CONFERENCE OF EUROPEAN STATISTICIANS

Workshop on Statistical Data Collection

10-12 October 2018, Geneva, Switzerland

Session 5

08 October 2018

Adaptive data collection at Statistics Netherlands with an application to the Health Survey

Kees van Berkel (Statistics Netherlands)

cam.vanberkel@cbs.nl

Abstract and Paper

Challenges that surveys are facing are increasing data collection costs and declining budgets. During the past years, many surveys at Statistics Netherlands were redesigned to reduce cost and to increase or maintain response rates. Currently, alternative approaches are investigated to produce more accurate estimates within the same budget. Adaptive data collection is proposed for achieving this goal.

Research into the effect of reducing face to face observation in mixed mode surveys on quality and costs was carried out in 2017. Reducing face to face observation can be done in various ways. It can be done through random selection, but also through stratified selection of nonrespondents eligible for face to face follow-up. By using the latter method, nonresponse bias can potentially be reduced. The key decisions to be made are how to divide the population into strata and how to compute the allocation probabilities for face to face follow-up in the different strata.

In this presentation the adaptive data collection is elaborated for the Health Survey as it is conducted by Statistics Netherlands since 2018. Attention is paid to the choice of the strata, the choice of the mixed mode observation strategy, the optimization problem with corresponding constraints and the effect of the adaptive data collection on most important survey estimates.

Key words: balanced response, nonresponse bias, accuracy, data collection costs.

Adaptive Data Collection at Statistics Netherlands with an application to the Health Survey

Kees van Berkel, Suzanne van der Doef en Barry Schouten

Statistics Netherlands

Abstract

Challenges that surveys are facing are increasing data collection costs and declining budgets. During the past years, many surveys at Statistics Netherlands were redesigned to reduce cost and to increase or maintain response rates. Currently, alternative approaches are investigated to produce more accurate estimates within the same budget. Adaptive data collection is proposed for achieving this goal.

Research into the effect of reducing face to face observation in mixed mode surveys on quality and costs was carried out in 2017. Reducing face to face observation can be done in various ways. It can be done through random selection, but also through stratified selection of nonrespondents eligible for face to face follow-up. By using the latter method, nonresponse bias can potentially be reduced. The key decisions to be made are how to divide the population into strata and how to compute the allocation probabilities for face to face follow-up in the different strata.

In this paper the adaptive data collection is elaborated for the Health Survey as it is conducted by Statistics Netherlands in 2018. Attention is paid to the mixed mode observation strategy, the choice of the strata, the calculation of the follow-up sampling fractions per stratum and the effect of the adaptive data collection on most important survey estimates.

Key words: balanced response, nonresponse bias, accuracy, data collection costs.

1. Introduction

Adaptive data collection assumes that differentiation of effort over relevant population subgroups is either effective in improving survey quality or efficient in reducing survey costs. The designs have received a lot of interest over the last decade in response to budget pressure due to gradual but persistent declines of response rates, e.g. Chun, Heeringa and Schouten (2018).

National Statistical Offices have the task of publishing reliable and coherent statistical information that responds to the needs of society. In order to maintain a good balance between quality, efficiency and cost-effectiveness, continuous evaluation and improvement of processes and working methods is necessary. In 2016, four data collection policy decisions were made at Statistics Netherlands in order to arrive at a more efficient data collection strategy: incentives were used to increase overall response rates, a second supplier of telephone numbers was deployed so that more telephone observation is possible, follow-up sample sizes for CATI and CAPI were fixed in order to stabilize interviewer workload, and adaptive data collection became a standard design choice in sequential mixed-mode surveys. These four changes were implemented to varying degrees for a large number of surveys since 2017. Here, the application of adaptive data collection is elaborated.

The paper reads as follows: Section 2 describes the methodology behind the adaptive data collection. Section 3 discusses the application to the Dutch Health survey. Section 4 ends with the effect of adaptive data collection on most important survey estimates.

2. Methodology

In this section, the four main elements of adaptive data collection are discussed: quality indicator for bias of survey estimates, design features, stratification of the target population, and interfering in the process of data collection.

