***EGE AKADEMİK BAKIŞ** / EGE ACADEMIC REVIEW*

Cilt 24 • Sayı 2 • Nisan 2024
SS. 165/178
Doi: 10.21121/eab.1268873
Başvuru Tarihi: 21.03.2023 • Kabul Tarihi: 01.03.2024

***Article Type:** Research Article*

The Internet Usage Rate in Turkey: A Machine Learning Approach

Mustafa İNCEKARA[1], Cemal ÖZTÜRK[2]

ABSTRACT

Most studies on measuring coverage bias in internet surveys use internet access as a critical measurement variable. However, access to the internet does not mean that individuals are using it. Therefore, using the internet usage rate as a key variable is crucial to get an accurate overview of the internet coverage of a population. This study closes these gaps by using a better indicator for measuring the internet usage rate. It is the first study measuring the internet usage rate in Turkey by using the real internet usage rate of the population and applying a machine learning algorithm. The results exposed significant differences in socio-demographic characteristics when internet users were compared with non-users. Furthermore, the coverage bias associated with internet users remained different for several demographic categories. The results of web-based surveys based on the actual internet usage rate are crucial for the scientific community and marketers.

Keywords: Internet Usage Rate, Socio-Economic Factors, Turkey, Machine Learning.

JEL Classification Codes: C45, C55, C83, D19, M30, L86

Referencing Style: APA 7

INTRODUCTION

According to the World Bank, 45.8% of the world's population uses the internet (Worldbank, 2018). However, home access penetration has not reached 100% even within the EU, ranging from 97% in the Netherlands to 45% in Italy, with an overall mean penetration of 70% for the EU (Eurobarometer, 2017). Furthermore, even in nations with high-level internet coverage, access is unequally dispersed over the inhabitants, with very well-educated and younger people more likely to have an internet connection (Mohorko et al., 2013). Thus, Hwang and Fesenmaier (2004) conclude that internet-centred surveys can only represent active users.

Therefore, the population with no internet access and usage is still a major issue in internet surveys, leading to a central issue in web-based surveys: under coverage. This is a consequence of the „digital divide", the discrepancy in the rate of internet access between demographic groups, e.g., differences related to gender, age, or education level (Couper, 2000). Researchers must consider digital inequality's impact on involvement when using internet surveys. Researchers applying online surveys should examine issues regarding respondents' representation of the target population, specifically nonresponse and coverage error (Robinson et al., 2015; Couper et al., 2007).

Based on this limited review, it is evident that the coverage error of internet surveys is closely related to access to the internet (Schaefer and Dillman, 1998). Thus, some people consider that if most of the population actively uses the internet, then internet users could be assumed to represent the general public (i.e. Yun and Trumbo, 2000). However, assuming that all individuals with internet access are capable and have sufficient facilities to participate in an internet survey can lead to an underestimation of potential coverage errors (Sterret et al., 2017).

Besides simple internet access, users must also have a requisite skill level to complete web-based tasks (van Deursen and van Dijk, (2009). Internet proficiency or adeptness can differ between socio-demographic groups (Hargittai and Hsieh, 2012; Mossberger et al., 2010; Stern et al., 2009). Previous researchers have shown that

[1] Assoc. Prof., Pamukkale University, Faculty of Economic and Administrative Sciences, Department of Business Administration, Denizli, Turkey, mincekara@pau.edu.tr

[2] Res. Asst.,Pamukkale University, Faculty of Economic and Administrative Sciences, Department of Economics, Denizli, Turkey, cemalo@pau.edu.tr

individuals without the skill to finish a web-based survey tend to be socially, politically, and economically different from those with sufficient access and proficiency to finish an internet-based survey (Selwyn, 2004; Mossberger; Sterret et al., 2017).

Coverage of internet surveys significantly increases as more individuals have the opportunity to get online; however, this coverage is still far from complete population coverage since internet access does not necessarily mean that people are using the internet. Actual usage is generally lower than the internet access rate (TUIK, 2017). Therefore, measuring internet usage at the point of access is important. Regardless of the means of data gathering, e.g., web surveys, social media surveys, or online panels, a lack of comprehensive understanding of internet adoption can lead to imprecise estimations (Robinson et al., 2015).

This research aims to highlight significant socio-demographic differences between people who use the internet and those who do not by comparing these two groups in an emerging country like Turkey. This research aims to evaluate different socio-demographic characteristics using the internet usage rate. Therefore, the findings of this study will indicate that there is a possibility that coverage error might be reduced by weighting the information obtained from web-based surveys conducted in Turkey concerning certain socio-demographic characteristics.

To address this question, we used the Information and Communication Technology (ICT) Usage Survey on Households and Individuals from TUIK (Turkish Statistical Institute) for the period 2011–2017, which provides data on the demographic variables of both internet users and non-internet users. The findings are important for researchers and users of internet surveys since no previous study has examined the effects of actual internet usage and internet survey bias in the case of an emerging market such as Turkey.

We concentrate on two critical gaps in the literature: From a research standpoint, we move beyond the typical dominant focus on digitized welfare in industrialized countries to investigate the influence of socio-demographic factors on internet usage. In practice, our approach reacts to market and governmental decision-makers requests to decrease the determinants, increasing the digital divide in society while promoting activities to reduce digital disparities.

The originality of this study will pique the interest of managers, politicians, and researchers. From a scientific standpoint, this study will be a management and social research trailblazer. We will combine traditional statistical methods with machine learning models to build a framework for policymakers, managers, and the scientific community. From this standpoint, this study will be one of the first to use machine learning technology in a socio-economic setting. The findings will allow public officials to modify policies to increase internet adoption in society and reduce the digital divide. Furthermore, firms' market researchers and managers can apply to their marketing activities the consequences of actual internet use and the bias of online surveys in the case of a growing market like Turkey. This study will provide a foundation for governments, businesses, and scientific researchers.

In the following sections, we estimate the fraction of Internet users in Turkey and analyze the difference between Internet access and regular internet users. Then, we will examine to what magnitude internet users differ from the overall population of Turkey. By applying binary logistic regression, we will identify the portion of socio-demographic characteristics that distinguish internet users from non-internet users within the socio-demographic categories. Furthermore, we apply a machine learning technique to get deeper insights into socio-demographic effects on internet usage rates. Finally, we will discuss the challenges of the internet usage rate in the context of the practical use of web-based surveys.

RELATED LITERATURE REVIEW

The process of gathering primary data has radically transformed within the last two decades, a change that can be seen in internet survey research (Robinson et al. 2015). The era of global connectivity and increased internet access has made web-based surveys a popular sampling technique for scientists and businesses.

