Adaptive list sequential sampling method for population-based observational studies

Michel H Hof1,*, Anita CJ Ravelli2 and Aeilko H Zwinderman1

Abstract

Background: In population-based observational studies, non-participation and delayed response to the invitation to participate are complications that often arise during the recruitment of a sample. When both are not properly dealt with, the composition of the sample can be different from the desired composition. Inviting too many individuals or too few individuals from a particular subgroup could lead to unnecessary costs or decreased precision. Another problem is that there is frequently no or only partial information available about the willingness to participate. In this situation, we cannot adjust the recruitment procedure for non-participation before the recruitment period starts.

Methods: We have developed an adaptive list sequential sampling method that can deal with unknown participation probabilities and delayed responses to the invitation to participate in the study. In a sequential way, we evaluate whether we should invite a person from the population or not. During this evaluation, we correct for the fact that this person could decline to participate using an estimated participation probability. We use the information from all previously invited persons to estimate the participation probabilities for the non-evaluated individuals.

Results: The simulations showed that the adaptive list sequential sampling method can be used to estimate the participation probability during the recruitment period, and that it can successfully recruit a sample with a specific composition.

Conclusions: The adaptive list sequential sampling method can successfully recruit a sample with a specific desired composition when we have partial or no information about the willingness to participate before we start the recruitment period and when individuals may have a delayed response to the invitation.

Keywords: List sequential sampling, Sample representativeness, πps sample, Population-based observational studies

Background

Population-based observational studies are frequently used to measure the prevalence of characteristics such as diseases by means of a sample from a population [1]. Two important problems that arise when a sample is recruited is that (i) not everyone in the population has the same willingness to participate in the study [2-6], and (ii) after inviting an individual, it might take some time before we receive a response.

Variation in the willingness to participate may bias the results of the study. To deal with this problem, we could invite more individuals from groups related to a low willingness to participate [7]. However, this approach requires that the participation probability per person or group is known before the sampling procedure starts. Unfortunately, this detailed knowledge on the willingness to participate among sub-groups in the population is often not available. If the willingness to participate is less than assumed we will invite too few individuals, which leads to a too small sample and a decreased precision. On the other hand, inviting too many individuals will lead to extra costs. Generally, we invite too many individuals when we underestimate the willingness to participate and there is delayed response to the invitation. In general, not accounting for delayed response will lead an unexpected number of extra individuals in the sample at the end of the recruitment period.

*Correspondence: m.h.hof@amc.uva.nl
1Department of Clinical Epidemiology, Bioinformatics, and Biostatistics, Academic Medical Center - University of Amsterdam, Meibergdreef 9, 1105 AZ, Amsterdam, The Netherlands
Full list of author information is available at the end of the article

© 2014 Hof et al.; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.
An example of a complex sampling problem is observed in the HELIUS study [8]. One objective of the HELIUS study is to measure ethnic inequalities in the incidence and prognosis of major diseases in the population of Amsterdam. The desired sample should have approximately 5000 individuals in each ethnic group, and should be representative for the population of Amsterdam. This is achieved by stratifying on the auxiliary variables: place of residence (spatial), age, (continuous), gender (categorical), and social economic status (categorical) available from municipal registries.

Unfortunately, it is not straightforward to implement stratification when we have a large number of auxiliary variables of mixed types [9]. In this case, too small or even empty strata might be obtained when we cross all strata from all variables. An alternative variance reduction technique, proposed by Graffström et al., is to obtain a well spread set of participants [10,11]. Basically, a set of participants is well spread when the number of participants is close to what is expected on average, for every set of auxiliary variables. Graffström et al. showed that the variance of commonly used estimators is usually low with a well spread set of participants.

In this paper, we use the list sequential method, developed by Bondesson and Thorburn [12] to obtain a well spread set of participants without replacement from a finite population. Instead of trying to cross all strata from all auxiliary variables, our approach is based on a distance function between individuals. Similar or almost similar individuals should seldom be invited both to participate in the study. In its current form, the list sequential sampling method assumes that (i) everyone participates in the study and that (ii) there is no delayed response to the invitation.

We developed approaches to correct for non-participation and delayed response to the invitation when we use a list sequential sampling method. The list sequential sampling method evaluates individuals from the population in a sequential order, and uses a random process to decide whether or not an individual should be invited to participate in the study. In this decision we have to correct for any non-participation. An approach is to weigh the probability of being invited with the (estimated) participation probability. When there is no or partial a-priori knowledge on the participation probability, we can estimate this probability during the recruitment period using the information from already invited individuals. To combine both prior information and information that is generated during the recruitment period, we developed a Bayesian approach to estimate the participation probabilities. Moreover, to deal with the delayed response to the invitation, we use the expected response of an individual when we have no answer yet.

We performed a simulation study to illustrate the performance of the adapted list sequential sampling method, when we have unknown heterogeneous participation probabilities and delayed response to the invitation.

Methods

Problem description

We consider a finite population D containing n individuals, where each individual i is described by a vector x_i of auxiliary variables. The auxiliary variables x_i are known for each individual before the recruitment period starts. Usually x_i is available from municipal or national person-registries. Examples of these variables are gender, age, place of residence, and social economic status. In addition to x_i, each individual i has an unobserved outcome of interest y_i. The goal of this paper is to obtain a sample of size m (m < n) from D, in which we can observe y_i.

