Strathprints Institutional Repository

Sfikas, Nikolaos and Greenhalgh, David and Huo, Wenwen and Mortimer, Janet and Robertson, Chris (2014) Quantifying unrecognised replication present in reports of HIV diagnoses. Statistics in Medicine, 33 (16). pp. 2774-2796. ISSN 0277-6715 , http://dx.doi.org/10.1002/sim.6121

QUANTIFYING UNRECOGNISED REPLICATION PRESENT IN REPORTS OF HIV DIAGNOSES

Nikolaos Sfikas[1], David Greenhalgh[2], Wenwen Huo[2], Janet Mortimer[3] and Chris. Robertson[2,4]

(1) Novartis Pharma AG, Basel, Switzerland.
(2) Department of Mathematics and Statistics, University of Strathclyde, Livingstone Tower, 26 Richmond Street, Glasgow G1 1XH, UK.
(3) Health Protection Agency – Communicable Disease Surveillance Centre (HPA – CDSC), Colindale, London, U.K, retired.
(4) Health Protection Scotland, 5 Cadogan Street, Glasgow G2 6QE, UK.

New diagnoses of HIV infection were reported confidentially to the Public Health Laboratory Service (PHLS) AIDS Centre under a national voluntary surveillance scheme. Two sets of data drawn from the national datasets were made available to us for analysis, the first in 1991, the second in 1994, by which time the replication of reports had been reduced. The data used in the analyses consisted of the numbers of replications of the reported full date of birth in the individual records (one, two, three and so on), for each year of birth. This paper uses a non-parametric maximum likelihood estimation method for quantifying the amount of replication in the data. The estimated amount of replication was 3.37% (95% confidence interval (0.98%,11.83%)) in the 1991 and 0.58% (95% confidence interval (0%,2.64%)) in the 1994 dataset.

Keywords: HIV; AIDS; replication; maximum likelihood estimation; database

1. BACKGROUND

Acquired Immune Deficiency Syndrome (AIDS) is a severe, life-threatening clinical condition. It was first recognised in 1981 and the virus which causes the disease, the Human Immunodeficiency Virus (HIV) was discovered in 1983 [1, 2]. HIV is commonly spread by sexual contact, by injecting drug use, from mother to child and by blood transfusion. Because the largest proportion (about three quarters) of all HIV infections happen through sexual contact, HIV is considered a sexually transmitted infection (STI) [3]. It is estimated that a total of 2.7 million people globally acquired HIV in 2010, down from 3.1 million in 2001 and that by the end of 2010 an estimated 34 million people in the world were living with HIV [4].

The incubation period from HIV infection to AIDS is usually long and variable with a mean of around ten years [1] and with a 95% confidence interval for this mean being [8.4,11.2] years [5]. This means that HIV infected individuals may remain ignorant of their infection for long periods of time, during which they may unknowingly transmit the virus. As HIV is an STI, there is a social stigma attached to a diagnosis of HIV or AIDS. These factors make it difficult to obtain both reliable estimates of the scale of the epidemic and data suitable for analysis [3].

Since the mid 1990s there have been combinations of antiretroviral drugs available which delay the onset of AIDS and increase the lifespan of HIV infected individuals, but a cure for the disease is yet to be found. Data on HIV and AIDS cases in England, Wales and Northern Ireland used to be collected by the Public Health Laboratory Service (PHLS), Colindale, London and for Scotland by the Scottish Centre for Infection and Environmental Health (SCIEH), Glasgow. Because of the sensitivity concerning knowledge of an individual's HIV status, the name of the patient was not held in the databases. On the other hand, because of the serious social and economic cost to both the individual and the nation, it is important that the information available on the number of diagnosed HIV infections is as accurate as possible [6]. In Section 2 we shall discuss the problem of recognising the extent of undetected repeated reporting of the same individual and the previous work of Greenhalgh, Doyle and Mortimer which is very relevant to the work presented here.

2. THE REPLICATION PROBLEM AND PREVIOUS WORK

2.1 The replication problem

A report of all individuals diagnosed as HIV positive was requested by the PHLS, usually through the completion of a short form. The name of the individual was not recorded on this report but the inclusion of both the date of birth and the "Soundex" code (a four character alphanumeric code of the surname) was requested [7]. In 1990 only a third of the records on the database held both Soundex code and full date of birth and at that time the Soundex code was available for only about half the database (Mortimer, 1996, personal communication). Often a local identifier, such as a clinic number, which allowed follow up for the missing information, had been given instead. Even when one or both of the Soundex code or date of birth were available there is still the possibility of mis-recording or transcription errors. It was also likely that some individuals were being repeatedly counted in the database due to having two or more HIV positive tests reported.

There are at least two reasons why an individual already diagnosed as HIV positive may be tested again. Firstly an individual may be unwilling to believe the result and may seek independent confirmation of it by being tested again elsewhere. Secondly, an individual who reports to a new GP or a new clinic as HIV positive is usually retested before receiving any HIV related treatment. The PHLS did all that it could to eliminate multiple counting of such individuals in the database but it wished to be aware of evidence of any multiple counting that still existed. This is not just a hypothetical problem; there is at least one example of an individual having five independent positive HIV tests reported, and it is known that mistakes in Soundex coding and name changes result in records of the same individual remaining unmatched despite the presence of a Soundex code on both records.

The PHLS was interested in a statistical method to test whether individuals were being repeatedly counted in the database from the date of birth data available and in 1991 sent us relevant information from the database as it stood then. This consisted, for each birth year, of the number of birth dates in that year for which there was at least one record in the database and the number of records corresponding to that birth date in the database. No information on Soundex codes or the lack of them for these records was sent to us at that time. The information sent to us in 1991 is displayed in Table 1 in Appendix A. If a given birth date occurs twice in the database it is not possible to tell from this information whether this multiple recording corresponds to one individual recorded twice, or two distinct individuals with the same birth date, and birth dates that occur three times or more in the database have a greater number of possible similar ambiguities. However from the observed statistical distribution of recorded birth dates it is possible to make inferences on whether statistically significant replication of individuals is present and to attempt to estimate the amount of such replication. Between 1991 and 1994, the PHLS was able to improve the quality of their database, both prospectively by ensuring that both date of birth and Soundex code were available for as many new cases as possible, and retrospectively by obtaining Soundex code or date of birth or both to complete reports received previously. This has allowed the elimination from the data of further multiple recording of individuals. In 1994 information from the entire database as it stood then was sent to us for further statistical analysis. This data is displayed in Table 2 in Appendix A.

HIV surveillance systems have of course advanced since 1994. As well as the New Diagnoses reporting system described above the Health Protection Agency (HPA) and Public Health England (PHE), the successors to the PHLS, have cross-sectional annual surveys of prevalent diagnosed HIV infections (SOPHID). These collect reports of all individuals in a calendar year, including those who move area, taking up HIV services offered by the National Health Service (NHS) in England, Wales and Northern Ireland. Neither surveillance system collects names but a Soundex code of the surname, sex and date of birth are held as identifiers [8,9].

2.2 Literature Review

We now give a brief review of work by other authors in the general area of this problem. Larsen [10] discusses estimation of the number of people in a register from the number of birth dates when unique identifiers are not available and when multiple entries can occur. A method for estimating the number of registered people is presented when dates of birth (day, month and year) are available. Registration of people who are HIV positive is cited as an appropriate example.

The problem is clearly related to classical occupancy theory, where *r* balls are placed at random in *n* boxes and the probability of *m* empty boxes is studied. Here *r* has a fixed value whereas Larsen is interested in estimating *r*. Larsen defines *n* to be the number of consecutive days in a sequence of possible birth dates; *r* is the number of registered people born in the sequence; *b* is the number of occupied birth dates in this sequence; and $m=n-b$ is the number of empty birth dates in this sequence. Larsen chooses the approximate maximum likelihood estimate,

$$\hat{r}_0(m,n) = n\log_e(n/m)$$

with approximate variance

$$V(r_0) \sim ne^{-r/n}.$$

An alternative approach is discussed in which *r* is taken to be a random variable reflecting the stochastic nature of registration.

Numerical calculation showed a negative bias for the true maximum likelihood estimator $\hat{r}$ and a small positive bias for the approximate maximum likelihood estimator $\hat{r}_0$. For fixed *r* the exact maximum likelihood was seen to be very near the approximate maximum likelihood estimate, at least for values of *n* near 100. As an example registrations of Chlamydia infections were considered.

The problem which we are studying can be thought of as related to a record linkage problem [11,12]. This addresses the problem of matching two files of individual data under conditions of uncertainty. Individual record linkage involves two files: file A and file B with records pertaining to individual cases [13]. Each individual file is assumed to contain no duplicate records. Obviously, one or more fields on file A must have equivalent fields on file B. For example, in order to match on surname and age, both files must contain fields with this information. The objective of the record linkage process is to decide for each pair of records whether it is a matched or unmatched pair.

A linking variable is a single criterion (such as birth date) utilized to establish or partially establish record linkage [14]. There are two basic methods for record linkage: deterministic linkage, which is effected only when there is an exact match on all linking variables, and the more complex probabilistic linkage, which affects linkage through evaluation of frequencies for linking variable values. Each time that a new set of records is added to the database we are in effect linking two datafiles, the old database and the new records [15]. In general probabilistic data linkage methods are useful in that they can help us rank agreement between different matching variables and they can also be used to incorporate effects such as data transcription errors [13]. However they are not very relevant to the problem of assessing residual duplication in the dataset supplied to us which had just one linking variable (the date of birth), particularly when each value of that linking variable is equally likely.

The efficiency of probabilistic data linkage can be measured by the positive predictive value which is the fraction of linked records which are actually true positives. Blakely and Salmond [16] describe a "duplicate method" to calculate this statistic when each record can be involved in only one match (for example linking population files to death files). Elmagarid, Ipeiritos and Verykios [17] survey methods for deterministic and probabilistic record linkage. They describe algorithms for field matching. They point out that probabilistic data linkage can be regarded as a Bayesian inference problem and describe a likelihood ratio based Bayes decision rule that gives minimum error. They also discuss supervised and unsupervised learning to classify data linkage and other methods.

Ades et al. [18] describe how an unlinked anonymous neonatal seroprevalence survey was used with electronic record linkage to assess HIV prevalence in the UK. In more recent work Rice et al. [9] created a cohort of HIV-diagnosed adults by deterministically linking records across the 1998 to 2007 SOPHID database. The records were also linked to the New Diagnoses database and to Office for National Statistics death records. This was done to examine HIV-service attendance. This is related to the problem studied in this paper as the problem of residual duplication may still exist in the HIV databases used in Rice et al. which would lead to biased results.

Methods to estimate the extent of duplication would therefore be of huge help in providing information on potential bias corrections.

Sometimes it is accepted that the data are imperfect and alternative statistical approaches are used to compensate. For example Goubar et al. [19] estimate HIV prevalence and proportion diagnosed in England and Wales. They take a Bayesian approach with informative priors to synthesise different sources of surveillance information, including the SOPHID database, using Markov Chain Monte Carlo methods. They find that there are inconsistencies in the data but these can be resolved by bias correction. Presanis et al. [20] use this approach to estimate prevalence and incidence of HIV amongst men who have sex with men in England and Wales. This is also related to the problem discussed here for similar reasons as discussed above in relation to the HIV databases of Rice et al.

The problem considered here is not the design of a data linkage mechanism, but rather to assess the efficiency of the data linkage already done in eliminating duplicate reports. Technically our problem is a data linkage one, but there is only one linking variable, date of birth, whereas normally there are several. Deterministic linkage would simply match all records with the same birth date, which would overlink. On the other hand there is not enough information contained in the linking variable for probabilistic linking to be useful. Thus classical data linkage techniques are not appropriate. We shall use maximum likelihood methods to estimate the percentage overcounting present.

2.3 Previous work

We briefly summarise the work of Greenhalgh, Doyle and Mortimer which shows that replication is present in the datasets. The first of these papers [21] examined statistical methods for deciding whether there is a greater amount of replication of birth dates in the sample than expected by chance alone. Greenhalgh and Doyle [15] discussed a statistical method to detect repeatedly counted individuals in the dataset based on the number of matching pairs in the sample. Five of the sixteen birth years tested from the 1991 dataset show evidence of more replication than would be expected by chance alone, using a 5% level test.