Most research assessing coverage bias in internet surveys utilizes internet access as a key measurement variable. Couper et al. (2018) investigated the socio-demographics of mobile phone and internet coverage in combination and independently. They examined the impact of different coverage levels. Their research has consequences for potential coverage biases that could appear when changing to an internet-based data-gathering method, either for follow-up investigations or to replace the primary in-person data collection.

García-Mora and Mora-Rivera (2023) aimed to estimate the effect of internet access on poverty for a section of rural Mexicans residing in distinct regions. They used a quasi-experimental methodology to discover that Internet access is an additional method for poverty reduction. Additionally, they showed that internet access could help increase the proportion of rural residents living above the poverty line.

Built on a Propensity Score Matching method, Mora-Rivera and -Mora (2021) indicated that policy measures should be taken to resolve topics that restrict Internet access for persons and households with greater social defenselessness, thereby causative to a decrease in the shortage stages practised by a significant portion of households with high levels of poverty.

Valentín-Sívico et al. (2023) showed that Internet access at home improves the value of life for households and expands their social and economic chances. Their study resulted in two main conclusions: first, variations in internet use for education, employment, and health could not be straight accredited to internet interference; and second, the internet interference was linked with paybacks from the capability to apply several devices.

Martínez-Domínguez and Mora-Rivera (2020) seek to identify the socio-economic and demographic factors that encourage the rural population to acquire and utilize the internet. Using an econometric model to account for the possibility of selection bias, findings suggest that the likelihood of Internet use is greater among people with digital abilities and women. Internet usage patterns alter by level of education, age, nature of employment, and location. Young populations are more prone to engage in virtual actions for enjoyment, whereas adults use the internet for communication, information, and e-business. These results offer verification of the extant digital split regarding Internet access and utilization.

Byaro et al. (2023) applied a generalized quantile regression approach to observe the association between health effects and internet use. The results demonstrate the diverse impact of health expenditures, income, and market on health results and carbon dioxide releases within quantiles. These indicate a diminishing return on investment or increased health outcomes at a particular level.

The internet has a significant effect on research methods. Its use for systematic data gathering is expanding because it offers cost efficiency, time savings, and access to different and large populations (Hays, Liu, and Kapteyn, 2015). Data acquisition via internet surveys continues to become more popular as rates of internet access rise (Sterret et al., 2017; Worldbank, 2018).

Respondent anonymity has peaked with Amazon's Mechanical Turk, and the idea of paid surveys has exploded with the birth of internet panel websites. Though these Web sites require memberships, fake demographic profiles may be created while registering. Knowing who responds to such surveys is virtually impossible, even if e-mail addresses are collected.

With the emergence of paid surveys, researchers began to be concerned about professional respondents providing lower-quality data, based on the assumption that professional respondents' extrinsic motivation—getting paid—would lead them to respond with minimal cognitive effort. This assumption was recently tested by de Leeuw and Mathijsse (2016), who could not find empirical evidence that professional respondents produce data of lower quality.

The quality of a survey depends on its complete measurement of the probability sample (Groves, 2006). Though Internet infiltration into households maintains a fast pace in the EU, the penetration is nevertheless far from complete and varies widely from country to country, even within the EU (Eurobarometer, 2017). Internet-based surveys may only frame internet users and cannot be generalized to the general public.

In recent years, some scholars have researched population coverage in internet surveys. However, the analysis of internet coverage rates in developed countries has primarily been examined for highly industrialized nations such as the United States and members of the European Union (e.g., Mohorko et al., 2013; Sterrett et al., 2017; Vicente & Reis, 2012; Yeager et al., 2011; Heerwegh & Loosveld, 2008). Thus, the importance of digital inequality for response rates in internet surveys is still unclear in newly industrialized countries and emerging markets such as Turkey (Boddin, 2016; Robinson et al., 2015; Stern, Bilgen, & Dillman, 2014). In recent years, survey research has faced a massive drop in participation across different survey modes, especially in interview-based measurement methods such as telephone surveys. In the early days of the world wide web, internet-based surveys seemed like an optimal solution to this issue. Several scholars thought that internet or social media surveys would substitute paper and pencil and telephone surveys and solve the problem of recruiting participants via traditional mail, telephone, or in-person surveying (Robinson et al. 2015).

Studies exploring the discrepancies between the populations of internet users and non-users have found some attitudinal, demographic, and behavioral variations between the two. For instance, in the Netherlands, non-internet users are older, live in a single household, and have a migration background. Similarly, German non-internet users are likely less educated and slightly older than internet users (Eckman, 2016). Furthermore, some personality differences were also identified between German internet users and non-users (Eckman, 2016). Differences in age, income, race, college education, and urbanity between Internet and non-user households have also been reported in the US (Couper, 2000).

Pew Research Center phone studies have noted a growth in internet acceptance, which has increased from 14% of US residents in 1995 to 89% in 2015. (Pew Research Center, 2015). Thus, one in ten adults has no internet access, leaving uncertainty in the survey's results.

The report's authors conclude that this is an outcome of the fact that the non-web survey group constituted a small part of the target population (Pew Research Center, 2015). However, it could also result from the nature of the researched topic, where the differences between internet- and non-internet-using respondents are insignificant. Moreover, even the Pew Center's researchers conclude that though 90% of Americans use the internet, web surveys are not free from a modest bias (Pew Research Center, 2015).

Not to be confused with other types of non-observations, it should be highlighted that not all bias from non-observations is equivalent. The reasons for not having internet access may differ from those that impact involvement amongst chosen test members. Therefore, bias as a result of undercoverage may be extremely distinct in extent and direction from bias due to nonresponses. Further, the motives for exclusion from the structure and nonresponse can be distinct, and the two may also vary in their demographic (Peytchev et al., 2011) and behavioral characteristics.

MODEL SPECIFICATIONS AND DATA ANALYSIS

Since 2004, studies on information and communication technology (ICT) usage by individuals and households have been conducted by TUIK (the Turkish Statistical Institute). This is the primary database on ICT usage in Turkey; the questionnaires are adapted and modified from model questions from Eurostat regarding conditions and needs in Turkey (TUIK, 2017). Data is collected through computer-assisted face-to-face interviews containing questions about internet usage and demographic variables (TUIK, 2017). Face-to-face interview processes are more effective than other survey modes at achieving coverage of a high percentage of the populace (e.g., Groves and Lyberg, 2010).

