Public Health
England

Protecting and improving the nation’s health

Producing modelled estimates of the size of the lesbian, gay and bisexual (LGB) population of England

Technical Report 2. Methodology for synthesis

About Public Health England

Public Health England exists to protect and improve the nation’s health and wellbeing, and reduce health inequalities. We do this through world-class science, knowledge and intelligence, advocacy, partnerships and the delivery of specialist public health services. We are an executive agency of the Department of Health, and are a distinct delivery organisation with operational autonomy to advise and support government, local authorities and the NHS in a professionally independent manner.

Public Health England
Wellington House
133-155 Waterloo Road
London SE1 8UG
Tel: 020 7654 8000
www.gov.uk/phe
Twitter: @PHE_uk
Facebook: www.facebook.com/PublicHealthEngland

Prepared by:
Sanne van Kampen, Mauro Fornasiero, William Lee, Kerryn Husk

This project was supported by the National Institute for Health Research (NIHR) Collaboration for Leadership in Applied Health Research and Care South West Peninsula at the Royal Devon and Exeter NHS Foundation Trust. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

Published January 2017
PHE publications gateway number: 2016534

Contents

Introduction

This technical report outlines a methodology for estimating the size of the LGB population within England using data generated and reported in the Final Report: 'Producing modelled estimates of the size of the lesbian, gay and bisexual (LGB) population of England'. The aims of this report are to:

- be informed by an in-depth understanding of the existing population measurement tools and an in-depth understanding of population modelling methodologies
- be based on and extend an existing methodology, considering the impact of respondents who decline to answer the question on sexual orientation
- provide a new approach to synthesize survey estimates of the LGB population

In-depth understanding of existing population measurement tools and modelling methodologies

Population data are used by health planners to assess need for health services and demographic modelling allows for a certain amount of prediction. However, populations are heterogeneous with subgroups of relatively high or low need, with disparate ease of access. Lesbian, bisexual and gay communities have been shown to be groups with relative high need. We propose a process created for other high need groups (minority ethnic groups) to use available data and estimate more accurately the prevalence of LGB in the English population. This allows for more accurate health need prediction and, for the first time, a synthesised national estimate of the total size of LGB groups in England.

We combine multiple surveys found through a systematic review in the first stage of the project (Technical Report 1) to derive weights for an aggregated estimate of the LGB population of England. The reference dataset for the national estimates is the 2011 England and Wales Census, or more accurately the most recent census-based population estimates from the Office of National Statistics (ONS). The pooling of specific datasets taking into account their quality and generalisability, yet applying the result to the Census, enables broad yet more robust population estimates.

Quantitative secondary data sources specifically referring to LGB populations were identified in the first part of this project and were from a diverse range of sources:

- inclusion of datasets in the final weights will be based on study quality, generalisability, sampling and applicability to the population of interest
- actual weights will include terms for and thus be sensitive to sample size and proportion of missing data

Our review of existing surveys and measurement tools on sexual identity (Technical Report 1) found a number of factors contributing to the robustness of estimates of the size of the LGB population, reported in Annex A. Ideally, each of these factors would be taken into account when combining surveys into a single estimate, by applying different weights to included surveys.[1]

[1] For example, surveys with smaller samples may be less representative of the general population and would therefore receive a lower weight and contribute less to the synthesized estimate.

However, it is not feasible or desirable to assign weights to all factors influencing survey quality. Fundamental factors such as sampling method or question formation determine the inclusion of a dataset in the final synthesis and are not robustly translated into a weight figure. Conversely, it is feasible to assign weights based on sample size and question non-response rate. Logically then, surveys with a higher sample size and higher question response rate will receive a higher weight in the synthesis.

It was acknowledged that mode of question administration could have an important influence on question response rate and accuracy of data, and thereby on the size of the LGB estimate derived from the survey. The general thought is that self-completed online (and postal) surveys give higher LGB estimates than face-to-face or telephone interviews, because the latter leads to social desirable answers.

The project considered how to design weights that would reflect the proportion of people who could potentially have underreported their sexual orientation. There were two problems with this however. First, evidence on the proportion of people that answers a question differently depending on the mode of administration was limited. Second, even if such a proportion could be arrived from the literature, this would mean that an average group weight would be applied to each survey rather than a survey-specific weight. After all, we do not know how many people actually misreported their sexual orientation in any given survey. While the other weights are survey-specific and based on real figures (eg sample size), a weight for mode of administration would be subjective and was therefore not included.

