Survey Research Methods (2017)
Vol. 11, No. 2, pp. 105-123
doi:10.18148/srm/2017.v11i2.6803

© European Survey Research Association
ISSN 1864-3361
http://www.surveymethods.org

Differences in General Health of Internet Users and Non-users and Implications for the Use of Web Surveys

Rainer Schnell
University of Duisburg-Essen
Germany

Marcel Noack
University of Duisburg-Essen
Germany

Sabrina Torregroza
University of Duisburg-Essen
Germany

Web surveys have become popular in many fields of research. To compensate persisting undercoverage and nonresponse problems of web surveys, weighting strategies are used. However, the underlying assumptions of weighting are rarely tested. If the probability of missing data depends on the missing data itself (missing not at random, MNAR), no standard weighting method will correct for nonresponse or undercoverage bias. We postulate a MNAR selection effect due to health conditions. Using real data from large scale non-internet surveys in different countries (European Social Survey (ESS), $n \approx 55,000$, Behavioral Risk Factor Surveillance System (BRFSS), $n \approx 492,000$), large differences in general subjective health between Internet users and non-users can be observed. Weighting by calibration on age, gender, ethnic background, urban residence, education and household income does not eliminate the observed health differences. Therefore, the underlying missing data mechanism might be considered as an example of MNAR. If this holds, no weighting strategy will be able to eliminate health bias in web surveys.

Keywords: MNAR; Bias; ESS; BRFSS; Weighting; Calibration

1 Introduction

Web surveys are increasingly used in various fields of research, such as psychology (Batterham, 2014), sociology (Lee, 2006), election studies (Berrens, Bohara, Jenkins-Smith, Silva, & Weimer, 2003; Vavreck & Rivers, 2008) and health (Liu et al., 2010; Russell, Boggs, Palmer, & Rosenberg, 2010). In commercial settings in developed countries, the amount of spending on web surveys exceeds all other modes of data collection (ESOMAR, 2014). This is hardly surprising since traditional modes of data collection such as F2F and CATI surveys require more and more time, effort and money to counter the increasing proportion of nonrespondents, while web surveys appear to offer fast results with lower costs (Czaja & Blair, 2005, p. 40).

However, web surveys seem to suffer under higher rates of undercoverage and nonresponse than traditional data collection modes (Bethlehem & Biffignandi, 2012). To compensate the effects of these problems, different weighting strategies are used. In addition to standard methods such as raking, post-stratification and GREG (Kalton & Flores-Cervantes, 2003), propensity score weighting (Rosenbaum & Rubin, 1983) is increasingly used for web surveys (Lee, 2006; Valliant & Dever, 2011). All of these methods are based on the assumption that the missing data is either missing completely at random (MCAR) or the missing data can be explained adequately using observed data (missing at random, MAR) (Zhou, Zhou, Liu, & Ding, 2014). The success of post-hoc bias reduction by statistical methods depends on the correlation between response probability and the variable of interest (Schnell, 1993). A closed form equation is presented by Bethlehem and Biffignandi (2012). If the target variable is related to the cause of missing data and can not be explained by observed data, resulting estimates will be biased. Accordingly, simulations (Bethlehem, 2009) and empirical studies (Schonlau et al., 2004; Yeager et al., 2011) suggest that a strong correlation between target variables and response mechanisms will result in biased estimates despite weighting.

Contact information: Rainer Schnell, Universität Duisburg-Essen, Methodology Research Group, Forsthausweg 2, 47057 Duisburg (email: Rainer.Schnell@uni-due.de)

This article will show that respondent health is a variable strongly correlated with Internet usage. Hereby, surveys on health related topics will suffer from nonresponse and undercoverage caused by health issues. Therefore, this nonsampling error of web surveys might be an example of a MNAR missing data generating mechanism which cannot be corrected by weighting methods.

This is demonstrated using high quality large scale noninternet surveys conducted in 28 countries. For Europe, we used European Social Survey (ESS) data; for the US, we used Behavioral Risk Factor Surveillance System data (BRFSS). For the intended analysis, each survey is considered as pseudo-population representing the target population. Respondents who reported internet use are considered as respondent in a pseudo-web survey with a 100% response rate. If internet usage is unrelated to health status, then parameter differences between pseudo-populations (estimates for the full samples) and pseudo-web surveys (estimates for the subsample of internet users) should be unsystematic, small and insignificant.

Our contribution differs from previous studies in several ways. First, we compare 29 countries and not only one country. Second, these surveys are not restricted to certain regions within a country. Third, neither are our samples restricted to special subpopulations. Fourth, the results are based on data collected after 2010 and therefore more recent by a decade than most other publications. Fifth, we argue that health related bias in web surveys is due to a missing not at random (MNAR) process and therefore cannot be corrected by any weighting procedure.

Since hypothesis on MNAR cannot be tested directly with the available data (Graham, 2012), the plausibility of the argument is based on indirect evidence. The paper starts with an explanation of sampling in web surveys and the persisting undercoverage and nonresponse problems in section 2. In the following section 3 we will explain bias reduction by weighting procedures and their dependency on the missing data mechanism. This mechanism depends on health related variables, therefore we summarize previous research on internet usage and health. In section 4 we discuss the data sets and the design used for the study. Section 5 reports on the results before and after weighting. Section 6 concludes.

2 Sampling for Web Surveys

For web surveys of clearly delimited special populations (for example, company employees or students of a specific university) sampling frames may pre-exist or may be constructed with little effort (Bethlehem & Biffignandi, 2012; Couper, 2007). However, for the general population such sampling frames for web surveys do not exist (Couper, 2000, p. 467). Since design-based inferences about a target population are only valid for probability samples (Cassel, Särndal, & Hakanwretman, 1977), the extent and nature of undercoverage and nonresponse in web surveys mainly depends on the recruitment (Couper, 2007).

As no suitable sampling frame for general population web surveys exists, most web surveys are based either on individuals recruited via websites or online recruited webpanels (Yeager et al., 2011). More expensive are surveys where respondents are selected offline, for example by a RDD sample or an address based F2F survey. This approach may result in overall low response rates as shown by an example reported by Bandilla, Kaczmirek, Blohm, and Neubarth (2009): 11% of the respondents of an address based F2F survey answered a web based follow-up survey. Additionally, these offline recruited surveys usually omit persons not using the Internet.

Therefore, a few panel studies provide Internet access to previously offline sample members, for example the Dutch LISS Panel (Scherpenzeel, 2011). More often, those who are unable or unwilling to complete the survey online are given other survey mode options (Bethlehem & Biffignandi, 2012; Blom et al., 2016; Couper, 2000). Since these types of web surveys are more expensive and require more time for field work, online panels and self-recruited surveys are more common. To simplify the discussion, we will therefore use single-mode web surveys as the reference model. Furthermore, we will concentrate on noncoverage and nonresponse problems and refer to Bethlehem and Biffignandi (2012) for an overview on measurement errors in web surveys.

2.1 Undercoverage in Web Surveys

In the context of web surveys, undercoverage usually refers to whether or not the target population has Internet access. Although the proportion of households with Internet access has increased rapidly, there are still large differences even between industrialized countries (Chinn & Fairlie, 2007; Mohorko, de Leeuw, & Hox, 2013; Pick & Nishida, 2015). 74,4% of households in the USA had Internet access in 2013 (File & Ryan, 2014). In Europe between 57% (Bulgaria) and over 90% of the households (in Denmark, Luxembourg, the Netherlands, Sweden and the United Kingdom) have access to the Internet (Eurostat, 2015).[1]

However, the bias caused by undercoverage does not only depend on the proportion of those excluded from Internet access ($\frac{N_{NI}}{N}$), but also on the differences regarding the target variable Y between people with (I) and without Internet (NI):

$$B(\bar{y}_I) = \frac{N_{NI}}{N}(\bar{Y}_I - \bar{Y}_{NI}) \tag{1}$$

(Bethlehem & Biffignandi, 2012).

Data for the USA as well as for Europe suggest that those with Internet access differ from those without access in regard to socio-demographic characteristics such as education,

[1] Due to the different types of Internet access, the construction of survey questionnaires on Internet penetration is becoming increasingly difficult (Nylander, Lundquist, & Brännström, 2009).

income, age and gender (Chinn & Fairlie, 2007) as well as ethnicity (Hoffman, Novak, & Schlosser, 2001).

2.2 Nonresponse in Web Surveys

Regardless of the mode of data collection, survey response rates have been declining in all industrialized western countries (de Leeuw & de Heer, 2002) and nonresponse rates are widely considered as rising (Brick & Williams, 2013; Meyer, Mok, & Sullivan, 2015). In general, response rates for web surveys are even lower than those of other collection methods (Lozar Manfreda, Bosnjak, Berzelak, Haas, & Vehovar, 2008; Shih & Fan, 2008).[2]

Of course, the bias in estimates due to nonresponse will be higher with increasing proportions of nonrespondents and increasing correlations between the target variable and the causes for nonresponse.