A sample is described by the vector s = (s_1, ..., s_n), where s_i takes the value 1 if individual i is in the sample and 0 otherwise [13,14]. With this representation there are 2^n possible samples. Before the recruitment period starts we need to determine \( \pi_i \), which is the probability that individual i is included in s (i.e. \( p(s_i = 1) = \pi_i \)). We want to recruit a sample of m individuals and therefore \( \sum_{i=1}^{n} \pi_i = m \) where m is a positive integer.

Different choices can be made for the inclusion probabilities \( \pi_i \). For instance, we can assign equal inclusion probabilities to all individuals, i.e. \( \pi_i = m/n \). In this case, the sample s is expected to be a ‘miniature’ version of the population D, because we expect s to have approximately the same composition of auxiliary characteristics as D. In this case, the sample is referred to as a representative sample [11]. However, \( \pi_i \) is frequently chosen to be proportional to x_i. For example, by oversampling a rare subgroup we could increase the precision of the result for that particular subgroup [15].

List sequential sampling method

To obtain the sample we use the list sequential method based on sampling without replacement developed by Bondesson and Thorburn [12]. To illustrate the list sequential method, we first consider the situation in which all invited individuals will participate in the study.

During the recruitment period, we sequentially decide for each individual i from D whether we include this individual in the sample (\( s_i = 1 \)) or not (\( s_i = 0 \)). After this decision, the probability of being included in the sample for the remaining non-invited individuals from D is updated. Let \( \pi^{(0)} = (\pi_1^{(0)}, ..., \pi_n^{(0)}) \) be the vector of initial inclusion probabilities which is determined before the sampling procedure starts, i.e. \( \pi_i^{(0)} = \pi_i \). We sequentially
evaluate each individual \( i \) from the population and update
the inclusion probabilities of all non-evaluated individ-
uals after each evaluation. For the first individual, we
have \( p(s_1) = \pi_1(0) \). Depending on whether individual 1
is included in the sample or not, the inclusion prob-
abilities of all other, non-evaluated, individuals are updated.
This gives us the vector \( \pi^{(1)} \), from which we use \( \pi_2^{(1)} \)
to determine \( s_2 \); i.e. decide whether to include the
second individual in the sample or not. The updating sche-
mework may be represented as

\[
\pi^{(0)} = \begin{pmatrix}
\pi_1^{(0)} \\
\pi_2^{(0)} \\
\pi_3^{(0)} \\
\vdots \\
\pi_n^{(0)}
\end{pmatrix},
\pi^{(1)} = \begin{pmatrix}
\pi_1^{(1)} \\
\pi_2^{(1)} \\
\pi_3^{(1)} \\
\vdots \\
\pi_n^{(1)}
\end{pmatrix},
\pi^{(2)} = \begin{pmatrix}
\pi_1^{(2)} \\
\pi_2^{(2)} \\
\pi_3^{(2)} \\
\vdots \\
\pi_n^{(2)}
\end{pmatrix},
\pi^{(3)} = \begin{pmatrix}
\pi_1^{(3)} \\
\pi_2^{(3)} \\
\pi_3^{(3)} \\
\vdots \\
\pi_n^{(3)}
\end{pmatrix}
\]

Generally, when we evaluate individual \( i \), we will use the
inclusion probability \( \pi_i^{(j-1)} \) to determine \( s_j \). After the eval-
uation of individual \( i \), we update all probabilities \( \pi_j^{(i)} \), for
\( j > i \) with

\[
\pi_j^{(i)} = \pi_j^{(i-1)} - (s_i - \pi_i^{(i-1)})w_{j-i}^{(i)},
\]

where \( w_{j-i}^{(i)} \) are weights that may depend on \( s_1, s_2, \ldots, s_{j-1} \).
Note that \( w_{j-i}^{(i)} \) determines how \( \pi_j^{(i)} \) is affected by the sam-
pling outcome from the individual \( i \), since \( w_{j-i}^{(i)} \) influences
the second order inclusion probability \( p(s_i = 1, s_j = 1) \).

The sampling scheme has a sample of size \( m \), when the
weights are restricted to sum up to one, i.e. \( \sum_{i=1}^{N} w_{j-i}^{(i)} = 1 \).
To guarantee that \( 0 \leq \pi_j^{(i)} \leq 1 \), all weights should satisfy

\[
-\min \left( \frac{1 - \pi_j^{(i-1)}}{1 - \pi_i^{(i-1)}}, \frac{\pi_j^{(i-1)} - \pi_i^{(i-1)}}{\pi_j^{(i-1)} - \pi_i^{(i-1)}} \right) \leq w_{j-i}^{(i)} \leq
\min \left( \frac{\pi_i^{(i-1)} - \pi_j^{(i-1)}}{1 - \pi_i^{(i-1)}}, \frac{\pi_i^{(i-1)} - \pi_j^{(i-1)}}{\pi_i^{(i-1)} - \pi_j^{(i-1)}} \right)
\]

Within these bounds, we can impose different restrictions
on \( w_{j-i}^{(i)} \), resulting in samples with certain characteristic.

Generally, when \( w_{j-i}^{(i)} > 0 \) we have \( \text{corr}(s_i = 1, s_j = 1) < 0 \)
(i.e. a negative correlation between the sampling indica-
tors of individuals \( i \) and \( j \)), whereas with \( w_{j-i}^{(i)} < 0 \), we
have \( \text{corr}(s_i = 1, s_j = 1) > 0 \). For more detail about the
list sequential method, we refer the reader to respectively
theorem 1 and remark 1 from Bondesson and Thorburn
[12].

