AUSTRALIAN
SCHOOL OF BUSINESS
THE UNIVERSITY OF NEW SOUTH WALES

The University of New South Wales
Australian School of Business

School of Economics Discussion Paper: 2010/29

Effect of Internet Health Information on Health Care Use

Agne Suziedelyte

School of Economics
Australian School of Business
UNSW Sydney NSW 2052 Australia
http://www.economics.unsw.edu.au

ISSN 1837-1035
ISBN 978-0-7334-2999-6

Effect of Internet health information on health care use

Agne Suziedelyte*

School of Economics, University of New South Wales, Australia

Australian School of Business Building, Level 4, 430B, UNSW Sydney NSW 2052, Australia

This version 25 October 2010

Abstract

This study estimates the effect of Internet health information on health care utilisation. The causal variable of interest is a binary variable that indicates whether or not an individual has used the Internet to search for health information. Health care utilisation is measured by an individual's number of visits to a health professional. I use the variation in telecommunication laws of U.S. states as a novel instrument to identify the causal effect. The analysis results show that, on average, using the Internet as health information source increases the utilisation of health care. The effect is quantitatively large and precisely estimated. An ordinary least squares regression underestimates the effect, even after controlling for a number of observed individual characteristics.

JEL classification: I1

Keywords: Health care; Health information; Internet

*Tel.: + 61 2 9385 6569. *E-mail address:* agne@unsw.edu.au

1 Introduction

The aim of this paper is to analyse how the health information that people obtain from the Internet affects their health care utilisation. To capture this effect, I construct a binary variable that indicates whether or not an individual has recently searched for health information on the Internet. Health care utilisation is measured by the number of visits to a health professional in the past 12 months. Since the probability of using the Internet to search for health information is likely to endogenous, I use instrumental variable estimation methods. Survey data shows that an individual's probability of searching for health information online is related to having high-speed Internet access at home (Fox and Jones, 2009, p.8). As an instrument, therefore, I use U.S. state telecommunication policies that are shown to affect the supply of high-speed Internet services. I find that Internet health information has a positive economically and statistically significant effect on health care utilisation. In particular, using the Internet to search for health information increases the annual number of visits to a health professional by approximately 3.6 visits, holding other factors fixed.

This research is motivated by the observation that a large proportion of population in developed countries use the Internet as a health information source. According to Pew Internet and American Life Project estimates, 61% of the U.S. adult population looked online for information about a health or medical issue in 2008 (Fox and Jones, 2009, p.2). The health information that people access on the Internet is likely to influence their health related decisions, including the use of health care services. Indeed, 53% of the respondents who looked for health information online reported that this information had a major or minor impact on their own health care or the way they care for someone else (Fox and Jones, 2009, p.27). Within this group, 35% say that the information obtained online affected their decision to see a doctor (Fox and Jones, 2009, p.28). Nevertheless, this effect is not necessarily causal. Other factors may have played a role in an individual's decision to visit a health professional.

I analyse the relationship between Internet health information and health care consumption in a framework introduced by Grossman (1972). In this model, the demand for health care is derived from the demand for good health. Consumers produce health by combining their own time and medical care. Individuals with different characteristics may have different marginal products of these inputs. Therefore, the demand for medical care and health may vary with personal characteristics. If an individual's health knowledge affects the (perceived)

marginal products of time and/or medical care, people who use the Internet to search for health information will have different health care consumption from those who do not.

The relationship between health knowledge and the demand for health care depend on the model assumptions. Both positive and negative effects are plausible. If we assume that health knowledge increases the marginal products of health care and own time, the people with more health knowledge would demand more health but less health care. I find a positive relationship between health knowledge and health care consumption, which implies that health knowledge reduces the perceived marginal product of health care. People with more health knowledge believe that an additional visit to a health professional produces less health compared to the people with less health knowledge.

In medical sociology, there are two opposing hypotheses related to technological advancement and physician-patient contact, summarised by Lee (2008). One hypothesis is that by diffusing health knowledge previously available only to health professionals, technological advancement weakens health professionals' control over their knowledge base; thus, the Internet might reduce people's dependence on health professionals as a source of health information and lower the frequency of physician-patient contact (Lee, 2008, p.451). The second hypothesis states that, despite people's access to physicians' knowledge base, the knowledge gap between the general public and health professionals remains, since new information constantly emerges and is first available to health professionals. Furthermore, health and medical information involves uncertainty and error; therefore, people rely on health professionals for the interpretation and application of health information. As a result, increasing access to health information on the Internet might increase the frequency of health professional contact (Lee, 2008, p.452). The results of my study provide support for the second hypothesis.