Denoting the average probability of participation in the target population as $\bar{p}$, the covariance between the values of the variable of interest and the probability of participation as S_{pY}, the corresponding correlation as R_{pY} and S_p and S_Y the variance of the probability of participation and the target variable respectively, Bethlehem (2009) shows that the resulting bias can be estimated as

$$B(\bar{y}_R) = \tilde{Y} - \bar{Y} = \frac{S_{pY}}{\bar{p}} = \frac{R_{pY} S_p S_Y}{\bar{p}} \tag{2}$$

with

$$\tilde{Y} = \frac{1}{N} \sum_{k=1}^{N} \frac{p_k}{\bar{p}} Y_k \approx E(\bar{y}_r) \tag{3}$$

and the population mean $\bar{Y}$ (Bethlehem, 2009, p. 222). However, as most web surveys lack at least a clearly defined sampling frame, it is difficult to determine if eventually observed biased estimates are due to coverage or nonresponse errors (AAPOR, 2010, p. 26).

3 Bias Reduction by Weighting Procedures

Weighting techniques are frequently used to adjust for both undercoverage and nonresponse. The most basic form is the use of cell weights (poststratification) using known marginal distributions under the assumption of conditional independence. Academic research as well as Official Statistics often use calibration approaches, such as generalized regression estimation and raking ratio estimation (Särndal & Lundström, 2005). However, auxiliary variables producing homogeneous strata in regard to the target variable are rarely available (Bethlehem & Biffignandi, 2012, p. 292). Finally, variants of propensity weights (Rosenbaum & Rubin, 1983) have been adapted for web surveys (Lee, 2006; Taylor, 2000; Valliant & Dever, 2011). The propensity score – the conditional probability of participating – is most often estimated using both the web survey as well as a reference survey conducted in a traditional survey mode or an auxiliary frame information (Bethlehem & Biffignandi, 2012).

More often than not, these weighting strategies are used without making the underlying assumptions plausible. Since the success of bias reduction by weighting depends critically on the mechanism resulting in missing data, the classification of missing data generating mechanisms according to Rubin (1976) has to be reviewed briefly.

3.1 Missing Data Mechanisms

Let $\boldsymbol{Y}_i = (Y_{i1}, Y_{i1}, \ldots, Y_{ip})^T$ be a set of p variables of interest for respondents $i = 1, \ldots, n$. For each respondent $\boldsymbol{Y}_i$ can be partitioned in two parts, a part of observed data ($\boldsymbol{Y}_{i,\text{obs}}$) and a part that is missing ($\boldsymbol{Y}_{i,\text{mis}}$) (Carpenter & Kenward, 2013). Furthermore, $\boldsymbol{R}_i = (R_{i1}, R_{i1}, \ldots, R_{ip})^T$ is a set of binary variables, indicating if the data on variable p for case i is missing ($\boldsymbol{R}_{ip} = 0$) or observed ($\boldsymbol{R}_{ip} = 1$).

Then, a missing data mechanism is defined by the conditional distribution of $\boldsymbol{R}_i$ given $\boldsymbol{Y}_i$. If the missing of $\boldsymbol{Y}_i$ does not depend on $\boldsymbol{Y}_i$ we can write

$$Pr(\boldsymbol{R}_i|\boldsymbol{Y}_i) = Pr(\boldsymbol{R}_i) \tag{4}$$

and the missing data mechanism is called *Missing Completely at Random* (MCAR). In this case $\boldsymbol{Y}_i$ can be considered as a simple random sample from the population. For analysis under MCAR no adjustments are needed.

The second missing data mechanism is *Missing at Random* (MAR). Here missingness is independent from unobserved data $\boldsymbol{Y}_{i,\text{mis}}$, but dependent on the observed data $\boldsymbol{Y}_{i,\text{obs}}$. Therefore, MAR can be written as

$$Pr(\boldsymbol{R}_i|\boldsymbol{Y}_i) = Pr(\boldsymbol{R}_i|\boldsymbol{Y}_{i,\text{obs}}). \tag{5}$$

To be clear, MAR implies that the probability of observing a variable is dependent on its value (Carpenter & Kenward, 2013, p. 12).[3] But since the missingness depends only on the observed data $\boldsymbol{Y}_{i,\text{obs}}$ and not on the unobserved data $\boldsymbol{Y}_{i,\text{mis}}$, the unobserved data does not contain any information about the probability of a response (Longford, 2005, p. 31).

Under MAR, the unobserved data $\boldsymbol{Y}_{i,\text{mis}}$ is not the result of simple random sampling. However, using the observed data $\boldsymbol{Y}_{i,\text{obs}}$ the unobserved data can be estimated. Therefore, unbiased estimation under MAR requires the use of imputation techniques or weighting procedures.

The third form of missing data mechanism is *Missing Not at Random* (MNAR). Missing data is said to be MNAR if

[2] The calculation of nonresponse rates requires information on the sampling frame. However, these are not available for self-recruited or convenience sampling. Therefore, useful information on nonresponse rates for web surveys of the general population is rare.

[3] Please note that covariates can be included in Y_i.

the probability of a value being missing depends on the underlying missing value itself, even given the observed data (Carpenter & Kenward, 2013, p. 17):

$$Pr(\boldsymbol{R}_i|\boldsymbol{Y}_i) \neq Pr(\boldsymbol{R}_i|\boldsymbol{Y}_{i,\text{obs}}). \tag{6}$$

Analysis under MNAR is more difficult than under MCAR and MAR, because in order to compensate the missing information, the missing data mechanism has to be modeled explicitly. In practice, this can be awkward (Carpenter & Kenward, 2013, p. 17).

3.2 Internet and Health

In statistical discussion of survey nonresponse in the general population, health status is rarely mentioned as a MNAR generating mechanism. Differences between Internet users and non-users in health related variables have been mentioned in the literature before. The most comprehensive review of these studies has been given by Tourangeau, Conrad, and Couper (2013). Table 1 reports the characteristics of the studies mentioned by Tourangeau et al. (2013).[4]

All of these studies are covering only the US (or specific US states), are more than 10 years old and some of them report only on restricted age ranges. Since internet usage has increased in the last 10 years and differences in internet use between the US and Europe are not entirely unlikely, an additional study seemed to be appropriate.

However, research on Internet access using other modes suggests that a persons' health might affect nonresponse and undercoverage in web surveys. Based on data on respondents 50 years and older from the American Health and Retirement Study, Schonlau, van Soest, Kapteyn, and Couper (2009) reported: "The prevalence rates suggest that respondents with Internet access have lower prevalence of chronic diseases, fewer symptoms of mental health problems, and fewer limitations in their activities of daily living than the rest of the population aged 50 and older" (Schonlau et al., 2009, p. 309). Using the same survey, Couper, Kapteyn, Schonlau, and Winter (2007) confirmed significant group differences in regard to health variables even after controlling for sociodemographics. Adams and White (2008) compared a web survey with a CAPI survey and reported significant differences for health behavior related variables (such as obesity and physical activity) even after weighting the web survey with socio-demographic variables. After a mode choice survey in a panel study of African-American women, Russell et al. (2010) concluded: "Web responders were less likely to be current smokers, to have children, or to have a chronic disease" (Russell et al., 2010, p. 1288). For Britain, Erens et al. (2014) compared four different non-probability web surveys (using respondents aged 18-44) with large-scale non-Internet surveys as well as official data. The authors concluded that most of the web survey estimates differed significantly from the reference data even when using quotas on demographic variables.

Our contribution to the discussion: First, we cover the European Union by using data from the ESS as well as the USA by using the BRFSS. So we provide a vast comparison between 29 countries and not only one solitary country, like Yeager et al. (2011). Second, these surveys we use are considered as high quality large scale surveys covering the entire area of their countries. Therefore we are not restricted to certain regions as the studies by Schonlau et al. (2004), Dever, Rafferty, and Valliant (2008) and Lee, Brown, Grant, Belin, and Brick (2009), which are limited to Michigan and California. Third, neither are our samples restricted to certain subpopulations as for example the studies of Schonlau, van Soest, and Kapteyn (2007) and Schonlau et al. (2009), where the respondents are older than 40 and 55 years. So we can investigate the interdependence of age, health and internet-usage across all ages. Fourth, the results are based on more recent data (ESS 2010; BRFSS 2013). Given the information in table 1, the results we report for the USA are based on data which is about one decade newer compared to all studies in table 1.

4 Data and Method

All studies mentioned above are limited to the USA or the UK, focus on population subsets (special age groups, african-american women) and/or are based on panel studies. Therefore, an international comparative study of health differences within the general population not restricted to subsets is lacking up to now. Such a comparative study of potential undercoverage effects in different countries is much easier if a multinational survey designed for comparisons can be used. The European Social Survey (ESS, Schnaudt, Weinhardt, Fitzgerald, and Liebig, 2014) is such a survey. The ESS round 5 was conducted in 2010 as face-to-face survey in 28 European countries. To include the United States, we used CATI data from the Behavioral Risk Factor Surveillance System (BRFSS, 2013). The ESS contains data on about 55, 000, the BRFSS on about 492, 000 respondents.

