80)</td>
</tr>
<tr>
<td>Unemployed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>332/3244 (10.2)</td>
<td>1</td>
</tr>
<tr>
<td>Yes</td>
<td>105/446 (23.5)</td>
<td>2.77 (2.09–3.66)</td>
</tr>
<tr>
<td>Ever had HIV test</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>51/841 (6.1)</td>
<td>–</td>
</tr>
<tr>
<td>Yes</td>
<td>388/2861 (13.6)</td>
<td>2.43 (1.80–3.29)</td>
</tr>
<tr>
<td>Attended GUM clinic in past year</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>136/2044 (6.7)</td>
<td>1</td>
</tr>
<tr>
<td>Yes</td>
<td>294/1614 (18.2)</td>
<td>2.32 (1.81–2.97)</td>
</tr>
<tr>
<td>STI diagnosed in past year</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>254/2848 (8.9)</td>
<td>1</td>
</tr>
<tr>
<td>Yes</td>
<td>184/843 (21.8)</td>
<td>1.57 (1.22–2.01)</td>
</tr>
<tr>
<td>Unprotected anal intercourse in past year</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>173/1804 (9.6)</td>
<td>1</td>
</tr>
<tr>
<td>Yes</td>
<td>266/1898 (14.0)</td>
<td>1.39 (1.11–1.73)</td>
</tr>
<tr>
<td>10+ male sexual partners in past year</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>210/2348 (8.9)</td>
<td>1</td>
</tr>
<tr>
<td>Yes</td>
<td>229/1354 (16.9)</td>
<td>1.54 (1.23–1.92)</td>
</tr>
</tbody>
</table>

*aAdjusted ORs provided for those factors selected in the prediction model.*
factors were selected as predictors in the model for being undiagnosed.

The prevalence of undiagnosed HIV infection in the UA is 5.1% (420/8219) in London and 1.8% (57/3161) outside London.

The naïve estimated prevalence of undiagnosed infection is 5.1% [95% confidence interval (95% CI 3.7–6.4)], top row of Table 3. However, the comparison with UA data shows that our naïve model overestimates this prevalence in GUM attendees by a similar amount both within and outside London. The basic bias effect and additional London bias effect are estimated to be $\hat{\beta}_{\text{basic}} = 0.78$ (95% CI 0.31–1.30) and $\hat{\beta}_{\text{lon}} = 0.08$ (95% CI −0.73 to 0.96), respectively. The estimated undiagnosed prevalence after initial bias correction (Scenario 1, Table 3) is 2.4% (95% CI 1.7–3.0), with a corresponding number of undiagnosed MSM in E&W 5500 (95% CI 3390–7180). The unadjusted prevalence for MSM resident in London who attend and do not attend GUM clinics is estimated to be 5.1% (95% CI 0.9–8.0) and 2.1% (95% CI 1.3–2.7), respectively; outside London, the corresponding estimates are 1.8% (95% CI 0.8–2.5) and 2.4% (95% CI 1.5–3.2), respectively. The HPA reported 11,769 diagnosed MSM in the UK in 2003 aged 15–44 years, and from this we estimate 11,200 aged 16–44 years in E&W. The undiagnosed HIV fraction is therefore estimated to be 33% (95% CI 23–39).

Table 3 also includes the sensitivity analysis. Scenario 10 leads to the highest prevalence of undiagnosed infection, at 3.3% (95% CI 2.4–4.1), resulting from all four sensitivity parameters [see Equations (4–6)] taking their minimum values. Scenario 9 represents the opposite extreme with undiagnosed prevalence 1.6% (95% CI 1.1–2.0). The parameter $\hat{\beta}_{\text{basic}}$, determining the overall bias in the UA data, has the greatest impact, as its value affects all four subgroups of MSM. When this parameter takes value 0.22 (Scenario 2), and other sensitivity parameters equal zero, the prevalence of undiagnosed infection decreases to 1.9%.

### Discussion

Our methodology enables the estimation of characteristics of hard-to-reach groups by reconciling different data sources. In our application to undiagnosed HIV in MSM in E&W, we made fuller use of the NATSAL and GMSHS HIV infection data than earlier work, and thereby relied on fewer subjective assumptions. Our approach may therefore be less susceptible to bias but there is a trade-off, in that the uncertainty we acknowledge through our confidence intervals and sensitivity analysis is appreciably greater than that seen previously. The methodology is relatively simple and intuitive with a straightforward approach to estimate our bias parameters. Conversely, the approach is relatively computer intensive, as bootstrapping is required.