Several studies analyzed the coverage error of internet surveys in Europe and the United States by using a dataset that used face-to-face interviews to examine Internet access (Heerwegh and Loosveldt, 2008; Vicente and Reis, 2012; Mohorko et al., 2013; Tourangeau et al., 2013). Face-to-face interviews linked with address-based sampling as a survey method has led to the greatest extent of population coverage (Vicente and Reis, 2012; Sterrett et al., 2017).

The ICT applied an address-based sampling method, using in-person interviews to ask about the internet usage rates of individuals (TUIK, 2017). Identical questions regarding internet usage rates administered in in-person interviews every year offered us a unique opportunity to compare internet users and non-internet users based on socio-demographic factors over time. Most studies on coverage bias in web-based surveys have used online data (Vicente & Reis, 2012). In contrast, this study uses data gathered in person, which offers theoretical coverage of the entire population. Therefore, it allows us to estimate the possible coverage bias of non-internet users.

Furthermore, we applied a machine learning algorithm to get deep insights regarding the socio-demographic classifications of internet users. We used a decision tree approach that also applied a classification system. A classification represents a relationship between input data and output data. As a supervised technique in data mining, classification determines the proper class labels for an unlabeled test case from a training dataset with associated training labels (Aggarwal, 2014). Classification involves target marketing, credit approval, systematic medical analysis, fraud detection, and scientific research (Mitchell et al., 1990; Hastie et al., 2013).

Decision trees classify samples by ordering them from the tree's root to the leaves. Each node on the tree contains the test of the example's feature (attribute), and each branch from that node relates to a value of that feature. A sample is classified by starting from the root and going to different branches according to the value of the feature at each node and reaching the leaves. Each leaf specifies a target value.

Given training vectors $x_i \in R^n, i = 1, \ldots, l,$ and a label vector $y \in R^l$, a decision tree recursively separates the feature space so that samples with identical labels or similar target values are clustered. Let Q_m represent the data at node with samples. If a target is a classification result taking on values $0,1,\ldots,K\text{-}1,$ for node m, let

$$p_{mk} = \frac{1}{n_m} \sum_{y \in Q_m} I(y = k),$$

be the ratio of class k observations in node m (Zhang, 2021). We use Entropy measures of impurity as the following,

$$H(Q_m) = -\sum_k p_{mk} \log(p_{mk}).$$

RESULTS

Estimates of Internet Access and User Rates in Turkey, 2011–2017

Data from a total of 130,723 respondents from 2013–2017 is analyzed. We start our analysis by comparing internet access and accurate internet usage rates during the study period (Figure 2).

The ratio of internet users increased from 48.9% in 2013 to 66.8% in 2017, a change of 17.9%. In contrast, internet access increased from 49.1% in 2013 to 80.7% in 2017, a mean increase of 31.6%. The increase in the internet usage rate implies a decrease in coverage error in web-based surveys based on the sampling frame. The distinctions between internet usage and household internet access in terms of internet coverage seem to be an obvious source of potential coverage errors in internet-based surveys focusing only on household internet access.

Disparities Between the Internet-User and Non-Internet-User Populations

For survey researchers, the low internet user rates reported in Table 1 are not a problem if the covered population delivers the same results as the general population (Fricker, 2008). However, data sampling may not cover the general population; therefore, we evaluate the variations between internet users and non-users.

Given the variables of the TUIK questionnaire regarding socio-demographics, we analyze age, gender, education degree, employment status, region of residence, family income, and family size. The results suggest noticeable differences between the population using mobile internet and those without mobile internet access. The internet user population in Turkey for the study period was noticeably younger than non-users.

While 25.6% of internet users were 16 to 24 years old, the applicable coverage rate in the non-user group was 5.7%. Furthermore, internet users were more likely to live in western Turkey. For instance, in region TR1

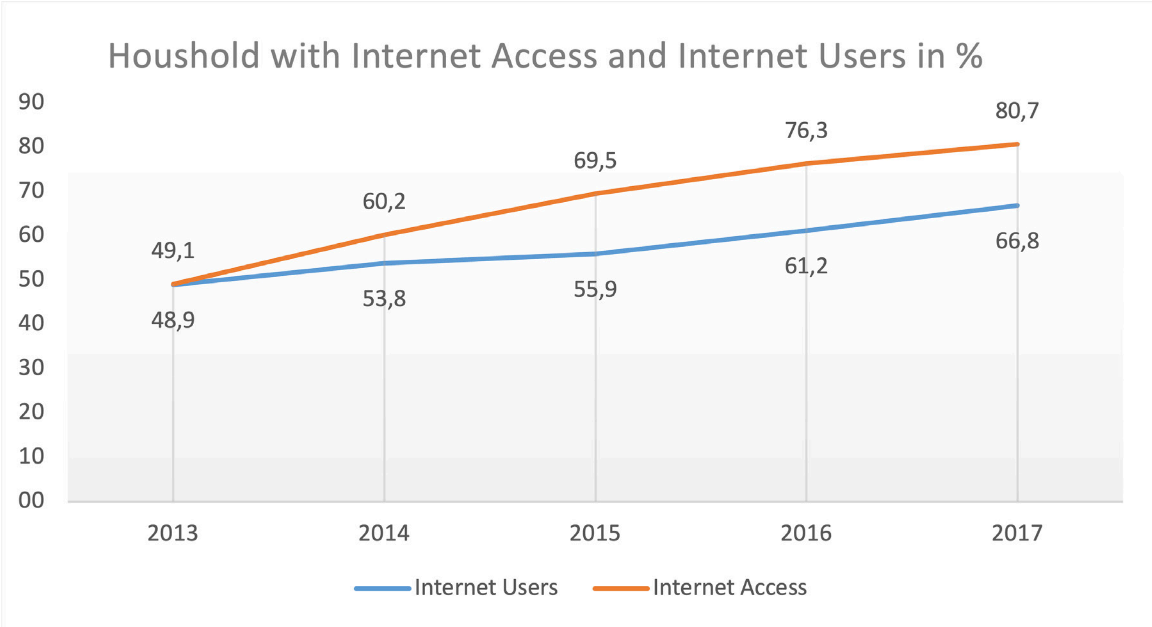


Figure 1: Household internet access and internet user rates in Turkey from 2011 to 2017.