2.1 Quality indicator

The adaptive data collection is focussed on optimizing balance of response through the coefficient of variation (CV) of response propensities for relevant population subgroups. See Schouten, Cobben, Bethlehem (2009), De Heij, Schouten, Shlomo (2015) and Moore, Durrant and Smith (2018). The CV is based on the desire to limit the risk of nonresponse bias over a range of variables.

The random response model is adopted. It is assumed that each population unit *k* has a response probability ρ_k which is only known to unit *k*, and response of the unit is independent of other population units. For the survey, a single random sample without replacement of size *n* is drawn from the target population. The size of the target population is N.

The number of respondents r in the sample survey is a random variable $r = \sum_{k=1}^{N} a_k r_k$ with expected value $n\bar{\rho}$, where $\bar{\rho}$ denotes the average response rate. An estimator of the population mean $\bar{Y}$ of target variable *Y* is the response mean,

$$\bar{y} = \frac{1}{r} \sum_{k=1}^{r} y_k.$$

The response mean $\bar{y}$ is in general a biased estimator for the population average $\bar{Y}$. If $\rho_k = \bar{\rho}$ for all *k*, then $\bar{y}$ is unbiased, but this is generally not true. Bethlehem (1988) shows that

$$B(\bar{y}) = E(\bar{y}) - \bar{Y} \approx \frac{1}{\bar{\rho}N} \sum_{k=1}^{N} (\rho_k - \bar{\rho})\, Y_k = \frac{1}{\bar{\rho}}\, cov(\rho, Y).$$

Here $cov(\rho, Y)$ is the population covariance between the response probabilities and the values of the target variable. So there is no bias if there is no correlation between response propensity and the target variable. Introduce Pearson's correlation coefficient:

$$R(\rho, Y) = \frac{cov(\rho, Y)}{S_\rho S_Y},$$

where S_ρ is the standard deviation of the response probabilities and S_Y is the standard deviation of the values of the target variable. Then the bias approximation formula can be written as

$$B(\bar{y}) \approx \frac{R(\rho, Y)\, S_\rho\, S_Y}{\bar{\rho}}.$$

From this expression it follows:

1. $B(\bar{y}) = 0$ if there is no linear relationship between ρ and *Y*.
2. The stronger the linear relationship between ρ and *Y*, the larger $B(\bar{y})$.
3. $B(\bar{y}) = 0$ if there is no variation of response rates or no variation in the values of the target variable.
4. The smaller the variation of response rates, the smaller $B(\bar{y})$.
5. The smaller the variation in the values of the target variable, the smaller $B(\bar{y})$.
6. The greater the average response rate, the smaller $B(\bar{y})$.

Since the absolute value of Pearson's correlation coefficient is maximum 1, an upper limit for the bias can be given:

$$|B(\bar{y})| \leq \frac{S_\rho \, S_Y}{\bar{\rho}} = CV(\rho) \, S_Y \, .$$

Here $CV(\rho)$ denotes the coefficient of variation of the response probabilities. In the remainder of this paper, an attempt is made to minimize this coefficient of variation by interfering in the process of data collection.

2.2 Design features

The focus is on the mix of survey modes. It is assumed that a sequential mixed-mode design is used with CAWI (Computer-Assisted Web Interviewing) as the starting mode. Follow-up of CAWI nonresponse is done through interviewer modes. Here, it is assumed that the follow-up is done by CAPI (Computer-Assisted Personal Interviewing). The design feature to adapt is the CAPI follow-up.

In the sequential mode strategy, all sample people are first asked by letter to participate in the survey by completing a questionnaire on the Internet. People who have not responded to this request after no more than two reminders are visited at home to conduct an interview. The observation strategy of the face-to-face interviews is adjusted as follows. To reduce the variation of response rates, more CAPI is used for groups that respond badly via the Internet than for groups that respond well. However, the entire sample does start with CAWI. The identification of these so-called target groups is carried out using cluster analysis.

It is assumed that the answers obtained are the same in different observation modes, i.e. mode-specific measurement bias is absent and can be ignored. This is a simplification as such biases are conjectured to exist and should then be incorporated in the design decisions within the adaptive data collection. Therefore, the effect of adaptation on survey estimates is quantified by means of bootstrapping.