Finally, Greenhalgh, Doyle and Mortimer [22] outline a partial ranking method suitable for small sample sizes. This uses a natural partial ordering on the sample space to test whether there are individuals repeatedly counted in the sample.

The partial ranking method cannot be used for larger sample sizes. It is applied to the five birth years in the 1991 dataset. One of those five years shows evidence of more replication of individuals than would be expected from independent random sampling from the population. The results were compared with an alternative maximum likelihood based test which reached the same conclusions. Finally, maximum likelihood methods were further used to estimate the percentage of underlinking of individuals in the sample.

The above papers by Greenhalgh, Doyle and Mortimer conclude that there is a significant amount of replication of individuals in the 1991 dataset. However it is of much more practical interest to the public health authorities to quantify the amount of replication, which we shall do so here. The Day Report [23] gives an accuracy of within 5% when quoting levels of HIV and AIDS incidence so anything smaller than this can be ignored in practice.

3. AVAILABLE DATA

In Tables 1 and 2 we have the data which was provided from the PHLS database in 1991 and 1994 respectively. The reported HIV positive individuals were divided according to their year of birth. In the 1991 dataset, the birth year of those individuals included in the dataset ranged from 1929 to 1944. For every given year we have a sample of size *r*. This is the number of records of individuals who were born in this year and are included in our data.

The sample consists of S_1 singletons, which represent a single birth date, S_2 doubletons, S_3 tripletons and so on up to S_r *r*-tuples with $\sum_{i=1}^{r} iS_i = r$. An *i*-tuple is a birth date which appears in exactly *i* records in the dataset. A singleton represents a unique birth date. A doubleton represents a birth date repeated twice, i.e. two actual records which might or might not correspond to the same individual. Similarly a tripleton represents a birth date repeated three times. The three records having this birth date in our dataset which could correspond to one, two or three distinct individuals. A four-tuple represents a birth date repeated four times and so on. If for example we consider the year 1939, then r=99, s_1=69, s_2=13, s_3=2 and all other s_i's are zero. The same notation applies also to the 1994 dataset, which is presented in Table 2. We adopt the usual convention of using a capital letter (such as S_i) for a random variable and the corresponding lower case letter s_i for a realisation of that random variable.

The problem is that there is no direct way of telling whether there is replication of individuals in the year in question. This is because for instance a tripleton in a year may record three individuals with co-incident birth dates or two individuals, with one reported twice and the other with a co-incident birth date with that individual or finally a single individual reported three times. Similarly a doubleton records either one or two distinct individuals, a four-tuple one, two, three or four distinct individuals and so on.

4. THE MAXIMUM LIKELIHOOD METHOD

For a given birth year the sample size of birth records is r and the replication vector is $\boldsymbol{S}_B = (S_1, S_2, \ldots S_r)$. Here the subscript B denotes the birth year. We denote a typical outcome for $\boldsymbol{S}_B$ by $\boldsymbol{s}_B = (s_1, s_2, \ldots s_r)$. This means that the collection of r birth records contains exactly s_1 unique birth dates, s_2 birth dates repeated exactly twice, s_3 birth dates repeated exactly three times, ... and s_r birth dates repeated exactly r times.

However some individuals may have more than one birth record in the database (due to having more than one HIV positive test result). If we know which birth records correspond to which distinct individuals then we can calculate the true replication vector $\boldsymbol{T}_B = (T_1, T_2, ..., T_r)$ of birth records corresponding to distinct individuals (with no multiple counting of individuals). A typical outcome is denoted $\boldsymbol{t}_B = (t_1, t_2, \ldots ,t_r)$. For example if the observed birth record replication vector is

$$s_1 = 10,\ s_2 = 5 \text{ and } s_3 = 2$$

(ten singleton birth records, five doubleton birth records and two tripleton birth records) and we know that for two of the five doubleton birth records this doubleton actually corresponds to an individual who has had two HIV positive tests and one of the tripleton birth records actually corresponds to an individual who has had one HIV positive test and a second individual who has had two HIV positive tests, and all other birth records correspond to individuals who have had just one test

$$t_1 = 12,\ t_2 = 4 \text{ and } t_3 = 1.$$

Suppose that n denotes the number of days in a year, that all individuals are independent and each day of the year is equally likely to be a birth date. Then

Theorem 1 in Appendix B gives the probability of occurrence of the birth record replication vector $\boldsymbol{T_B}$ where all birth records correspond to distinct individuals.

As a matter of fact there is a small but statistically significant seasonal variation in the birth rate with births being more likely in the summer than the winter. It is possible to modify Theorem 1 to take this into account. However our previous work on statistical tests for whether replication was present in the dataset found that it made no significant difference to the likelihood function [15]. Hence for simplicity and because we were not given data on the actual calendar days on which birth dates were repeated we ignore it here. Theorem 1 also assumes that individuals in a given birth year are sampled randomly without replacement from the population consisting of all people born in that birth year. The size of the database is very small compared to this population so the fact that individuals are sampled without replacement can be ignored.

If the observed birth record replication vector is $\boldsymbol{s_B}$ then there are at most $r = 1s_1 + 2s_2 + \dots + rs_r$ distinct individuals (this will be the case if every birth record in the observed replication vector corresponds to a distinct individual) and at least $\sigma = s_1+s_2+ \dots +s_r$ distinct individuals (this will be the case if every birth date in the observed birth record replication vector corresponds to one distinct individual repeated an appropriate number of times). So if $r^* = \sum_{i=1}^{r} it_i$ denotes the true number of distinct individuals in the observed birth record replication vector then $r \geq r^* \geq \sigma$. We assume that the probability distribution for the number of reported positive HIV tests of a random individual recorded on the database is given by the unknown probability distribution p_i, $i = 1,2,3, \dots$. We write $\boldsymbol{p} = (p_1,p_2,p_3, \dots)$. Of course $p_i \geq 0$ for $i \geq 1$ and $\sum_{i=1}^{\infty} p_i = 1$.

For each of the possible values of r^* we can calculate the likelihood function

$$L(r^*,\boldsymbol{p}|\boldsymbol{s_B}).$$

The exact calculation is complicated and detailed in Appendix B. This likelihood function is then maximised over r^* and $\boldsymbol{p}$, to give the maximum likelihood estimators $\widehat{r^*}$ of r^*, the true number of distinct individuals in the sample, and $\widehat{\boldsymbol{p}}$ of $\boldsymbol{p}$ the distribution of the number of reported HIV tests that an individual has had.

Once the maximum likelihood estimates are obtained we are able to calculate the percentage of replication in a given birth year. This can be estimated as either

$$100\frac{r-\widehat{r^*}}{\widehat{r^*}}\ \%$$

where r is the observed sample size and $\widehat{r^*}$ is the estimated sample size or

$$100\sum_{i=2}^{\widehat{r^*}}(i-1)\hat{p}_i\ \%$$

where $\hat{\boldsymbol{p}}$ is the estimated distribution of the number of reported HIV tests that an individual has had. In practice both of these methods give very similar results so we present the results only for the first one. For instance for the birth year 1925 the percentages of replication calculated by the two methods are to two decimal places 6.45% and 6.44% respectively.

As an example for the birth year 1930 the observed birth record replication vector is $\boldsymbol{s}_{1930}$=(23,1) and r=25. We have two possible values of r^* that should be examined. The first is if there is no replication present, r^*=25 and the true replication vector of birth records of distinct individuals is $\boldsymbol{t}_{1930}$=(23,1), and the second is if the doubleton is actually a single individual tested twice. In that case if we get rid of the replication present we have r^*=24 and $\boldsymbol{t}_{1930}$=(24,0).

In the first case the likelihood is

$$P(\boldsymbol{s}_{1930}|\boldsymbol{p},r^*=25) = P_0p_1^{25}$$

where $P_0 = P(t_1=23,\ t_2=1,\ t_3=t_4=t_5= \ldots =t_{25}=0)$ and in the second case the likelihood is

$$P(\boldsymbol{s}_{1930}|\boldsymbol{p},r^*=24) = 24P_1p_1^{23}p_2$$

where $P_1 = P(t_1=24,\ t_2=t_3= \ldots =t_{25}=0)$. P_0 and P_1 can be calculated from Theorem 1.

Hence the overall likelihood function is

$$L(r^*,\boldsymbol{p}|\boldsymbol{s}_{1930}) = \begin{cases} P_0p_1^{25}, & r^* = 25, \\ 24P_1p_1^{23}p_2, & r^* = 24. \end{cases} \tag{1}$$

This needs to be maximised over $r^*$and $\boldsymbol{p}$ to find the maximum likelihood estimators, $\widehat{r^*}$ and $\hat{\boldsymbol{p}}$. Once the maximum likelihood estimates are obtained we are able to calculate the amount of replication in a given birth year.

Suppose that the data observed for that birth year is $\boldsymbol{s}_B = (s_1,s_2,\ldots,s_r)$ and that there are actually r^* distinct individuals in the database for that year. We can then calculate the likelihood function

$$L(r^*,\boldsymbol{p}|\boldsymbol{s}_B)$$

of the unknown parameters $\boldsymbol{\theta} = (r^*,\boldsymbol{p})$ given the data $\boldsymbol{s}_B$ where $\boldsymbol{p} = (p_1,p_2,p_3, \ldots)$ and after that obtain the maximum likelihood estimates $\widehat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$ given the observed data. To find the maximum likelihood value Lemma 1 [22], presented in Appendix B is sometimes useful for simple cases but for more complicated situations numerical maximisation is needed.

For example for the birth year 1930 discussed earlier if r^*=25 the likelihood function in (1) is maximised over $\boldsymbol{p}$ at p_1=1, p_2=p_3 ... =p_{25}=0 and if r^*=24 using Lemma 1 it is maximised over $\boldsymbol{p}$ at p_1=$\frac{23}{24}$, p_2=$\frac{1}{24}$, p_3=p_4= ... =p_{25}=0. The overall maximum likelihood is the maximum of these two values.

As a second example for birth year 1934 of the 1991 dataset and for r^*=36 the likelihood function

$$L(r^*,\boldsymbol{p}|\boldsymbol{s}_{1934}) = 637{,}086{,}601\ p_1^{22} p_2^{14}$$

and by using Lemma 1, we find that this is maximised over $\boldsymbol{p}$ at p_1= 0.6111 and p_2= 0.3889 and the rest of the p_i 's are zero. In Section 5 we present in detail the results for the year 1934 and some of the results for the year 1935 of the 1991 dataset in order to show how the maximum likelihood method works.

5. REPLICATION AND OVERCOUNTING

In Section 6 we present the summary results for both datasets. But because it is quite difficult to understand how the maximum likelihood method works, we present first in detail how we obtained the results for two years. As stated before for $i \geq 1$, p_i is defined as the probability that an HIV positive individual who has had at least one positive test recorded in the dataset has actually had exactly i positive HIV tests recorded. Of course $p_i \geq 0$ for $i \geq 1$ and $\sum_{i=1}^{\infty} p_i = 1$.

In Tables 3 and 4 (see Appendix A) we present the results for the years 1934 and 1935 for the 1991 dataset. We give the results for all the possible outcomes for the true replication vector $\boldsymbol{t}_B$, for a given value of r^*, when we eliminate individuals repeatedly recorded in the sample, where r^* is the true number of individuals in the sample (given in column one). The second column contains the true replication vector or vectors corresponding to this r^* and the third column the likelihood function $L(r^*,\boldsymbol{p}|\boldsymbol{s}_B)$ of r^* and $\boldsymbol{p}$ given the observed data. Then the following column presents the probabilities which maximise the likelihood function (for that value of r^*). The last column has the maximised likelihood function $L(r^*,\hat{\boldsymbol{p}}(r^*)|\boldsymbol{s}_B)$ (given r^*) .