Table 1: Socio-Demographic Characteristics of Internet Users and Non-Internet Users

	2013		2017		Differences 2013-2017	
	Internet	Non-Internet	Internet	Non-Internet	Internet	Non-Internet
Socio-Demographic characteristics	48.90%	51.1%	66.8%	33.2%	**17.9%**	**-17.9%**
Age categories						
16 - 24	31.17%	8.68%	25.6%	5.7%	-5.62%	-2.97%
25 - 34	33.04%	16.10%	28.9%	8.4%	-4.12%	-7.70%
35 - 44	21.86%	20.08%	24.8%	15.5%	2.89%	-4.54%
45 - 54	9.94%	23.35%	13.8%	23.6%	3.89%	0.28%
55 - 64	3.31%	19.70%	5.6%	27.2%	2.29%	7.48%
65 - 74	0.68%	12.09%	1.4%	19.5%	0.69%	7.45%
Total	100.00%	100.00%	100.0%	100.0%	0.00%	0.00%
Gender						
Male	60.09%	39.51%	56.0%	37.5%	-4.07%	-2.00%
Female	39.91%	60.49%	44.0%	62.5%	4.07%	2.00%
Total	100.00%	100.00%	100.0%	100.0%	0.00%	0.00%
Education Status						
Did not finished school	1.09%	28.06%	2.6%	34.1%	1.50%	6.05%
Primary school	16.60%	53.26%	21.1%	51.5%	4.53%	-1.79%
Secondary School	26.06%	10.93%	25.8%	8.8%	-0.31%	-2.08%
High School	32.24%	6.47%	26.9%	4.6%	-5.32%	-1.85%
Higher Education	24.01%	1.29%	23.6%	1.0%	-0.40%	-0.33%
Total	100.00%	100.00%	100.0%	100.0%	0.00%	0.00%
Working Status						
Working	56.07%	29.96%	52.8%	24.8%	-3.30%	-5.16%
Not Working	43.93%	70.04%	47.2%	75.2%	3.30%	5.16%
Total	100.00%	100.00%	100.0%	100.0%	0.00%	0.00%
Region of Residence						
TR1	23.09%	13.91%	23.4%	12.8%	0.27%	-1.10%
TR2	5.11%	4.48%	4.3%	4.8%	-0.80%	0.32%
TR3	13.59%	13.15%	13.1%	13.5%	-0.52%	0.30%
TR4	11.58%	8.97%	10.2%	9.6%	-1.38%	0.61%
TR5	11.77%	7.99%	11.1%	7.2%	-0.70%	-0.77%
TR6	11.01%	13.59%	13.0%	12.5%	1.98%	-1.10%
TR7	4.83%	5.39%	4.8%	5.2%	-0.03%	-0.20%
TR8	5.03%	7.27%	5.4%	6.8%	0.33%	-0.45%
TR9	2.95%	4.15%	3.0%	4.0%	0.09%	-0.16%
TRA	1.64%	3.14%	1.7%	3.6%	0.07%	0.42%
TRB	2.93%	6.23%	3.4%	6.1%	0.47%	-0.12%
TRC	6.46%	11.73%	6.7%	14.0%	0.23%	2.24%
Total	100.00%	100.00%	100.0%	100.0%	0.00%	0.00%
Family Income (Monthly)						
0-999	17.37%	47.98%	3.5%	13.5%	-13.87%	-34.46%
1000-1999	33.33%	33.82%	25.9%	42.9%	-7.45%	9.08%
2000-3999	36.39%	16.12%	43.6%	34.6%	7.24%	18.52%
4000-5999	9.63%	1.61%	17.3%	7.1%	7.63%	5.53%
6000-7999	1.91%	0.24%	5.4%	1.1%	3.46%	0.88%
8000-9999	0.36%	0.06%	1.5%	0.2%	1.10%	0.13%
10000 and above	1.02%	0.17%	2.9%	0.5%	1.89%	0.31%
Total	100.00%	100.00%	100.0%	100.0%	0.00%	0.00%
Household size						
1	3.52%	5.42%	4.0%	5.6%	0.52%	0.17%
2	39.96%	34.74%	12.4%	21.2%	-27.61%	-13.52%
3	24.05%	23.21%	21.6%	15.6%	-2.40%	-7.64%
4	19.61%	18.67%	27.3%	15.5%	7.66%	-3.14%
5	7.51%	8.59%	15.7%	12.0%	8.21%	3.42%
>6	5.35%	9.37%	19.0%	30.1%	13.62%	20.71%
Total	100.00%	100.00%	100.0%	100.0%	0.00%	0.00%

(Istanbul), 23.4% are internet users, and 12.8% are non-users (2017). There is an inequality between gender (male and female) and working status (working vs. not working) in the internet usage rate (56% vs. 44%). The internet-using population was also, on average, far more educated than the remaining portion of the general population. In summary, internet users in Turkey tend to be under 45 years of age, more male and better educated than non-users, residing in western Turkey, and living in a household of 3–4 persons with a monthly income between 2,000 and 4,000 TL.

The Effect of Socio-Demographic Variables on Internet Usage Rates

One of the objectives of this study is to identify to what extent the absence of non-internet users from internet-based surveys can generate a bias regarding demographic variables. In this step, we apply a binary logistic regression model to evaluate the influence of socio-demographic variables on internet usage rates. A logistic regression examination identifies the most important socio-demographic and economic factors separating internet users and non-users by analyzing the key drivers of the independent variables: age, gender, educational status, working status, region of residence, monthly household income, and household size relative to the internet use rate.

The research model is estimated using SPSS version 25 for a binary logistic regression evaluation. The binary logistic regression model is executed using a dichotomous dependent variable. The dependent variable was coded as 0 for “non-internet user” or 1 for “internet user”. Table 2 shows the effects of the binary logistic regression on variables predicting internet usage rate.

Several demographic and socio-economic features are coded as independent variables of the logistic regression model, including gender (1 = male, 2 = female), age coded from 1 to 6 to categorize different age ranges from 16 on (1 = 16–24, 2 = 25–34, 3 = 35–44, 4 = 45–54, 5 = 55–64, and 6 = 65–74), education status (from “Did not attend School”, “Primary School”, “Secondary School”, “High School” and “Higher Education”, coded as 1 to 5), and working status (1 = working, 2 = not working). The different regions in Turkey (Appendix) are categorized from 1–12. The number of living people in the household is coded from 1–6, with 7 used for all households with 7 or more members. The socio-economic factor of household income was coded from 1–7 for different income ranges (1 = 0-999, 2 = 1000–1999, 3 = 2000–3999, 4 = 4000–5999, 5 = 6000–7999, 6 = 8000–9999, and 7 = 10000 and above).

