Disease prevalence estimations based on contact registrations in general practice

RUDOLF HOOGENVEEN¹*, GERT WESTERT¹, MARCEL DIJKGRAAF², FRANÇOIS SCHELLEVIS³ AND DINNY DE BAKKER³

¹National Institute of Public Health and the Environment (RIVM); P.O. Box 1; 3720 BA Bilthoven; The Netherlands
²Academic Medical Center (AMC) KEB J2.224; P.O. Box 22700; 1100 DE Amsterdam; The Netherlands
³Netherlands Institute of Primary Health Care (NIVEL); P.O. Box 1568; 3500 BN Utrecht; The Netherlands

*Correspondence to: Rudolf Hoogenveen, National Institute of Public Health and the Environment (RIVM), P.O. Box 1, 3720 BA Bilthoven, The Netherlands †E-mail: rudolf.hoogenveen@rivm.nl

SUMMARY
This paper describes how to estimate the prevalence of chronic diseases in a population using data from contact registrations in general practice with a limited time length. Instead of using only total numbers of observed patients adjusted for the length of the observation period, we propose the use of (i) the time of the first contact of patients, (ii) the joint total numbers of patients and contacts, and (iii) the sets of patients in distinct time intervals, to generate prevalence rate estimates. The three new prevalence rate estimators have been developed assuming either a homogeneous or a parameterized heterogeneous patient population. Systematic and stochastic components of the estimators have been analysed by cross-validation for five chronic diseases using data from the Dutch ‘Study on Chronic Conditions’. The results show that the first two estimators work well for diseases with a relatively structured visiting behaviour, such as hypertension and diabetes mellitus, assuming a time-constant contact rate and homogeneous patient population. For diseases such as ischaemic heart disease, chronic non-specific respiratory diseases and osteoarthritis, that do not satisfy these assumptions, the methods generally result in underestimations.

INTRODUCTION
There are several approaches to estimate the prevalence of chronic diseases in a population. One could use health interview surveys or health examination surveys [1; 2]. In several countries – for example in the Netherlands – disease-specific registers are available, for example, for cancer [3]. Also, one could describe morbidity in a given population via the providers of care. In most cases medical specialists register the diagnoses and treatments of their patients. However, these data are specialist-dependent (ears, hearts and legs are treated by different specialists) and for less severe diseases only a few patients visit these specialists. In health care systems with regulated access to specialist care, such as in the Netherlands, this selection of morbidity is even greater. To explore the whole range of diseases, the use of general practice registrations seems most promising. In the Netherlands nearly everybody is listed in a general practice. Hence, at least theoretically, disease prevalence numbers in general practice can be estimated by counting the number of patients that are registered by the practitioner [4]. This way of estimating prevalence numbers has two important
advantages: all patients are covered, and the ‘epidemiological denominator’ [5] is known. However, the method assumes that each patient history is well known and registered uniformly. Most registration projects in general practice actually derive their data from contact registrations. For some regional projects in the Netherlands the registration time period is long; for the one national project so far the time period was restricted [6].

One important disadvantage of using contact registrations with limited time duration is that prevalent cases can be missed. These are persons that have had at least one contact before the registration period started, but no contact during the registration period. If no other data sources are available, as in our case, the number of these persons cannot be calculated from a ‘gold standard’. The research question that has been addressed in this paper was how to estimate disease point prevalence numbers from contact registrations in general practice over a relatively short time period in the absence of a ‘gold standard’. In other words, to estimate the number of prevalent cases with no contact during the registration period from the information available from those patients with at least one contact. Answering this research question helps improve existing methods of estimating disease prevalence.