We used the full survey of each of the 29 countries (EU+USA) as pseudo-population for the corresponding country. The estimates for the full samples serve as pseudo-population parameters to which the estimates from the subgroups of internet users (the pseudo-web surveys) are compared, similar to the approach of Dever et al. (2008) used the 2003 Michigan BRFSS. It should be mentioned that this design assumes unbiased full-sample estimates for the ESS and BRFSS and that 100% of internet users take part in the web survey. The latter assumption is unrealistic, since non-

[4] Tourangeau et al. (2013) reported two additional studies: Berrens et al. (2003) and Lee (2006). Since they do not contain health variables, these studies are not included here.

Table 1
Previous studies on health bias in web surveys (based on Tourangeau, Conrad, and Couper, 2013)

Article	Correction Method[a]	Data (PP vs. WS)[b]	Age	Area[c]	n	Health Items
Dever, Rafferty, and Valliant (2008)	GREG	BRFSS[d] 2003 (Full sample vs. Internet user subsample)	18+	MI	3445	25
Lee, Brown, Grant, Belin, and Brick (2009)	PSc + GREG	BRFSS[d] 2003 (Full sample vs. Internet user subsample)	18+	MI	3410	5
Schonlau et al. (2004)	PSc	RDD-Survey vs. Harris Interactive 2000	18+	CA	4089 + 8195	37
Schonlau, van Soest, and Kapteyn (2007)	PSc	RDD-Survey 2004 vs. American Life Panel 2003	40+	USA	516 + 1128	2
Schonlau, van Soest, Kapteyn, and Couper (2009)	PSc	HRS[e] 2002 (Full sample vs. Internet user subsample)	55+	USA	16698	33
Yeager et al. (2011)	Raking	NHIS[f] 2004 vs. Probability Sample, RDD 2004	18+	USA	966	4
	Raking	NHIS[f] 2004 vs. Probability Sample, Internet 2004	18+	USA	1175	4
	Raking	NHIS[f] 2004 vs. Nonprobability Sample, Internet 1 2004	18+	USA	1841	4
	Raking	NHIS[f] 2004 vs. Nonprobability Sample, Internet 2 2005	18+	USA	1101	4
	Raking	NHIS[f] 2004 vs. Nonprobability Sample, Internet 3 2004	18+	USA	1223	4
	Raking	NHIS[f] 2004 vs. Nonprobability Sample, Internet 4 2004	18+	USA	1103	4
	Raking	NHIS[f] 2004 vs. Nonprobability Sample, Internet 5 2004	18+	USA	1086	4
	Raking	NHIS[f] 2004 vs. Nonprobability Sample, Internet 6 2004	18+	USA	1112	4
	Raking	NHIS[f] 2004 vs. Nonprobability Sample, Internet 7 2004	18+	USA	1075	4

[a] PSc=Propensity scoring [b] PP=Pseudo-population; WS=Websurvey [c] MI=Michigan; CA=California
[d] Behavioral Risk Factor Surveillance System [e] Health and Retirement Study [f] National Health Interview Survey

response will occur in practice. Most likely, the results reported here are the lower bound of differences between users and non-users.

4.1 Measures

In the ESS, *general subjective health* is measured with the question

> *How is your health in general? Would you say it is... (1) very good, (2) good, (3) fair, (4) bad, or, (5) very bad?*

A similar item is used in the BRFSS:

> *Would you say that in general your health is: (1) Excellent, (2) Very good, (3) Good, (4) Fair, (5) Poor.*

Both surveys also asked about Internet use. The ESS asked

> *Now, using this card, how often do you use the Internet, the World Wide Web or e-mail - whether at home or at work - for your personal use? (0) No access at home or work, (1) Never use, (2) Less than once a month, (3) Once a month, (4) Several times a month, (5) Once a week, (6) Several times a week, (7) Every day.*

In the BRFSS this was asked using the question

> *Have you used the Internet in the past 30 days?*
> *(1) Yes (2) No.*

Although different questions were used, they both seem to at least allow a separation of Internet users and non-users. For analysis, the ESS answers *(0)* and *(1)* were considered as indicating a non-user and in the BRFSS the answer *(2)*.

4.2 Analysis Method and Weighting

Internet use is strongly related to age in many countries. Naturally, age is related to health. Therefore, age has to be controlled for in the analysis. To explore the different non-linear relationships of age, Internet usage and health between countries, a separate nonparametric regression (Loess, see Cleveland, 1979) for each country seems to be appropriate.

Not considering nonresponse and noncoverage, the estimator for respondents (r)

$$\hat{Y} = \sum_r d_k y_k \tag{7}$$

with design weights $d_k = 1/\pi_k$ is an appropriate estimator for the total of the target population (U)

$$Y = \sum_U y_k. \tag{8}$$

Nonresponse and noncoverage might require corrections of the design weights d_k. One approach is the calibration estimator

$$\hat{Y}_W = \sum_r w_k y_k. \tag{9}$$

The weights w_k are said to be calibrated to the information input $\boldsymbol{X}$ if they satisfy the so-called calibration equation

$$\sum_r w_k \boldsymbol{x}_k = \boldsymbol{X} \tag{10}$$

where $\boldsymbol{x}_k$ is a vector of auxiliary variables (Särndal & Lundström, 2005, pp. 57-58). The calibrated weights w_k themselves are a product of the initial weights d_k and a correction factor v_k: $w_k = d_k v_k$. To obtain the correction factor v_k Särndal and Lundström (2005, p. 58) suggest the form

$$v_k = 1 + \boldsymbol{\lambda}' \boldsymbol{x}_k \tag{11}$$

which leads to

$$\boldsymbol{\lambda}'_r = \left(\boldsymbol{X} - \sum_r d_k \boldsymbol{x}_k \right)' \left(\sum_r d_k \boldsymbol{x}_k \boldsymbol{x}'_k \right)^{-1} \tag{12}$$

when (11) gets substituted into formula (10) and solved for $\boldsymbol{\lambda}'$ (assuming that $\left(\sum_r d_k \boldsymbol{x}_k \boldsymbol{x}'_k\right)$ is invertible). The resulting calibrated weight is

$$w_k = d_k + d_k \boldsymbol{\lambda}'_r \boldsymbol{x}_k. \tag{13}$$

In general, most weighting procedures of web surveys are based on basic demographic variables such as age and gender (see table 1). Schonlau et al. (2007) tested the use of attitudinal and behavioral variables in addition to demographic variables. However, weighting a survey by an unreliable auxiliary such as an attitude (“do you often feel alone”, Schonlau et al., 2007) or volatile conditions (“had a sunburn in the past 12 months”, Lee et al., 2009) will increase sampling variance. Furthermore, these kind of variables are rarely available in practice. To the best of our knowledge, this approach has rarely been used.

In this study we restrict ourselves to calibration on standard demographic variables: age (14–24, 25–34, 3-5-44, 45–54, 55–64, 65+), gender, ethnic background, urban residence, education (ISCED 1–2, 3–4, 5–6) and household income (quintiles).[5] Due to item nonresponse, household income was imputed by stochastic regression imputation using education, living with partner, household size, employment and retirement status. Mean R^2 over all countries is 0.37.[6]

5 Results

Figures 1–4 show the resulting estimated nonparametric regressions of subjective health depending on age, grouped by Internet usage, for each of the 28 countries considered (for the size of the samples, see table A2). In each plot, the solid line represents the regression estimate for Internet users and the dashed line the regression estimate for non-users. Each regression line is enclosed by its 95%-confidence band.[7]

The subjective health reported by the respondents is clearly worse for non-users of the Internet in 28 of 29 countries. The irregular pattern in Sweden may be due to sampling errors as indicated by the wide confidence band in the plot.

Furthermore, in 2/3 of the countries the worse health status of non-internet users can clearly be observed across all ages (ungrouped data). If crossing of the loess-lines is used as criterion, only Cyprus, Germany, Greece, Israel, Netherlands, Poland, Spain, Sweden, Ukraine and United Kingdom ($10/29 \approx 0.34$) may be considered as exceptions.

[5] Additionally, we estimated CART-Models to identify potential interaction effects. The inclusion of interaction effects did not improve the models. For the sake of simplicity, we report only the models containing main effects.

[6] For the computation of the weights, we scaled the weights so that $\sum w_k = n$. We followed the ESS policy to limit the weights. We used $w_k \leq 10$.

[7] The confidence bands were computed with ggplot2 (Wickham, 2016). This is a simple bootstrap, not taking the sampling design into consideration. The large number of observations (491,773 records) in the BRFSS 2013 are computationally challenging for confidence bands of nonparametric regressions. Therefore, the plot for the BRFSS is based on a random 5% subsample.