**Well spread samples**

We are interested in recruiting a well spread sample
with the list sequential sampling method. Usually, a well
spread sample leads to parameter estimates with low vari-
ances. Before we can introduce the definition of a well
spread sample, we require the concept of coherent sub-
sets. Let \( d(i, k) \) be the distance between individuals \( i \) and
\( k \). A subset \( D' \) from the population \( D \) is coherent if the
following holds. First, let some individual \( i \in D' \). Individ-
ual \( k \) is included in \( D' \) if and only if \( d(i, k) \leq r \), where
\( r \geq 0 \). Consequently, \( D' \) can be constructed by including
all individuals within a ball of radius \( r \) around individual \( i \).

Grafström and Schelin considered a sample to be well
spread with respect to the inclusion probabilities \( \pi \) when,
for every coherent subset \( D' \subset D \),

\[
h' \approx \sum_{i \in D'} \pi_i.
\]

A smaller distance to individual \( i \) increases the proba-
bility of being included in the coherent subset \( D' \). To satisfy
(3), it is clear that the inclusion probability of individual \( i \)
should be more influenced by the sampling indicators \( s \) of
(individuals with a smaller distance. We propose to mea-
sure distance between individuals with the auxiliary vari-
ables \( x \), where \( d(x_i, x_k) \) is the distance between individual
\( i \) and \( k \). Based on the types of auxiliary variables, we can
choose, for instance, the Mahalanobis or the Manhattan
distance.

To obtain a well spread sample with the list sequential sampling method, we will use preliminary weights which
are specified before the recruitment period starts. The pre-
liminary weight \( w'^{(i)}_k \) reflects the effect of \( s_k \) from individ-
ual \( k \) on the inclusion probability of individual \( i \). The
weights are referred to as preliminary because the upper
bound from (2) has an effect on the conditional inclusion
probabilities.

The preliminary weights are constructed in the follow-
ing way. Let \( c^{(i)}_k \) be the rank of the distance of the \( k^{th} \)
individual to individual \( i \), where \( k \neq i \). We rank the dis-
tances in ascending order, where we assign \( c^{(i)}_i = 1 \) to the
closest individual, \( c^{(i)}_2 = 2 \) to the second closest individ-
ual, and so on. To construct the preliminary weights, we
could use the linear function

\[
\hat{w}'^{(i)}_k = \mu + c^{(i)}_k \lambda,
\]

where the weights \( \mu \) and \( \lambda \leq 0 \) are arbitrarily chosen
weights. The sampling indicator \( s_k \) of individual \( k \) has a
larger effect on individuals at smaller distance, whereas
it has less effect on individuals at further distance. To
recruit a set of approximately \( m \) individuals, we restrict
the weights to satisfy \( \sum_{k \neq i} w'^{(i)}_k = 1 \).

**Heterogeneous participation probabilities**

A problem of sampling from population \( D \) is that indi-
viduals that are invited to participate in the study
can decline the invitation. Let \( b = (b_1, \ldots, b_n) \) be
the vector that indicates whether an individual \( i \) is
invited to participate \((b_i = 1)\) or not \((b_i = 0)\). When individual \(i\) refuses to participate in the study, we have \(s_i = 0\) and we do not observe \(y_i\). Let \(\phi = (\phi_1, \ldots, \phi_n)\) be the vector that contains the participation probability per person in the population, where \(\phi_i = p(s_i = 1|b_i = 1)\). Note that when every invitee participates (i.e. \(\phi_i = 1\), for \(i = 1, \ldots, n\), we have \(s = b\).

Let \(\pi_i^{(i-1)}\) be the inclusion probability \(\pi_i^{(i-1)}\) corrected for non-participation, i.e. the probability of being invited to participate in the study for individual \(i\) from \(D\). When \(\phi_i\) is known before the recruitment period starts, non-participation can be dealt with by using \(\hat{\pi}_i^{(i-1)} = \pi_i^{(i-1)}/\phi_i\) as probability to invite individual \(i\). Moreover, we can use the updating rule from (1) to update the inclusion probabilities of the non-evaluated individuals \(\pi_i^{(j)}, j > i\), after individual \(i\) responded to the invitation. This will give us a sample that approximately satisfies the inclusion probabilities \(\pi\).

The following small sampling problem illustrates this modification. Consider that, for the first individual, we have \(\pi_1^{(0)} = 0.25\) and \(\phi_1 = 0.5\). The probability to invite this individual is therefore \(\hat{\pi}_1^{(0)} = 0.5\). Using this strategy there might be some individuals \(i\) with \(\hat{\pi}_i^{(i-1)} > 1\). This means that the participation probability of individual \(i\) is too low with respect to \(\pi_i^{(i-1)}\); the desired probability to be included in \(s\) for individual \(i\) cannot be reached. For instance, this would happen in the example above for individual 1 when \(\phi_1 = 0.1\) and consequently \(\hat{\pi}_1^{(0)} = 2.5\). This means that we have to invite individual 1 two and a half times to satisfy \(\pi_1^{(0)}\). Because we can only invite an individual once, we restrict all values \(\hat{\pi}_i^{(i-1)}\) to be one or lower.

**Adaptive list sequential sampling method**

Usually, \(\phi_i\) is not known before the recruitment period starts. In this section we suggest how \(\phi_i\) can be estimated adaptively during the recruitment period. In addition, we consider delayed response to the invitation.