The general problem of estimating disease prevalence numbers from registrations with limited time duration has been specified and analysed in different ways in the literature. One way is to derive a mathematical relation between incidence numbers, prevalence numbers and duration time, and to use two of these variables to obtain the unknown third [7]. However, in our case only one variable is known from the registration data – the number of new cases. Another way is the evaluation of different possible population screening strategies [8]. This problem is relevant for chronic disorders with relatively short duration times. In most situations the main aim of contact registrations in general practice is to analyse the episodes of health care use, and so one is not interested in the number of prevalent cases without any contact. Occasionally the problem of estimating prevalence numbers from contact registrations with short time of duration has been mentioned [9]. The problem is known to be complicated. Only for some specific disease characteristics can prevalence numbers can be calculated simply. In all other cases prevalence numbers are generally calculated using ad hoc expert knowledge [4].

However, data from contact registrations generally yield more information than only a list of persons having had at least one contact during the registration period. This extra information can be used to estimate the number of persons with no contact. Examples of this extra information are the timing of the contacts, the number of contacts for each patient, and the scheduled time period until the next appointment. We have developed prevalence estimators that make use of this extra information. These estimators have been based on the theory of stochastic processes. The prevalence estimators are related to those described by Jonsson [8]. He has described how prevalence numbers can be estimated from repeated screening programmes over limited time periods and has compared the results to those from continuous screening. However, he considered screening tests instead of contact registrations, he did not consider the case of heterogeneous patient populations, and he did not validate his model by comparing results for different observation time periods.

The prevalence estimators that we have developed have been applied for the chronic disorders hypertension, chronic ischaemic heart diseases (IHD), diabetes mellitus (DM), asthma and COPD (chronic non-specific respiratory diseases, CNSLD), and osteoarthritis of knee and/or hip. We have used data from the Dutch Study on Chronic Conditions [9].

STATISTICAL MODEL

Three disease prevalence estimators have been developed that make use of the extra information that is available from contact registrations in general practice. The three estimators have been derived from a statistical model that relates numbers of contacts to disease prevalence numbers. These estimators follow from fitting the model to:

1. the first contact time-points of all patients during the registration period;
2. the joint total numbers of patients and contacts;
3. the sets of patients in two distinct time periods.

First, we present the main model assumptions that have been made. We have assumed a closed patient population. Therefore the estimations only relate to prevalent patients, that is, those who already have had at least one contact before the starting time of the observation period (that is, with
their type of first contact coded or would have been coded as a ‘repeated contact’). ‘Incident’ cases have been excluded from our analyses. Methods to adjust disease prevalence estimations for incidence during the observation period can be found in the literature [4]. We have assumed no changes in the population due to death or migration. Moreover, all patients have been assumed independent.

We have used the inter-contact time intervals to find proper model assumptions on the visiting behaviour of patients (see Appendix A). Based on the frequency distribution of the inter-contact time intervals we have calculated the contact rates as a function of the time since the last contact. For most chronic disorders, the calculated contact rate was almost constant or slightly decreasing over time. This result suggested to us to assume time-constant individual contact rates. The case of time-constant population rate values implies a homogeneous patient population and the case of decreasing population values implies a heterogeneous population. The gamma distribution is the conjugate distribution to a time-constant contact rate and has been used to model the population heterogeneity [10]. The result of fitting the model has been used to specify the heterogeneity of the patient population. Two limiting cases of the gamma distribution function have been distinguished: the degenerate distribution function with all probability mass on the mean value, and the exponential distribution that results in Pareto-distributed population contact times, respectively (see Appendix B).

[ APPENDIX A-B ]

Two main characteristics of the prevalence estimators have been analysed; (i) one systematic component – whether or not they describe the means of the registration data correctly for registration periods with different lengths; (ii) one stochastic component – whether or not they describe the variability of the registration data correctly for different registration periods with the same length. These characteristics have been analysed by cross-validation.

We have estimated prevalence numbers, not prevalence rates. When the population size (the ‘denominator’ [5]) is known, prevalence numbers can be transformed to prevalence rates. The terms contacts, events and arrivals have been used as synonyms. The term ‘contact’ relates to the application of the estimations in health care research and the terms ‘event’ and ‘arrival’ relate to the mathematical theory on stochastic processes.