2.3 Stratification of the target population

Determining target groups is also called segmentation or clustering of the target population. The target groups are composed by means of response propensities of people per mode. This may mean that two target groups have approximately the same response rate at CAWI, but that their CAPI response rates differ. It is also possible that the total response rates of two target groups are approximately the same, but that their response rates differ per mode.

Clustering is can be carried through algorithms generating a classification tree. Such an algorithm is implemented in the R package rpart. The reference manual and package source can be found on the Internet site https://CRAN.R-project.org/package=rpart. Demographic and regional characteristics of people can be used that are known to have different response behaviour. Examples are ethnicity, age, income, degree of urbanization, educational level, and marital status. The algorithm determines which characteristics are used to split the groups and in which order. The characteristic with the largest differences in response behaviour is used first to split the population. For categorical variables, the algorithm also determines where to split. This ensures that for variables such as age, a classification can be made that best matches the response behaviour.

To ensure that reliable response rates per mode can be estimated for each group, it is important that the groups do not become too small. To prevent this, a minimum size per target group can be set. The final groups are called target groups. Within each target group there is little variation in response behaviour per mode, but between two target groups there is a big difference in response behaviour for at least one mode.

2.4 Interfering in the process of data collection

Let G be the set of groups used to determine the target groups. Each target group is the union of one or more groups from G. For each $g \in G$, let $N(G)$ denote the population size of group g. For a simple random sample of size n, it is assumed that the size of the sample in group g equals $n(g) = n \cdot N(g)/N$.

Furthermore, for each group $g \in G$ it is assumed that all people have the same CAWI-response probability $p_w(g)$, the same probability $p_h(g)$ of being eligible for face-to-face follow-up and the same CAPI-response probability $p_p(g)$. The latter probability is the chance of CAPI-response in the face-to-face approached sample of group g. Let $f_h(g)$ be the fraction of people who are approached face-to-face in the CAWI-nonrespondents who are eligible for face-to-face follow-up of group g. For the total response probability in group g, the following applies

$$p(g) = p_w(g) + f_h(g)p_h(g)p_p(g).$$

This allows the average response probability and the population variance of the response probabilities to be calculated:

$$\bar{\rho} = \frac{1}{N}\sum_{g\in G} N(g)\,p(g) \text{ and } S_\rho^2 = \frac{1}{N}\sum_{g\in G} N(g)\,(p(g) - \bar{\rho})^2 .$$

The following problem needs to be solved.

Minimize $CV(\rho) = S_\rho/\bar{\rho}$ under a specified number of constraints.

Different types of constraints can be used:

- *Budget*. This can be done at different levels, such as an available budget for the total observation or per observation mode.
- *Capacity*. An upper limit can be specified for the sample size to be approached face-to-face. This can be at national or regional level.
- *Precision*. This concerns requirements for the number of respondents or the number of respondents per subpopulation.
- *Response rate*s. For example, a minimum response rate, or minimum response rates per mode or per subpopulation.

One CAPI sampling fraction is used per target group. This leads to the extra constraint:

For each target group d and all groups $g_1, g_2 \subset d : f_h(g_1) = f_h(g_2)$ applies.

The decision variables for which the minimum can be found, are the sample size n and the sampling fractions $f_h(g)$ for face-to-face interviews in the groups $g \in G$.

The optimization problem requires a search for the numbers of people to be approached by target group and observation mode. The lower the CAWI response propensity of a target group is, the more face-to-face observation is applied. This may lead to a smaller variation of response rates, and the ratio of the target groups in the response may be more similar to the ratio of the target groups in the population. This may, however, be at the expense of the overall response rate.

The minimization problem can be solved with the Auglag function of the Alabama R package. The reference manual and package source can be found on the Internet site https://CRAN.R-project.org/package=alabama. Auglag is the abbreviation of Augmented Lagrangian Adaptive Barrier Minimization Algorithm. In general, the Auglag function can be used for optimizing smooth nonlinear objective functions with constraints. Our minimization problem can also be solved in Excel. The

solver in Excel uses the generalized reduced gradient method. Since our problem is nonlinear, the algorithms can end up in a local minimum. So it is recommended to use different random starting values for the sample size and the sampling fractions, and select the solution that fits best.