For 1934 the observed replication vector is $\boldsymbol{s}_{1934}$ = (22,14). Hence the true number of individuals is somewhere between 36 (if each of the doubletons correspond to exactly one distinct individual) and 50 (if each of the doubletons correspond to exactly two distinct individuals). If r^*=36 then the true replication vector after individuals repeatedly counted have been eliminated is (36,0), if r^*=37 then the true replication vector is (35,1) and so on up to r^*=50 when the true replication vector is (22,14).

If r^*=50 then the probability that all 50 individuals have had only one birth date recorded is p_1^{50}. The independent probability that there are 22 singleton birthdays and 14 doubleton birthdays is calculated to be 4.26794x10^{-9} with the use of Theorem 1. So in that case

$$L(r^*=50,\boldsymbol{p}|\boldsymbol{s}_{1934}) = 4.26794\text{x}10^{-9}p_1^{50}$$

which has a maximum value of 4.26794x10^{-9}at $\hat{p}_1 = 1, \hat{p}_2 = \hat{p}_3 = ... = 0$.

In the second case of r^*=49 we get the observed replication vector precisely when exactly one individual has had two recorded positive tests and this individual has a distinct birthday from everyone else. The rest of the individuals must have had exactly one test. So after the elimination of the double counting, the true replication vector is (23,13). The probability of 49 randomly chosen individuals giving rise to the true replication vector of birth dates is calculated using Theorem 1 to be 3.7929x10^{-8}. The probability that exactly one out of the 23 singleton birth dates corresponds to an individual who has had two recorded HIV tests and everyone else has had exactly one is $23p_1^{48}p_2$. So the likelihood function is:

$$L(r^*=49,\boldsymbol{p}|\boldsymbol{s}_{1934}) = 3.7929\text{x}10^{-8} \text{ x } 23\, p_1^{48}p_2 = 8.72366\text{x}10^{-7}\, p_1^{48}p_2.$$

By Lemma 1, this has a maximum value of 6.61714x10^{-9}, when $\hat{p}_1 = 0.9796, \hat{p}_2 = 0.0204, \hat{p}_3 = \hat{p}_4 = ... = 0$.

In the same way, for r^*=48, the probability of 48 randomly chosen individuals, each counted once giving rise to the replication vector (24,12) is 3.06076x10^{-7}. But in order to get this replication vector out of the original data, we must have exactly two individuals out of the 24 singleton birth dates corresponding to an individual who has had two recorded tests and all the other 46 individuals in the sample having had just one recorded test. The likelihood is then:

$$L(r^* =48, \boldsymbol{p}\, |\boldsymbol{s}_{1934}) = 3.06076\text{x}10^{-7}\, \frac{23.24}{2!}\, p_1^{46}p_2^2 = 8.44771\text{x}10^{-5}\, p_1^{46}p_2^2.$$

The maximum value for this function is 2.07052x10^{-8} when $\hat{p}_1 = 0.9583, \hat{p}_2 = 0.0417, \hat{p}_3 = \hat{p}_4 = ... = 0$.

Finally after calculating all the possible outcomes for the true replication vector and the respective maximum likelihood functions, we estimate that for the year 1934 in the 1991 dataset, there were 14 individuals reported exactly twice and the rest of the individuals were reported exactly once. So the corresponding true replication vector would be t_1=36, t_2=t_3=t_4=0, and the maximised likelihood function was 0.0227 for $\hat{p}_1 = 0.6111, \hat{p}_2 = 0.3889, \hat{p}_3 = \hat{p}_4 = ... = 0$.

For the birth year 1935 the calculations are more complicated because of the presence of a four-tuple in the observed data. For example we can see that if the true value of r^*=53, then we have two possible true replication vectors corresponding to the same value of r^*. That means that the maximum likelihood function for r^*=53 is $L(r^* = 53, \boldsymbol{p} | \boldsymbol{s}_{1935}) = 41P(41,4,0,1)$ x $p_1^{52}p_2 + 3P(40,5,1,0)$ x $p_1^{52}p_2 = 0.0750\, p_1^{52}p_2$, the maximised value of which is calculated to be 0.000525, when the estimated probabilities are $\hat{p}_1 = 0.9811, \hat{p}_2 = 0.0189, \hat{p}_3 = \hat{p}_4 = ... = 0$. Here the notation $P(t_1,t_2,t_3..., t_r)$ denotes the probability of observing the birth date replication vector $\boldsymbol{t}_B$ = $(t_1,t_2,t_3..., t_r)$ if there are $r^* = \sum_{i=1}^{r} it_i$ distinct individuals in the sample and each has only one recorded HIV test.

It is obvious from the next cases for the year 1935 that we can have different replication vectors corresponding to the same value of r^* and even different factors corresponding to the same replication vectors or vice versa. Different factors corresponding to the same true replication vector occurs for r^*=47 where both $p_1^{41}p_2^5p_3$ and $p_1^{42}p_2^4p_4$ correspond to the true replication vector (45,1,0,0). We presented the situation of different replication vectors corresponding to the same value of r^* for r^*=53.

Finally after calculating all the possible outcomes for the true replication vector and the respective maximum likelihood functions, we came to the conclusion that for 1935, it is most likely that r^*=51 (and hence there was one person counted four times and fifty one people counted exactly once). The true replication vector was estimated as (t_1,t_2,t_3,t_4) = (41,5,0,0) and the maximised likelihood for this vector was 0.0288, which occurred when $\hat{p}_1 = 0.9804$, $\hat{p}_2 = 0.0$, $\hat{p}_3 = 0.0$, $\hat{p}_4 = 0.0196$, $\hat{p}_5 = \hat{p}_6 = ... = 0$. As the number of observations in a year becomes even larger the number of combinations that have to be taken into account increases dramatically.

6. SUMMARY RESULTS

We used two different programs for these calculations. Two versions were written, one in Fortran, one in C. For each possible value of r^*, the first program calculated all the possibilities for the true replication vector $\boldsymbol{T}_B= (T_1, T_2, \ldots T_{11})$ giving the same value for r^* and the relative probabilities for those vectors. Then for each value of r^*, the second program calculated the likelihood function $L(r^*,\boldsymbol{p}|\boldsymbol{s}_B)$ using the probabilities from the first program, and maximised it with respect to $\boldsymbol{p}$, calculated the maximising value $\hat{\boldsymbol{p}}$ and the partially maximised likelihood function $L(r^*, \hat{\boldsymbol{p}}\,(r^*)|\boldsymbol{s}_B)$ (given r^*). Further details of the programs and algorithm used are given in Sfikas [24].

We present for the 1991 dataset, wherever it is possible, all the possible outcomes for the replication vector together with the respective calculated maximum likelihood estimators. For this dataset for the birth years 1934-1941, $\hat{p}_i$ was always zero for $i \geq 5$ so we give only $\hat{p}_1$, $\hat{p}_2$, $\hat{p}_3$ and $\hat{p}_4$. For the birth years and 1942-1944, $\hat{p}_i$ was always zero for $i \geq 7$, so we give only $\hat{p}_1$, $\hat{p}_2$, $\hat{p}_3$, $\hat{p}_4$, $\hat{p}_5$ and $\hat{p}_6$. For the 1994 dataset it was difficult to give as much detail for the vast majority of birth years, due to the large number of HIV positive records in the observed data and the large size of the subsequent possible outcomes. So for this dataset we just present $\widehat{r^*}$ and $\hat{\boldsymbol{p}}(\widehat{r^*})$ and the maximised likelihood function $L(\widehat{r^*},\hat{\boldsymbol{p}}(\widehat{r^*})|\boldsymbol{s}_B)$. As for this dataset $\hat{p}_i$ was always zero for $i \geq 12$ we give only $\hat{p}_1, \hat{p}_2,\ldots, \hat{p}_{10}$ and $\hat{p}_{11}$.

In Tables 5 and 6 we present the results for the 1991 dataset and in Table 7 the results for the 1994 dataset. The first column of Table 5 gives the year of birth and the second one all the possible outcomes for the true number r^* of distinct individuals in the sample born during that year. The third column shows for each possible outcome for r the true replication vector (or vectors) that correspond to the specific value of r. The fourth column gives the likelihood function for each value of r and $\boldsymbol{p}$. The maximised likelihood function is given in the last column. For each year we also give the values $\widehat{r^*}$ of r^* and $(\hat{p}_1, \hat{p}_2, \hat{p}_3, \hat{p}_4)$ that maximise the likelihood function.

We should also mention that for the years 1929-1933 of the 1991 database the results were presented in full by Greenhalgh, Doyle and Mortimer [22]. So they are briefly summarised in Table 6 together with the results for the birth years 1942-1944 which could not be analytically presented in full due to their length.

In the 1991 dataset we estimated that there was some replication in five out of the sixteen birth years for which we had data. The years where replication was estimated to be present were: 1931, 1934, 1935, 1943 and 1944. The amount of replication that we estimated to be present was 37 records out of 1,097 which is equivalent to a 3.37% proportion of replication.

For the 1994 dataset the results are summarised in Table 7. The first column again gives the year of birth of the individuals found HIV positive, the second column shows the observed total number of HIV positive records in this birth year and the third column the estimated number of distinct individuals by using the maximum likelihood method. The replication is the "observed number" r minus the "estimated number" $\widehat{r^*}$. Column four gives us the probabilities which maximise our likelihood estimator. Further details could not be presented due to lack of space. Finally, the last column gives the maximised likelihood estimator (with the estimated $\widehat{r^*}$ as shown in the third column).

We found that there was replication present in 16 years out of the 64 years examined. The amount of replication present in this dataset was 100 records out of a sample of 17,272 which is the actual total number of individuals tested according to our research after we get rid of the replication present. This gives us a replication of 0.58% of the total number of distinct individuals present. Thus the replication found in the 1994 dataset is less than one fifth of the one found in the 1991 dataset. This can be attributed to the fact that the PHLS was able to eliminate much but not all of the replication when collecting the 1994 dataset, by improving their methods of identifying repeatedly counted individuals.

Hence we have obtained a point estimate of the percentage replication in the PHLS HIV test data. However point estimates by themselves are of limited value and it is more useful to have some indication of the amount of uncertainty associated with these estimates. To do this we calculated 95% bootstrap confidence intervals based on the percentile method. However because the bootstrap distribution of the estimated percentage amount of replication was skew, rather than use the simple percentile method we use a more appropriate method which reflected an adjusted version of the bootstrap distribution about the estimated value [25]. This is sometimes called the centred bootstrap percentile interval. To be more precise for a given birth year with r

observed individual records, we estimated $\widehat{r^*}$, the number of distinct individuals in the sample, as above, and $\widehat{\boldsymbol{p}}(\widehat{r^*})$, the probability distribution of the number of reported HIV tests that an individual has had. Next we took a random bootstrap sample of $\widehat{r^*}$ independent individuals whose birth dates were chosen at random and each of whom was independently assigned a number of HIV tests according to the distribution $\widehat{\boldsymbol{p}}(\widehat{r^*})$. Then for this simulated observed replication vector the maximum likelihood method was again applied to estimate the amount of replication corresponding to that observed replication vector and the percentage replication calculated. This was repeated with 100 random bootstrap samples to give a distribution of the amount of replication in a sample and a bootstrap confidence interval calculated as follows:

Suppose that η denotes the true percentage replication in our sample. For each bootstrap sample we calculate $\eta^* - \hat{\eta}$ where η^* is the estimated percentage replication in the bootstrap sample and $\hat{\eta}$is the estimated percentage replication. From these we find the empirical values δ^L and δ^U such that 2.5% of the adjusted observations lie below δ^L and 2.5% lie above δ^U. Hence we deduce the 95% bootstrap confidence interval for the true percentage replication η as

$$(\hat{\eta} - \delta^U, \hat{\eta} - \delta^L).$$

The estimated percentage replication for the PHLS 1991 dataset is shown in Table 8 together with the associated 95% bootstrap confidence intervals. These were calculated using our C program. From this table it is clear that the birth years with the wider confidence intervals tend to have relatively high estimated probabilities of an individual having two or more HIV tests. Note that the point estimates for the percentage replication always lie within the corresponding 95% bootstrap confidence interval supporting the validity of the method. From the confidence intervals we deduce that a 95% confidence interval for the overall amount of replication in this dataset is (0.98%,11.83%).