The first contact time-points
The likelihood function of the model parameters given the first contact (event, arrival) timepoints is

\[
L(N, \alpha; (t_i)_{i \in M}) \propto \left( \frac{N}{M} \right) \left\{ \prod_{i \in M} f(t_i; \alpha) \right\} S(T_{\text{max}}; \alpha)^{N-M}
\]

with \( N \) total prevalence number, \( \alpha \) parameter(s) of first contact time probability distribution, \( i \) index over population, \( t_i \) first contact time-point, \( M \) set/number of patients with at least one contact, \( f; S \) probability density and survival function of first contact time-point, respectively, and \( T_{\text{max}} \) observation time length. The maximum likelihood (ML) parameter estimates are found by maximizing the likelihood function. By elaborating the binomial formula, the ML estimator of \( N \) can be approximated by, conditional on the parameter(s) \( \alpha \)

\[
\hat{N}(\alpha) = \frac{M}{F(T_{\text{max}}; \alpha)}
\]

with \( F(T_{\text{max}}; \alpha) \) the probability of at least one contact before time \( T_{\text{max}} \). Assuming a time-constant contact rate, the ML estimator of the parameter \( \alpha \) is, conditional on the parameter \( N \)

\[
\hat{\alpha}(N) = \frac{\sum_{i \in M} t_i + (N-M)T_{\text{max}}}{M}
\]

This is the ML hazard rate estimator for given population number \( N \) [11].

The number of first contacts and total number of contacts
Assuming a homogeneous population of independent individuals, the number of registered patients is binomially distributed, and the total number of contacts is Poisson distributed:

\[
M|N, p \sim \text{Bin}(N, p^*), \quad K|N, p \sim \text{Poisson}(Np)
\]

with \( K \) total number of contacts, \( p^* = 1 - \exp(-p) \) the probability of at least one contact, and \( M \) and \( N \) as above. The model parameters \( N \) and \( p \) can be estimated in two ways, using the method of moments or by maximum
likelihood. Notice that the variables $M$ and $K$ are correlated. Ignoring this correlation in maximum likelihood estimation may give strongly biased results. It can be shown that $M$ is multinomially distributed given $K$ [12], but this result is not of practical use here. The probability distribution function of $K$ conditional on $M$ is very complex, but can be approximated. The ‘rest’ number of contacts $(K - M)$ is made by the $M$ persons with at least one contact. The exposure time for this rest number of contacts is the time period after the first contact time-point. The mean rest number of contacts is approximately equal to the product of the number of first contacts $M$, the contact rate and the expected rest time period, respectively (see Appendix C):

$$M x T_{\text{max}} \left( \frac{1 - e^{-\theta T_{\text{max}}}}{\theta T_{\text{max}}} \right) = M \left( 1 - e^{-\theta} \right)^{-1} - 1$$

Notice that all terms of the resulting likelihood function are only dependent on $\alpha$ through the parameter $p = \alpha T_{\text{max}}$. In the case of a heterogeneous population the likelihood function is too complex to be used in practice, therefore we have chosen the method of moments for the parameter estimations.

[APPENDIX C]

Relating the contacts in two different subperiods

The registrations of patients in different subperiods can be interpreted as different ‘captures’ of the patient population. The problem of estimating the size of closed populations using data from different captures is known as the capture-recapture problem. Estimation methods have extensively been described in the literature [12; 13]. Although mostly used in ecology and demography, they have also been applied in epidemiology in general and in disease prevalence estimations in particular [14; 15].