3. Application of adaptive data collection to the Dutch Health Survey

3.1 About the Dutch Health Survey

The aim of the Dutch Health Survey is to provide as complete an overview as possible of developments in health, medical contacts, lifestyle and preventive behaviour of the population in the Netherlands. The target population consists of all people living in the Netherlands who do not belong to the institutional population. The sample is a stratified two stage sample in which people with equal probabilities are selected. This sampling design is approximately the same as the simple random sampling design. The observation starts with CAWI and the follow-up mode is CAPI. As a response increasing measure, iPads are raffled among the sampled people.

3.2 Stratification of the target population

Stratification is carried out through the R package rpart, using the dataset of the Health Survey, January – June 2017. The personal characteristics used for explaining the response behaviour are included in Annex 1. The results of the classification tree are the characteristics used for the Health Survey to record the target groups: ethnicity (NL resident, migrant with a Western background, migrant with a non-Western background), age (in years), income (in quintiles) and degree of urbanisation of the municipality in which the person lives (very strongly urban, strongly urban, moderately urban, few urban and non-urban). The classification tree ensures that the characteristics are merged into larger groups. Ethnicity is divided into two groups, namely Western (NL residents and migrants with a Western background) and non-Western (migrants with a non-Western background). Age is divided into four categories: 0-11, 12-24, 25-64 and 65+. The income used is the standardised household income and the classification is into two categories, with the low income category consisting of the lowest 20% and the high income category consisting of the remaining 80%. The degree of urbanization is reduced to two categories, namely very strongly urban and all others. Figure 1 shows the classification tree. The tree is read from top to bottom. In each node a division is made based on a characteristic and the group is split. At the bottom of the tree the ultimate target groups can be found, together with the response rates of CAWI and CAPI.

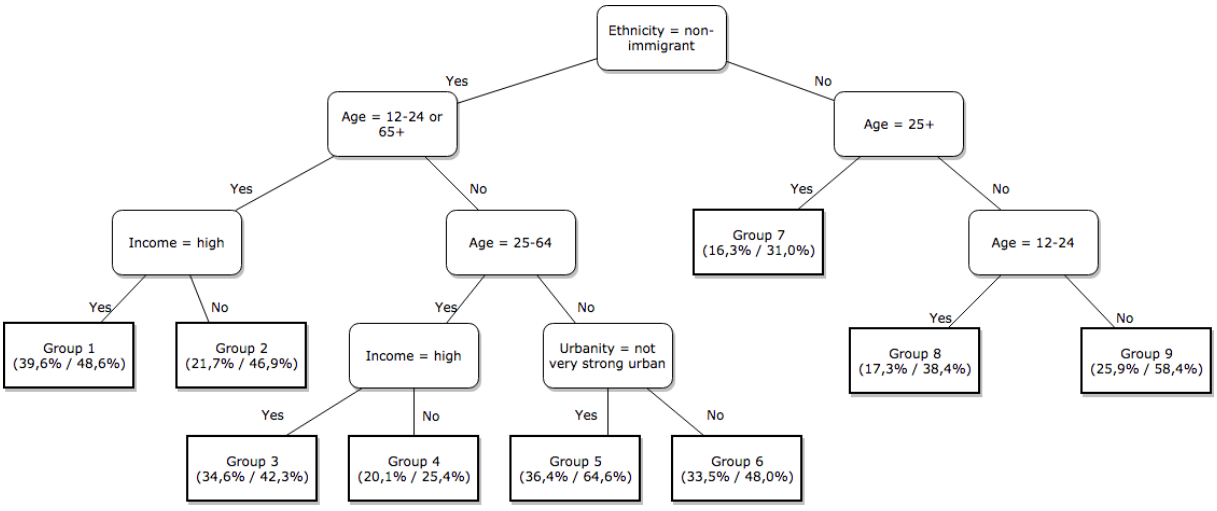

Figure 1: Classification tree for the Health Survey based on data collection January – June 2017.