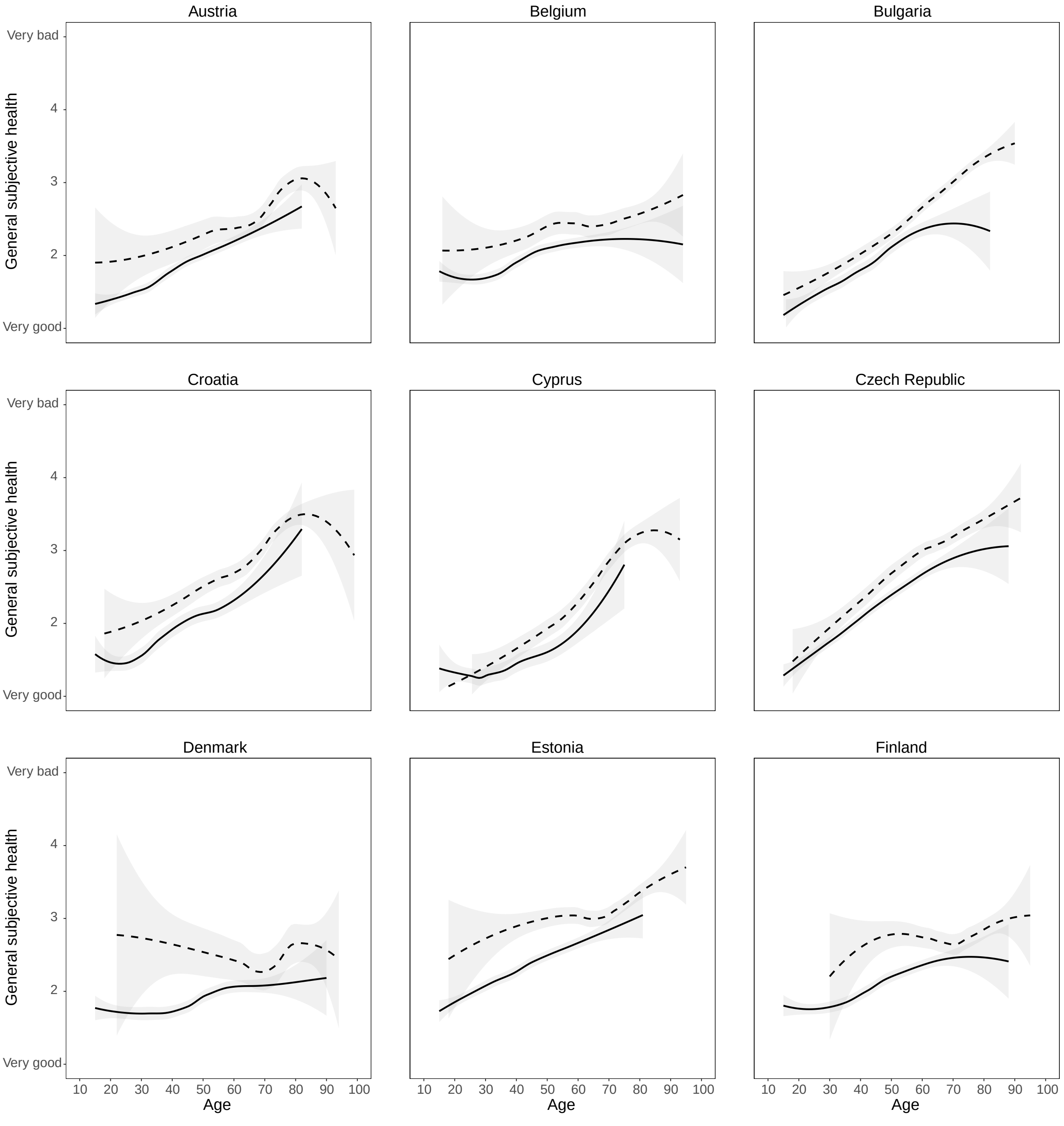

Figure 1. Nonparametric regression (Lowess, bandwith 0.66, 95% confidence bands): age vs. health by Internet use (solid line: Internet usage, dashed line: no Internet usage), ESS round 5: Austria – Finland

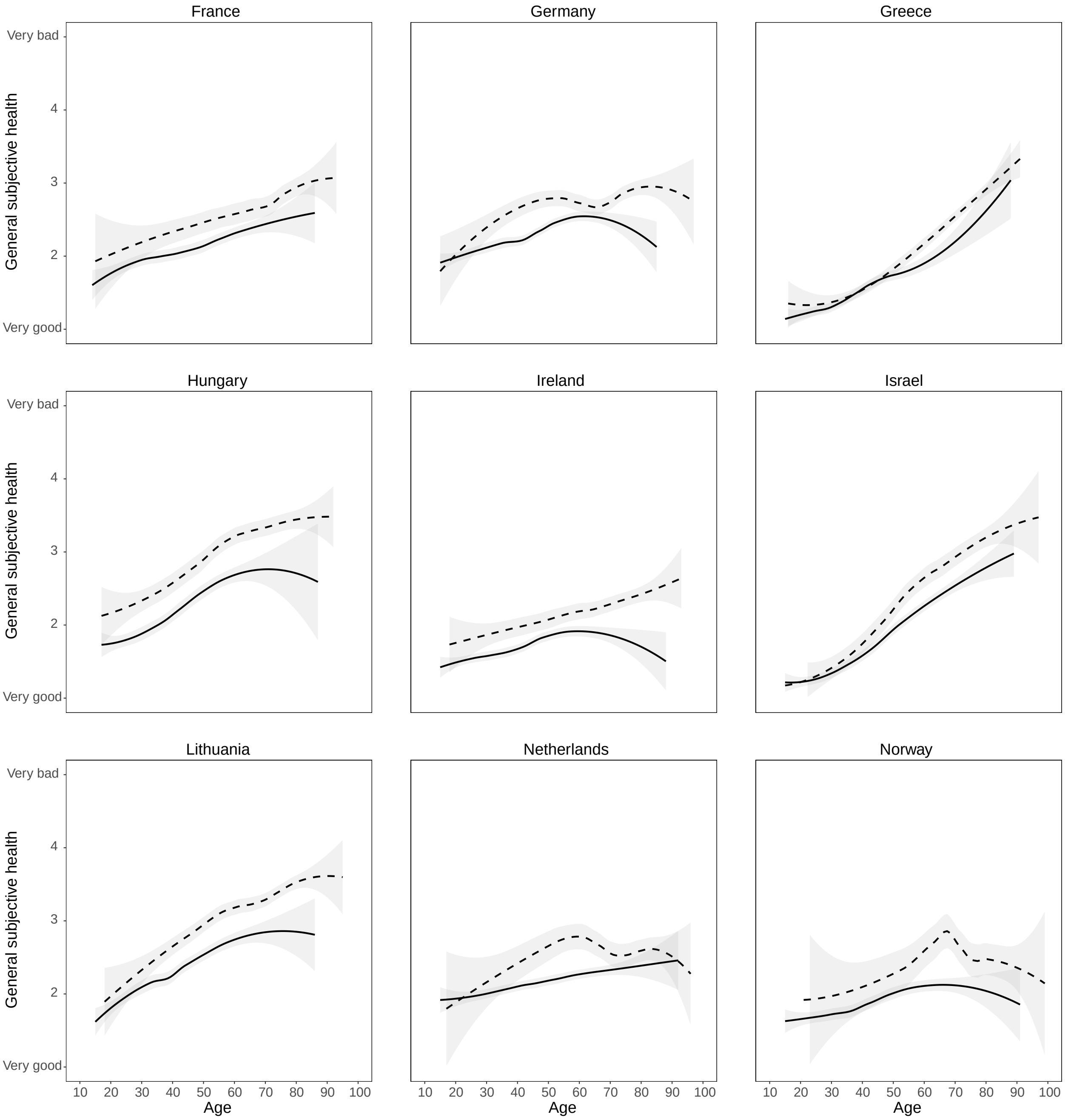

Figure 2. Nonparametric regression (Lowess, bandwidth 0.66, 95% confidence bands): age vs. health by Internet use (solid line: Internet usage, dashed line: no Internet usage), ESS round 5, France – Norway