Using the C program, we used the parametric bootstrap method to generate 95% bootstrap confidence intervals for the 1994 dataset. These are shown in Table 9. The birth years with the narrower confidence intervals again tend to have relatively low estimated probabilities of an individual having two or more HIV tests and relatively high numbers of birth records, thus relatively high numbers of individuals. This is what would be expected. Again point estimates for the percentage replication always lie within the appropriate 95% bootstrap confidence interval. From these confidence intervals we deduce that a 95% confidence interval for the overall amount of replication in this dataset is (0%,2.64%).

7. VALIDATION OF THE METHOD

We take a simple example to test the validity of the method. We take the artificial replication vector (9,1) as an example so that nine individuals have had exactly one HIV test and one individual has had exactly two HIV tests. We allocate the birth dates at random to each individual to construct the observed birth record replication vector then use the maximum likelihood method to estimate the amount of replication in this vector. The process was repeated 10,000 times with results as shown in Table 10.

We had to perform the simulation a large number of times to get an accurate distribution for the observed birth record replication vector and the program took a long time to run. Hence we had to choose a small true number of individuals (10). Consequently the resulting bootstrap confidence intervals were quite wide. Nonetheless the method is validated quite well. In 87.8% of cases the sample size was estimated correctly and in a further 11.7% of cases it was just one out. In the 87.8% of cases where the sample size was estimated correctly the probability vector was also estimated correctly.

The 95% bootstrap confidence intervals associated with the observed bootstrap replication vector were typically quite large. For example for the replication vector (9,1) this was (0%,31.75%) and for (8,0,1) it was (0%,66.67%). Note that the simulations to calculate the observed bootstrap replication vector were conditional on there being exactly one individual in the dataset who had had exactly two reported HIV tests and that the simulations to calculate the 95% bootstrap confidence intervals were conditional on both the estimated sample size and the number of reported HIV tests that an individual had had following the estimated probability distribution.

8. DISCUSSION

We were given two datasets by the PHLS AIDS Surveillance Centre, containing only numbers of repeated birth dates by year of birth for individuals whose HIV positive test had been reported. As, for various reasons, many people with HIV infection are tested more than once, there was potential for some replication present in these datasets. Our aim was to quantify the replication. A maximum likelihood estimation method was used along with the parametric bootstrap method to construct the corresponding 95% confidence intervals. We estimated that the replication present in the 1991 dataset was 3.37% (95% confidence interval (0.98%,11.83%)) and for the 1994 dataset 0.58% (0%,2.64%). We noted before that we expected the replication to be smaller in the 1994 database because both the current and retrospective rates of inclusion in records of surname Soundex code had improved by that date. This made for more efficient elimination of replicate reports of the same individual from the national database than had been possible earlier. This improvement must be offset by people presenting for testing under different names, a problem that can only increase over time. Of course, the accuracy of the estimation process relies on the available information, and in this sense any estimate of replication would provide a minimum value for the true replication. Underlinking may particularly affect certain subgroups, such as women who are, through marriage, more likely than men to change their surname, and thus its Soundex code, and patients of foreign origin for whom surname may be less consistently distinguished from forename than is the case for those born in the UK. This would tend to distort the view of the epidemic in the UK. Female patients and patients from abroad with heterosexually acquired HIV may historically have been differentially overrepresented in the database due to overlinking, but this differential contribution has yet to be estimated.

In 2011 there were an estimated 6,280 people in the UK newly diagnosed as HIV infected [26]. In view of the numbers and problems already described it is obvious that there will continue to be underlinking of reports which in fact relate to the same individual.

The work presented here ignores the possible inconsistencies in the recording of dates of birth for individuals reported more than once; the presence of these will lead to unquantifiable numbers of unrecognised repeat records. Those making the reports were fully aware of the importance of accuracy and double entry was used at the PHLS AIDS Surveillance Centre with the aim of minimising such errors.

A preliminary version of the method used in this paper was used in [22] to analyse the amount of replication in five birth years from the 1991 dataset where the sample sizes were very small. The current paper very substantially extends this work as it analyses the entire 1991 and 1994 datasets where the sample sizes are very substantially increased and the method is much more complex. It also extends the work in [22] by calculation of bootstrap confidence intervals for the percentage replication.

Importantly for the 1991 dataset, the years where we found replication by the method used here were the same as the ones identified by the matching pairs method [15]. The same conclusion is arrived at by two distinct methods, re-inforcing our confidence in the results.

REFERENCES

1. Anderson RM, May RM. *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press: Oxford, 1991.
2. Benenson S. *Control of Communicable Diseases in Man*. 16th ed., American Public Health Association: Washington, D.C., 1990.
3. Doyle MT. *Modelling Some Aspects of the AIDS Pandemic*. Ph.D. thesis, University of Strathclyde: Glasgow, UK, 1994.
4. Chan, M., Sidibé, M. And Lake, A. Global HIV/AIDS response – Epidemic update and health sector progress towards Universal Access – Progress Report 2011, Joint World Health Organisation, UNAIDS, UNICEF Report, World Health Organisation, Geneva, Switzerland, 2011.
5. Longini IM, Clark WS, Byers RH, Ward JW, Darrow WW, Lemp GF, Hethcote HW. Statistical analysis of the stages of HIV infection using a Markov model. *Statistics in Medicine* 1989; **8**:831-843.
6. Evans G, Howitt DJ, Mortimer J. Surveillance of HIV infection and AIDS in the U.K.: an overview from the PHLS AIDS Centre. *PHLS Microbiology Digest* 1993; **10**:141-143.
7. Mortimer J, Salathiel JA. 'Soundex' codes of surnames provide confidentiality and accuracy in a national HIV database. *Communicable Disease Report* 1995; **5**:183-186.
8. McHenry A, Macdonald N, Sinka K, Mortimer J, Evans B. National assessment of prevalent diagnosed HIV infections, *Communicable Disease and Public Health* 2000;**3**:277-281.
9. Rice BD, Delpech DC, Chadborn TR, Elford J. Loss to follow-up among adults attending Human Immunodeficiency Virus Services in England, Wales and Northern Ireland, *Sexually Transmitted Diseases*, 2011:**38**:685-690.
10. Larsen SO. Estimation of the number of people in a register from the number of birthdates. *Statistics in Medicine* 1994; **13**:177-183.
11. Acheson ED. *Medical Record Linkage*. Oxford University Press: London, 1976.
12. Newcombe HB. *Handbook of Record Linkage Methods for Health and Statistical Studies, Administration and Business.* Oxford University Press: London, 1988.
13. Jaro MA. Probabilistic linkage of large public health data files. *Statistics in Medicine* 1995; **14**:491-498.
14. Shevenko IP, Lynch JT, Mattie AS, Reed-Fourquet LI. Verification of information in a large medical database using linkages with an external database. *Statistics in Medicine* 1995; **14**:511-530.
15. Greenhalgh D, Doyle MT. A test to detect replication in HIV serological data labelled by birthdate based on the number of matching pairs in a sample. *Statistics in Medicine* 1999; **18**:1641-1656.

16. Blakely T, Salmond C. Probabilistic record linkage and a method to calculate the positive predictive value, *International Journal of Epidemiology* 2002;**31**:1246-1252.
17. Elmagarid AK, Ipeiritos PG, Verykios VS. Duplicate record detection: a survey. *IEEE Transactions on Knowledge and Data Engineering* 2007;**19**:1-16.
18. Ades AE, Walker J, Botting B, Parker S, Cubitt, D, Jones R. Effect of the worldwide epidemic on HIV prevalence in the United Kingdom: record linkage in anonymous neonatal seroprevalence surveys. *AIDS* 1999;**13**:2437-2443.
19. Goubar A, Ades AE, De Angelis D, McGarrigle CA, Mercer CH, Tookey PA, Fenton K, Gill ON. Estimates of human immunodeficiency virus prevalence and proportion diagnosed based on Bayesian multiparameter synthesis of surveillance data. *J. R. Statist. Soc. A* 2008;**171**:541-580.
20. Presanis AM, De Angelis D, Goubar A, Gill ON, Ades AE. Bayesian evidence synthesis for a transmission dynamic model for HIV among men who have sex with men. *Biostatistics* 2011;**12**:666-681.
21. Doyle MT, Greenhalgh D, Mortimer J. Three statistical tests for detecting overcounting of individuals in serological test data, *Applied Stochastic Models and Data Analysis*. 1998; **13**:307-314.
22. Greenhalgh D, Doyle MT, Mortimer J. A partial ranking method for identifying repeated inclusion of individuals in anonymised HIV infection reports. *Biometrics*. 1999; **55**:165-173
23. Day NE. The incidence and prevalence of AIDS and other severe HIV disease in England and Wales for 1992-1997. *Communicable Disease Report*. 1993; **3(Supplement 1)**:S1-S17.
24. Sfikas N. *Mathematical Models for Vaccination Programs and Statistical Analysis of Infectious Diseases of Humans*. Ph.D. thesis, University of Strathclyde: Glasgow, UK, 1999.
25. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. Chapman and Hall: New York, 1993.
26. Health Protection Agency. HIV in the United Kingdom: 2012 Report, London, Health Protection Services, Colindale, London, 2012.

APPENDIX A.

Year of Birth	r	S_1	S_2	S_3	S_4	S_5	S_6
1929	**28**	26	1	-	-	-	-
1930	**25**	23	1	-	-	-	-
1931	**26**	19	2	1	-	-	-
1932*	**27**	23	2	-	-	-	-
1933	**44**	38	3	-	-	-	-
1934	**50**	22	14	-	-	-	-
1935	**54**	40	5	-	1	-	-
1936*	**52**	48	2	-	-	-	-
1937	**68**	57	4	1	-	-	-
1938	**78**	66	6	-	-	-	-
1939	**99**	67	13	2	-	-	-
1940*	**87**	71	8	-	-	-	-
1941	**83**	63	10	-	-	-	-
1942	**124**	86	13	4	-	-	-
1943	**113**	69	17	2	1	-	-
1944*	**176**	104	24	6	-	-	1

Table 1. 1991 dataset sent to us by the PHLS AIDS Centre. Birth years are tabulated in ascending order and leap years are indicated with an asterisk.