The first six target groups partition the NL residents and migrants with a Western background. Figure 2 contains an overview of these target groups. Here urbanity = 1 means very strongly urban and urbanity = 2-5 means the union of the remaining categories.

	income	high		low	
age	urbanity	1	2-5	1	2-5
0-11		5	6	5	6
12-24		1	1	2	2
25-64		3	3	4	4
65+		1	1	2	2

Figure 2: Partition of NL residents and migrants with a Western background into target groups

The migrants with a non-Western background are divided into three target groups by age: 25+ in target group 7, 12 – 24 in target group 8 and 0 – 11 in target group 9.

3.3 The Dutch Health Survey minimization problem

The set G of groups used to determine the target groups consists of 32 groups: ethnicity(2) × age(4) × income(2) × degree of urbanization(2). Minimising $CV(\rho) = S_\rho/\bar{\rho}$ is carried out under the constraints:

- $n \leq n_{max}$ {CAWI sample size does not exceed n_{max}},
- $n \cdot \bar{\rho} \geq R$ {expected response size is at least R},
- $\sum_{g \in G} f_h(g) p_h(g) n(g) \leq C$ {total CAPI-sample size is at most C },
- For each target group d and all groups $g_1, g_2 \subset d$: $f_h(g_1) = f_h(g_2)$ applies {one CAPI sampling fraction per target group}.

Here n_{max}, R and C are constants to be filled in. The parameters with which the minimum can be found are the sample size n and the sampling fractions $f_h(g)$ for face-to-face observation in the groups $g \in G$. Note that it follows from the first two constraints that $\bar{\rho} \geq R/n_{max}$.

In the case of the Health Survey 2018, the target groups with corresponding response rates and probabilities of re-approachable CAWI nonresponse have been determined with data from the results of the Health Survey in January-June of 2017. As a response increasing measure, in 2018 iPads are raffled among the sampled people. Therefore, the CAWI response rates were increased by three percentage points per target group. The maximum CAWI sample size n_{max} has been set to 18,000 people, the minimum expected response size R is 9,628 people, and for the maximum CAPI sample size C is 8,039 addresses. The average response rate must therefore be at least 9,628 / 18,000 = 53.5%.

3.4 Mathematical minimization

The minimization problem is solved with the solver in R, with different random starting values for the sample size n and the sampling fractions $f_h(g)$, $g \in G$, because the problem is nonlinear allowing the algorithm to stop in a local minimum. The optimal solution is the solution with the lowest coefficient of variation. In 100 hours the algorithm has found 11 solutions. The coefficients of variation of the different solutions are between 0.1123 and 0.1190. Except for one solution that had a coefficient of variation of 0.21. This solution was not considered further. It cannot be guaranteed that 0.1123 is the overall minimum of the coefficient of variation.

The remaining 10 solutions have almost the same coefficients of variation, but with different sampling fractions for face-to-face observation per target group. Not all solutions use the maximum allowable number of people to be approached face-to-face. For the solution with the least use of CAPI, 7406 people are approached face-to-face with a coefficient of variation of 0.1123. The solution with the most use of CAPI, 8039 people are approached face-to-face with a coefficient of variation of 0.1158. This solution was ultimately chosen because the variation coefficients hardly differ from each other, but the use of face-to-face observation is fully utilised.

With the adaptive data collection, the average response rate decreases compared to a sequential CAWI-CAPI design in which all CAWI non-respondents eligible for follow-up are visited at home. However, both the standard deviation and the variation coefficient of the response probabilities are smaller for adaptive data collection.

Table 1 shows the results of the chosen solution. The column *n CAWI* contains the CAWI sample size, the column *r CAWI* the expected number of CAWI respondents and *p CAWI* shows the expected response rate for CAWI. Column *n elig* shows the number of CAWI- nonrespondents eligible for face-to-face follow-up. Columns *n CAPI, f CAPI, r CAPI* and *p CAPI* represent the CAPI sample size, the CAPI sampling fraction n CAPI / n elig, the expected number of CAPI respondents and the expected CAPI response rate. The columns *r tot* and *p tot* indicate the total number of expected responses and the total response rates per target group. These response rates have been estimated with results of the Health Survey, January - June 2017, with an adjustment to the CAWI response rates due to the raffle of iPads among the sampled people.