Assuming a homogeneous population, the captures in different subperiods in our problem are independent. The nearly unbiased estimator of the total prevalence number is [13]

$$\hat{N} = \frac{(M_1 + 1)(M_2 + 1)}{M_1 + 1} - 1$$

with $M_1; M_2; M_{11}$ the number of prevalent persons in the first and second capture and in both captures, respectively. In the case of a heterogeneous population, the captures are dependent, because persons having been captured the first time have higher contact rates than those not having been captured, and so are more likely to be captured again. In this case an approximate estimator can be derived by working out the standard ‘capture-recapture’ population estimator following the specified correlation structure (see Appendix D). The approximate estimator can be written as the standard ‘capture-recapture’ estimator multiplied by a correction factor. However, the correction factor cannot be estimated from the numbers of persons in both captures only, and is very sensitive to the degree of heterogeneity. Therefore we have applied the capture-recapture estimator only under the assumption of a homogeneous patient population.

[APPENDIX D]

Cross-validation

The total registration period of the Dutch Study on Chronic Conditions of 600 days has been divided by 1, 2, 4, 6 or 8 to create non-overlapping subperiods with length 600, 300, 150, 100 and 75 days, respectively. For each chronic disease and subperiod the disease prevalence number was estimated using the three methods described above and assuming a homogeneous or heterogeneous patient population. Since there is no ‘gold standard’, the estimation based on the full time period of 600 days has been used as our reference value. Note that our reference value depends on the method used. In the case of the ‘capture-recapture’ analysis (method 3) the reference value chosen is the estimation based on method 2 (contact numbers). By plotting the differences between the values obtained for the subperiods with the reference values, one gets information on the validity of the systematic component of the prevalence estimators.

Confidence intervals and sensitivity analysis results

Confidence intervals of the disease prevalence estimations have been found by parametric bootstrapping. Bootstrap samples ($n = 500$) have been generated for every subperiod considered. The
parameter values used are the geometric mean values of the parameter estimations for all different subperiods with the same length. In two cases, where clear outliers were found (IHD and CNSLD using method 1), we used the median values. In every bootstrap sample for every patient a random first contact time-point (in the case of method 1), a random number of contacts (in the case of method 2), or a random ‘success outcome’ (in the case of method 3), respectively, have been generated according to the specific model assumptions. The confidence intervals follow from the series of parameter values that have been re-estimated from these samples. These confidence intervals provide information on the stochastic component of the estimators.

To study the sensitivity of the disease prevalence estimators to the registration design parameters (total number of patients and length of registration period), the asymptotic variances of the prevalence estimators also have been approximated analytically (see Appendices B and C for methods 1 and 2, respectively).

DATA
Data were derived from the subproject on chronic diseases from the 1987/1988 Dutch National Survey of General Practice [16]. For this project, called the Dutch Study on Chronic Conditions (CC) [9], 15 general practitioners registered data on all contacts with their patients during 21 months for a set of specific diseases. The number of patients listed was approximately 23,500. Only contacts were registered for the following diseases: hypertension; chronic ischaemic heart disease (IHD); diabetes mellitus (DM); asthma and COPD (chronic non-specific respiratory diseases, CNSLD); and osteoarthritis of knee and/or hip. Identification of these patients took place on the occasion of the first contact during the 3-months registration period for the Dutch National Survey preceding the 21 months follow-up. Hypertension and diabetes mellitus can be considered as examples of chronic diseases that are mainly managed by general practitioners in the Netherlands and for which a systematic monitoring schedule is usual, that is, follow-up contacts on a regular basis. CNSLD and osteoarthritis are examples of diseases for which no regular management plan exists and patients only contact their general practitioner when the need arises. IHD may be managed by the general practitioner, a cardiologist, or both, depending on the stage of the disease and its stability. The completeness of the registered data has been checked.

For our analyses all contact time-points were specified by patient, general practitioner, disease and type of contact. The contact time-points have been redefined relative to the registration starting time-point for each general practitioner. All analyses were restricted to prevalent cases, that is, patients who already have had a contact before the starting time of the registration period (that is, with their type of first contact coded or would have been coded as a ‘repeated contact’).

RESULTS
In Table I some characteristics of the disease data have been presented. The sum of numbers of patients defined as incident and prevalent is smaller than the total number of patients due to missing values. The epidemiological differences between the diseases are shown by the ratios of the numbers of prevalent patients to the total number of patients, see for example hypertension and osteoarthritis.