Overall model fit is estimated using the Nagelkerke R2 statistic and the Hosmer–Lemeshow test. The latter defines the accuracy of the distribution of the detected measures, comparing the observed figures with the expected figures (Hosmer et al., 2013). Its estimates follow a chi-square (χ2) distribution; its results indicate that all non-significant p values fit our model well. The Wald test is used for a significance test of each variable. Finally, following a conservative recommendation, we estimate the rate of correct case categorizations and determine values above the threshold of 60% as acceptable and values over 70% as good (Hair, Black, Babin, & Anderson, 2014).

The overall model is statistically significant (χ2 (34) = 88221.808, p < 0.0001). Hence, the model effectively differentiates respondents between internet users and non-users. The Nagelkerke R2 equals 0.656; our models can describe around 65% of the variance. The result of the Hosmer–Lemeshow test is significant; regarding model accuracy, our model can predict nearly 84% of cases correctly (see Table 2). We further examine the data by applying logistic regressions to determine the likelihood of each predictor impacting the likelihood of being an internet user.

Estimations of the binary logistic regression model show the probability of being an internet user (Table 2). The estimated model proposes that the possibility of being an internet user varies by age, gender, education status, working status, region of residence, household income, and household size. The likelihood of being an internet user is higher among younger age classes. Assuming all variables are constant in odds estimates, for every Internet user aged 65–74, we measured almost 100 individuals between 16 and 25 (odds ratio = 98.833:1) and more than three (3.292:1) individuals aged 55–64 years. The pattern is linear: the odds ratio of being an internet user declines as the age category increases. An individual’s educational status is crucial to the odds of being an internet user—the odds of being an internet user increase with an increase in education level. The working status also significantly affects internet user status (odds ratio = 1.3:1).

The region of residence is also relevant. Living in regions TR1 through TR9 has a significant positive association with being an active internet user. The odds ratio for these regions is at least 1.2 times higher than the reference region TRC. Residence in region TRB has a negative influence on being an internet user; the odds ratio is smaller than 1 and, therefore, the probability of being an internet user is lower than living in region

Table 2: Effect of Socio-demographic Variables on Internet Usage Rate

VARIABLES OF INTERNET USER	β Estimate	Standard error	P value	Odds Ratio
AGE CATEGORIES (ref. 65-74)				
16-24	4.593***	0.056	0.000	98.833
25-34	3.919***	0.053	0.000	50.334
35-44	3.364***	0.052	0.000	28.909
45-54	2.222***	0.051	0.000	9.222
55-64	1.192***	0.052	0.000	3.292
GENDER (ref. female)				
Male	0.773***	0.020	0.000	2.166
EDUCATION STATUS (ref. Higher Education)				
Did not finished school	-4.489***	0.054	0.000	0.011
Primary school	-3.112***	0.045	0.000	0.045
Secondary School	-2.067***	0.047	0.000	0.127
High School	-1.219***	0.047	0.000	0.296
WORKING STATUS (ref. Not Working)	0.263***	0.020	0.000	1.301
Working				
IBBS_1_REGION (ref. TRC)				
TR1	0.568***	0.037	0.000	1.765
TR2	0.377***	0.045	0.000	1.458
TR3	0.405***	0.038	0.000	1.500
TR4	0.517***	0.040	0.000	1.677
TR5	0.393***	0.040	0.000	1.481
TR6	0.392***	0.038	0.000	1.480
TR7	0.260***	0.043	0.000	1.297
TR8	0.239***	0.042	0.000	1.270
TR9	0.182***	0.047	0.000	1.199
TRA	-0.091	0.047	0.052	0.913
TRB	-0.132**	0.042	0.002	0.877
HOUSEHOLD MONTHLY INCOME (ref. 10000 and above)				
0-999	-2.639***	0.125	0.000	0.071
1000-1999	-1.647***	0.124	0.000	0.193
2000-3999	-0.958***	0.124	0.000	0.384
4000-5999	-0.421**	0.127	0.001	0.656
6000-7999	0.026	0.150	0.863	1.026
8000-9999	0.157	0.223	0.481	1.170
HOUSEHOLD SIZE (ref. 7 and above)				
1	1.597***	0.063	0.000	4.938
2	0.843***	0.037	0.000	2.323
3	0.888***	0.036	0.000	2.431
4	0.887***	0.035	0.000	2.428
5	0.737***	0.037	0.000	2.090
6	0.468***	0.042	0.000	1.598
Constant	-0.369**	0.138	0.007	0.691
*p < 0.05; ** p < 0.01; *** p < 0.001				
Nagelkerke R Square:	0.656			
Hosmer and Lemeshow Test	χ2 (8) = 127.948 (Sig.:0.000)			
Correct Classification	83.9			

TRC (odds ratio: 0.877:1), though living in TRA is not significant. Household income is a significant predictor of belonging to the internet user population. Lower-income significantly negatively affects internet users. The odds ratio is lower than 1 for all income categories lower than 6000 TL; incomes higher than 6000 TL have no significant effect on being an internet user. Finally, the number of members in the household is also correlated with being an active web user. Individuals living in a single-person household are nearly five times more likely to be internet users than members of households with seven or more members. A positive association is also found for other household categories.

The Classification of Socio-Demographic Variables on Internet Usage Rates

The machine-learning analysis decision tree underscores the importance of education. Education is the most critical variable for distinguishing between internet and non-internet users. The decision tree's first (and thus most important) branch directly predicts Internet usage classification. In the first path, if the age is below 46.5, it reinforces an internet user classification. Furthermore, the decision tree analysis reinforces the importance of internet access.

Conversely, if education is low and internet access is given, then age does not influence being an internet user. The second path is that if education is high, internet access is given, and the age is over 43.5, internet usage is forced to be yes. The classification is still an internet user if internet access is not given. If internet access is provided and the person is over the age of 43,5 years, the person is classified as a non-internet user. Gender, working status, region, household income, and size are not decision tree nodes.

DISCUSSION

The present study focuses on understanding internet usage rates in newly industrialized countries or emerging markets, such as Turkey (Boddin, 2016), where no previous study has estimated the effect of socio-demographic variables on internet usage.

Based on the TUIK ICT data, this investigation made it possible to analyze coverage errors. The results indicate that Internet surveys are increasingly attractive as more Turkish people have access to the internet. The ICT Usage Survey on Households and Individuals reveals that the rate of adults in Turkey who have internet access increased from 49.1% in 2013 to 80.7% in 2017; in the same time frame, the internet usage rate increased from 48.9% to 66.8%. These disparities could lead to coverage errors involving those using the internet and non-users, leading to major concerns for those administering Internet surveys.