Table 1: Results of adaptive data collection.

stratum	n CAWI	r CAWI	p CAWI	n elig	n CAPI	f CAPI	r CAPI	p CAPI	r tot	p tot
			%			%		%		%
1	4,550	1,939	42.6	2,567	1,448	56,4	703	48.5	2,642	58.1
2	786	195	24.8	553	526	95.1	248	47.1	443	56.4
3	7,373	2,770	37.6	4,767	3,405	71.4	1,441	42.3	4,211	57.1
4	728	168	23.1	551	551	100.0	140	25.4	308	42.3
5	1,474	580	39.3	933	406	43.5	263	64.8	843	57.2
6	333	121	36.3	221	147	66.5	70	47.6	191	57.4
7	1,276	246	19.3	1,040	1,040	100.0	320	30.8	566	44.4
8	411	83	20.2	341	341	100.0	132	38.7	215	52.3
9	363	105	28.9	266	175	65.8	102	58.3	207	57.0
total	17,295	6,208	35.9	11,238	8,039	71.5	3,420	42.5	9,626	55.7

Table 2 shows quality measures, in which the situations without and with adaptive data collection are compared. This table shows that the use of adaptive data collection causes the overall response rate to decrease, but the variation of the response rates is improving and the ultimate quality measure $CV(\rho)$ is also improving and therefore there is less bias due to nonresponse.

Table 2: Quality indicators for adaptive data collection.

Adaptive data collection	$\bar{\rho}$	S_ρ	$CV(\rho)$
	%		
No	64.4	10.2	15.8
Yes	55.7	6.4	11.6

3.5 Method effects for the Dutch Health Survey

To get an idea of the effect of the adaptive data collection on the results of the Health Survey, simulations were carried out using bootstrapping. To this end, samples were drawn with replacement from the sample of the past year with the correct numbers for CAWI and matching numbers per target group for CAPI. The response data and the survey answers are then linked to these samples. For each sample, the corresponding response was weighted using the weighting model of the Health Survey. Then estimates were made for the most important target variables and these were compared with the regular estimates.

For the bootstrapping, 1000 samples with replacement were drawn from the 2016 sample. Each sample has the right CAWI size and the right CAPI size per target group. By chance, the numbers of responses may vary per sample. The sample numbers are taken from the sampling design with adaptive data collection for the Health Survey 2018. After weighting the response per sample, target variables were estimated for both the entire population and subpopulations. These estimates were compared with the results of the Health Survey 2015 and 2016. One of the assumptions to use CAPI in an adaptive data collection is that the respondents' answers do not depend on the mode in which they respond. This is a strong assumption that is not always true in practice.

The target variable smoking status is known to have mode effects. The proportion of smokers in CAWI respondents is smaller than in CAPI respondents. So if relatively more is observed via CAWI and less via CAPI, the number of smokers is expected to decrease. With adaptive data collection, more migrants with a non-Western background are approached face-to-face and fewer NL residents or migrants with a Western background. Therefore, it is expected that the proportion of smokers among migrants with a non-Western background will increase and decrease among NL residents and migrants with a Western background.

The results of the bootstrapping are in line with this, see Figure 3. The left part of figure 3 shows the smoking status for migrants with a non-Western background. The estimate with the corresponding 95% confidence interval of the Health Survey 2015 is shown in blue, the estimate with the corresponding 95% confidence interval of the Health Survey 2016 is in red. The histogram represents the results for this variable in the 1000 samples of the bootstrapping. The right part of figure 3 shows the smoking status for NL residents and migrants with a Western background. For NL residents and migrants with a Western background the percentage of smokers seems to decrease when adaptive data collection is used and for migrants with a non-Western background the percentage of smokers seems to increase compared to the measurement from 2016.