Here, we build on previous research by one of the team which sought to estimate the prevalence of a ‘write-in’ ethnic category on the 2001 Census and quantify the degree of undercounting. Weighting methods were developed for the purpose,[1,2] where improved Census estimates were derived from weighted means, themselves resulting from aggregated secondary sources which better enumerated that group (ie a Census estimate ‘n’, altered by the secondary source mean ‘x’, to give ‘n± x’). In that particular instance, secondary sources were few and so all were pooled with caveats around outputs.

Three linked approaches were used. The first was simply deriving an aggregated mean of the raw pooled sources. The second used expansion weighted means amended by sample size and best described by Clarke and Cook [3], where a weighted average across categories is taken and weighted according to the total numbers in the category (sample size). In this way larger samples are prioritised in the weight. The third was created including terms for the range and precision using an adapted version of Heyl’s[4] methods described by Hedges and Olkin.[5] However, unlike ethnicity measures where variance is central, in this instance we would amend this to reflect the central problem facing LGB measures; missing data.

In the previous work around ethnic minority groups, differing breakdowns of the population in question were produced based on improved estimates of the group. Thus applying these unstandardized breakdowns to Census population data enabled a more accurate estimate of the prevalence of the group. We were then able to explore key health inequalities with the group to assess the impact of a ‘hidden’ classification on an already marginalised population; something which we might extend the current analyses to include though not within the scope of this project.

This approach can be employed to estimate the LBG population where over-counting is as much of an issue as the reverse, however clearly amendment of the way in which sources are pooled and weighted is necessary in line with known faults in LGB survey tools as described in Technical Report 1.

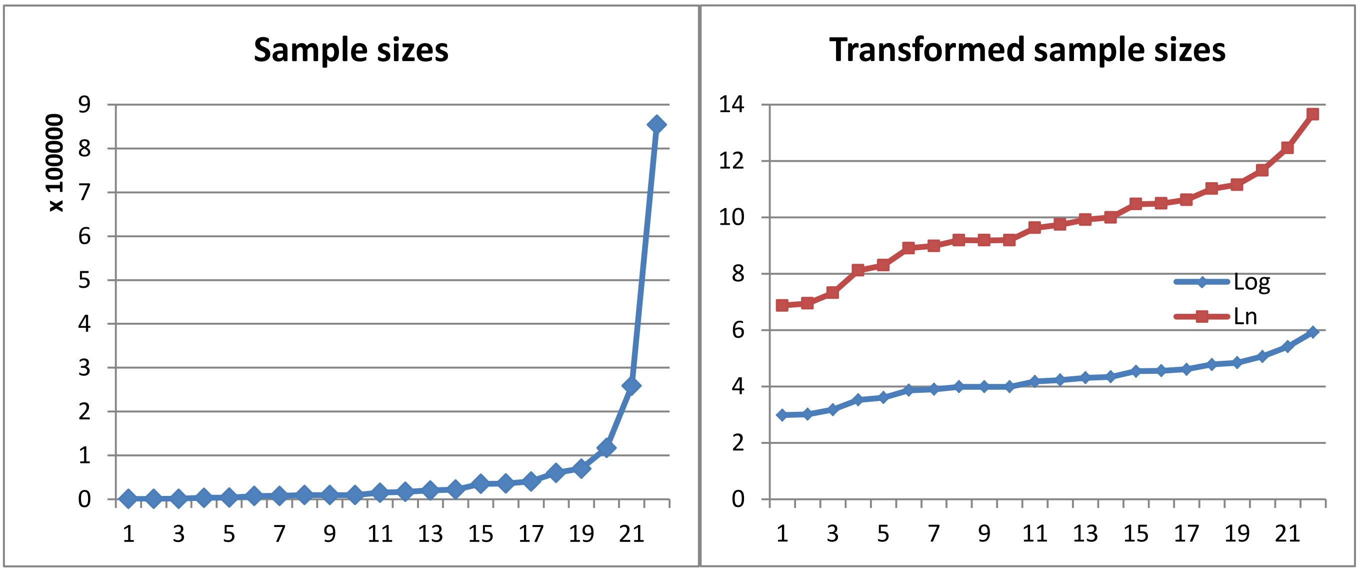
Adopt and extend an existing methodology to synthesize survey estimates of the LGB population

There are four important aspects that require an extension of the existing approach.

First, the most recent 2011 Census did not include a question on sexual identity, so there is no national estimate of the LGB population to use as a base for the newly aggregated means. In this case we simply project the aggregated means onto the national population to derive numbers of the LGB population. In addition, we apply the distribution of LGB individuals across age, gender and ethnicity from the broadest (base) survey to these numbers to generate stratified LGB estimates. The base survey is the broadest and most representative of the population of England. We also explored the production of these estimates for sub-national geographies where possible, however these are contingent on data quality and availability in the broadest survey.