Lee (2008) also investigates the relationship between health information search on the Internet and health professional contact. He addresses endogeneity by exploring panel nature of the data rather than using instrumental variable methods. The results of Lee's analysis are consistent with those of my study; he finds that Internet use for health information has a positive effect on health professional contact. The consistency of the two sets of results suggests that the positive causal relationship between Internet health information and health care consumption is a robust finding.

To my knowledge, this is the first study in economics literature that focuses on the effect of Internet health information on health care utilisation. The other related studies only look

at how health care consumption is affected by health knowledge or information in a general sense. I briefly summarise their findings. Kenkel (1990) and Hsieh and Lin (1997) find that the knowledge about the symptoms of certain diseases has a significant positive effect on the probability of visiting a physician. Kenkel (1990), however, finds no significant effect on the demand for physician visits conditional on at least one visit. Wagner et al. (2001b) show that access to new health information is negatively related to the number of phone calls to physicians. On the other hand, the authors do not find any statistically significant effect of health information on the number of physician visits (Wagner et al., 2001a).

2 Empirical strategy

To determine the effect Internet health information on health care utilisation, I estimate the following reduced form demand for health care model:

$$\begin{aligned} HC_i &= E(HC_i|eHI_i, \mathbf{x_{1i}}, \mathbf{t_i}, c_i; \beta) + \varepsilon_i \\ &= f(eHI_i, \mathbf{x_{1i}}, \mathbf{t_i}, c_i; \beta) + \varepsilon_i, \end{aligned} \tag{1}$$

where HC_i denotes health care consumption, measured by an individual's number of visits to a health professional in the past 12 months; eHI_i indicates if an individual uses the Internet to obtain health information or not; $\mathbf{x_{1i}}$ is the $1 \times k$ vector, containing unity and socio-demographic characteristics; $\mathbf{t_i}$ is the $1 \times l$ vector of time variables that account for the changes in health care utilisation over time; and c_i denotes the omitted heterogeneity. Function f maps eHI_i, $\mathbf{x_{1i}}$, $\mathbf{t_i}$, and c_i to HC_i; β is the $m \times 1$ parameter vector ($m = 1+k+l$). The regression error ε_i is mean-independent of eHI_i, $\mathbf{x_{1i}}$, $\mathbf{t_i}$, and c_i.

Omitted heterogeneity c_i could contain such characteristics as an individual's health status, education, household income, trust in his physician, concern about his health, and confidence in his ability to care for own health. These factors are omitted from the model for two reasons. First, I do not observe some of the variables (such as concern about health or confidence in one's ability to care for own health) in the data. Second, the variables such as health status, education, or household income are likely to be endogenous. If these characteristics are also correlated with the probability of searching for health information online, their inclusion in the model will lead to inconsistent estimation of the effect of Internet health information. The data used in this analysis shows that this is the case: individuals that are more educated,

have higher household income, or are less healthy are more likely to use the Internet to search for health information.

To treat eHI_i as an exogenous variable, we must assume that it is uncorrelated with c_i. It is unlikely to be true. As mentioned above, we known that some of the individual characteristics in c_i (health status, education and household income) are correlated with the probability of using the Internet to search for health information. The unobserved characteristics, included in c_i, may also be correlated with eHI_i. For example, people who are more concerned about their health might be more likely to search for health information on the Internet.

To address the endogeneity of eHI_i, I use instrumental variable (IV) estimation methods. This approach requires data on at least one variable excluded from equation (1) that is correlated with eHI_i, but not with c_i and ε_i. This variable must affect the probability of searching for health information on the Internet (the relevance assumption) and must not be correlated with the omitted heterogeneity (the exogeneity assumption). Let $\mathbf{z_{1i}}$ denote the $1 \times p$ vector of these variables ($p \geq 1$).

First, I specify the function f in (1) to be linear and estimate the model using the two-stage least squares (2SLS) estimator. In the second specification, f is assumed to be exponential to account for the fact that the dependent variable, which is measured in the number of health professional visits, takes on only nonnegative values:

$$HC_i = \exp(\mathbf{x_i}\beta)\eta_i + \varepsilon_i, \tag{2}$$

where $\eta = \exp(c_i)$ and $\mathbf{x_i} = (eHI_i, \mathbf{x_{1i}}, \mathbf{t_i})$.