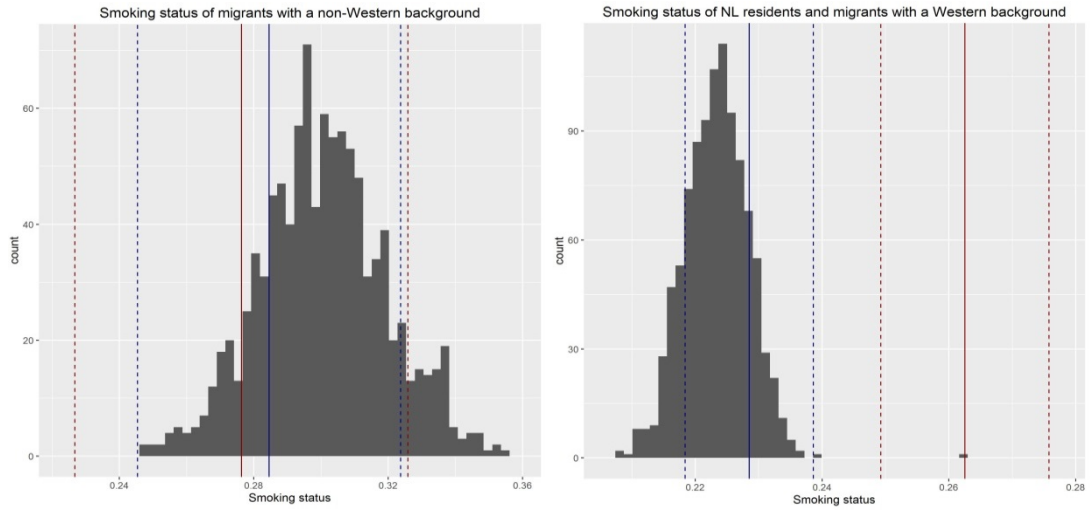

Figure 3: Smoking status by ethnicity

Questions about alcohol, drug use and sexual health are asked in the face-to-face approach via Computer Assisted Self Interviewing. Therefore, fewer mode effects are expected for these variables. Using the sample data from the bootstrapping, estimates were made for the eleven core variables: contact with general practitioner, contact with dentist, non-prescribed use of medication, experienced health, diabetes, mental health problems, disabilities, informal care, smoking, obesity and drug use, see table 3. Columns 2016 and SE 2016 show the estimates and standard errors for the core variables from the Health Survey 2016. The last two columns contain the estimates from the bootstrapping samples.

Table 3: Estimates and standard errors for the 11 core variables.

	2016	*SE 2016*	*bootstrap*	*SE bootstrap*
	%			
General practitioner contact	70.9	0.6	71.2	0.8
Dentist contact	79.0	0.5	79.4	0.7
Unprescribed use of medication	39.7	0.6	39.1	0.8
Experienced health	76.3	0.5	76.1	0.7
Diabetes	5.8	0.3	5.7	0.4
Psychologically unhealthy (MHI-5 score)	11.8	0.4	12.4	0.6
At least 1 OESO-restriction	12.2	0.4	12.0	0.5
Informal care	13.8	0.4	13.8	0.6
Smoking	23.4	0.5	23.2	0.7
Obesity	13.6	0.4	13.8	0.4
Drug use in the last 30 days	5.1	0.3	4.9	0.4

On the basis of the bootstrapping, it is expected that the results for experienced health, informal care and diabetes with adaptive data collection do not differ much from the results without adaptive data collection. The greatest shifts can be seen in non-prescribed use of medication and psychologically unhealthy. Figures 5 and 6 show the estimates of non-prescribed use of medication and psychologically unhealthy. The blue and red lines show estimates and confidence intervals for the Health Survey 2015 and 2016 respectively.

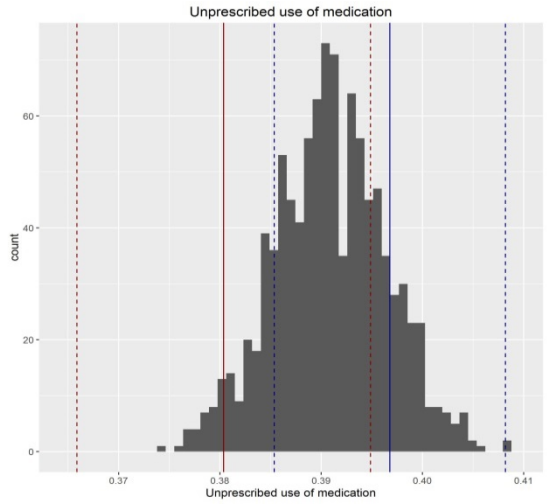

Figure 5: Unprescribed use of medication.