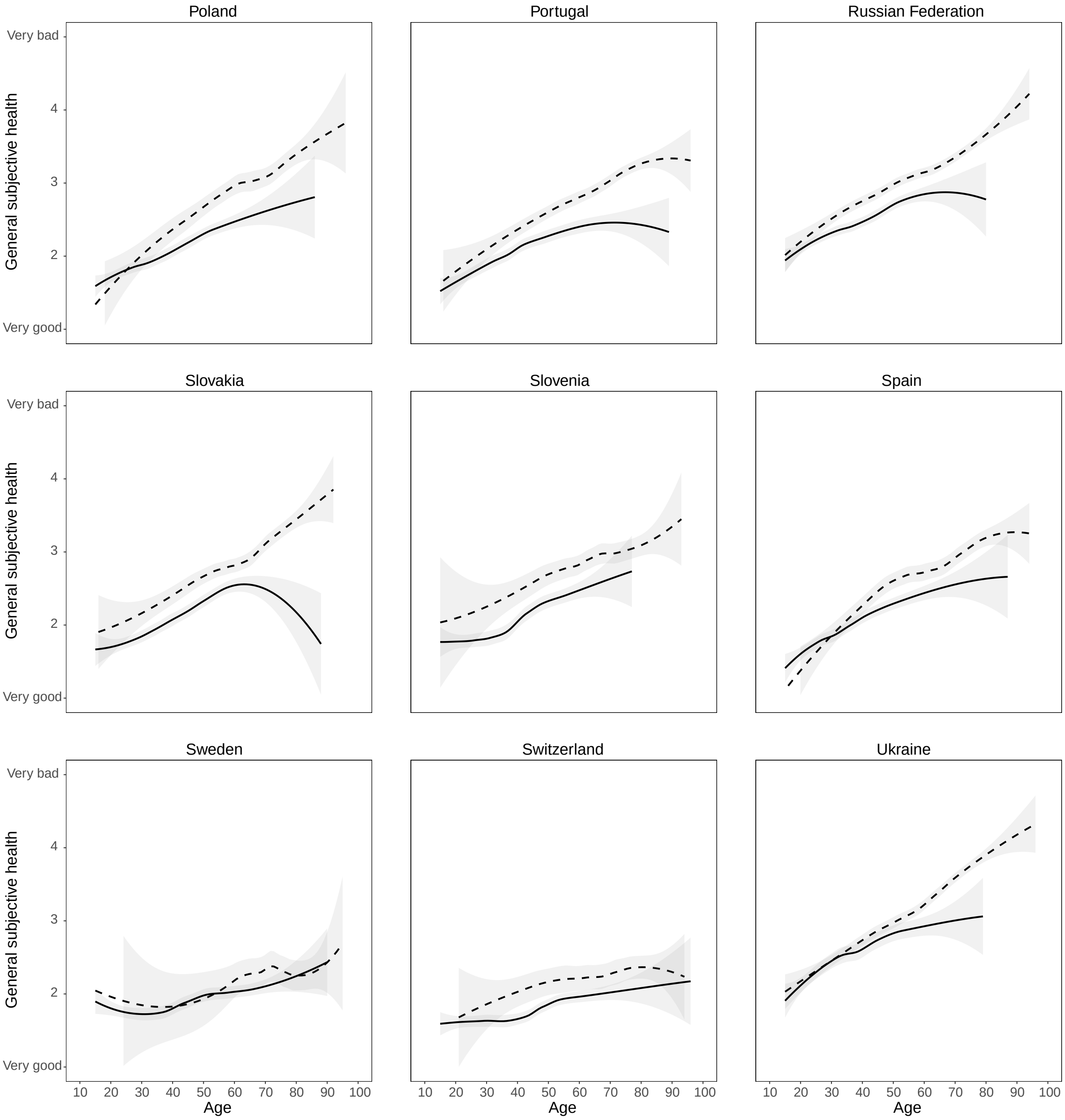

Figure 3. Nonparametric regression (Lowess, bandwidth 0.66, 95% confidence bands): age vs. health by Internet use (solid line: Internet usage, dashed line: no Internet usage), ESS round 5, Poland – Ukraine

Finally, the plots show increasing differences in reported health between Internet users and non-users with increasing age for the majority of the countries ($16/29 \approx 0.55$). However, given the problems of small numbers of observations in the subgroups, possible different nonresponse bias in different countries and the problems of bootstrapping confidence bands for non-parametric regressions (Givens & Hoeting, 2013), the results concerning the exceptions should be considered with caution. Therefore, even though deviating patterns can be observed, the general tendency seems to be the same in all countries.

To check for additional covariates, different multilevel mixed-effects linear regression models (Snijders & Bosker, 2012) for general health using the ESS data were fitted. Independent variables were Internet usage, age, years of full-time education, household's total net income and gender. The weighted models are weighted on both levels, the respondent level as well as the country-level.

Table 2 shows the estimates. With regard to the R^2-values, all three models show acceptable fits given the small number of independent variables. The most important result in the table is the persistence of the supposed effect: People who use the Internet tend to be more healthy than people who do not use the Internet, even after controlling for age, income, gender and education. It should be noted that linear, quadratic and cubic effects of age and an interaction effect of age with Internet use were included in all models. Since the interaction effect of age and Internet use is significant in all models, it can be stated that older respondents who do use the Internet are significantly more healthy than older persons who do not use the Internet. Therefore, participating older respondents are not a random sample from all old people. Thus, they do not represent all old people, but only the healthy ones. Since this effect is significant in all three models, this result is independent from weighting the data.

However, as can be seen in Figures 1–3, health differences between Internet users and non-users seem to be present in nearly all countries, independent of age. The multi-level regressions in Table 2 show that this effect is significant despite controlling for demographics.[8]

For practical applications, the size of differences is more import than their statistically significance. Therefore, for the evaluation of the supposed selection effect in web surveys, the differences in health should be quantified with a measure of effect size such as the widely used Cohen's D (Cohen, 1988, p. 67). A version corrected for unequal group sizes (Rosnow, Rosenthal, & Rubin, 2000, pp. 448-449) is given by

$$D = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2)}{n_1+n_2-2}}} \sqrt{\frac{\frac{n_1+n_2}{2}}{\frac{2n_1n_2}{n_1+n_2}}}. \tag{14}$$

Although disputed, D values greater than 0.5 are considered by Cohen as 'medium' effect, values greater than 0.8 as 'large' (Ellis, 2010). For the computations, means and standard deviations were based on proper survey design adjusted estimates.

The estimates of D for the differences in reported health between Internet users and non-internet users for each country are shown in Table 3.

In the table, D ranges roughly between 0.63 for Sweden and 1.26 for Estonia. The average D is about 0.91. Given the classification of effect sizes described, 'large' effect sizes are observed for 20 of 28 countries, the remaining effect sizes are 'medium'. The supposed difference in subjective health between Internet users and non-users seems to be considerable.

To quantify the possible impact of this difference between reported users and non-users on estimates based on a web survey, we used the standardized bias, SB, defined as

$$\text{SB} = 100 * \frac{\bar{x} - \mu}{\hat{\sigma}_{\bar{x}}} \tag{15}$$

where $\bar{x}$ is the sample estimate, μ the population value and $\hat{\sigma}_{\bar{x}}$ the estimated standard error (Collins, Schafer, & Kam, 2001, p. 340; Graham, 2012, p. 19). SB is therefore the difference between the estimate and the population value in standard error units. For example, SB = 50 would indicate a difference between the estimate and the population value of half a standard error. Since the standard error of the estimate is directly affected by n, larger surveys yield larger values of SB given the same difference. Absolute values of SB larger than 40 are regarded by Collins et al. (2001, p. 340) and Graham (2012, p. 19) as of practical importance.

Table 4 shows the standardized bias between the subsample of Internet users and the total sample in the ESS and the BRFSS (USA) for 28 countries. The smallest standardized biases are observed for Sweden and the Netherlands, but even here SBs of −195 and −257 indicate differences with practical importance. The obvious outlier USA is due to the large number of observations in the BRFSS: If this survey would have had the medium number of observations of the ESS (n=2000), SB would be about −679 (between Belgium and France). In general, the standardized bias seem to decrease with increasing gross domestic product per capita (Pearson r between SB and GDP for 2010: −0.64).[9]

To sum up the results so far, the differences in general health between Internet users and non-users in all countries examined here are statistical significant, have medium to large effect sizes and seem to be of practical importance for

[8] The reduction in sample size is due to missing poststratification weights (Austria) or missing household income due to the fact that the question was not part of the questionnaire (Portugal) and item nonresponse. Household-income for the remaining countries was imputed as described in section 4.2.