Similar to the study conducted by Couper et al. (2018), which examined the socio-demographic factors associated with Internet and smartphone coverage, the present results also exposed substantial dissimilarities in socio-demographic variables when Internet users were compared with non-users. Age is the strongest predictor of internet use; gender, working status, educational level, the region's residence, household income, household size, and marital status are (in order) the strongest predictors of internet use. Therefore, internet users cannot be an absolute variable since their characteristics vary with socio-economic and demographic determinants.

The effect on coverage bias size is another challenge when measuring internet survey efficiency since the extent of coverage bias varies across socio-demographic

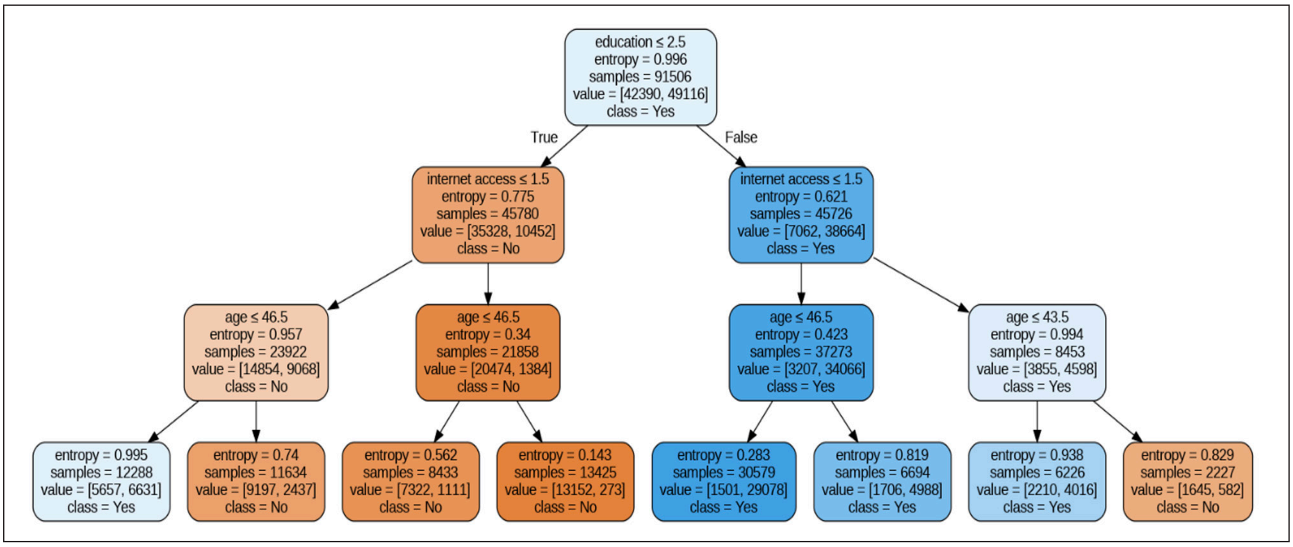


Figure 2: Decision Tree

measures. These results indicate that web-based surveys' precision level is related to the population distribution, which is connected to these socio-demographic variables in the sample frame.

In agreement with other studies on coverage error in Internet surveys in other countries (e.g., Mohorko, de Leeuw, & Hox, 2013; Sterrett et al., 2017), this paper shows possible issues regarding Internet-based surveys in Turkey. Like Vicente & Reis (2012), the present study shows that scrutinizing non-internet users' representation is vital to ensure survey quality. The dissimilarities in socio-demographic and economic appearances among internet and non-internet users indicate possible coverage errors in internet-based surveys relating to gender, age, education, working situation, region of residence, income, and household size. Consequently, using internet users as the sampling population for surveys may lead to systematic under-coverage, for instance, of older people aged 65–74 and females without a university degree. Therefore, similar to Couper et al. (2018) results, this study's findings have significant implications for potential coverage biases that could emerge during the transition to a Web-based data collection method, whether for subsequent surveys or as a substitute for the initial in-person data collection process.

These factors implicate various behaviors and attitudes, so excluding such groups can produce different results. Our outcomes reveal potential sub-populations that are underrepresented in internet-only sampling. Furthermore, the results recommend that weighting the information from web-based surveys in Turkey for specific socio-demographic factors could decrease coverage error. In the next step, an examination of the differences in survey results based on a weighting of the vulnerable socio-demographic variables for the general public group and subgroup should be conducted.

The results of this study also expose substantial dissimilarities in socio-demographic variables when internet users are compared with non-internet users. Age is the strongest predictor of internet use; gender, working status, educational level, the region's residence, household income, household size, and marital status are other strong predictors of internet use. Thus, internet users cannot be used as an absolute variable since their characteristics vary with socio-economic and demographic determinants.

Our results are similar to Valentín-Sívico et al. (2023) findings. They did not find a direct causal relationship between internet intervention and changes in internet use for employment, education, and health.

Drops in coverage errors related to educational status and age are comparable to those detected in Europe and the United States (Mohorko et al., 2013; Sterrett et al., 2017). This study shows that the relative Internet coverage error related to gender, age, education, household size, and income in Turkey declined considerably from 2013 to 2017, and the proportion of decline differs across the explanatory variables.

Furthermore, the study is aligned with the outcome of Martínez-Domínguez and Mora-Rivera (2020). They revealed that the utilization patterns of the internet exhibit variations based on factors such as age, educational attainment, occupational characteristics, and geographical location. The propensity for young individuals to partake in online activities primarily for entertainment is contrasted with adults' inclination to utilize the internet for information acquisition, communication, and engaging in electronic commerce.

CONCLUSION

Our research paper explores the potential coverage biases that may occur during the transition from traditional face-to-face data collection to a web-based mode. This includes both follow-up surveys and the complete replacement of in-person data collection. The findings above underscore the existing regional variations within Turkey and imply that governments must formulate more effective and focused public policies that tackle the unequal distribution of Internet adoption. Implementing these policy enhancements would enable governments to optimize the potential advantages of the internet, particularly in nations such as Turkey. The results offer empirical support for a digital divide in Turkey, explicitly concerning Internet penetration and usage. The impact of internet utilization and adoption on health effects is also examined. It is recommended to prioritize strategies, policies, or laws about internet use and adoption that guarantee the accessibility of digital tools, such as computers and mobile phones, for internet connectivity. The study's policy implications are expected to provide valuable guidance for policymakers to enhance internet connectivity, encouraging further evaluation research. The necessity for communities to maintain economic competitiveness is progressively reliant on the presence of high-speed broadband infrastructure. Robust evaluations play a pivotal role in ensuring the efficacy of government funds and providing valuable insights for future allocations of infrastructure expenditure. Community-level assessments serve as valuable tools for local elected representatives and decision-makers, enabling them to identify the potential effects of internet access on their community and make informed projections.