Year of Birth	r	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	S_{11}
1901	**0**	-	-	-	-	-	-	-	-	-	-	-
1902	**0**	-	-	-	-	-	-	-	-	-	-	-
1903	**1**	1	-	-	-	-	-	-	-	-	-	-
1904*	**0**	-	-	-	-	-	-	-	-	-	-	-
1905	**2**	2	-	-	-	-	-	-	-	-	-	-
1906	**0**	-	-	-	-	-	-	-	-	-	-	-
1907	**0**	-	-	-	-	-	-	-	-	-	-	-
1908*	**1**	1	-	-	-	-	-	-	-	-	-	-
1909	**0**	-	-	-	-	-	-	-	-	-	-	-
1910	**0**	-	-	-	-	-	-	-	-	-	-	-
1911	**2**	2	-	-	-	-	-	-	-	-	-	-
1912*	**4**	4	-	-	-	-	-	-	-	-	-	-
1913	**5**	5	-	-	-	-	-	-	-	-	-	-
1914	**10**	10	-	-	-	-	-	-	-	-	-	-
1915	**5**	5	-	-	-	-	-	-	-	-	-	-
1916*	**4**	2	1	-	-	-	-	-	-	-	-	-
1917	**7**	5	1	-	-	-	-	-	-	-	-	-
1918	**6**	6	-	-	-	-	-	-	-	-	-	-
1919	**10**	8	1	-	-	-	-	-	-	-	-	-
1920*	**6**	6	-	-	-	-	-	-	-	-	-	-
1921	**3**	3	-	-	-	-	-	-	-	-	-	-
1922	**11**	7	2	-	-	-	-	-	-	-	-	-
1923	**13**	13	-	-	-	-	-	-	-	-	-	-
1924*	**19**	17	1	-	-	-	-	-	-	-	-	-
1925	**33**	28	1	1	-	-	-	-	-	-	-	-
1926	**20**	17	-	1	-	-	-	-	-	-	-	-
1927	**24**	22	1	-	-	-	-	-	-	-	-	-
1928*	**30**	26	2	-	-	-	-	-	-	-	-	-
1929	**41**	39	1	-	-	-	-	-	-	-	-	-
1930	**43**	35	4	-	-	-	-	-	-	-	-	-
1931	**52**	37	6	1	-	-	-	-	-	-	-	-
1932*	**59**	51	4	-	-	-	-	-	-	-	-	-
1933	**74**	68	3	-	-	-	-	-	-	-	-	-
1934	**82**	60	8	2	-	-	-	-	-	-	-	-
1935	**78**	59	8	1	-	-	-	-	-	-	-	-
1936*	**95**	69	10	2	-	-	-	-	-	-	-	-
1937	**118**	86	16	-	-	-	-	-	-	-	-	-
1938	**129**	96	12	3	-	-	-	-	-	-	-	-
1939	**156**	94	25	4	-	-	-	-	-	-	-	-
1940*	**143**	106	17	1	-	-	-	-	-	-	-	-
1941	**149**	105	19	2	-	-	-	-	-	-	-	-
1942	**212**	101	43	7	1	-	-	-	-	-	-	-
1943	**202**	115	28	9	1	-	-	-	-	-	-	-
1944*	**280**	127	49	14	2	1	-	-	-	-	-	-
1945	**279**	118	55	11	2	2	-	-	-	-	-	-
1946	**320**	130	56	18	6	-	-	-	-	-	-	-
1947	**411**	133	69	35	6	1	1	-	-	-	-	-
1948*	**392**	128	66	27	9	3	-	-	-	-	-	-
1949	**418**	150	66	29	11	1	-	-	-	-	-	-
1950	**430**	131	68	36	10	3	-	-	-	-	-	-
1951	**444**	137	78	33	6	3	1	1	-	-	-	-
1952*	**515**	131	91	41	13	3	2	-	-	-	-	-
1953	**485**	133	75	35	11	6	1	-	1	1	-	-
1954	**591**	110	78	53	28	7	2	1	-	-	-	-
1955	**624**	88	104	57	24	11	1	-	-	-	-	-
1956*	**648**	130	92	61	16	8	3	3	1	-	-	-

Year of Birth	*r*	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	S_{11}
1957	**724**	103	99	57	30	15	2	4	1	1	-	-
1958	**770**	84	107	62	40	15	5	3	-	-	-	-
1959	**798**	86	91	75	37	18	4	5	1	-	-	-
1960*	**890**	87	92	71	37	25	8	7	2	1	-	1
1961	**858**	79	96	72	48	19	9	2	2	-	-	-
1962	**929**	68	107	55	47	29	13	4	3	1	1	-
1963	**880**	82	80	75	47	25	12	4	-	-	-	-
1964*	**856**	80	87	76	42	18	12	5	-	1	-	-
1965	**703**	107	92	71	28	9	7	-	-	-	-	-
1966	**639**	109	96	52	22	15	2	1	-	-	-	-
1967	**508**	114	88	38	16	8	-	-	-	-	-	-
1968*	**380**	123	68	24	11	1	-	-	-	-	-	-
1969	**294**	112	63	12	2	1	-	1	-	-	-	-
1970	**221**	107	41	9	-	1	-	-	-	-	-	-
1971	**125**	92	12	3	-	-	-	-	-	-	-	-
1972*	**76**	62	7	-	-	-	-	-	-	-	-	-
1973	**35**	31	2	-	-	-	-	-	-	-	-	-

Table 2. 1994 dataset sent to us by the PHLS AIDS Centre. Birth years are tabulated in ascending order and leap years are indicated with an asterisk.

r^*	True Replication Vector $\boldsymbol{t}_{1934}$	$L(r^*,\boldsymbol{p}\|\boldsymbol{s}_{1934})$	$(\hat{p}_1, \hat{p}_2, \hat{p}_3, \hat{p}_4)$ (given r^*)	$L(r^*,\hat{\boldsymbol{p}}(r^*)\|\boldsymbol{s}_{1934})$
50	(22,14)	$4.26794\text{x}10^{-9}\ p_1^{50}$	(1.0,0,0,0)	$4.27\text{x}10^{-9}$
49	(23,13)	$8.72366\text{x}10^{-7}\ p_1^{48}p_2$	(0.9796,0.0204,0,0)	$6.62\text{x}10^{-9}$
48	(24,12)	$8.44771\text{x}10^{-5}\ p_1^{46}p_2^{2}$	(0.9583,0.0417,0,0)	$2.07\text{x}10^{-8}$
47	(25,11)	$5.13903\text{x}10^{-3}\ p_1^{44}p_2^{3}$	(0.9362,0.0638,0,0)	$7.34\text{x}10^{-8}$
46	(26,10)	$0.2195020\ p_1^{42}p_2^{4}$	(0.9130,0.0870,0,0)	$2.75\text{x}10^{-7}$
45	(27, 9)	$6.9668027\ p_1^{40}p_2^{5}$	(0.8889,0.1111,0,0)	$1.06\text{x}10^{-6}$
44	(28, 8)	$169.52553\ p_1^{38}p_2^{6}$	(0.8636,0.1364,0,0)	$4.15\text{x}10^{-6}$
43	(29, 7)	$3{,}214.3802\ p_1^{36}p_2^{7}$	(0.8372,0.1628,0,0)	$1.63\text{x}10^{-5}$
42	(30, 6)	$47{,}748.497\ p_1^{34}p_2^{8}$	(0.8095,0.1905,0,0)	$6.27\text{x}10^{-5}$
41	(31, 5)	$553{,}276.23\ p_1^{32}p_2^{9}$	(0.7805,0.2195,0,0)	$2.36\text{x}10^{-4}$
40	(32, 4)	$4{,}925{,}507.9\ p_1^{30}p_2^{10}$	(0.75,0.25,0,0)	$8.39\text{x}10^{-4}$
39	(33, 3)	$32{,}687{,}461\ p_1^{28}p_2^{11}$	(0.7179,0.2821,0,0)	0.002745
38	(34, 2)	$152{,}960{,}558\ p_1^{26}p_2^{12}$	(0.6842,0.3158,0,0)	0.007803
37	(35, 1)	$452{,}069{,}665\ p_1^{24}p_2^{13}$	(0.6486,0.3514,0,0)	0.017301
36	**(36, 0)**	$\mathbf{637{,}086{,}601}\ \boldsymbol{p_1^{22}p_2^{14}}$	**(0.6111,0.3889,0,0)**	**0.022720**

Table 3. Results for year 1934 of the 1991 dataset. The maximum likelihood estimator for the true replication vector is given in bold letters.

r^*	True Replication Vector $\boldsymbol{t}_{1935}$	$L(r^*,\boldsymbol{p}\vert\boldsymbol{s}_{1935})$	$(\hat{p}_1, \hat{p}_2, \hat{p}_3, \hat{p}_4)$ (given r^*)	$L(r^*,\hat{\boldsymbol{p}}(r^*)\vert\boldsymbol{s}_{1935})$
54	(40,5,0,1)	0.000504 p_1^{54}	(1.0,0,0,0)	0.000504
53	(41,4,0,1) (40,5,1,0)	0.074989 $p_1^{52}p_2$	(0.9811,0.0189,0,0)	0.000525
52	(42,3,0,1) (41,4,1,0) (40,6,0,0)	3.755868 $p_1^{50}p_2^2$+ 0.563380 $p_1^{51}p_3$	(0.981,0, 0.019,0)	0.004024
51	**(43,2,0,1)** **(42,3,1,0)** **(41,5,0,0)**	**92.27157** $\boldsymbol{p_1^{48}p_2^3}$**+** **39.54496** $\boldsymbol{p_1^{49}p_2p_3}$**+** **3.954496** $\boldsymbol{p_1^{50}p_4}$	**(0.9804,0,0, 0.0196)**	**0.028808**
50	(44,1,0,1) (43,2,1,0) (42,4,0,0)	1,226.411 $p_1^{46}p_2^4$+ 1,132.072 $p_1^{47}p_2^2p_3$+ 283.0179 $p_1^{48}p_2p_4$	(0.96, 0.02,0, 0.02)	0.015985
49	(45,0,0,1) (44,1,1,0) (43,3,0,0)	8,546.151 $p_1^{44}p_2^5$+ 16,528.24 $p_1^{45}p_2^3p_3$+ 8,264.122 $p_1^{46}p_2^2p_4$	(0.9385,0.041,0,0.0205)	0.015424
48	(45,0,1,0) (44,2,0,0)	24,623.71 $p_1^{42}p_2^6$+ 123,118.6 $p_1^{43}p_2^4p_3$+ 123,118.6 $p_1^{44}p_2^3p_4$	(0.9167, 0.0625,0, 0.0208)	0.013704
47	(45,1,0,0)	374,485.6 $p_1^{41}p_2^5p_3$+ 936,214.0 $p_1^{42}p_2^4p_4$	(0.8936, 0.0851,0, 0.0213)	0.009279
46	(46,0,0,0)	2,908,237.5 $p_1^{40}p_2^5p_4$	(0.8696, 0.1087,0, 0.0217)	0.003581

Table 4. Results for year 1935 of the 1991 dataset. The maximum likelihood estimator for the true replication vector is given in bold letters.