[9] USA omitted. GDP data taken from United Nations, http://hdr.undp.org/en/data

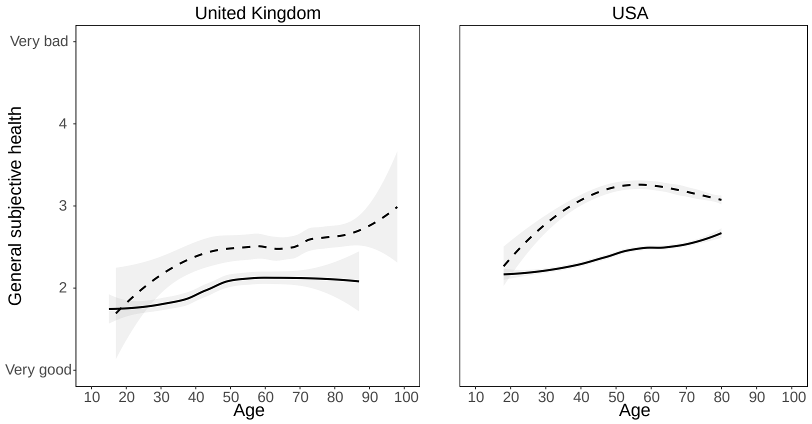

Figure 4. Nonparametric regression (Lowess, bandwidth 0.66; 95% confidence bands): age vs. health by Internet use (solid line: Internet usage, dashed line: no Internet usage), ESS round 5: United Kingdom; BRFSS 2013: USA

Table 2
General subjective health (ESS, round 5), Linear Multilevel Model[a]

	Model 1[b]		Model 2[c]		Model 3[d]	
Variable	Coef.	Std. Err.	Coef.	Std. Err.	Coef.	Std. Err.
Age	0.295**	0.023	0.296**	0.023	0.298**	0.023
Age^2	−0.070**	0.010	−0.074**	0.008	−0.072**	0.009
Age^3	0.008	0.006	0.004	0.007	0.007	0.006
Netuse	−0.027**	0.005	−0.027**	0.006	−0.027**	0.006
Age × Netuse	−0.071**	0.010	−0.069**	0.008	−0.069**	0.006
Years of Education	−0.018**	0.003	−0.018**	0.003	−0.018**	0.003
HH-income (Quincile)	−0.077**	0.007	−0.073**	0.006	−0.071**	0.006
Female	0.085**	0.024	0.076**	0.025	0.082**	0.023
Constant	2.949**	0.061	2.939**	0.073	2.938**	0.077
σ_u^2	.079		.077		.073	
σ_e^2	.607		.605		.598	
ρ	.115		.112		.109	
$R^2_{\text{Level 1}}$	.236		.219		.236	
$R^2_{\text{Level 2}}$	.168		.031		.020	
R^2_{Overall}	.229		.202		.217	
n	49321		49321		49321	

Age and Internet use z-transformed

[a] Level 1: Individuals, Level 2: Countries [b] Not weighted at level 1, population weight at level 2

[c] Design weight at level 1, population weight at level 2

[d] Poststratification weight at level 1, population weight at level 2

* $p < 0.05$, ** $p < 0.01$

the estimation of general population health using Internet users only.

It should be noted that the estimates shown are based on the initial weights of the surveys, no additional weighting has been applied. However, one might argue that this bias will be diminished by weighting procedures. Therefore, starting with the initial poststratification weights for the group of Internet users, the weights were re-calibrated with the Stata macro “calibrate” (D’Souza, 2010) to the population totals estimated using the full sample (Internet users and non-users) for sex, age and years of education. Estimated Cohen’s D and standardized biases before and after this calibration for each of the 28 countries considered are shown in Table 5.

Before calibration, the average of Cohen’s D was 0.91, after calibration 0.73. Therefore, calibration reduced the effect from a ’large’ effect to a ’medium’ effect. However,

Table 3
Cohen's D (equation 14) for general subjective health for 28 countries, groups for Cohen's D by Internet usage (yes/no)

Country	Cohen's D
Estonia	1.255
Austria	1.186
Lithuania	1.177
Poland	1.125
Norway	1.079
Hungary	1.046
Czech Republic	1.045
Finland	1.027
Croatia	0.991
Slovakia	0.964
Denmark	0.960
Bulgaria	0.957
Cyprus	0.940
Slovenia	0.933
Russian Federation	0.907
Israel	0.875
Netherlands	0.858
USA	0.843
Spain	0.828
Belgium	0.814
Greece	0.780
Ukraine	0.779
Switzerland	0.718
France	0.716
Germany	0.684
United Kingdom	0.675
Ireland	0.668
Sweden	0.626
Mean (D)	0.909
Std. Dev. (D)	0.171

Table 4
Standardized bias (equation 15) between Internet users and the full sample for general subjective health in 28 countries, weighted using initial ESS weights

Country	Standard. Bias
USA	-4501
Greece	-1752
Bulgaria	-1560
Russian Federation	-1458
Poland	-1451
Hungary	-1444
Czech Republic	-1239
Croatia	-1210
Cyprus	-1108
Spain	-1083
Israel	-1054
Estonia	-1054
Slovakia	-1037
Lithuania	-1004
Slovenia	-972
Ukraine	-893
Germany	-777
Finland	-723
Belgium	-715
France	-620
Ireland	-605
Austria	-556
Switzerland	-510
United Kingdom	-499
Denmark	-353
Norway	-316
Netherlands	-257
Sweden	-195
Mean (SB)	-1034
Std. Dev. (SB)	797

7 countries (Austria, Denmark, Estonia, Finland, Lithuania, Norway and Poland) still show 'large' effects with D>0.8. Overall, the effect size diminished from 'large' ($D \geq 0.8$) to 'medium' ($0.5 \leq D < 0.8$) for 20 out of 28 countries. But not one single country shows an effect less than 'medium' after calibration. To put it in other words: Calibration failed in all countries to reduce the magnitude of the difference between internet users and nonusers to a negligible or at least 'small' effect ($D < 0.2$). The effect of calibration on standardized bias is also remarkable: The average SB before calibration is about -1034, after calibration about -341. If the USA is considered as outlier and omitted from the analysis, the average SB before calibration is about -905, after calibration about -306. Although the amount of bias reduction by calibration is impressive, the remaining bias is not even close to zero. An average bias of 250 is equivalent to a bias of the size of 2.5 standard errors. Even after calibration, 24 of 28 countries have $|SB| > 100$, 19 of 28 countries have $|SB| > 200$. The remaining differences after calibration are still worrisome and can hardly be ignored.

It is worth noting that the reduction of the difference between Internet users and non-users is negatively correlated with GDP per capita (-0.68 for Cohen's D and -0.64 for SB without USA). In countries with high GDP per capita, large health differences remain after calibration. The calibration variables compensate for gender, educational and age differences regarding Internet access, but not for additional variables affecting differential Internet access caused by health related variables. This limited success of calibration can be attributed to the proposed MNAR response process for web

Table 5
Cohen's D and Standardized Bias before calibration and after calibration for 28 countries

Country	Cohen's D before Calibration	Cohen's D after Calibration	Std. Bias before Calibration	Std. Bias after Calibration
Austria	1.186	1.023	−556	−158
Belgium	0.814	0.693	−715	−302
Bulgaria	0.957	0.756	−1560	−749
Croatia	0.991	0.709	−1210	−314
Cyprus	0.940	0.702	−1108	−376
Czech Republic	1.045	0.766	−1239	−283
Denmark	0.960	0.857	−353	−138
Estonia	1.255	1.050	−1054	−375
Finland	1.027	0.802	−723	−95
France	0.716	0.556	−620	−164
Germany	0.684	0.575	−777	−300
Greece	0.780	0.537	−1752	−497
Hungary	1.046	0.767	−1444	−390
Ireland	0.668	0.579	−605	−286
Israel	0.875	0.561	−1054	−72
Lithuania	1.177	0.962	−1004	−463
Netherlands	0.858	0.781	−257	−92
Norway	1.079	0.999	−316	−146
Poland	1.125	0.857	−1451	−392
Russian Federation	0.907	0.726	−1458	−740
Slovakia	0.964	0.745	−1037	−367
Slovenia	0.933	0.743	−972	−292
Spain	0.828	0.615	−1083	−288
Sweden	0.626	0.538	−195	−15
Switzerland	0.718	0.604	−510	−179
USA	0.843	0.710	−4501	−1288
Ukraine	0.779	0.583	−893	−407
United Kingdom	0.675	0.627	−499	−387
Mean	0.909	0.729	−1034	−341

surveys with regard to health.

6 Discussion

Health estimates of Internet users and Internet non-users reported here show that people who are less healthy tend to use the Internet less frequently than healthy people. This result was observed in all 28 European ESS-countries as well as the United States of America. After controlling for age, the differences in health remained for most countries. These observed health differences between Internet users and non-users are of interest by themselves.

However, if the subset of the Internet users in the ESS and BRFSS samples can be seen as a random sample of the frame population of a web survey, the observed health differences are relevant for web surveys in general. In this case, web surveys will cause biased estimates of health related variables.

To estimate the bias of a surveys, two principal designs are common (Bound, Brown, & Mathiowetz, 2001, p. 3741):

1. survey micro data is compared with external micro data for each survey respondent or
2. external population parameters are compared with survey estimates.