The outcomes of this research have a variety of implications for academic scientists. Even though the proportion of Turkish people with home internet access has increased in recent years, internet usage is still far behind the access rate. Researchers still must consider possible coverage bias when carrying out internet-only surveys.

The possible coverage error of important socio-demographic variables essential to internet-based surveys makes it crucial to use various survey methods to frame the sample and assess a broader and more representative population. Therefore, postal, phone or face-to-face surveys offer a chance to contact the non-internet user and reduce coverage errors across the population.

Furthermore, measuring the advantages and disadvantages of applying weight adjustments that reflect dissimilarities in internet usage rates within a demographic is valuable. Turkish non-internet users continue to be a separate segment of the population, which must be considered during survey design to allow scientists to make valid assumptions about the general population.

The findings of this study have significant implications for informing the design of future studies conducted through internet-based platforms. This paper provides recommendations for the identification of suitable outcome variables, the implementation of effective recruitment strategies, and the selection of optimal timing for surveys.

Moreover, future studies should explore the factors of the digital divide in society to reduce the coverage bias in internet-based surveys. It is suggested that researchers could investigate the correlation between internet accessibility and various health outcomes. This can be achieved by incorporating additional health indicators and examining different time frames. Additionally, employing diverse econometric methodologies and accounting for additional confounding factors could enhance the validity of the findings. In order to provide decision-makers with more precise implications that are specific to each country, it is recommended that future research incorporates nation-specific studies.

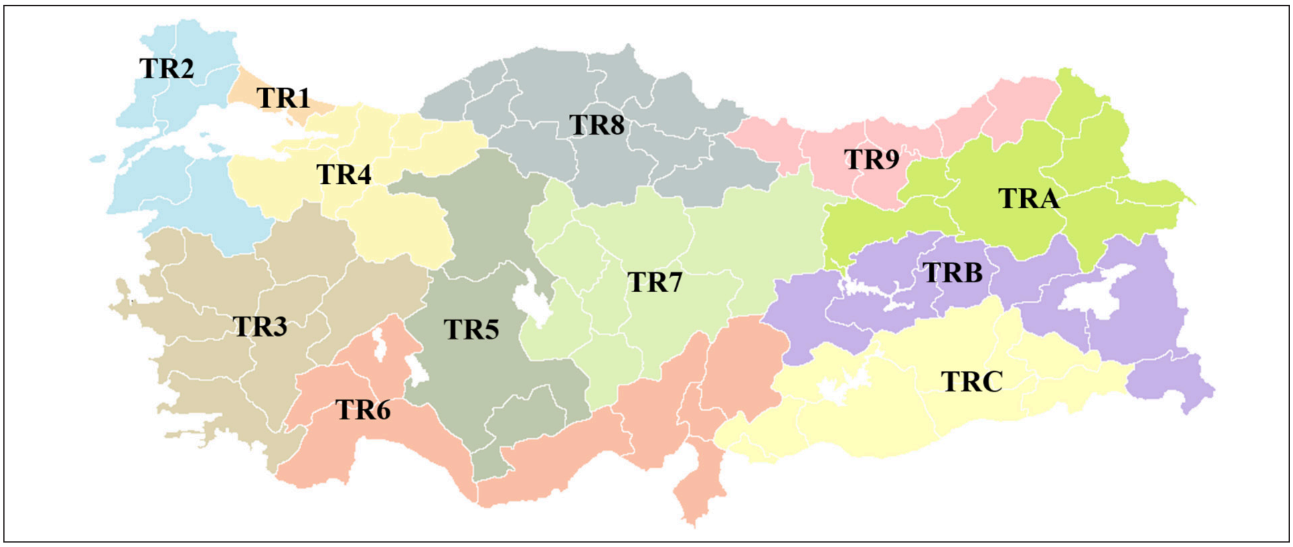
One notable strength of this study is its utilization of logistic regression analysis and a machine learning methodology. Furthermore, a significantly large sample size was incorporated into each respective methodology. Regarding limitations, it is essential to acknowledge that this study employed a cross-sectional survey design, which introduces the possibility of recall and social desirability biases. Moreover, the machine learning techniques utilized in this research can be effectively employed, for example, in analyzing product preference and predicting demand. Hence, investigating applications utilizing current and real-time data, particularly in E-commerce and related industries, represents distinct avenues of research emphasis, with internet connectivity assuming a crucial role.

ACKNOWLEDGEMENT

The authors gratefully acknowledge the support of TUIK, the Turkish Statistical Institute, for providing data from the Information and Communication Technology (ICT) Usage Survey on Households and Individuals.

APPENDIX

NUTS Statistical Regions of Turkey



TR1	Istanbul Region	TR7	Central Anatolia Region
TR2	West Marmara Region	TR8	West Black Sea Region
TR3	Aegean Region	TR9	East Black Sea Region
TR4	East Marmara Region	TRA	Northeast Anatolia Region
TR5	West Anatolia Region	TRB	Central East Anatolia Region
TR6	Mediterranean Region	TRC	Southeast Anatolia Region

REFERENCES

Aggarwal, C.C. (2014) Data Classification: Algorithm and Applications, CRC Press, New York.

Boddin, Dominik. 2016. "The Role of Newly Industrialized Economies in Global Value Chains." IMF Working Papers 16 (207): 1. https://www.imf.org/external/pubs/ft/wp/2016/wp16207.pdf (Accessed August 15, 2018).

Couper, Mick P. 2000. "Review: Web Surveys: A Review of Issues and Approaches." Public Opinion Quarterly 64 (4): 464–94. https://academic.oup.com/poq/article-pdf/64/4/464/5414708/640464.pdf.

Couper, Mick P. 2000. "Web Surveys." Public Opinion Quarterly 64 (4): 464–94.

Couper, Mick P., Arie Kapteyn, Matthias Schonlau, and Joachim Winter. 2007. "Noncoverage and nonresponse in an Internet survey." Social science research 36 (1): 131–48.

de Leeuw, Edith, and Suzette Mathijsse. 2016. "Professional Respondents in Online Panels | Insights Association." https://www.insightsassociation.org/article/professional-respondents-online-panels (August 6, 2018).

Eckman, Stephanie. 2016. "Does the Inclusion of Non-Internet Households in a Web Panel Reduce Coverage Bias?" Social Science Computer Review 34 (1): 41–58.