[ TABLE 1 ]

We have used the statistical model of the contact time-points assuming time-constant individual contact rates that are gamma distributed to specify the population heterogeneity. The results suggested that for hypertension and DM the patient population could be assumed homogeneous, resulting in negative-exponentially distributed population contact time-points. For IHD, CNSLD and osteoarthritis the population heterogeneity could be described by negative exponentially distributed contact rate values resulting in Pareto-distributed population contact time-points. In the next step we have applied the disease prevalence estimators to the data for all subperiods. The model results have been presented in graphical form. The figures show the differences of the disease prevalence estimations based on the shorter subperiods relative to the reference values based on the full period of 600 days. The lines describe the asymptotic 95 per cent confidence intervals based on parametric bootstrapping. Notice that in the case of the ‘capture-recapture’ analysis the x-axis describes the total length of the pair of
capture periods. In Figure 1 the results have been given for hypertension and diabetes mellitus for methods 1 and 2 (using first contact time-points and contact numbers, respectively) assuming a homogeneous population. In Figure 2 the results have been presented for IHD, CNSLD and osteoarthritis for methods 1 and 2, using the specification of the population heterogeneity given above. For all conditions Figure 3 shows the results for method 3 (capture-recapture) assuming a homogeneous population.

[FIGURE 1-3]

We illustrate the model results with some specific comments on the figures. In all figures it can be seen that the errors of the prevalence estimations decrease for increasing registration time length. Figure 1 shows that for hypertension and diabetes for short time periods the errors are distributed skewed to the right, and the prevalence numbers have been underestimated. Figure 1 also shows that method 2 (contact numbers) results in smaller errors than method 1 (first contact time-points), again for these two diseases. In Figure 2 the results have been shown for the other three disease types, under the assumption of a heterogeneous population. The confidence intervals are generally larger than those in Figure 1. In most cases we find clear systematic differences between the prevalence estimations over time duration. From Figure 3 it can be seen that method 3 (capture-recapture) and assuming a homogeneous population may result in biased estimates for relatively long registration time lengths.

DISCUSSION

The results show that disease prevalence estimations based on contact registration periods with limited duration can be made by including extra available information. The simple models based on the assumption of identical time-constant individual contact rates work fairly well for chronic conditions with relatively structured visiting behaviour and long disease duration, such as hypertension and diabetes mellitus. For conditions that do not satisfy these assumptions the prevalence numbers are generally underestimated for small time periods. For these diseases more complex models than the ones used seem to be necessary to describe the visiting behaviour of patients.

The five diseases have been selected to represent the group of chronic conditions with relatively low mortality rates and long disease periods. Treatments by the general practitioner in the Netherlands are different for these diseases. For hypertension and diabetes mellitus the treatment is relatively structured with regular time periods between visits. For the other diseases the treatment is relatively unstructured: patients generally visit their general practitioner only when the need arises. These different treatment characteristics were confirmed by the results of statistical analyses on the specification of the population heterogeneity.

One might expect that within a registration period of approximately 2 years, all prevalent patients would have visited their general practitioner at least once. However, this seems not to be true, since for all five diseases the cumulative number of registered patients is still increasing with time after 2 years. Thus the total number of patients with at least one contact during these 2 years only represents a lower boundary of the total number of patients. Veltman [4] found an optimal registration time period of 4 years. However, his prevalence estimation was based on data on the total number of patients with at least one contact only, not on the extra information we have used.

The results have been analysed by distinguishing a systematic and a stochastic component of the prevalence estimators. The systematic component relates to whether or not the means of the estimators are equal (within uncertainty bounds) for time intervals of different lengths. The stochastic component relates to whether or not the variability of the estimators is described well. The stochastic and systematic components are not independent, see for example Figure 2.