For each individual, we have some knowledge about the willingness to participate before the recruitment period starts. For example, we might have participation estimates from a small pilot study or from previously performed studies. In addition, information from the invited individuals becomes available during the recruitment period. Therefore, we propose to use a Bayesian method to estimate the participation probability of individual \(i\) during the recruitment period, in which we use both the available prior knowledge and the information that becomes available during the recruitment period.

Let \(z_i\) be the vector of all observed characteristics of individual \(i\), which are related to the participation probability. We assume a missing at random type of mechanism for the participation probabilities, where the participation probability of individual \(i\) only depends on observed characteristics \(z_i\), i.e. \(p(s_i = 1|b_i = 1, z_i)\). The participation probability can be written as

\[
p(s = 1|b_i = 1, z_i, \alpha, \beta) = \frac{\exp(\alpha + f(z_i, \beta))}{1 + \exp(\alpha + f(z_i, \beta))},
\]

where \(\alpha\) is the intercept term, and \(f()\) is a function of the observed characteristics \(z_i\) and the regression weights \(\beta\). Because more information becomes available during the recruitment period, the participation probability estimates become more accurate. The vector of estimated participation probabilities of all \(n\) individuals after the evaluation of individual \(i\) is denoted as \(\hat{\phi}_i = (\hat{\phi}_1^{(0)} , \ldots, \hat{\phi}_n^{(0)})\). We then adapt the inclusion probabilities as \(\hat{\pi}_i^{(i-1)} = \hat{\pi}_i^{(i-1)}/\hat{\phi}_i^{(i-1)}\).

After an invitation has been sent to an individual, it might take some time to get a response. Let \(u_j^{(i)}\) be the indicator whether individual \(j\) has responded to the invitation before individual \(i\) is evaluated, where \(u_j^{(i)} = 1\) when we observe \(s_j\) and \(u_j^{(i)} = 0\) when we do not observe the participation indicator \(s_j\) during the evaluation of individual \(i\). Note that when individual \(j\) has not been invited (i.e. \(b_j = 0\), \(s_j = 0\) since individual \(j\) is not included in the set of participants. A problem of delayed response is that we cannot use the update rule from (1) to determine \(\pi_j^{(i)}\), when the participation indicator of the previous individual is not observed. Consequently, we cannot update \(\pi_j^{(i)}\) which means that our sampling method is less successful in recruiting a well spread sample. As a solution, we propose to use the data from all previously invited individuals, and replace the non-observed participation indicators with their estimated expected value. We use this approach in step 1 of the adaptive list sequential sampling method listed below.

Before we start the adaptive list sequential sampling method, we specify the vector \(\pi^{(0)} = \pi\), which contains the initial probabilities of being included in \(s\) for every individual \(i\) in \(D\). The desired number of individuals in \(s\) is \(m = \sum_{i=1}^{n} \pi_i^{(0)}\), where \(m\) is a positive integer. The first individual from \(D\) is invited with the probability \(\pi_1^{(0)} = \pi_1^{(0)}/\hat{\phi}_1^{(0)}\), where \(\hat{\phi}_1^{(0)}\) is an initial guess of the participation probability of the first individual. All other individuals from \(D\) are invited in a sequential way, where the steps of the adaptive list sequential sampling method for individual \(i = 2, \ldots, n\) are

1. **Calculate** \(\pi_i^{(i-1)}\)

To deal with delayed response to the invitation, we propose to use a modified version of the column-wise updating rule proposed by Bondesson and Thorburn [12].
We calculate $\pi_i^{(i-1)}$ by iterating over $k = 1, 2, \ldots, i - 1$, where
\[ \pi_i^{(k)} = \pi_i^{(k-1)} - \left( s_k - \pi_i^{(k-1)} \right) w_k^{(i)}, \] 
and $w_k^{(i)}$ is calculated as
\[ w_k^{(i)} = \min \left( \frac{w_k^{(i-1)}}{1 - \pi_i^{(k-1)}}, \frac{1 - \pi_i^{(k-1)}}{\pi_i^{(k-1)}} \right). \]

The weight $w_k^{(i)}$ determines the effect of $s_k$ on $\pi_i^{(k)}$ and therefore also $\pi_i^{(i-1)}$. The choice of preliminary weights $\tilde{w}_k^{(i)}$ is discussed in the previous section. Because (6) still requires the observed indicators $s_1, s_2, \ldots, s_{i-1}$, we modify (6) to deal with delayed response to the invitation. When $u_k^{(i)} = 0$, we replace $s_k$ with its estimated expectation $\hat{s}_k^{(i-1)}$, where $\hat{s}_k^{(i-1)}$ is the participation probability estimate of individual $k$ from the previous evaluation $i - 1$. The delayed response adjusted column-wise updating rule from (6) is
\[
\pi_i^{(k)} = \begin{cases} 
\pi_i^{(k-1)} - \left( s_k - \pi_i^{(k-1)} \right) w_k^{(i)} & \text{if } u_k^{(i)} = 1, \\
\pi_i^{(k-1)} - \left( \hat{s}_k^{(i-1)} - \pi_i^{(k-1)} \right) w_k^{(i)} & \text{if } u_k^{(i)} = 0.
\end{cases}
\]

2. Calculate $\tilde{p}_i^{(i-1)}$

Decide whether individual $i$ should be invited to participate in the study, where $b_i = 1$ if the individual is invited and $b_i = 0$ if not. This decision is based on the probability of being invited,
\[ \tilde{p}_i^{(i-1)} = \frac{\pi_i^{(i-1)}}{\hat{s}_i^{(i-1)}}, \]
where $\hat{s}_i^{(i-1)}$ is the participation probability estimated from the previous evaluation $i - 1$. We draw the decision to invite individual $i$ from a Bernoulli distribution with $p(b_i = 1) = \tilde{p}_i^{(i-1)}$.