Secondly, a revision and addition is made to the second weighting method that uses an aggregated mean corrected for sample size. Our analysis of the 22 key surveys in Technical Report 1 shows that sample size increases exponentially from the smallest to the largest surveys. Using a 1:1 weight for sample size would pull a mean aggregated estimate considerably towards the largest survey. To avoid overweighting for sample size, we use the logarithmic transformation of sample size instead (Figure 5).

Figure 5: Logarithmic transformation of sample sizes of key surveys



Legend. Log: logarithm; Ln: natural logarithm

Third, in surveys, the size of the sample is effectively a result of the sampled population multiplied by the response rate. When the response rate is low, the study population will most probably be less representative of the target population and therefore survey results will be of lower quality. To account for these variations, the second weighting method calculates an additional aggregated mean weighted by firstly simply the sample size (2a), and also by the sample size and response rate combined (ie total respondents multiplied by the % response rate) (2b). This gives a weight where large, high response surveys are prioritised rather than simply large surveys.

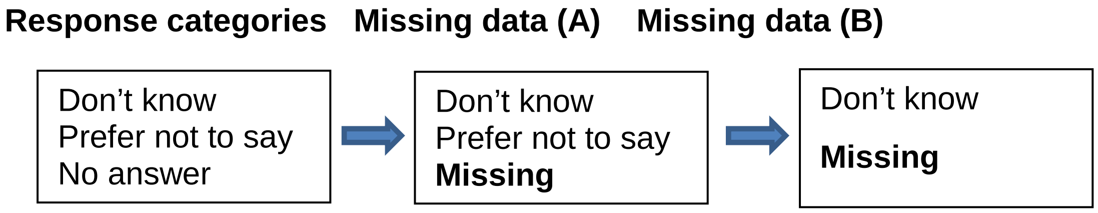
Finally, an amendment is required in the third weighting method, which was relevant for surveys examining Cornish ethnicity where variance was of importance. However, for surveys on sexual identity the most important element of variation is the ratio of missing data and question non-response. Therefore, the third method is adjusted to calculate an aggregated mean weighted by the ratio of question non-response.

Our review of key surveys found four main categories of non-substantive answers: ‘prefer not to say’, ‘refused’, ‘don’t know’, and ‘no answer’ (or a variation of these). ‘Prefer not to say’ and ‘refused’ can be grouped together into ‘prefer not to say’, since they both indicate a respondent not willing to answer the question. There is some evidence breakdown of people answering ‘prefer not to say’ or ‘refused’: being older, female (particularly if the interviewer was male); having no qualifications; belonging to a lower socio-economic group; living in London; and coming from a deprived neighbourhood.[6] Whether ‘prefer not to say’ is an informative answer or represents missing data can be debated. ‘Don’t know’ implies that the respondent does not know their sexual orientation or, perhaps, that the question is not well understood.

There is some suggestion in the literature that a small proportion of heterosexuals may not understand the question and may therefore answer ‘don’t know’.[7] Don’t know is an informative answers and is not considered missing data in our weighting. ‘No answer’ is provided as answer category for any case where an answer is not provided by the respondent, could not be obtained by the interviewer, or is missing altogether. Regardless of the reason, this answer does not provide any information on the respondent group and is therefore classified as missing.

Taking the above into account, the third weight derivation for this project uses two differing approaches to missing data: using either ‘no answers’ (A) or ‘no answers’ and ‘prefer not to say’ (B) as missing data.

This is visualized in the following diagram:



Response categories **Missing data (A)** **Missing data (B)**

Don’t know
Prefer not to say
No answer

Don’t know
Prefer not to say
Missing

Don’t know
Missing

It should be noted that for one survey that grouped ‘don’t know’ and ‘refused’ into a combined category (Integrated Household Survey), these answers were regarded as missing data (‘prefer not to say’) in the weighting methodology.

New approach to synthesize survey estimates of the LGB population

Calculate single survey estimates of the LGB population

Data extracted from each survey provide the proportion of the study population that self-identifies as either heterosexual, lesbian, gay, bisexual or other. The proportion of LGB is the sum of the proportions of gay/lesbian, bisexual and other (ie the sexual minority group). We enumerate 'other' under this heading because the group include people that are unsure about their sexuality, have no sexual feelings, or are against categorisation of gender in general.[8] Thus the generic proportion formula for each dataset is:

Estimate = % lesbian + % gay + % bisexual + % other

The denominator consists of all subjects that were eligible to respond to the question on sexual orientation. This includes the substantive categories mentioned above, plus all non-substantive categories: 'prefer not to say', 'refused', 'don't know', and 'no answer'.