It can be concluded that the assumption of a homogeneous patient population works well for the two diseases with relatively structured visiting behaviour, that is, hypertension and diabetes mellitus. Both methods 2 (contact numbers) and 3 (capture-recapture) generate better results than method 1 (first contact time-points). The twist in the curve for smaller time values can be explained by the large number of prevalent cases having inter-contact time intervals of approximately 3 months. For the other diseases the patient population is clearly heterogeneous, although our specification of the heterogeneity seems to be inappropriate for small time duration values. Method 3 (capture-recapture)
is useful when assuming independent captures. It must be stressed that all prevalence estimator are sensitive to the specification of the population heterogeneity for small time duration values, and method 3 even for longer time periods. The capture-recapture method is not the best method for each disease separately, but seems more robust in general than the other methods.

For small ‘capture probabilities’ the variance of the population number estimator for both methods 1 (time-points) and 2 (contact numbers) can be approximated analytically. The variance is approximately proportional to the population number and inversely proportional to the ‘capture probability’ with order 3 and 2 for methods 1 and 2, respectively (Appendices B and C). As a result, to increase the precision of the population estimator it is better to increase \( p \) (by extending the observation period) than to proportionally increase \( N \) (by extending the sample size), (at least) for relatively short observation times. This result has been confirmed by our empirical model results.

The assumption that the series of visits to the general practitioner can be described using time-constant individual contact rates seems to be invalid for chronic conditions with relatively irregular visiting behaviour, that is, with visits that strongly cluster within disease episodes. The assumption is essential to prevalence estimation method 2 (contact numbers). For method 1 (first contact time-points) in fact it has only been assumed that the first contact rate is time-constant. In order to improve the results for conditions that do not satisfy the assumptions of homogeneity and/or time-constant contact rates, the statistical model to describe the contacts has to be more complex and consequently data on a less aggregated level have to be used. For example, in the case of contact rates that are time-dependent, the inter-arrival and rest-arrival time probability distribution functions are different (see Appendix A).

In conclusion, the methods described to estimate disease prevalence numbers from registrations in general practice with limited time length work well for diseases with relatively structured visiting behaviour of all patients. The results might be presented in the form of correction factors on the number of registered patients. The methods have to be improved to find better results for diseases that do not satisfy this characteristic.

### APPENDIX A: THE REST-ARRIVAL TIME

The patient registration period can be interpreted as a time-window over parallel arrival processes. The probability distribution function of the time period until the next arrival starting from any given time-point depends on the inter-arrival time probability distribution and the last arrival time-point. We assume here the arrival processes to be stationary, meaning that the rest time period is independent on the first arrival time-point in the past. Then the probability of ‘drawing’ a time-interval \( t \) between two arrivals is proportional to the time-length \( t \) and its probability density value

\[
f_r(t) \propto t f(t)
\]

with \( f \) the probability density function of the inter-arrival time period being ‘drawn’ and \( f_r \) the probability density function of inter-arrival time period. The proportionality factor equals the mean inter-arrival time period. The cumulative probability function and the probability density function of the rest-time period \( U \) until the next arrival are

\[
F_r(u) = \int_{t \leq u} f_r(t) \, dt + \frac{u}{t} f(t) \, dt
\]

\[
f_r(u) = \frac{d}{du} F_r(u) = \mu^{-1} S_r(u)
\]

The expected value of the rest-time period is

\[
E_r(U) = \mu^{-1} \int S_r(u) \, du = 0.5 \mu + 0.5 \frac{1}{\mu^2} \sigma^2
\]

with \( \mu \) and \( \sigma^2 \) the mean and variance of inter-arrival time period, respectively. It is well known and can be shown easily that in the case of a negative-exponential distribution, both the rest-arrival and inter-arrival time are identically distributed. This result does not hold for other distributions.