3. Update the vector $\phi^{(i)}$

Let $R^{(i)} = \{ r, b = 1, u^{(i)} = 1, r \in D \}$ be the set of all $m_i$ individuals that responded to the invitation to participate. Each individual from $R^{(i)}$ is described by $r = (s, z)$, where $s = 1$ when invitee $r$ participates and $s = 0$ otherwise, and $z$ is a vector of known characteristics. The participation probability of individual $k$ is defined as (5). Because we might have some a-priori knowledge about the intercept $\alpha$ and the regression weights $\beta$, we use Bayesian inference to estimate the posterior distribution $g(\alpha, \beta | R^{(i)})$, i.e.
\[
g \left( \alpha, \beta | R^{(i)} \right) = \frac{h \left( R^{(i)} | \alpha, \beta \right) f(\alpha, \beta | \theta)}{\int_{(\alpha, \beta)} h(R^{(i)} | \alpha, \beta) f(\alpha, \beta | \theta) \, d(\alpha, \beta)},
\]
where $\theta$ is a vector of parameters, and $f()$ is the prior distribution of $(\alpha, \beta)$. The likelihood of $R^{(i)}$ given $(\alpha, \beta)$ is
\[
h \left( R^{(i)} | \alpha, \beta \right) = \prod_{\ell=1}^m p(s_{\ell} = 1 | z_{\ell}, \alpha, \beta)^{z_{\ell}} \left( 1 - p(s_{\ell} = 1 | z_{\ell}, \alpha, \beta) \right)^{1-z_{\ell}}
\]
where $p(s_{\ell} = 1 | z_{\ell}, \alpha, \beta)$ is given by (5). Following (8) we update the vector of estimated participation probabilities $\hat{\phi}_k^{(i)}$, where for individual $k = 1, \ldots, n$
\[
\hat{\phi}_k^{(i)} = \int_{(\alpha, \beta)} p(s_k = 1 | z_k, \alpha, \beta) g(\alpha, \beta | R^{(i)}) \, d(\alpha, \beta).
\]

To estimate $\hat{\phi}_k^{(i)}$, we can use quadrature or MCMC methods. The values of $\theta$ depend on the amount of prior knowledge that is available before the recruitment period starts. For instance, we can assume that $(\alpha, \beta)$ is sampled from some flat distribution with large variance when no prior knowledge is available.

Simulations

We illustrated the performance of the adaptive list sequential sampling method with two simulations. In these two simulations, we created populations with unknown heterogeneous willingness to participate and delayed response to the invitation. The first simulation was focused on recruiting a well spread, representative set of participants. In the second simulation, we investigated stratified sampling from a population in which some subgroups were over-represented.

Simulation 1

Consider a population $D$ of size $n = 4000$ from which we drew a random sample without replacement of size $m = 400$ with the adaptive list sequential sampling method. To recruit a representative sample from the population, we assigned equal inclusion probabilities to all individuals from the population; i.e. $\pi_i^{(0)} = m/n = 0.1$ for $i = 1, \ldots, n$. When the sample is well spread, the distribution of the auxiliary characteristics $x$ should be approximately similar in the population and the sample.

The data was generated as follows. The vector $z_i$ was drawn from a multivariate normal distribution with means zero, and covariances zero. The probability of positively responding to the invitation was $p(s_i = 1 | b_i = 1, z_i) = \text{invlogit}(\alpha + z_i \beta)$, where invlogit denotes the inverse logit transformation, $\alpha = 1$, and $\beta = (0.3, -0.7, 0.1, 0.4)$. The response was drawn from
a Bernoulli distribution with \( p(s_i = 1|h_i = 1, z_i) \). In addition, for individual \( i \), delayed response to the invitation was simulated by drawing time \( t_i \) from a Poisson distribution with expectation 15. Individual \( i \) responded to the invitation after the evaluation of individual \( i + t_i \). Thus if \( t_i = 0 \), individual \( i \) responded immediately to the invitation.

For individual \( i \), the characteristics \( x_i \) were drawn from a multivariate normal distribution with means zero, variances one, and covariance matrix

\[
\begin{pmatrix}
1.00 & 0.20 & -0.50 & 0.30 \\
0.20 & 1.00 & 0.20 & -0.40 \\
-0.50 & 0.20 & 1.00 & -0.20 \\
0.30 & -0.40 & -0.20 & 1.00
\end{pmatrix}.
\]

To obtain a well spread and representative sample, we used the adaptive list sequential method. To satisfy (3), we assumed that \( \pi_i \) had a large impact on how well the sample was spread.

We assumed that we had no prior knowledge about the participation probability before the recruitment period started. Therefore flat, non-informative priors were used for \( \alpha \) and all regression weights \( \beta \) by assuming they followed normal distributions with means zero and variance 100. Because we assumed zero means, the initial estimated participation probabilities were 50%, i.e. \( \hat{\phi}_i^{(0)} = 0.5 \) for \( i = 1, \ldots, n \).