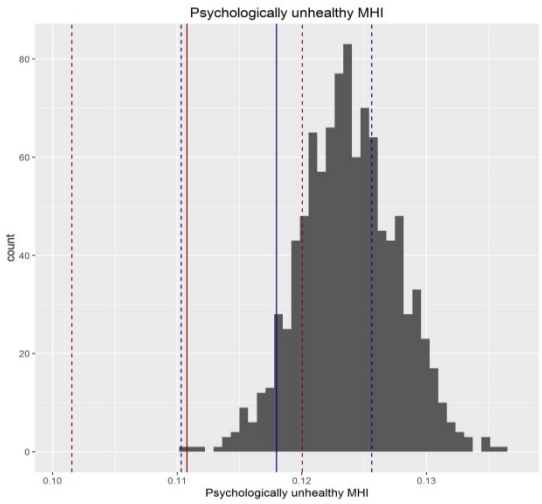

Figure 6: Psychologically unhealthy MHI.

The number of people taking non-prescribed medicines is expected to rise compared to the 2016 estimate. It is likely that the introduction of adaptive data collection will increase the number of people who are mentally unhealthy.

4. References

Bethlehem, J.G. (1988): Reduction of Nonresponse Bias Through Regression Estimation, Journal of Official Statistics, vol. 4. No. 3. Pp. 251 – 260.

Chun, A.Y., Heeringa, S.G., Schouten, B. (2018): Responsive and adaptive design for survey optimization, Journal of Official Statistics, 34 (3), 581 – 597.

De Heij, V., Schouten, B., Shlomo, N. (2015): RISQ 2.1 manual. Tools in SAS and R for the computation of R-indicators and partial R-indicators, available at www.risq-project.eu

Moore, J.C., Durrant, G.B., Smith, P.W.F. (2018): Data set representativeness during data collection in three UK social surveys: generalizability and the effects of auxiliary covariate choice, Journal of the Royal Statistical Society, Series A, 181 (1), 229 – 248.

Schouten, J.G., Cobben, F., Bethlehem, J. (2009): Indicators for the representativeness of survey response, Survey Methodology, 35 (1), 101 – 113.

Annex 1: Characteristics for segmentation of the population

1. Wealth of household: 1% groups.
2. Home ownership: owner, rent without rent subsidy, rent with rent subsidy.
3. Income: 1% groups of standardised disposable household income.
4. Socio-economic category: employee of private company, government employee, director or large shareholder, self-employed, employed other, claiming unemployment benefit, claiming income support benefit, claiming other social provision, disabled, pensioner younger than 65 years, pensioner 65 years or older, unemployed other.
5. Household size: number of people in the household.
6. Household status: child living at home, single person, partner without children, partner with children, parent in single parent household, reference person in other household, other household member.
7. Type of household: Single person household, unmarried couple without children, married couple without children, unmarried couple with children, married couple with children, married couple with children, single parent household, other household.
8. Gender: male, female.
9. Marital status: unmarried, married, partnership, divorced, widowed.
10. Age: in years.
11. Age of eldest child: in years.
12. Age of youngest child: in years.
13. Duration of stay in the Netherlands: in years.
14. Part of the country: north, east, south, west.
15. Province: the 12 provinces of the Netherlands.
16. G32: the largest 32 municipalities, other.
17. G4: the largest 4 municipalities, other.
18. Ethnicity: NL residents, migrants with a Western background, migrants with a non-Western background, unknown.
19. Ethnicity of mother: same.
20. Ethnicity of father: same.
21. Generation: NL residents, first generation migrants, second generation migrants.
22. Highest attained educational level: primary education, secondary general education, secondary vocational education, higher professional education, university.
23. Highest level of education: same.
24. Degree of urbanisation of municipality: very strongly urban, strongly urban, moderately urban, few urban and non-urban.
25. Degree of urbanisation of neighbourhood: same.