### APPENDIX B: THE POPULATION ESTIMATOR USING THE FIRST CONTACT TIME-POINTS

The probability of no contact before \( t \) assuming a gamma distributed time-constant contact rate has the following distribution:

\[
F(t) = \int_0^t f(t) \, dt = 1 - \frac{\mu^t}{t^t}
\]
\[ P(T \geq t) = \int P(T \geq t|u)f(u; \alpha)du = \int_n e^{-\alpha u} - u \Gamma(n+1) - 1 du = \frac{1}{(1 + \alpha u)^{n+1}} \]

with \( T \) the first arrival (contact) time, \( u \) the contact rate, \( \alpha, n \) the parameters of the contact rate distribution function. A formula for the asymptotic variance of the patient population estimator can be derived from the log-likelihood function \([16–18]\). We have applied the Sterling formula to elaborate the binomial coefficients, assuming the patient numbers to be sufficiently large. In the case of a homogeneous patient population, the log-likelihood function becomes approximately (omitting the constants):

\[ l(N, p; \{t_i\}_{i \in M}) \approx N(\log(N) - 1) - (N - M)(\log(N - M) - 1) + M \log(p) - p \sum_{i \in M} t_i^* -(N - M)p \]

with \( M; N; p; t_i \) as defined before and \( t_i^* = t_i/T_{\text{max}} \). The asymptotic variance of the patient population estimator can be approximated by calculating the inverse of the Hessian matrix. Assuming small contact probabilities \((p^\ast = \exp{-p})\) the following proportionality relation has been found:

\[ \text{var}(N) \propto Np^{-2} \]

In the case of a heterogeneous patient population, the log-likelihood function can be approximated in the same way:

\[ l(N, p; \{t_i^*\}_{i \in M}) \approx N(\log(N) - 1) - (N - M)(\log(N - M) - 1) + M \log(p) - 2 \sum_{i \in M} (1 + pt_i^*) -(N - M)\log(1 + p) \]

In this case we do not get rid of the variables \( t_i^* \), when taking the second-order derivatives. However, when we assume that the contacts take place halfway on the time interval, the same proportionality relation holds for the prevalences estimator.

**APPENDIX C: THE POPULATION ESTIMATOR USING TOTAL CONTACT NUMBERS**

In the case of a negative-exponentially distributed first contact time-point, the expected first contact time-point for those having at least one contact is

\[ \int_0^{T_{\text{max}}} x dF(x) = sF(T_{\text{max}}) = \alpha^{-1} - T_{\text{max}}(e^{\alpha T} - 1)^{-1} \]

with \( F \) the cumulative probability distribution function. The expected rest period is the complement to the maximum time-point \( T_{\text{max}} \). Analogously, the expected first contact time-point can be calculated in the case of the Pareto distribution. To calculate the correlation between the number of first contacts and total number of contacts, the number of registered patients is written as the sum of indicator variables defined on the number of contacts

\[ M = \sum_{i \in N} I(K_i > 0) \]

with \( I \) the indicator function \((I(\text{true}) = 1, I(\text{false}) = 0)\) and \( K_i \) the number of contacts of patient \( i \). Then we find

\[ E(KM) = \sum_{i \in N} E(K_i I_i) + \sum_{i \neq j} EK_i E_I_j = E(K)E(M) + (1 - p^\ast)E(K) \]

with \( p^\ast \) the probability of at least one contact. Asymptotic methods can be used to calculate the asymptotic variances of the prevalence estimators:

\[ \text{var}(N) \approx \left( \frac{\partial N}{\partial M} \right)^2 \text{var}(M) + \left( \frac{\partial N}{\partial K} \right)^2 \text{var}(K) + 2 \left( \frac{\partial N}{\partial M} \right) \left( \frac{\partial N}{\partial K} \right) \text{cov}(M,K) \]

In the case of a homogeneous population, the population estimator using the method of moments is only implicitly defined through the moment equations. We have calculated the second-order approximation of this estimator and then worked out the formula of the variance, resulting in the following proportionality relation for the variance for small contact probabilities:

\[ \text{var}(N) \propto Np^{-2} \]