We quantified how well a sample was spread with the following measure based on Voronoi polytopes, suggested by Grafström and Lundström [10]. Let individual \( i \in s \), i.e. individual \( i \) is included in the set of participants \( s \). The Voronoi polytope \( v_i \) consists of all individuals \( j \) from the population \( D \) for which \( d(x_i, x_j) \leq d(x_i, x_k) \), for all other individuals \( k \in s \). Note that when \( d(x_i, x_j) = d(x_i, x_k) \), individual \( j \) is included in both polytopes \( v_i \) and \( v_k \), but weighted with 1/2.

Let \( q_i \) be the sum of initial inclusion probabilities of the individuals in \( v_i \),

\[ q_i = \sum_{j \in v_i} \pi_j. \]

Grafström and Lundström showed that a sample can be considered to be well spread if \( q_i \) is one or close to one for all polytopes \( v_i \). Therefore, a measure to quantify how well a sample is

\[ R = \frac{1}{n} \sum_{i \in s} (q_i - 1)^2, \]

where a low \( R \) corresponds to well spread sample. To investigate how well the adaptive list sequential sampling methods performed in recruiting a well spread sample, the simulation was performed 1000 times. We calculated the mean and variance of \( R \), and the average sum of recruited participants. Note that the best adaptive list sequential sampling method should give us a set of approximately 400 participants with a low \( R \) in every simulation.

**Simulation 2**

In simulation 2, we considered a population \( D \) of size \( n = 5000 \), in which each individual was described by a categorical auxiliary variable \( x_i \) and an unobserved binary outcome of interest \( y_i \). The auxiliary variable \( x_i \) had five possible values \( g \). The main goal of this simulation was to estimate the sum of the outcome \( y \) in the population, denoted as \( Y = \sum_{i=1}^{n} y_i \), with a set of participants in which we can measure \( y \). Moreover, we had resources to measure \( y \) in a set of participants of size \( m = 500 \). The set of participants was obtained with an adaptive list sequential sampling method where we dealt with non-participating during the recruitment period.
Individuals in different subgroups had different participation probabilities and different frequencies of the outcome $y$. The characteristics of the populations were

$$g = (1, 2, 3, 4, 5)$$

$$p(x_i = g) = (40\% 20\% 20\% 10\% 10\%)$$

$$p(s_i = 1|b_i = 1, x_i = g) = (50\% 60\% 70\% 80\% 90\%)$$

$$p(y_i = 1|x_i = g) = (10\% 20\% 30\% 40\% 50\%)$$

where $p(s_i = 1|b_i = 1, x_i = g)$ was the participation probability of individual $i$ given $x_i = g$, i.e., for individual $i$ the probability of participating depended on $x_i$. The response to an invitation was drawn from a Bernoulli distribution with probability $p(s_i = 1|b_i = 1, x_i = g)$. Moreover, $E(Y) = n \sum g p(y_i = 1|x_i = g)p(x_i = g) = 1150$.

The individuals in the set of participants $s$ were used to estimate $Y$, denoted as $Y_{HT}$, where we used the Horvitz-Thompson estimator and its variance \cite{14-16} to determine $Y_{HT}$. The estimate $Y_{HT}$ was calculated as

$$\hat{Y}_{HT} = \frac{1}{n_s} \sum_{i \in s} y_i$$

(9)

where $\pi_i^{(0)}$ was the desired probability of being included in the set of participants $s$, specified before the recruitment period started. The variance of $Y_{HT}$ was approximated with

$$\hat{V}(\hat{Y}_{HT}) = \frac{1}{n_s} \sum_{i \in s} \frac{\pi_{ij}^{(0)} - \pi_i^{(0)} \pi_j^{(0)}}{\pi_i^{(0)} \pi_j^{(0)}}$$

where $\pi_{ij}^{(0)}$ is the second order joint-inclusion probability of the $i^{th}$ and $j^{th}$ individuals in $s$, i.e., $\pi_{ij}^{(0)} = p(s_i = 1, s_j = 1)$. To determine $\pi_i^{(0)}$, we used the sample based approximation technique proposed by Hájek \cite{17,18}.

The set of participants $s$ was obtained with the adaptive list sequential sampling method. Before the recruitment period started, we specified the vector $\pi^{(0)}$. We considered a vector $\pi^{(0)}$, in which the probability of being included in $s$ was proportional to the size of group $g$ in the population. Because not all groups were observed with the same frequency in $D$, we oversampled the smaller subgroups in such a way that each group $g$ was observed with similar frequency in $s$. For each invited individual with $x = 1$, we have to invite 2, 2, 4, and 4 individuals with respectively $x = 2, 3, 4, 5$ to obtain an equal number of individuals from each group in $s$. Therefore, depending on the value of $x_i$, we used the following probabilities for individual $i$

$$\pi_i^{(0)} = \begin{cases} 
0.05 & \text{if } x_i = 1 \\
0.10 & \text{if } x_i = 2 \text{ or } x_i = 3 \\
0.20 & \text{if } x_i = 4 \text{ or } x_i = 5
\end{cases}$$

Note that we could also use stratified sampling to get our desired set of participants because we only have five disjoint groups. However when we have a large number of groups, stratification becomes impracticable. A large number of groups is no problem for the (adaptive) list sequential sampling design, if it is possible to specify a distance measure between individuals (see (3)). With $\pi^{(0)}$, we expected to have an equal number of individuals for each subgroup $g$ in the set of participants.

We considered two adaptive list sequential methods to recruit the sample.