### Table 3  Sensitivity analysis for the estimated prevalence of undiagnosed HIV infection and numbers of undiagnosed MSM aged 16–44 years in E&W, 2003

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Sensitivity parameters</th>
<th>Estimated total MSM (95% CI)</th>
<th>Estimated overall prevalence % (95% CI)</th>
<th>Estimated overall undiagnosed fraction % (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive</td>
<td></td>
<td>11,610 (7270–15070)</td>
<td>5.1 (3.7–5.4)</td>
<td>51 (39–57)</td>
</tr>
<tr>
<td>1</td>
<td>$\beta_{\text{basic}}$</td>
<td>10,7 (2.8–16.6)</td>
<td>5.1 (0.9–8.0)</td>
<td>2.1 (1.3–2.7)</td>
</tr>
<tr>
<td>2</td>
<td>$\beta_{\text{lon}}$</td>
<td>4040 (2950–5320)</td>
<td>5.1 (0.9–8.0)</td>
<td>2.1 (1.3–2.7)</td>
</tr>
<tr>
<td>3</td>
<td>$\beta_{\text{GM}}$</td>
<td>3021 (21–36)</td>
<td>5.1 (0.9–8.0)</td>
<td>2.1 (1.3–2.7)</td>
</tr>
<tr>
<td>4</td>
<td>$\beta_{\text{GM}}$, $\beta_{\text{lon}}$</td>
<td>12.9–19.8)</td>
<td>5.1 (0.9–8.0)</td>
<td>2.1 (1.3–2.7)</td>
</tr>
<tr>
<td>5</td>
<td>$\beta_{\text{GM}}$, $\beta_{\text{lon}}$, $\beta_{\text{GM}}$</td>
<td>4.4 (2.8–6.3)</td>
<td>5.1 (0.9–8.0)</td>
<td>2.1 (1.3–2.7)</td>
</tr>
</tbody>
</table>

---

1364 INTERNATIONAL JOURNAL OF EPIDEMIOLOGY
Our findings are broadly consistent with the most recent Bayesian evidence synthesis, which estimated undiagnosed prevalence among all MSM, including ‘past MSM’, whereas here we employ a definition of MSM active in the past year. Those authors estimated the undiagnosed fraction in E&W in 2001 to be 40% [95% credible interval (95% CrI) 32–47%], with a similar 42% (95% CrI 35–50%) fraction in 2003, falling to 33% (95% CrI 26–40%) in 2005. These correspond to estimates of 6300 (95% CrI 4600–8600) MSM aged 15–44 years with undiagnosed HIV in 2001, 8200 (95% CrI 6100–11100) in 2003 and 6300 (95% CrI 4600–8800) in 2005.

This methodological work not only confirms the results of Bayesian evidence syntheses, but also potentially opens the way for further methodological development to use NATSAL and GMSHS data more fully in future. Indeed the most recent synthesis uses the ratio presented here of the prevalence outside of London of undiagnosed infection in MSM attending GUM clinics compared with MSM not attending clinics in a sensitivity analysis. In contrast to a Bayesian data synthesis, in our approach, it is difficult to construct single intervals to simultaneously reflect sampling variability and uncertainty in all assumptions, though in further work we could investigate this possibility following the ideas of Greenland. There may be a loss of efficiency through our use of a weighted approach to estimation of the prevalence, and our choice not to model the distribution of the predictors within NATSAL. We have not made any explicit consideration of biases in the NATSAL data though these are possible. Our approach relied on multiple imputation for the NATSAL MSM, which is valid only under the assumption that data are ‘missing at random’.

The synthesis was made possible by ensuring that questions in different surveys were matched. Any future synthesis would benefit from enhanced surveillance in clinics, e.g., the collection of behavioural information, to allow more cross-checking. The difference seen between our naïve undiagnosed prevalence estimate and that after bias correction demonstrates that whereas convenience surveys have an important role to play in research, their biases must be acknowledged. The relatively wide uncertainty we attribute to our estimate demonstrates the importance of research to define (and narrow) the range for bias parameters considered in our sensitivity analysis, such as those affecting the UA data. The large uncertainty also arises because, whereas a substantial number of individuals from a hard-to-reach group such as MSM may be recruited from convenience sources, only modest numbers will be recruited in representative surveys, such as NATSAL, and these latter individuals play a central role in the synthesis. Methods to efficiently boost the numbers from hard-to-reach groups in representative surveys without introducing bias would be invaluable.

Funding
This methodological work was not supported by external funding, but the NATSAL 2000 survey data collection was supported by the Medical Research Council, UK (grant G9811620) and the GMSHS survey was supported by the Health Protection Agency, UK.

Conflict of interest: None declared.

KEY MESSAGES
- Convenience surveys of ‘hard-to-reach’ groups in the population such as MSM may be large but are susceptible to bias.
- We have developed a straightforward method to synthesize evidence from a convenience survey with evidence from other sources such as a probability survey and clinic surveillance data.
- Making fuller use of a bar survey of MSM, and the 2001 British NATSAL probability survey, we estimated the prevalence of undiagnosed HIV infection in E&W in 2003 among MSM aged 16–44 years and the undiagnosed fraction to be, respectively, 2.4% and 33%, findings consistent with previous syntheses.
- Reconciling data from multiple sources can validate estimates derived from potentially biased samples, and our methodology could be applied in other settings to take full advantage of multiple imperfect data sources.

References