Eurobarometer. 2017. "E-Communications and Digital Single Market: Report.: Special Eurobarometer 432." [en]. Publications Office of the European Union. July 27. https://publications.europa.eu/en/publication-detail/-/publication/57889a55-8fb6-11e8-8bc1-01aa75ed71a1/language-en (September 26, 2018).

Fricker, Ronald D. 2008. "Sampling Methods for Web and E-mail Surveys." In The SAGE handbook of online research methods, eds. Nigel Fielding, Raymond M. Lee and Grant Blank. 1 Oliver's Yard, 55 City Road, London England EC1Y 1SP United Kingdom: Sage pub, 195–216.

Groves, R. M., and L. Lyberg. 2010. "Total Survey Error: Past, Present, and Future." Public Opinion Quarterly 74 (5): 849–79.

Groves, Robert M. 2006. "Nonresponse Rates and Nonresponse Bias in Household Surveys." Public Opinion Quarterly 70 (5): 646–75.

Hair, Joseph F., William C. Black, Barry J. Babin, and Rolph E. Anderson. 2014. Multivariate data analysis. Always learning. Harlow, Essex: Pearson.

Hastie, T., Tibshirani, R. and Friedman, J. (2013) The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer, New York.

Hargittai, Eszter, and Yuli P. Hsieh. 2012. "Succinct Survey Measures of Web-Use Skills." Social Science Computer Review 30 (1): 95–107.

Hays, Ron D., Honghu Liu, and Arie Kapteyn. 2015. "Use of Internet Panels to Conduct Surveys." [eng]. Behavior research methods 47 (3): 685–90.

Heerwegh, D., and G. Loosveldt. 2008. "Face-to-Face versus Web Surveying in a High-Internet-Coverage Population: Differences in Response Quality." Public Opinion Quarterly 72 (5): 836–46.

Hosmer, David W., Stanley Lemeshow, and Rodney X. Sturdivant. 2013. Applied logistic regression [eng]. 3rd ed. Wiley series in probability and statistics. Hoboken, NJ: Wiley.

Hwang, Yeong-Hyeon, and Daniel R. Fesenmaier. 2004. "Coverage Error Embedded in Self-Selected Internet-Based Samples: A Case Study of Northern Indiana." Journal of Travel Research 42 (3): 297–304.

Mitchell, T., Buchanan, B., DeJong, G., Dietterich, T., Rosenbloom, P., & Waibel, A. (1990) 'Machine learning', Annual Review of Computer Science, Vol. 4, No. 1, pp.417–433.

Mohorko, Anja, Edith de Leeuw, and Joop Hox. 2013. "Internet Coverage and Coverage Bias in Europe: Developments Across Countries and Over Time." Journal of Official Statistics 29 (4): 609–22.

Mossberger, Karen, Caroline J. Tolbert, and Ramona S. McNeal. 2010. Digital citizenship: The internet, society, and participation. 2nd ed. Cambridge, Mass.: MIT Press.

Pew Research Center. 2015. "Coverage Error in Internet Surveys: Who Web-Only Surveys Miss and How That Affects Results." Pew Research Center. 20150922. http://www.pewresearch.org/2015/09/22/coverage-error-in-internet-surveys/ (July 31, 2018).

Peytchev, Andy, Lisa R. Carley-Baxter, and Michele C. Black. 2011. "Multiple Sources of Nonobservation Error in Telephone Surveys: Coverage and Nonresponse." Sociological Methods & Research 40 (1): 138–68.

Robinson, Laura, Shelia R. Cotten, Hiroshi Ono, Anabel Quan-Haase, Gustavo Mesch, Wenhong Chen, Jeremy Schulz, Timothy M. Hale, and Michael J. Stern. 2015. "Digital inequalities and why they matter." Information, Communication & Society 18 (5): 569–82.

Schaefer, David R., and Don A. Dillman. 1998. "Development of a Standard E-Mail Methodology: Results of an Experiment." Public Opinion Quarterly 62 (3): 378.

Selwyn, Neil. 2004. "Reconsidering Political and Popular Understandings of the Digital Divide." New Media & Society 6 (3): 341–62.

SPSS Statistics, version 25 (IBM Corp., Armonk, NY, USA)

Stern, Michael J., Alison E. Adams, and Shaun Elsasser. 2009. "Digital Inequality and Place: The Effects of Technological Diffusion on Internet Proficiency and Usage across Rural, Suburban, and Urban Counties." Sociological Inquiry 79 (4): 391–417.

Stern, Michael J., Ipek Bilgen, and Don A. Dillman. 2014. "The State of Survey Methodology." Field Methods 26 (3): 284–301.

Sterrett, David, Dan Malato, Jennifer Benz, Trevor Tompson, and Ned English. 2017. "Assessing Changes in Coverage Bias of Web Surveys in the United States." Public Opinion Quarterly 81 (S1): 338–56.

Tourangeau, Roger, Frederick G. Conrad, and Mick Couper. 2013. The science of web surveys [eng]. Oxford: Oxford Univ. Press.

TUIK: Turkish Statistical Institute. Information and Communication Technology (ICT) Usage Survey on Households and Individuals, 2017.

van Deursen, A.J.A.M., and J.A.G.M. van Dijk. 2009. "Improving digital skills for the use of online public information and services." Government Information Quarterly 26 (2): 333–40.

Vicente, Paula, and Elizabeth Reis. 2012. "Coverage Error in Internet Surveys: Can Fixed Phones Fix It?" International Journal of Market Research 54 (3): 323–45.

Worldbank. 2018. "Individuals using the Internet (% of population) - World Bank Open Data." https://data.worldbank.org/indicator/IT.NET.USER.ZS (August 7, 2018).

Yeager, David S., Jon A. Krosnick, LinChiat Chang, Harold S. Javitz, Matthew S. Levendusky, Alberto Simpser, and Rui Wang. 2011. "Comparing the Accuracy of RDD Telephone Surveys and Internet Surveys Conducted with Probability and Non-Probability Samples." Public Opinion Quarterly 75 (4): 709–47.

Yun, Gi W., and Craig W. Trumbo. 2000. "Comparative Response to a Survey Executed by Post, E-mail, & Web Form." Journal of Computer-Mediated Communication 6 (1): 0.

Zhang, Y. (2021). An interactive machine learning approach to integrating physician expertise into delirium prediction model development (Order No. 28770345). Available from ProQuest Dissertations and Theses Global. (2607315462). Retrieved from https://www.proquest.com/dissertations-theses/interactive-machine-learning-approach-integrating/docview/2607315462/se-2