Regarding the supposed bias of web surveys as suggested here, both approaches would be helpful. However, given the research obstacles imposed by the European data protection jurisdiction, obtaining micro data for a population is difficult at best. For example, using administrative data of survey nonrespondents requires special authorizations in most European countries. Therefore, comparing individual micro data of surveys with administrative data is hardly an option in comparative research. In contrast, the second design has no data protection problem. However, for a comparative study, general population health survey estimates are needed for each country, to be considered. Hence, the use of surveys with different designs and fieldwork details seems to be unavoidable. Using different surveys has to take the different elements of the *Total Survey Error* (Weisberg, 2005) into account. Doing that for independent surveys for many countries in a way that is methodologically sound will require a study of its own. Finally, to the best of our knowledge, no European multi-country web survey with a common design containing health indicators is currently available. Therefore, different web surveys would have to be compared to different health surveys in other modes. This will seriously increase the number of methodological problems for such a comparative design.

The approach taken in this paper is different from the two designs mentioned above. The results are based on subgroup differences within a face-to-face survey. The results here assume 100% nonresponse of non-internet users in a web survey, 100% response of Internet users in a web survey and unbiased population estimates of the ESS. With regard to the supposed bias of web surveys, the assumption of unbiased ESS estimates can be regarded as uncritical. Albeit 100% nonresponse of non-internet users in a web survey is not likely, the percentage will be higher than 65%, since nonresponse in web surveys for Internet users in general seems to be on average above 65% (Shih & Fan, 2008) and we see no reason to expect lower nonresponse rates for non-users. Therefore, the only plausible mechanism which might yield lower bias in a web survey than reported here is differential nonresponse bias. This is a nonresponse mechanism which causes different signs of bias in different subgroups. Although it is mathematically possible that differential nonresponse in a web survey can reduce its noncoverage error, it seems unlikely that in a web survey refusal is more likely for healthy than for unhealthy respondents.

Therefore, we consider the results reported here to be plausible. Of course, confirmation by studies using the two designs above is needed. Although single item measures of general health have been validated by several studies (for a review, see McDowell, 2006, p. 583), it could be argued that the reported differences concern reported general subjective health, but not objective health. To refute this objection, the analysis reported here should be repeated with objective measurements. Since neither micro-data with objective health measurements nor reported objective health indicators and Internet usage seems to be available for a multinational comparison, currently such an analysis is limited to a few countries. Using the BRFSS for the USA and the ALLBUS for Germany, comparable effects as reported here could be found for nearly all objective indicators (Schnell, Noack, & Torregroza, 2015). Therefore, the differences in health between Internet users and non-users do not seem to be limited to subjective indicators.

The most alarming finding is the fact that calibration does not eliminate the health differences. This may be due to unsuited weighting variables, but we used the standard variables usually available in survey research. Therefore, we consider the underlying missing data mechanism as an example of MNAR. If this holds true, the usual weighting techniques could not be used to correct for this bias, because the fundamental MAR-assumption shared by all these techniques would be violated. This would be a serious limitation of web surveys concerning health related variables.

Author contributions

The idea of the study was due to RS. He suggested the datasets, directed the data analysis and wrote the final version. ST wrote the initial draft. MN did the data analysis and contributed to the text. All authors approved the final version.

References

AAPOR. (2010). AAPOR report on online panels. *Public Opinion Quarterly*, *74*(4), 711–781.

Adams, J. & White, M. (2008). Health behaviours in people who respond to a web-based survey advertised on regional news media. *The European Journal of Public Health*, *18*(3), 335–338.

Bandilla, W., Kaczmirek, L., Blohm, M., & Neubarth, W. (2009). Coverage- and Nonresponse-effekte bei Online-Bevölkerungsumfragen. In N. Jackob, H. Schoen, & T. Zerback (Eds.), *Sozialforschung im Internet* (pp. 129–144). Wiesbaden: VS Verlag.

Batterham, P. J. (2014). Recruitment of mental health survey participants using internet advertising: content, characteristics and cost effectiveness. *International Journal of Methods in Psychiatric Research*, *23*(2), 184–191.

Berrens, R. P., Bohara, A. K., Jenkins-Smith, H., Silva, C., & Weimer, D. L. (2003). The advent of internet surveys for political research: a comparison of telephone and internet samples. *Political Analysis*, *11*(1), 1–22.

Bethlehem, J. (2009). *Applied survey methods: a statistical perspective*. Hoboken: Wiley.

Bethlehem, J. & Biffignandi, S. (2012). *Handbook of web surveys*. Hoboken: Wiley.

Blom, A. G., Bosnjak, M., Cornilleau, A., Cousteaux, A.-S., Das, M., Douhou, S., & Krieger, U. (2016). A comparison of four probability-based Online and mixed-mode panels in Europe. *Social Science Computer Review*, *34*(1), 8–25.

Bound, J., Brown, C., & Mathiowetz, N. (2001). Measurement error in survey data. In J. J. Heckman & E. Leamer (Eds.), *Handbook of econometrics, volume 5* (pp. 3705–3843). Amsterdam: Elsevier Science.

Brick, J. M. & Williams, D. (2013). Explaining rising nonresponse rates in cross-sectional surveys. *The ANNALS of the American Academy of Political and Social Science*, *645*, 36–59.

Carpenter, J. R. & Kenward, M. G. (2013). *Multiple imputation and its application*. Chichester: Wiley.

Cassel, C.-M., Särndal, C.-E., & Hakanwretman, J. (1977). *Foundations of inference in survey sampling*. New York: Wiley.

Chinn, M. D. & Fairlie, R. W. (2007). The determinants of the global digital divide: a cross-country analysis of computer and internet penetration. *Oxford Economic Papers*, *59*(1), 16–44.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, *74*(368), 829–836.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2th). Hillsdale: Erlbaum.

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*(4), 330–351.

Couper, M. P. (2000). Web surveys: a review of issues and approaches. *Public Opinion Quarterly*, *64*(4), 464–494.

Couper, M. P. (2007). Issues of representation in eHealth research: with a focus on web surveys. *American Journal of Preventive Medicine*, *32*(5, Supplement), S83–S89.

Couper, M. P., Kapteyn, A., Schonlau, M., & Winter, J. (2007). Noncoverage and nonresponse in an internet survey. *Social Science Research*, *36*(1), 131–148.

Czaja, R. & Blair, J. (2005). *Designing surveys: a guide to decisions and procedures*. Thousand Oaks: Pine Forge Press.

de Leeuw, E. & de Heer, W. (2002). Trends in household survey nonresponse: a longitudinal and international comparison. In R. M. Groves, D. A. Dillman, J. L. Eltinge, & R. J. Little (Eds.), *Survey nonresponse* (pp. 41–54). New York: John Wiley & Sons.

Dever, J. A., Rafferty, A., & Valliant, R. (2008). Internet surveys: can statistical adjustments eliminate coverage bias? *Survey Research Methods*, *2*(2), 47–62.

D'Souza, J. (2010). *Calibrate: stata module to calibrate survey datasets to population totals*. Stata Users Group.

Ellis, P. D. (2010). *The essential guide to effect sizes: statistical power, meta-analysis, and the interpretation of research results*. Cambridge: Cambridge University Press.

Erens, B., Burkill, S., Couper, M. P., Conrad, F., Clifton, S., Tanton, C., . . . Copas, A. J. (2014). Nonprobability web surveys to measure sexual behaviors and attitudes in the general population: a comparison with a probability sample interview survey. *Journal of Medical Internet Research*, *16*(12), e276.

ESOMAR. (2014). *Global Market Research 2014*. Amsterdam: ESOMAR.

Eurostat. (2015). Information and communication technologies: level of internet access - households data for 2014. http://ec.europa.eu/eurostat/cache/metadata/en/isoc_bde15c_esms.htm.

File, T. & Ryan, C. (2014). *Computer and internet use in the United States: 2013 – American Community Survey reports*. Economics and Statistics Administration: United States Census Bureau.

Givens, G. H. & Hoeting, J. A. (2013). *Computational statistics* (2th). Hoboken: Wiley.

Graham, J. W. (2012). *Missing data - analysis and design*. New York: Springer.

Hoffman, D. L., Novak, T. P., & Schlosser, A. E. (2001). The evolution of the digital divide: examining the relationship of race to internet access and usage over time. In B. M. Compaine (Ed.), *The digital divide: facing crisis or creating a myth* (pp. 47–97). Cambridge, USA: MIT Press.

Kalton, G. & Flores-Cervantes, I. (2003). Weighting methods. *Journal of Official Statistics*, *19*(2), 81–97.

Lee, S. (2006). Propensity score adjustment as a weighting scheme for volunteer panel web surveys. *Journal of Official Statistics*, *22*(2), 329–349.

Lee, S., Brown, E. R., Grant, D., Belin, T. R., & Brick, J. M. (2009). Exploring nonresponse bias in a health survey using neighborhood characteristics. *American Journal of Public Health*, *99*(10), 1811–1817.

Liu, H., Cella, D., Gershon, R., Shen, J., Morales, L. S., Riley, W., & Hays, R. D. (2010). Representativeness of the patient-reported outcomes measurement information system internet panel. *Journal of Clinical Epidemiology*, *63*(11), 1169–1178.