Year of Birth	r^*	True Replication Vector $\boldsymbol{t}_B$	$L(r^*,\boldsymbol{p}\mid\boldsymbol{s}_B)$	$L(r^*,\widehat{\boldsymbol{p}}(r^*)\mid\boldsymbol{s}_B)$
1934	50	(22,14)	4.26794x10^{-9} p_1^{50}	4.27x10^{-9}
	49	(23,13)	8.72366x10^{-7} $p_1^{48}p_2$	6.62x10^{-9}
	48	(24,12)	8.4477x10^{-5} $p_1^{46}p_2^2$	2.07x10^{-8}
	47	(25,11)	5.1390x10^{-3} $p_1^{44}p_2^3$	7.34x10^{-8}
	46	(26,10)	0.2195020 $p_1^{42}p_2^4$	2.75x10^{-7}
	45	(27, 9)	6.9668027 $p_1^{40}p_2^5$	1.06x10^{-6}
	44	(28, 8)	169.52553 $p_1^{38}p_2^6$	4.15x10^{-6}
	43	(29, 7)	3,214.3802 $p_1^{36}p_2^7$	1.62x10^{-5}
	42	(30, 6)	47,748.497 $p_1^{34}p_2^8$	6.27x10^{-5}
	41	(31, 5)	553,276.23 $p_1^{32}p_2^9$	2.35x10^{-4}
	40	(32, 4)	4,925,507.9 $p_1^{30}p_2^{10}$	8.39x10^{-4}
	39	(33, 3)	32,687,461 $p_1^{28}p_2^{11}$	0.002745
	38	(34, 2)	152,960,558 $p_1^{26}p_2^{12}$	0.007803
	37	(35, 1)	452,069,665 $p_1^{24}p_2^{13}$	0.017301
	36	**(36, 0)**	**637,086,601** $\boldsymbol{p_1^{22}p_2^{14}}$	**0.022720**
1934	$\widehat{r}^*$**=36, (** $\widehat{\boldsymbol{p}}_1,\widehat{\boldsymbol{p}}_2,\widehat{\boldsymbol{p}}_3,\widehat{\boldsymbol{p}}_4$**) = (0.6111,0.3889,0,0)**			
1935	54	(40,5,0,1)	0.000504 p_1^{54}	0.000504
	53	(41,4,0,1)	0.074989 $p_1^{52}p_2$	0.000525
		(40,5,1,0)		
	52	(42,3,0,1)	3.755868 $p_1^{50}p_2^2$+	0.004024
		(41,4,1,0)	0.563380 $p_1^{51}p_3$	
		(40,6,0,0)		
	51	**(43,2,0,1)**	**92.271572** $\boldsymbol{p_1^{48}p_2^3}$**+**	**0.028808**
		(42,3,1,0)	**39.54496** $\boldsymbol{p_1^{49}p_2p_3}$**+**	
		(41,5,0,0)	**3.954496** $\boldsymbol{p_1^{50}p_4}$	
	50	(44,1,0,1)	1,226.4107 $p_1^{46}p_2^4$+	0.015985
		(43,2,1,0)	1,132.07 $p_1^{47}p_2^2p_3$+	
		(42,4,0,0)	283.0179 $p_1^{48}p_2p_4$	
	49	(45,0,0,1)	8,546.1509 $p_1^{44}p_2^5$+	0.015424
		(44,1,1,0)	16,528.24 $p_1^{45}p_2^3p_3$+	
		(43,3,0,0)	8,264.122 $p_1^{46}p_2^2p_4$	
	48	(45,0,1,0)	24,623.712 $p_1^{42}p_2^6$+	0.013704
		(44,2,0,0)	123,118.6 $p_1^{43}p_2^4p_3$+	
			123,118.55 $p_1^{44}p_2^3p_4$	
	47	(45,1,0,0)	374,485.6 $p_1^{41}p_2^5p_3$+	0.009279
			936,214.0 $p_1^{42}p_2^4p_4$	
	46	(46,0,0,0)	2,908,237.5 $p_1^{40}p_2^5p_4$	0.003581
1935	$\widehat{r}^*$**=51, (** $\widehat{\boldsymbol{p}}_1,\widehat{\boldsymbol{p}}_2,\widehat{\boldsymbol{p}}_3,\widehat{\boldsymbol{p}}_4$**) = (0.98,0,0,0.02)**			
1936	**52**	**(48,2)**	**0.181447** $\boldsymbol{p_1^{52}}$	**0.181447**
	51	(49,1)	5.094471 $p_1^{50}p_2$	0.037113
	50	(50,0)	36.56032 $p_1^{48}p_2^2$	0.008244
1936	$\widehat{r}^*$**=52, (** $\widehat{\boldsymbol{p}}_1,\widehat{\boldsymbol{p}}_2,\widehat{\boldsymbol{p}}_3,\widehat{\boldsymbol{p}}_4$**) = (1,0,0,0)**			
1937	**68**	**(57,4,1)**	**0.045946** $\boldsymbol{p_1^{68}}$	**0.045946**
	67	(58,3,1)	3.452670 $p_1^{66}p_2$	0.019100
		(57,5,0)		
	66	(59,2,1)	96.7337 $p_1^{64}p_2^2$+	0.045276
		(58,4,0)	8.061138 $p_1^{65}p_3$	
	65	(60,1,1)	1,307.696 $p_1^{62}p_2^3$+	0.012565
		(59,3,0)	356.6443 $p_1^{63}p_2p_3$	
	64	(61,0,1)	8,678.344 $p_1^{60}p_2^4$+	0.005448
		(60,2,0)	6,008.085 $p_1^{61}p_2^2p_3$	
	63	(61,1,0)	22,843.24 $p_1^{58}p_2^5$+	0.001769
			45,686.48 $p_1^{59}p_2^3p_3$	
	62	(62,0,0)	132,345.7 $p_1^{57}p_2^4p_3$	0.000306
1937	$\widehat{r}^*$**=68, (** $\widehat{\boldsymbol{p}}_1,\widehat{\boldsymbol{p}}_2,\widehat{\boldsymbol{p}}_3,\widehat{\boldsymbol{p}}_4$**) = (1,0,0,0)**			

Year of Birth	r^*	True Replication Vector $\boldsymbol{t}_B$	$L(r^*,\boldsymbol{p}\vert\boldsymbol{s}_B)$	$L(r^*,\widehat{\boldsymbol{p}}(r^*)\vert\boldsymbol{s}_B)$
1938	**78**	**(66,6)**	$\mathbf{0.104460}\,\boldsymbol{p_1^{78}}$	**0.104460**
	77	(67,5)	$5.865815\,p_1^{76}p_2$	0.028208
	76	(68,4)	$139.02744\,p_1^{74}p_2^{2}$	0.013381
	75	(69,3)	$1{,}780.52689\,p_1^{72}p_2^{3}$	0.006029
	74	(70,2)	$12{,}997.8467\,p_1^{70}p_2^{4}$	0.002269
	73	(71,1)	$51{,}288.8067\,p_1^{68}p_2^{5}$	0.000621
	72	(72,0)	$85{,}481.2751\,p_1^{66}p_2^{6}$	0.000092
1938	$\boldsymbol{\widehat{r}^*}$**=78, (** $\boldsymbol{\widehat{p}_1,\widehat{p}_2,\widehat{p}_3,\widehat{p}_4}$ **) = (1,0,0,0)**			
1939	**99**	**(67,13,2)**	$\mathbf{0.0100108}\,\boldsymbol{p_1^{99}}$	**0.010011**
	98	(68,12,2) (67,14,1)	$1.4025257\,p_1^{97}p_2$	0.005292
	97	(69,11,2) (68,13,1) (67,15,0)	$96.02954\,p_1^{95}p_2^{2}+$ $1.6495871\,p_1^{96}p_3$	0.006289
	96	(70,10,2) (69,12,1) (68,14,0)	$4{,}367.806\,p_1^{93}p_2^{3}+$ $198.6307\,p_1^{94}p_2p_3$	0.006958
	95	(71,9,2) (70,11,1) (69,13,0)	$99{,}506.85\,p_1^{91}p_2^{4}+$ $11{,}044.95\,p_1^{92}p_2^{2}p_3$ $70.8010\,p_1^{93}p_3^{2}$	0.006239
	94	(72,8,2) (71,10,1) (70,12,0)	$1{,}999{,}201\,p_1^{89}p_2^{5}+$ $377{,}208\,p_1^{90}p_2^{3}p_3+$ $7{,}072.65\,p_1^{91}p_2p_3^{2}$	0.006573
	93	(73,7,2) (72,9,1) (71,11,0)	$30{,}410{,}620\,p_1^{87}p_2^{6}+$ $8{,}861{,}372\,p_1^{88}p_2^{4}p_3+$ $329{,}555\,p_1^{89}p_2^{2}p_3^{2}$	0.006625
	92	(74,6,2) (73,8,1) (72,10,0)	$357{,}270{,}106\,p_1^{85}p_2^{7}+$ $1.518\text{x}10^{8}\,p_1^{86}p_2^{5}p_3+$ $9{,}485{,}047\,p_1^{87}p_2^{3}p_3^{2}$	0.006321
	91	(75,5,2) (74,7,1) (73,9,0)	$3.274\text{x}10^{9}\,p_1^{83}p_2^{8}+$ $1.957\text{x}10^{9}\,p_1^{84}p_2^{6}p_3+$ $1.882\text{x}10^{8}\,p_1^{85}p_2^{4}p_3^{2}$	0.005630
	90	(76,4,2) (75,6,1) (74,8,0)	$2.344\text{x}10^{10}\,p_1^{81}p_2^{9}+$ $1.93\text{x}10^{10}\,p_1^{82}p_2^{7}p_3+$ $2.717\text{x}10^{9}\,p_1^{83}p_2^{5}p_3^{2}$	0.004609
	89	(77,3,2) (76,5,1) (75,7,0)	$1.303\text{x}10^{11}\,p_1^{79}p_2^{10}+$ $1.47\text{x}10^{11}\,p_1^{80}p_2^{8}p_3+$ $2.938\text{x}10^{10}\,p_1^{81}p_2^{6}p_3^{2}$	0.003415
	88	(78,2,2) (77,4,1) (76,6,0)	$5.538\text{x}10^{11}\,p_1^{77}p_2^{11}+$ $8.57\text{x}10^{11}\,p_1^{78}p_2^{9}p_3+$ $2.41\text{x}10^{11}\,p_1^{79}p_2^{7}p_3^{2}$	0.002230
	87	(79,1,2) (78,3,1) (77,5,0)	$1.749\text{x}10^{12}\,p_1^{75}p_2^{12}+$ $3.78\text{x}10^{12}\,p_1^{76}p_2^{10}p_3+$ $1.50\text{x}10^{12}\,p_1^{77}p_2^{8}p_3^{2}$	0.001241
	86	(80,0,2) (79,2,1) (78,4,0)	$3.902\text{x}10^{12}\,p_1^{73}p_2^{13}+$ $1.22\text{x}10^{13}\,p_1^{74}p_2^{11}p_3+$ $6.99\text{x}10^{12}\,p_1^{75}p_2^{9}p_3^{2}$	0.000600
	85	(80,1,1) (79,3,0)	$5.615\text{x}10^{12}\,p_1^{71}p_2^{14}+$ $2.73\text{x}10^{13}\,p_1^{72}p_2^{12}p_3+$ $2.37\text{x}10^{13}\,p_1^{73}p_2^{10}p_3^{2}$	0.000240
	84	(81,0,1) (80,2,0)	$4.276\text{x}10^{12}\,p_1^{69}p_2^{15}+$ $3.80\text{x}10^{13}\,p_1^{70}p_2^{13}p_3+$ $5.56\text{x}10^{13}\,p_1^{71}p_2^{11}p_3^{2}$	0.000074
	83	(81,1,0)	$2.48\text{x}10^{13}\,p_1^{68}p_2^{14}p_3+$ $8.05\text{x}10^{13}\,p_1^{69}p_2^{12}p_3^{2}$	0.000015
	82	(82,0,0)	$5.45\text{x}10^{13}\,p_1^{67}p_2^{13}p_3^{2}$	0.000002

Year of Birth	r^*	True Replication Vector $\boldsymbol{t}_B$	$L(r^*,\boldsymbol{p}\vert\boldsymbol{s}_B)$	$L(r^*,\widehat{\boldsymbol{p}}(r^*)\vert\boldsymbol{s}_B)$
1939	$\widehat{r}^*=99, (\widehat{p}_1,\widehat{p}_2,\widehat{p}_3,\widehat{p}_4) = (1,0,0,0)$			
1940	**87**	**(71,8)**	**0.083757** p_1^{87}	**0.083757**
	86	(72,7)	5.637743 $p_1^{85}p_2$	0.024258
	85	(73,6)	167.9523 $p_1^{83}p_2^{2}$	0.012886
	84	(74,5)	2,892.732 $p_1^{81}p_2^{3}$	0.006926
	83	(75,4)	31,510.11 $p_1^{79}p_2^{4}$	0.003434
	82	(76,3)	222,317.1 $p_1^{77}p_2^{5}$	0.001475
	81	(77,2)	992,293.6 $p_1^{75}p_2^{6}$	0.000510
	80	(78,1)	2,562,112 $p_1^{73}p_2^{7}$	0.000126
	79	(79,0)	2,930,416 $p_1^{71}p_2^{8}$	0.000017
1940	$\widehat{r}^*=87, (\widehat{p}_1,\widehat{p}_2,\widehat{p}_3,\widehat{p}_4) = (1,0,0,0)$			
1941	**83**	**(63,10)**	**0.056044** p_1^{83}	**0.056044**
	82	(64,9)	4.929215 $p_1^{81}p_2$	0.022250
	81	(65,8)	197.46915 $p_1^{79}p_2^{2}$	0.016704
	80	(66,7)	4,745.7607 $p_1^{77}p_2^{3}$	0.013190
	79	(67,6)	75,783.866 $p_1^{75}p_2^{4}$	0.010113
	78	(68,5)	840,337.56 $p_1^{73}p_2^{5}$	0.007219
	77	(69,4)	6,553,914.7 $p_1^{71}p_2^{6}$	0.004623
	76	(70,3)	35,505,437.3 $p_1^{69}p_2^{7}$	0.002539
	75	(71,2)	127,889,651 $p_1^{67}p_2^{8}$	0.001119
	74	(72,1)	276,620,586 $p_1^{65}p_2^{9}$	0.000312
	73	(73,0)	272,882,184 $p_1^{63}p_2^{10}$	0.000059
1941	$\widehat{r}^*=83, (\widehat{p}_1,\widehat{p}_2,\widehat{p}_3,\widehat{p}_4) = (1,0,0,0)$			

Table 5. Birth years 1934-1941. Summary results for the 1991 dataset.