From data already collated we anticipate LGB proportions for each survey potentially needing to be adjusted by one or more of the following:

- Denominator: ensuring the denominator only includes people that were asked the sexual identity question, by excluding cases for whom the item was not applicable[2] (in STATA, recode items not applicable into missing values, eg **recode sexuali2 -1 -2 = .** in the British Social Attitudes Survey);
- Location: ensuring both the numerator and denominator only include people that reside in England (in STATA, limit the tabulation of sexual identity by country of England, eg **tab sexuali2 if Country==1**);
- Adjust for survey design and non-response: adjust proportions by weights produced by investigators[3] (in STATA, tabulate sexual identity using the weighting variable as analytic weight, eg **tab sexuali2 if Country==1 [aweight = WtFactor]**).

[2] Item not applicable is different from no answer. The former means that respondents were not eligible to answer the question, eg because of a proxy respondent, while the latter means that respondents were eligible but did not answer the question.

[3] We are interested in combining estimated from surveys where each is the best possible assessment of the population by the original investigators. As such we will follow due process for each to replicate that, including required weights/adjustments.

Calculate aggregated means of survey estimates

Using each of the individual estimates, which are tabulated for reporting so as to graphically illustrate differences, we combine them in a final synthesis using the discussed approaches.

Method 1: Simple aggregated mean of survey estimates:

- sum all proportions of the LGB population
- divide by the number of surveys

Equation 1: $$\frac{estimate1+estimate2+estimate\ k}{number\ of\ surveys}$$

Method 2a: aggregated mean of survey estimates weighted by log sample size:

- multiply each LGB proportion (e) by log sample size (s) and sum results
- divide by the sum of all log sample sizes

Equation 2a: $$\frac{(e1\times s1)+(e2\times s2)+(ek\times sk)}{s1+s2+sk}$$

Method 2b: aggregated mean of survey estimates weighted by log sample size and response rate:

- multiply each LGB proportion (e) by log sample size (s) and response rate (r) and sum results;
- divide by the sum of all log sample sizes times response rate

Equation 2b: $$\frac{(e1\times s1\times r1)+(e2\times s2\times r2)+(ek\times sk\times rk)}{(s1\times r1)+(s2\times r2)+(sk\times rk)}$$

Method 3a: mean of survey estimates weighted by ratio of missing data (A):

- multiply each LGB proportion (e) by the weight of missing data (w) (=100-% *no answer)* and sum results;
- divide by the sum of all weights for missing data

Equation 3a: $$\frac{(e1\times w1)+(e2\times w2)+(ek\times wk)}{wa1+wa2+wak}$$

Method 3b: mean of survey estimates weighted by ratio of missing data (B):

- multiply each LGB proportion (e) by the weight of missing data (w) (=100-% *no answer + prefer not to say*) and sum results;
- divide by the sum of all weights for missing data

Equation 3b: $$\frac{(e1\times wb1)+(e2\times wb2)+(ek\times wbk)}{wb1+wb2+wbk}$$

Method 4: mean of survey estimates weighted by log sample size, response rate and ratio of missing data:

- multiply each LGB proportion (e) by log sample size (s) and response rate (r) and weight of missing data (w) (=100-% *no answer* + *prefer not to say*) and sum results
- divide by the sum of all log sample sizes times response rates times weights for missing data

Equation 4: $$\frac{(e1 \times s1 \times r1 \times wb1) + (e2 \times s2 \times r2 \times wb2) + (ek \times sk \times rk \times wbk)}{(s1 \times r1 \times wb1) + (s2 \times r2 \times wb2) + (sk \times rk \times wbk)}$$

Apply aggregated means to national population of England

Method 4 incorporates all weights and is considered the most robust method to estimate the size of the LGB population of England. This aggregated weighted mean is then applied to the national population using the latest population estimates by ONS based on from the 2011 England and Wales Census.

This is simply done by multiplying the national population numbers by the mean proportion derived from Method 4.

We also present the national LGB population estimates by: age, gender and ethnicity as well as sub-national geographies where possible, where the broadest and most representative survey population distribution is used as a standard. The distribution of LGB individuals across age, gender, ethnicity and region from the broadest survey is applied to the national population breakdown to get the actual numbers of LGB and ‘others’. After multiplying the base survey population number by the mean proportion derived from Method 4, the total number of LGB people is stratified by age, gender, ethnicity and region according to their distributions in the base survey.