In the case of the parameterized heterogeneous patient population, the moment equations result in an explicit formula for the patient population estimator. Working out the formula of the variance in the same way results in the same proportionality relation for small contact probabilities.
APPENDIX D: THE CAPTURE-RECAPTURE POPULATION ESTIMATOR IN THE CASE OF A HETEROGENEOUS POPULATION

The captured population numbers have been presented in the scheme below:

<table>
<thead>
<tr>
<th>Capture 1</th>
<th>drawn</th>
<th>not drawn</th>
</tr>
</thead>
<tbody>
<tr>
<td>Capture 2</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>D</td>
</tr>
</tbody>
</table>

In the case of a heterogeneous population, the captures are assumed independent conditional on the contact rate. Thus the expected values of the captured population numbers are

\[
E(A \mid N) = N \int_0^{T_{\text{max}}} (1 - e^{-\alpha t/2})^2 e^{-t} \, dt = N \left( 1 - \frac{2}{1 + p} + \frac{1}{1 + 2p} \right)
\]

\[
E(B \mid N) = E(C \mid N) = N \left( 1 + p - \frac{1}{1 + 2p} \right)
\]

\[
E(D \mid N) = N \left( \frac{1}{1 + 2p} \right)
\]

with \( p = 0.5 \alpha \) and \( T_{\text{max}} \) the expected number of contacts during the half of the total time length. Working out the ML estimator in the case of independent captures, and omitting the higher order terms, gives the following result:

\[
E \left( \frac{BC}{A} \mid N \right) \approx \frac{E(B \mid N)E(C \mid N)}{E(A \mid N)} \approx S(0.5T_{\text{max}}) \frac{1}{2} E(D \mid N)
\]

with \( S(0.5T_{\text{max}}) = (1 + p)^{-1} \) for any individual.

TABLES AND FIGURES

Table I. The number of contacts and patients.

<table>
<thead>
<tr>
<th></th>
<th>Contacts</th>
<th>Patients</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
<td>Prevalent</td>
</tr>
<tr>
<td>Hypertension</td>
<td>4366</td>
<td>896</td>
</tr>
<tr>
<td>IHD</td>
<td>812</td>
<td>287</td>
</tr>
<tr>
<td>DM</td>
<td>1167</td>
<td>253</td>
</tr>
<tr>
<td>CNSLD</td>
<td>1025</td>
<td>336</td>
</tr>
<tr>
<td>Osteoarthritis</td>
<td>506</td>
<td>127</td>
</tr>
</tbody>
</table>

A patient is defined as ‘prevalent’ (already known), when his first contact is coded as a repeated contact; numbers based on total registration period of 600 days; IHD, ischaemic heart disease; DM, diabetes mellitus; CNSLD, chronic non-specific respiratory diseases.
Figure 1. The relative differences of the prevalence number for hypertension and diabetes, together with the 95 per cent confidence limits, in the case of homogeneous patient population. x-axis is observation time length, y-axis is relative difference of patient population estimation with respect to estimation based on time length 600 days, columns are results for method 1 (first contact time-points) and 2 (total contact numbers), respectively, rows are results for hypertension and diabetes, respectively. Points are estimated values for empirical data and lines are 95 per cent confidence limits using the parametric bootstrapping method. Patient population is assumed homogeneous.
Figure 2. The relative differences of the prevalence number estimations for IHD, CNSLD and osteoarthritis, together with the 95 per cent confidence limits, in the case of heterogeneous patient population. X-axis is observation time length, y-axis is relative difference of patient population estimation with respect to estimation based on time length 600 days, columns are results for method 1 (first contact time-points) and 2 (total contact numbers) respectively, rows are results for ischaemic heart disease, chronic non-specific respiratory diseases and osteoarthritis, respectively. Points are estimated values for empirical data and lines are 95 per cent confidence limits using parametric bootstrapping method. Patient population is assumed heterogeneous.
REFERENCES