Year of Birth	Observed r	Estimated $\widehat{r^*}$	$(\hat{p}_1, \hat{p}_2, \hat{p}_3, \hat{p}_4, \hat{p}_5, \hat{p}_6)$	$L(\widehat{r^*}, \hat{\boldsymbol{p}}(\widehat{r^*}) \vert \boldsymbol{s}_B)$
1929	**28**	28	(1,0,0,0,0,0)	0.386
1930	**25**	25	(1,0,0,0,0,0)	0.379
1931	**26**	22	(0.864,0.091,0.045,0,0,0)	0.056
1932	**27**	27	(1,0,0,0,0,0)	0.170
1933	**44**	44	(1,0,0,0,0,0)	0.211
1942	**124**	124	(1,0,0,0,0,0)	0.007085
1943	**113**	102	(0.902,0.093,0.005,0,0,0)	0.001600
1944	**176**	171	(0.994,0,0,0,0,0.006)	0.001699

Table 6. Birth years 1929-1933 and 1942-1944. Summary results for the 1991 dataset.

Year of Birth	r	$\widehat{r^*}$	$(\hat{p}_1,\hat{p}_2,\hat{p}_3,\hat{p}_4,\hat{p}_5,\hat{p}_6,\hat{p}_7,\hat{p}_8,\hat{p}_9,\hat{p}_{10},\hat{p}_{11})$	$L(\widehat{r^*},\hat{\boldsymbol{p}}(\widehat{r^*})\|s_B)$
1903	**1**	1	(1,0,0,0,0,0,0,0,0,0,0)	1.0
1904	**0**	0	----------------------	---
1905	**2**	2	(1,0,0,0,0,0,0,0,0,0,0)	0.9973
1906	**0**	0	----------------------	---
1907	**0**	0	----------------------	---
1908	**1**	1	(1,0,0,0,0,0,0,0,0,0,0)	1.0
1909	**0**	0	----------------------	---
1910	**0**	0	----------------------	---
1911	**2**	2	(1,0,0,0,0,0,0,0,0,0,0)	0.9973
1912	**4**	4	(1,0,0,0,0,0,0,0,0,0,0)	0.9837
1913	**5**	5	(1,0,0,0,0,0,0,0,0,0,0)	0.9729
1914	**10**	10	(1,0,0,0,0,0,0,0,0,0,0)	0.8831
1915	**5**	5	(1,0,0,0,0,0,0,0,0,0,0)	0.9729
1916	**4**	3	(0.667,0.333,0,0,0,0,0,0,0,0,0)	0.4408
1917	**7**	6	(0.833,0.167,0,0,0,0,0,0,0,0,0)	0.3856
1918	**6**	6	(1,0,0,0,0,0,0,0,0,0,0)	0.9596
1919	**10**	9	(0.889,0.111,0,0,0,0,0,0,0,0,0)	0.3529
1920	**6**	6	(1,0,0,0,0,0,0,0,0,0,0)	0.9597
1921	**3**	3	(1,0,0,0,0,0,0,0,0,0,0)	0.9918
1922	**11**	9	(0.778,0.222,0,0,0,0,0,0,0,0,0)	0.2771
1923	**13**	13	(1,0,0,0,0,0,0,0,0,0,0)	0.8056
1924	**19**	19	(1,0,0,0,0,0,0,0,0,0,0)	0.3060
1925	**33**	31	(0.968,0,0.032,0,0,0,0,0,0,0,0)	0.1309
1926	**20**	18	(0.944,0,0.056,0,0,0,0,0,0,0,0)	0.2472
1927	**24**	24	(1,0,0,0,0,0,0,0,0,0,0)	0.3726
1928	**30**	30	(1,0,0,0,0,0,0,0,0,0,0)	0.2139
1929	**41**	41	(1,0,0,0,0,0,0,0,0,0,0)	0.2444
1930	**43**	43	(1,0,0,0,0,0,0,0,0,0,0)	0.1045
1931	**52**	47	(0.894,0.106,0,0,0,0,0,0,0,0,0)	0.1852
1932	**59**	59	(1,0,0,0,0,0,0,0,0,0,0)	0.1802
1933	**74**	74	(1,0,0,0,0,0,0,0,0,0,0)	0.0388
1934	**82**	78	(0.974,0,0.026,0,0,0,0,0,0,0,0)	0.0161
1935	**78**	78	(1,0,0,0,0,0,0,0,0,0,0)	0.0400
1936	**95**	95	(1,0,0,0,0,0,0,0,0,0,0)	0.0203
1937	**118**	118	(1,0,0,0,0,0,0,0,0,0,0)	0.0229
1938	**129**	129	(1,0,0,0,0,0,0,0,0,0,0)	0.0110
1939	**156**	156	(1,0,0,0,0,0,0,0,0,0,0)	0.0086
1940	**143**	143	(1,0,0,0,0,0,0,0,0,0,0)	0.0123
1941	**149**	149	(1,0,0,0,0,0,0,0,0,0,0)	0.0188
1942	**212**	212	(1,0,0,0,0,0,0,0,0,0,0)	4.3x10^{-4}
1943	**202**	202	(1,0,0,0,0,0,0,0,0,0,0)	0.0021
1944	**280**	280	(1,0,0,0,0,0,0,0,0,0,0)	7.3x10^{-4}
1945	**279**	279	(1,0,0,0,0,0,0,0,0,0,0)	5.5x10^{-5}
1946	**320**	320	(1,0,0,0,0,0,0,0,0,0,0)	4.4x10^{-4}
1947	**411**	411	(1,0,0,0,0,0,0,0,0,0,0)	2.9x10^{-5}
1948	**392**	392	(1,0,0,0,0,0,0,0,0,0,0)	5.2x10^{-5}
1949	**418**	418	(1,0,0,0,0,0,0,0,0,0,0)	2.1x10^{-5}
1950	**430**	430	(1,0,0,0,0,0,0,0,0,0,0)	4.2x10^{-5}
1951	**444**	439	(0.998,0,0,0,0,0.002,0,0,0,0,0)	5.2x10^{-6}
1952	**515**	515	(1,0,0,0,0,0,0,0,0,0,0)	2.2x10^{-5}
1953	**485**	469	(0.994,0,0.002,0,0,0,0,0.004,0,0,0)	2.7x10^{-7}
1954	**591**	591	(1,0,0,0,0,0,0,0,0,0,0)	1.3x10^{-7}
1955	**624**	624	(1,0,0,0,0,0,0,0,0,0,0)	7.4x10^{-8}
1956	**648**	632	(0.994,0,0,0,0.006,0,0,0,0,0,0)	2.1x10^{-9}

Year of Birth	r	$\widehat{r^*}$	$(\hat{p}_1, \hat{p}_2, \hat{p}_3, \hat{p}_4, \hat{p}_5, \hat{p}_6, \hat{p}_7, \hat{p}_8, \hat{p}_9, \hat{p}_{10}, \hat{p}_{11})$	$L(\widehat{r^*}, \hat{\mathbf{p}}(\widehat{r^*}) \vert s_B)$
1957	**724**	714	(0.997,0,0,0,0,0.003,0,0,0,0,0)	7.3x10^{-9}
1958	**770**	770	(1,0,0,0,0,0,0,0,0,0,0)	2.8x10^{-7}
1959	**798**	798	(1,0,0,0,0,0,0,0,0,0,0)	3.7x10^{-8}
1960	**890**	877	(0.998,0,0,0.001,0,0,0,0,0,0,0.001)	3.2x10^{-11}
1961	**858**	858	(1,0,0,0,0,0,0,0,0,0,0)	2.1x10^{-7}
1962	**929**	917	(0.998,0,0,0,0,0,0.002,0,0,0,0)	7.9x10^{-12}
1963	**880**	880	(1,0,0,0,0,0,0,0,0,0,0)	3.4x10^{-8}
1964	**856**	856	(1,0,0,0,0,0,0,0,0,0,0)	5.1x10^{-9}
1965	**703**	703	(1,0,0,0,0,0,0,0,0,0,0)	4.8x10^{-7}
1966	**639**	639	(1,0,0,0,0,0,0,0,0,0,0)	6.9x10^{-7}
1967	**508**	508	(1,0,0,0,0,0,0,0,0,0,0)	2.5x10^{-6}
1968	**380**	380	(1,0,0,0,0,0,0,0,0,0,0)	3.8x10^{-5}
1969	**294**	288	(0.997,0,0,0,0,0,0.003,0,0,0,0)	7.9x10^{-6}
1970	**221**	217	(0.995,0,0,0,0.005,0,0,0,0,0,0)	2.3x10^{-4}
1971	**125**	125	(1,0,0,0,0,0,0,0,0,0,0)	0.0136
1972	**76**	76	(1,0,0,0,0,0,0,0,0,0,0)	0.1115
1973	**35**	35	(1,0,0,0,0,0,0,0,0,0,0)	0.2653

Table 7. Summary results for 1994 dataset.

Year of Birth	Estimated Replication	95% Confidence Interval
1929	**0%**	**(0%,12.00%)**
1930	**0%**	**(0%,8.69%)**
1931	**18.18%**	**(0%,46.38%)**
1932	**0%**	**(0%,12.50%)**
1933	**0%**	**(0%,7.32%)**
1934	**38.89%**	**(28.57%,61.76%)**
1935	**5.88%**	**(0%,23.02%)**
1936	**0%**	**(0%,4.00%)**
1937	**0%**	**(0%,6.25%)**
1938	**0%**	**(0%,6.94%)**
1939	**0%**	**(0%,8.23%)**
1940	**0%**	**(0%,10.21%)**
1941	**0%**	**(0%,5.06%)**
1942	**0%**	**(0%,5.08%)**
1943	**10.78%**	**(0%,20.44%)**
1944	**2.09%**	**(0%,10.06%)**

Table 8. Summary results of estimated replication and 95% confidence intervals for 1991 dataset.

Year of Birth	Estimated Replication	95% Confidence Interval
1903	**0%**	**(0%,0%)**
1904	------	------
1905	**0%**	**(0%,0%)**
1906	------	------
1907	------	------
1908	**0%**	**(0%,0%)**
1909	------	------
1910	------	------
1911	**0%**	**(0%,0%)**
1912	**0%**	**(0%,0%)**
1913	**0%**	**(0%,0%)**
1914	**0%**	**(0%,0%)**
1915	**0%**	**(0%,0%)**
1916	**33.33%**	**(0%,66.67%)**
1917	**16.67%**	**(0,%50.00%)**
1918	**0%**	**(0%,0%)**
1919	**11.11%**	**(0%,25.00%)**
1920	**0%**	**(0%,0%)**
1921	**0%**	**(0%,0%)**
1922	**22.22%**	**(0%,55.56%)**
1923	**0%**	**(0%,0%)**
1924	**0%**	**(0%,11.76%)**
1925	**6.45%**	**(0%,19.35%)**
1926	**11.11%**	**(0%,33.33%)**
1927	**0%**	**(0%,17.29%)**
1928	**0%**	**(0%,7.14%)**
1929	**0%**	**(0%,10.81%)**
1930	**0%**	**(0%,10.47%)**
1931	**10.64%**	**(0%,28.26%)**
1932	**0%**	**(0%,7.27%)**
1933	**0%**	**(0%,7.35%)**
1934	**5.13%**	**(0%,17.81%)**
1935	**0%**	**(0%,9.95%)**
1936	**0%**	**(0%,6.74%)**
1937	**0%**	**(0%,4.00%)**
1938	**0%**	**(0%,2.38%)**
1939	**0%**	**(0%,4.73%)**
1940	**0%**	**(0%,4.38%)**
1941	**0%**	**(0%,4.20%)**
1942	**0%**	**(0%,2.91%)**
1943	**0%**	**(0%,3.64%)**
1944	**0%**	**(0%,2.19%)**
1945	**0%**	**(0%,5.20%)**
1946	**0%**	**(0%,2.18%)**
1947	**0%**	**(0%,1.46%)**
1948	**0%**	**(0%,1.82%)**
1949	**0%**	**(0%,1.56%)**
1950	**0%**	**(0%,1.86%)**
1951	**1.14%**	**(0%,3.54%)**
1952	**0%**	**(0%,1.55%)**
1953	**3.41%**	**(0%,4.84%)**
1954	**0%**	**(0%,1.69%)**
1955	**0%**	**(0%,1.28%)**
1956	**2.53%**	**(0%,4.04%)**
1957	**1.40%**	**(0%,2.17%)**

Year of Birth	Estimated Replication	95% Confidence Interval
1958	**0%**	**(0%,1.17%)**
1959	**0%**	**(0%,0.75%)**
1960	**1.48%**	**(0%,2.43%)**
1961	**0%**	**(0%,0.70%)**
1962	**1.31%**	**(0%,1.91%)**
1963	**0%**	**(0%,1.48%)**
1964	**0%**	**(0%,1.52%)**
1965	**0%**	**(0%,1.85%)**
1966	**0%**	**(0%,2.22%)**
1967	**0%**	**(0%,1.52%)**
1968	**0%**	**(0%,2.12%)**
1969	**2.08%**	**(0%,4.00%)**
1970	**1.84%**	**(0%,7.96%)**
1971	**0%**	**(0%,5.98%)**
1972	**0%**	**(0%,5.56%)**
1973	**0%**	**(0%,9.38%)**

Table 9. Summary results of 95% confidence intervals for 1994 dataset.