**Simple random sampling:** Assign zero to all weights $w_k^{(i)}$. Therefore $\pi_i^{(i-1)} = \pi_i^{(0)}$.

**Adjusted sampling:** To recruit a well spread sample, the inclusion probability of individual $i$ should only be influenced by individuals located in the same group. Therefore, we used the following preliminary weights

$$\tilde{w}_k^{(i)} = \begin{cases} 
1/(n_g - 1) & \text{if } x_i = g \text{ and } x_k = g, \\
0 & \text{otherwise},
\end{cases}$$

where $n_g$ is the number of individuals in group $g$.

For both adaptive list sequential sampling methods, we used the following model to describe the participation probability

$$p(s_i = 1|b_i = 1, x_i = g, \beta) = \frac{\exp[\beta_g I(x_i = g)]}{1 + \exp[\beta_g I(x_i = g)]}$$

where $\beta_g$ is the regression weight for group $g$. Because we assumed we had no a-priori information about the participation probabilities, we used non-informative priors for $\beta$ by sampling all five parameters $\beta_g$ from a normal distribution with mean zero and variance 100. For individual $i$, delayed response to the invitation was simulated by drawing time $t_i$ from a Poisson distribution with expectation 15. Individual $i$ responded to the invitation after the evaluation of individual $i + t_i$.

The simulations were performed 1000 times and we calculated the bias, MSE, and coverage of $Y_{HT}$ for both adaptive list sequential methods.

**Results**

**Simulation 1**

The results from simulation 1 have been summarized in Table 1. The results showed that the adaptive list sequential sampling method with the adjusted sampling 2 performed best. In this approach, the participation probability of individual $i$ was only influenced by the participation indicator of the 50 nearest neighbors. The recruited sets of participants better spread than with the other sampling approaches, reflected by the lower median and spread of $R$. 

---

[14-16]: Hof et al. BMC Medical Research Methodology 2014, 14:81
http://www.biomedcentral.com/1471-2288/14/81
Table 1 95% Confidence interval of $R$ and the number of participants in simulation 1

| Sampling method                  | Measure $R$ | Number of participants |
|----------------------------------|-------------|------------------------|
|                                  | 2.5%        | 50%        | 97.5%        | Mean | Standard deviation |
| Simple random sampling           | 0.192       | 0.238      | 0.304        | 401  | 18                   |
| Adjusted sampling 1              | 0.199       | 0.241      | 0.298        | 397  | 11                   |
| Adjusted sampling 2              | 0.157       | 0.189      | 0.225        | 397  | 11                   |

Estimated participation probability: $\text{invlogit}[\alpha + z_i \hat{\beta}]$

| Sampling method                  | Measure $R$ | Number of participants |
|----------------------------------|-------------|------------------------|
|                                  | 2.5%        | 50%        | 97.5%        | Mean | Standard deviation |
| Simple random sampling           | 0.188       | 0.230      | 0.304        | 405  | 18                   |
| Adjusted sampling 1              | 0.197       | 0.238      | 0.291        | 400  | 11                   |
| Adjusted sampling 2              | 0.154       | 0.184      | 0.225        | 400  | 11                   |

Using all the auxiliary characteristics $z_i$ to estimate the participation probability of individual $i$, the simple random sampling approach resulted in a median $R$ of 0.238 (95% confidence interval: 0.192–0.304). The mean number of participants with the simple random sampling approach was about 401 (95% confidence interval: 365 – 436). For the adjusted sampling 1 approach, approximately similar results were found for $R$, i.e. on average, the set of participants obtained with the simple random sampling approach and the adjusted sampling 1 approach were comparable in how well they were spread. With the adjusted sampling 1 approach, the average size of the set of participants was 397 (95% confidence interval: 376 – 418). However, compared to the simple random sampling approach, the variation in the size of the set of participants was considerably lower with the adjusted sampling 1 approach (respectively standard deviations of 18 and 11).

On average, a set of participants recruited with the adjusted sampling 2 approach was better spread than with the other two approaches. Not only was the median $R$ 0.189, the spread around the median was also smaller than with the other two approaches (95% confidence interval: 0.157–0.225). The mean size of the set of participants with the adjusted sampling 2 approach was 397 (95% confidence interval: 376 – 418), which was comparable to the adjusted sampling 1 approach.

Interestingly, the performances of all three approaches remained similar when we ignored the auxiliary characteristics $z_i$ in the estimation of the participation probability of individual $i$. Since fitting a model with just an intercept gave comparable results to the more complicated model where we also included $z_i$, the results suggested that the adaptive list sequential sampling method was robust to miss-specification of the participation probability model.

Simulation 2

The results from simulation 2 have been summarized in Table 2. Using the set of participants obtained with the simple random sampling approach resulted in a biased estimate of $\hat{Y}_{HT}$. With the adjusted sampling approach, $\hat{Y}_{HT}$ was more accurately estimated. This was reflected in the bias (+31 for simple random sampling and +1 for adjusted sampling), and the variance of the estimate (7995 for simple random sampling and 7817 for adjusted sampling). Consequently, the coverage of the 95% confidence interval was better when we used the adjusted sampling approach (0.86 for simple random sampling and 0.92 for adjusted sampling).

Discussion

In this paper, we developed an adaptive list sequential sampling method when a random sample from the population is required and the willingness to participate varies between individuals and is not known beforehand. Our adaptive list sequential sampling method requires that the characteristics that are related to the participation probability are known of all individuals. With simulations, we showed that the adaptive list sequential sampling method

Table 2 Estimated frequency $\hat{Y}$ derived from the set of participants in simulation 2

| Sampling method  | $E(\hat{Y}_{HT})$ | Bias | Variance | Mean squared error | Coverage of the 95% confidence interval |
|------------------|-------------------|------|----------|-------------------|----------------------------------------|
| Simple random sampling | 1181             | 31   | 7995     | 14457             | 0.86                                   |
| Adjusted sampling     | 1151             | 1    | 7817     | 10288             | 0.92                                   |
could successfully deal with unknown heterogeneous participation probabilities.