Longford, N. T. (2005). *Missing data and small-area estimation - modern analytical equipment for the survey statistician*. New York: Springer.

Lozar Manfreda, K., Bosnjak, M., Berzelak, J., Haas, I., & Vehovar, V. (2008). Web surveys versus other survey modes: a meta-analysis comparing response rates. *International Journal of Market Research*, *50*(1), 79–104.

McDowell, I. (2006). *Measuring health: a guide to rating scales and questionnaires* (3th). Oxford: Oxford University Press.

Meyer, B. D., Mok, W. K. C., & Sullivan, J. X. (2015). Household surveys in crisis. *Journal of Economic Perspectives*, *29*(4), 199–226.

Mohorko, A., de Leeuw, E., & Hox, J. (2013). Internet coverage and coverage bias in europe: developments across countries and over time. *Journal of Official Statistics*, *29*(4), 609–622.

Nylander, S., Lundquist, T., & Brännström, A. (2009). At home and with computer access: why and where people use cell phones to access the internet. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 1639–1642). CHI '09. Boston, MA, USA: ACM.

Pick, J. B. & Nishida, T. (2015). Digital divides in the world and its regions: a spatial and multivariate analysis of technological utilization. *Technological Forecasting and Social Change*, *91*, 1–17.

Rosenbaum, P. & Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55.

Rosnow, R. L., Rosenthal, R., & Rubin, D. (2000). Contrasts and correlations in effect-size estimation. *Psychological Science*, *11*(6), 446–453.

Rubin, D. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592.

Russell, C. W., Boggs, D. A., Palmer, J. R., & Rosenberg, L. (2010). Use of a web-based questionnaire in the black women's health study. *American Journal of Epidemiology*, *172*(11), 1286–1291.

Särndal, C.-E. & Lundström, S. (2005). *Estimation in surveys with nonresponse*. Chichester: Wiley.

Scherpenzeel, A. (2011). Data collection in a probability-based internet panel: how the LISS panel was built and how it can be used. *BMS: Bulletin of Sociological Methodology*, *109*(1), 56–61.

Schnaudt, C., Weinhardt, M., Fitzgerald, R., & Liebig, S. (2014). The European Social Survey: contents, design, and research potential. *Schmollers Jahrbuch*, *134*(4), 487–506.

Schnell, R. (1993). Die Homogenität sozialer Kategorien als Voraussetzung für 'Repräsentativität' und Gewichtungsverfahren. *Zeitschrift für Soziologie*, *22*(1), 16–32.

Schnell, R., Noack, M., & Torregroza, S. (2015, December). *Disease differences between internet users and nonusers in the USA and Germany*. Unpublished discussion paper, University of Duisburg-Essen.

Schonlau, M., van Soest, A., & Kapteyn, A. (2007). Are "'webographic"' or attitudinal questions useful for adjusting estimates from web surveys using propensity scoring? *Survey Research Methods*, *1*(3), 155–163.

Schonlau, M., van Soest, A., Kapteyn, A., & Couper, M. (2009). Selection bias in web surveys and the use of propensity scores. *Sociological Methods & Research*, *37*(3), 291–318.

Schonlau, M., Zapert, K., Simon, L. P., Sanstad, K. H., Marcus, S. M., Adams, J., ... Berry, S. H. (2004). A comparison between responses from a propensity-weighted web survey and an identical RDD survey. *Social Science Computer Review*, *22*(1), 128–138.

Shih, T.-H. & Fan, X. (2008). Comparing response rates from web and mail surveys: a meta-analysis. *Field Methods*, *20*(3), 249–271.

Snijders, T. A. B. & Bosker, R. J. (2012). *Multilevel analysis* (2th). London: Sage.

Taylor, H. (2000). Does internet research work? *International Journal of Market Research*, *42*(1), 51–63.

Tourangeau, R., Conrad, F. G., & Couper, M. P. (2013). *The science of web surveys*. New York: Oxford University Press.

Valliant, R. & Dever, J. A. (2011). Estimating propensity adjustments for volunteer web surveys. *Sociological Methods & Research*, *40*(1), 105–137.

Vavreck, L. & Rivers, D. (2008). The 2006 cooperative congressional election study. *Journal of Elections, Public Opinion and Parties*, *18*(4), 355–366.

Weisberg, H. F. (2005). *The total survey error approach: a guide to the new science of survey research*. Chicago: The University of Chicago Press.

Wickham, H. (2016). *ggplot2 - elegant graphics for data analysis.* (2nd ed.). Cham: Springer Nature.

Yeager, D. S., Krosnick, J. A., Chang, L., Javitz, H. S., Levendusky, M. S., Simpser, A., & Wang, R. (2011). Comparing the accuracy of RDD telephone surveys and internet surveys conducted with probability and non-probability samples. *Public Opinion Quarterly*, *75*(4), 709–747.

Zhou, X.-H., Zhou, C., Liu, D., & Ding, X. (2014). *Applied missing data analysis in the health sciences*. Hoboken: Wiley.

Appendix
Tables

(Appendix tables follow on next page)

Table A1
Covariates used in the articles in table 1

Article	Covariates
Dever, Rafferty, and Valliant (2008)	Age group, race/ethnicity, gender, education, presence of children in household, employment, and marital status
Lee, Brown, Grant, Belin, and Brick (2009)	General health, having health care coverage, having personal doctor/health care provider, cost prevented from doctor's visit in the past 12 months, participate in any physical activities other than regular job during the past month, ever told to have diabetes by a doctor, ever checked blood cholesterol, trying to lose weight, weight advice given by health professional in the past 12 months, ever told to have asthma by a doctor, had a flu shot in the past 12 months, ever had a pneumonia shot, had a sunburn in the past 12 months, age in years, education, household income, current weight, number of residential phone lines, gender, had any symptoms of pain, aching, or stiffness around joint in the past 30 days, limited in any activities because of physical, mental, or emotional problems, moderate activities for at least 10 minutes in a usual week when not working, ever served on active duty in the U.S. armed forces, have a cell or mobile phone, amount of alcohol consumption, household size, work full-time, marital status, race, amount of vegetable consumption
Schonlau et al. (2004)	Race, gender, age, income, health insurance status
Schonlau, van Soest, and Kapteyn (2007)	Demographic: gender, log10 income, log10 income squared, age/10, education, primary language is English, born in the US (race and ethnicity are not available), self assessed health status. Webographic: attitudinal variables (Do you often feel alone? Are you eager to learn new things? Do you take chances?), factual variables (In the last month have you traveled? In the last month have you participated in a team or individual sport? In the last month have you read a book?), privacy variables: (Which of these practices, if any, do you consider to be a serious violation of privacy? Please check all that apply: 1. Thorough searches at airport checkpoints, based on visual profiles. 2. The use of programs such as 'cookies' to track what an individual does on the Internet. 3. Unsolicited phone calls for the purpose of selling products or services. 4. Screening of employees for AIDS. 5. Electronic storage of credit card numbers by Internet stores.), variables related to knowing gay people (Do you know anyone who is gay, lesbian, bisexual, or transgender? Please check all that apply. 1. Yes, a family member, 2. Yes, a close personal friend, 3. Yes, a co-worker, 4. Yes, a friend or acquaintance (not a co-worker), 5. Yes, another person not mentioned, 6. No)
Schonlau, van Soest, Kapteyn, and Couper (2009)	Race/ethnicity, gender, dummies for several education levels, age (transformed into a small number of categorical dummy variables), marital status, personal income (transformed to log personal income, also included a dummy variable for whether income equals zero), an indicator of home ownership, self-assessed health.
Yeager et al. (2011)	Race, ethnicity, census region, cross-tabulation of sex by age (12 groups), cross-tabulation of sex by education (10 groups)

Table A2
Sample sizes for the 28 ESS countries and USA, separated by Internet usage

Country	No Internet	Internet	Total
Austria	432	1,802	2,234
Belgium	441	1,263	1,704
Bulgaria	1,587	842	2,429
Croatia	839	793	1,632
Cyprus	582	496	1,078
Czech Republic	881	1,500	2,381
Denmark	223	1,353	1,576
Estonia	574	1,218	1,792
Finland	442	1,436	1,878
France	540	1,187	1,727
Germany	862	2,168	3,030
Greece	1,523	1,187	2,710
Hungary	690	867	1,557
Ireland	804	1,769	2,573
Israel	713	1,572	2,285
Lithuania	849	809	1,658
Netherlands	283	1,545	1,828
Norway	186	1,362	1,548
Poland	618	1,128	1,746
Portugal	1,315	832	2,147
Russian Federation	1,401	1,180	2,581
Slovakia	915	932	1,847
Slovenia	529	873	1,402
Spain	730	1,155	1,885
Sweden	221	1,276	1,497
Switzerland	347	1,158	1,505
Ukraine	1,358	560	1,918
United Kingdom	679	1,743	2,422
USA	118,358	366,560	484,918
Total (ESS)	20,564	34,006	54,570