Observed Bootstrap Replication Vector	Frequency	Estimated Sample Size	Estimated Probability Vector
(9,1)	8,779	10	(0.9,0.1)
(8,0,1)	238	9	(0.889,0,0.111)
(7,2)	937	9	(0.778,0.222)
(6,1,1)	21	8	(0.75,0.125,0.125)
(5,3)	25	8	(0.624,0.376)

Table 10. Validation of the method.

APPENDIX B.

DISTRIBUTION OF THE REPLICATION VECTOR, MAXIMISATION LEMMA, AND CALCULATION OF THE LIKELIHOOD FUNCTION

We have the following theorem and lemma [22]:

THEOREM 1

$$\mathrm{P}(T_1=t_1,T_2=t_2,\ldots T_r=t_r) = \frac{n!}{t_1!t_2!..t_r!(n-\tau)!}\frac{r!}{(1!)^{t_1}(2!)^{t_2}...(r!)^{t_r}}\frac{1}{n^r}.$$

Here $\tau = t_1 + t_2 + \ldots + t_r$ is the total number of tuples in a given birth year.

PROOF (adapted from [22])

Suppose that there are n days in a year, that all birth records correspond to distinct individuals, and that all birth dates are equally likely. The birth record replication vector is $\boldsymbol{t} = (t_1,t_2,...,t_r)$. Then the probability of a given ordered sequence of r birth dates occurring is $1/n^r$. In total there are $\tau = \sum_{i=1}^{r} t_i$ different birth dates that are distinct. These can be chosen from the n days in a year in

$$\frac{n!}{t_1!t_2!\ldots t_r!(n-\tau)!}$$

ways. This gives rise to $r = r^* = \sum_{i=1}^{r} it_i$ birth dates (including repeated birth dates). But even when these birth dates have been chosen there will be several ordered sequences of them corresponding to the same $\boldsymbol{t}$.

As a simple example consider $t_1 = 2$, $t_2 = 1$ so $r = 4$. We can choose two singleton birth dates b_1 and b_2 and one doubleton birth date b_3 from n=365 days in a year in $\frac{1}{2}365.364.363$ ways. But several ordered birth date sequences, for example $\{b_1,b_2,b_3,b_3\}$, $\{b_3,b_2,b_3,b_1\}$ and $\{b_3,b_3,b_1,b_2\}$, give rise to the same $\boldsymbol{s_B}$. Hence there are four birth dates, one of which is repeated twice, which can be arranged in $\frac{4!}{1!1!2!}$ distinct ordered sequences each corresponding to $\boldsymbol{t}$ the birth record (or birth date) replication vector.

In general there are r birth dates, t_1 repeated once, t_2 repeated twice, t_3 repeated three times, ... and t_r repeated r times which gave rise to

$$\frac{r!}{(1!)^{t_1}(2!)^{t_2}...(r!)^{t_r}}$$

distinct ordered birth date sequences (each of which gives ***t*** as the replication vector).

Hence the total probability of ***t*** is

$$\frac{n!}{t_1!t_2!..t_r!(n-\tau)!}\frac{r!}{(1!)^{t_1}(2!)^{t_2}...(r!)^{t_r}}\frac{1}{n^r}$$

as required.

LEMMA 1

Suppose that $m \geq 1$ and k_1, k_2, $\ldots k_m$ are strictly positive real numbers. Then $p_1^{k_1}p_2^{k_2}...p_m^{k_m}$ is maximised over $p_i \geq 0$, $1 \leq i \leq m$, $p_1+p_2+\ldots+p_m=1$, at $\hat{p}_i = k_i/(\sum_{j=1}^m k_j)$

when its value is $$\frac{k_1^{k_1}k_2^{k_2}...k_m^{k_m}}{(\sum_{j=1}^m k_j)^{(\sum_{j=1}^m k_j)}}.$$

PROOF (adapted from [22])

The problem is equivalent to maximising $F(\boldsymbol{p}) = \sum_{j=1}^m k_j \; log_e p_j$ subject to $p_1+p_2+\ldots+p_m=1$. Introducing the Lagrange multiplier $\lambda > 0$ the Lagrangian is

$$\sum_{j=1}^m k_j \; log_e p_j + \lambda\left(1-\sum_{j=1}^m p_j\right).$$

This is maximised over the set $A = \{p_i \geq 0,\ 1 \leq i \leq m\}$ at $\tilde{p}_i = \frac{k_i}{\lambda}$. Choose λ so that $\sum_{j=1}^m \tilde{p}_j = 1$ so that $\lambda = \sum_{j=1}^m k_j$. Then $(\tilde{p}_1, \tilde{p}_2, \ldots \tilde{p}_m)$ is feasible and maximises $F(\boldsymbol{p})$ over A subject to $\sum_{j=1}^m p_j = 1$.

CALCULATION OF THE LIKELIHOOD FUNCTION

Suppose that the observed birth record replication vector of r birth records is $\boldsymbol{s} = (s_1, s_2, ..., s_r)$ with $\sum_{i=1}^r is_i = r$ whereas due to some individuals being repeated in the dataset the true replication vector corresponding to birth dates of distinct individuals is $\boldsymbol{t} = (t_1, t_2, ..., t_r)$ with $\sum_{i=1}^r i\, t_i = r^* \leq r$.

For $i+j \leq r$, $i \geq 1$, $j \geq 0$ let $x_{i,j}$ denote the number of observed $i+j$-tuple birth records in ***s*** that are actually true i-tuple individual records in ***t***. Thus for $k = 1,2, \ldots r$

$$t_k = \sum_{l=0}^{r-k} x_{k,l} \; . \tag{B.1}$$

We shall use ***x*** to denote the vector $(x_{1,1}, \ldots x_{1,r-1}, x_{2,1}, \ldots x_{2,r-2}, x_{3,1}, \ldots x_{r-1,1})$. By considering the number of k-tuples in ***s*** we see that for $k = 2,3, \ldots r$,

$$s_k \geq \sum_{l=1}^{k-1} x_{l,k-l}. \tag{B.2}$$

Moreover let $f_{i,j}$ denote the probability that an observed $i+j$-tuple in ***s*** is actually a true i-tuple in ***t***. Thus

$$f_{i,j} = \sum_{\Xi_i} p_{\xi_1} p_{\xi_2} \cdots p_{\xi_i},$$

so $f_{i,j}$ is the sum of all products of i p's whose subscripts sum to $i+j$. Here

$$\Xi_i = \{(\xi_1,\xi_2, \ldots , \xi_i) \in \boldsymbol{Z}^{+i}, \xi_k \geq 1, 1 \leq k \leq i \text{ and } \sum_{k=1}^{i} \xi_k = i + j\}$$

where $\boldsymbol{Z}^{+}$ denotes the set of positive integers.

The $f_{i,j}$ can be defined recursively by

$$f_{1,j} = p_{j+1}, \qquad 0 \leq j \leq r\text{-}1,$$

$$f_{i,j} = p_1 f_{i\text{-}1,j} + p_2 f_{i\text{-}1,j\text{-}1} + p_3 f_{i\text{-}1,j\text{-}2} + \ldots + p_{j+1} f_{i\text{-}1,0}, \qquad \text{for } i \geq 2,\ i+j \leq r.$$

Suppose that x_1 denotes the total number of observed doubletons s_2, tripletons s_3, four-tuples s_4, ... and r-tuples s_r which correspond to the true singletons t_1, x_2 denotes the total number of observed tripletons s_3, four-tuples s_4, ... and r-tuples s_r which correspond to either true singletons t_1 or true doubletons t_2, x_3 denotes the total number of observed four-tuples s_4, five-tuples s_5, ... or r-tuples s_r which correspond to either true singletons t_1, doubletons t_2 or tripletons t_3, ... and $x_{r\text{-}1}$ denotes the total number of observed r-tuples s_r which correspond to either true singletons t_1, doubletons t_2, tripletons t_3, ... or (r-1)-tuples $t_{r\text{-}1}$.

Then for k = 1,2, ... r-1,

$$x_k = \sum_{l=1}^{k} \sum_{j=l}^{r-k+l-1} x_{k-l+1,j}, \tag{B.3}$$

$$t_1 = s_1 + x_1,$$

for k = 2,3, ... r-1, $\qquad t_k = s_k + x_k - x_{k\text{-}1},$ (B.4)

and $\qquad t_r = s_r - x_{r\text{-}1}.$

Given the observed birth record replication vector $\boldsymbol{s}$ and $x_{i,j}$, $i+j \leq r$, $i \geq 1$, $j \geq 1$ satisfying inequalities B.2 then the true replication vector $\boldsymbol{t}$ is uniquely determined by equations B.3 and B.4 and will be feasible (i.e. $t_k \geq 0$, $1 \leq k \leq r$). In these circumstances we denote the true replication vector by $\boldsymbol{s}$ and $\boldsymbol{x}$ as $\boldsymbol{t}(\boldsymbol{s},\boldsymbol{x})$. The values of $x_{i,0}$, $1 \leq i \leq r$, will then follow from equations B.1.

The likelihood for a true replication vector $\boldsymbol{t}$ given $\boldsymbol{x}$ and $\boldsymbol{p}$ is

$$P(\boldsymbol{t}(\boldsymbol{s},\boldsymbol{x})|\boldsymbol{x},\boldsymbol{p}) = \left(\prod_{k=1}^{r-1}\left(t_k! \prod_{l=0}^{r-k} \frac{f_{k,l}^{x_{k,l}}}{x_{k,l}!}\right)\right) P(\boldsymbol{t}). \tag{B.5}$$

Here $P(\boldsymbol{t}) = P(T_1{=}t_1, T_2{=}t_2, \ldots T_r{=}t_r)$ is given by Theorem 1 (assuming that all individuals are distinct).

A given $\boldsymbol{x}$ will correspond to r^* distinct individuals if and only if

$$r - r^* = \sum_{i=1}^{r-1} \sum_{j=1}^{r-i} j\, x_{i,j}. \tag{B.6}$$

Let X_{r^*} denote the set of $\boldsymbol{x}$ satisfying B.2 and B.6. Then the likelihood given r^* and $\boldsymbol{p}$ is given by

$$P(\boldsymbol{s}|r^*,\boldsymbol{p}) = \sum_{\boldsymbol{x} \in X_{r^*}} P\,(\boldsymbol{t}(\boldsymbol{s},\boldsymbol{x})|\boldsymbol{x},\boldsymbol{p}). \tag{B.7}$$

Hence the likelihood function of r^* and $\boldsymbol{p}$ given the data is

$$L(r^*,\boldsymbol{p}|\boldsymbol{s}) = P(\boldsymbol{s}|r^*,\boldsymbol{p}).$$

This is maximised over $\boldsymbol{p}$ for each value of r^* separately and then the overall maximum calculated.