Calculate ranges around estimates

Because the LGB proportion estimates from UK national surveys will expectedly be low (in the range of 0-7%), any variation in the proportion of non-substantive answers will have an important effect on the proportion of LGB. As seen above, not a lot is known about what type of people answer ‘prefer not to say’, ‘don’t know’ or ‘no answer’. We could hypothesize how the mean estimate is affected when these people were either heterosexual or lesbian/gay/bisexual. By calculating the most extreme scenarios where ‘prefer not to say’, ‘don’t know’ or ‘no answer’ were either all heterosexual or all lesbian/gay/bisexual, we can produce maximum ranges around our weighted mean LGB proportion estimates as derived from Method 4.

Conclusions

This report has set out a new approach to synthesize survey estimates of the LGB population of England. Using an amendment of previously developed methods, this will result in five weighted mean proportions based on an aggregation of existing surveys that measure sexual orientation. Surveys are included in the synthesis based on study quality, generalisability, sampling and applicability to the population of interest. Aggregated means are calculated weighted by sample size, response rate, and proportion of missing data. Estimates are stratified by age, gender, ethnicity and sub-national geographies were possible. Lower and upper bounds are estimated based on further assumptions around missing data.

Annexes

Annex A. Scoring of survey methods for generalizability and comparability

Survey characteristics that influence sexual identity estimates	**Represen-tativeness**	**Item response**	**Truthful reporting**
Study population			
Adults in private households, all ages	+++		
Adults in private households, 16-74y	++		
Adults in current or recent employment	+		
Adults registered with a GP	+		
Adolescents, 16-21y	-		
Children in school, 14-15y	-		
All 42y olds born in 1 week	--		
Adult women	-		
Adult patients (cancer, mental health)	--		
Sampling method			
Multi-stage random sampling	?		
Single-stage random sampling	?		
Random digit dialling	-		
Complex stratification	++		
Less complex stratification	+		
No stratification	-		
Frame: Small user postcode address file	++		
Frame: Inter Departmental Business Register	-		
Frame: National Pupil Database	-		
Frame: HSCIC patient registration records	+		
Sample size			
>100,000	+++		
50,000-100,000	++		
10,000-50,000	+		
1,000-10,000	+-		
<1,000	-		
Response rate			
>80%	+++		
%60-80%	++		
%40-60	+		
20-40%	+-		
<20%	-		

Mode of administration			
Face-to-face interview using show card		++	+-
Face-to-face interview self-completion on laptop		+	+-
Face-to-face interview interviewer question		+-	+-
Telephone interview		?	+-
Paper-based self-completion questionnaire		-	+
Online self-completion questionnaire		-	++
Postal survey		-	+
Answered by proxy respondent		--	--
Answered through translator		--	--
Question format			
‘options to describe how you think of yourself’		++	++
‘do you consider yourself to be…’		+	+
‘how to describe your sexual orientation’		+-	+-
Question after religion question		-	
Question at the beginning of survey		+	
Response categories - substantive			
Heterosexual/Straight; Gay/Lesbian; Bisexual; Other		++	++
Heterosexual; Gay; Lesbian; Bisexual; Can’t choose		+	+
Entirely heterosexual; Mostly heterosexual; Bisexual; Entirely gay/lesbian; Mostly gay/lesbian		+-	+-
Response categories – non-substantive			
Prefer not to say (*respondent option*)		-	+
Don’t know (*respondent option*)		-	+
Refused (*interviewer option*)		+-	+
Refused; Don’t know		+	++
Refused; No answer		+-	+
Refused/Don’t know		+-	+-
Refused/No answer		+-	+

References

1. Husk K. Cornish ethnicity and undercounting: utilising the 2001 England and Wales Census to develop an accurate measurement methodology. Methodological Innovations Online. 2012;7(2): 1-12. 10.4256/mio.2012.007
2. Husk K. The legitimation of ethnicity: the case of the Cornish. Studies in Ethnicity and Nationalism. 2012;12(2): 249-67.
3. Clarke G, Cooke D. A basic course in statistics. 4 ed. London: Arnold; 1998.
4. Heyl P. A re-determination of the constant of gravitation.: National Bureau of Standards Journal of Research; 1930.
5. Hedges L, Olkin I. Statistical methods for meta-analysis. London: Academic Press; 1985.
6. Betts P. Developing survey questions on sexual identity: UK experiences of administering survey questions on sexual identity/orientation. Office for National Statistics, 2008.
7. Developing survey questions on sexual identity: Rationale and design of sexual identity questioning on the Integrated Household Survey (IHS). Office for National Statistics, 2008.
8. Joloza T, Evans J, O'Brien R, Potter-Collins A. Measuring sexual identity: an evaluation report. Newport: Office for National Statistics, 2010.