In our adaptive list sequential sampling method, we evaluate each individual from the population only once. Therefore we only have one opportunity to decide whether to invite an individual or not. When we overestimate the participation probability for all individuals from the population, we end up with a too small set of participants. A simple solution for this problem would be to re-evaluate non-invited individuals until the desired size of participants in the study has been reached.

The simulations suggested that the adaptive list sequential sampling method is robust to miss-specification of the participation probability model. Just using an intercept term to describe the participation probability seems to work quite well. However, to what extent the adaptive list sequential sampling method can deal with wrong participation probability estimates was not investigated in this paper. In addition, extreme delayed response to the invitation has influence on the performance of the list sequential sampling method. Further research is necessary to determine in which situations the adaptive list sequential sampling method succeeds and fails to recruit a well spread set of participants.

A problem that was not considered here was the use of multiple invitation techniques in sampling designs. For instance, there could be individuals in the population that have a low willingness to participate when they are invited by a letter, but a much larger willingness when invited by telephone. Our method can be adopted by introducing multiple participation probabilities by extending step 3 of our algorithm and estimate multiple logistic regression participation probabilities.

Conclusions

We showed that correcting for heterogeneity in the participation probability during the recruitment period is an effective approach when we have no or partial knowledge on the willingness to participate in population studies. By inviting individuals from the population in stages, the participation probability can be estimated and used in the sampling procedure.

Competing interests

The authors declare that they have no competing interests.

Authors’ contributions

All authors participated in the design of the study. MHH performed the statistical analysis and drafted the manuscript. All authors read and approved the final manuscript.

Author details

1Department of Clinical Epidemiology, Bioinformatics, and Biostatistics, Academic Medical Center - University of Amsterdam, Meibergdreef 9, 1105 AZ, Amsterdam, The Netherlands. 2Department of Clinical Informatics, Bioinformatics, and Biostatistics, Academic Medical Center - University of Amsterdam, Meibergdreef 9, 1105 AZ, Amsterdam, The Netherlands.

Received: 14 March 2014 Accepted: 12 June 2014

References

1. Rothman KJ: Epidemicology: An Introduction. Oxford: Oxford University Press; 2002.
2. Kruskal W, Mosteller F: Representative sampling, i: Scientific literature, excluding statistics. Int Stat Rev Rev Int Stat 1979, 47(2):111–127.
3. Kruskal W, Mosteller F: Representative sampling, ii: Non-scientific literature. Int Stat Rev Rev Int Stat 1979, 47(1):13–24.
4. Kruskal W, Mosteller F: Representative sampling, iii: The current statistical literature. Int Stat Rev Rev Int Stat 1979, 47(3):245–265.
5. Kruskal W, Mosteller F: Representative sampling, iv: The history of the concept in statistics, 1895-1939. Int Stat Rev Rev Int Stat 1980, 48(2):169–195.
6. Galea S, Tracy M: Participation rates in epidemiologic studies. Ann Epidemiol 2007, 17(9):693–698.
7. Pickery J, Carton A: Oversampling in relation to differential regional response rates. Surv Res Methods 2008, 2(2):83–92.
8. Stronks K, Snijder MB, Peters RJG, Prins M, Schene AH, Zwinderman AH: Unravelling the impact of ethnicity on health in Europe: the HELIUS study. BMC Public Health 2013, 13:402. doi:10.1186/1471-2458-13-402.
9. Sabio M: Population-based Cohort Studies. Epidemiol Rev 1998, 20(1):91–90.
10. Graffström A, Lundström NLP: Why well spread probability samples are balanced. Open J Stat 2013, 3(6):41.
11. Graffström A, Schelin L: How to Select Representative Samples. Scand J Stat 2014, 41(2):277–290. doi:10.1111/sjos.12016.
12. Bondeson L, Thorburn D: A list sequential sampling method suitable for real-time sampling. Scand J Stat 2006, 33(3):466–483.
13. Deville J, Tillé Y: Efficient balanced sampling: The cube method. Biometrika 2004, 91(4):893–912.
14. Horvitz DG, Thompson DJ: A generalization of sampling without replacement from a finite universe. J Am Stat Assoc 1952, 47(260):663–685.
15. Tillé Y: Sampling Algorithms. Springer Series in Statistics. New York: Springer; 2006.
16. Narain RD: On sampling without replacement with varying probabilities. J Indian Soc Agric Stat 1951, 3:169–175.
17. Hājek J: Asymptotic theory of rejective sampling with varying probabilities from a finite population. Ann Math Stat 1964, 35(4):1491–1523.
18. Bergers YG: Asymptotic consistency under large entropy sampling designs with unequal probabilities. Pak J Stat 2011, 27(4):407–426.

Cite this article as: Hof et al.: Adaptive list sequential sampling method for population-based observational studies. BMC Medical Research Methodology 2014 14:81.

Submit your next manuscript to BioMed Central and take full advantage of:

• Convenient online submission
• Thorough peer review
• No space constraints or color figure charges
• Immediate publication on acceptance
• Inclusion in PubMed, CAS, Scopus and Google Scholar
• Research which is freely available for redistribution

Submit your manuscript at www.biomedcentral.com/submit