I estimate the model (2) using a non-linear IV estimator developed by Mullahy (1997). Let $\mathbf{z_i}$ be the $1 \times s$ vector of the instrumental variables ($\mathbf{z_i} = (\mathbf{x_{1i}}, \mathbf{t_i}, \mathbf{z_{1i}})$, $s = k+l+p$). The required assumptions on $\mathbf{z_i}$ are that $E(HC_i|\mathbf{x_i}, \mathbf{z_i}) = E(HC_i|\mathbf{x_i})$, $E(\eta_i|\mathbf{z_i}) = 1$, and $E(\varepsilon_i|\mathbf{x_i}, \mathbf{z_i}, \eta_i) = 0$. To obtain a consistent IV estimator, the model has to be transformed by dividing both sides of equation (2) by $\exp(\mathbf{x_i}\beta)$:

$$\exp(-\mathbf{x_i}\beta)HC_i = \eta_i + \exp(-\mathbf{x_i}\beta)\varepsilon_i. \tag{3}$$

If we define $T(HC_i, \mathbf{x_i}; \beta) = \exp(-\mathbf{x_i}\beta)HC_i$ and $v_i = \eta_i - 1$, we can rewrite the equation (3) as:

$$T(\mathbf{x_i}; \beta) - 1 = v_i + \exp(-\mathbf{x_i}\beta)\varepsilon_i. \tag{4}$$

$T(\mathbf{x_i}; \beta) - 1$ is a residual function of the transformed model (3). Mullahy establishes the consistency of the estimator by showing that the residual function satisfies the conditional moment restriction $E[T(\mathbf{x_i}; \beta) - 1|\mathbf{z_i}] = 0$:

$$\begin{aligned} E[T(\mathbf{x_i}; \beta) - 1|\mathbf{z_i}] &= E[v_i|\mathbf{z}] + E[\exp(-\mathbf{x_i}\beta)\varepsilon_i|\mathbf{z_i}] \\ &= 0 + E_x[\exp(-\mathbf{x_i}\beta)E[\varepsilon_i|\mathbf{x_i}, \mathbf{z_i}]|\mathbf{z_i}] \\ &= 0. \end{aligned} \tag{5}$$

The first term of equation (5) is equal to zero because of the assumption $E(\eta_i|\mathbf{z_i}) = 1$; the second term is zero by iterated expectations and the assumption $E(\varepsilon_i|\mathbf{x_i}, \mathbf{z_i}, \eta_i) = 0$. We can obtain the estimates of β by using this conditional moment restriction as the basis for generalised methods of moments estimation. I estimate the model in Stata 11 using the user written command ivpois (Nichols, 2007).

3 Variables and data

3.1 Data

I use the U.S. Health Information National Trends Survey (HINTS) data for the analysis. The HINTS is a repeated cross-sectional survey of the U.S. civilian non-institutionalised adult population. The National Cancer Institute manages and funds the survey. The Institute is a part of the National Institutes of Health, a medical research agency of the U.S. Department of Health and Human Services. Although the main purpose of the survey is to collect data about the public's use of cancer-related information, the HINTS contains questions about the exposure to, and search for, general health information in different media, including the Internet. The survey additionally asks respondents about their health care utilisation. Availability of these key variables makes this data set suitable for analysing the relationship

between Internet health information and health care utilisation. To my knowledge, this is the first economic study that uses the HINTS data.

Until recently, the HINTS data have been collected mainly via telephone interviews. The telephone sample is drawn from all telephone exchanges in the U.S. One randomly selected adult (18 years or older) is interviewed in each household. In the last survey, mail questionnaires have supplemented telephone interviews to increase coverage and reach people who do not use a landline telephone. The mail sample is drawn from the national listing of addresses. All adult household members are asked to fill in the mail questionnaires. To produce reliable estimates for minority groups, stratified random sampling is used, with over-sampling of the units from the stratum with a higher proportion of black and Hispanic population[1]. The data is available for the years 2003, 2005, and 2008[2]. To increase the precision of the parameter estimates, I pool the data over all three years.

The HINTS response rates vary from 20.83% to 33.05%. Other surveys on health-related Internet use have similar response rates; for example, the response rate of the Pew Internet & American Life survey in 2008 was 21% for the landline telephone sample and 25% for the cellular phone sample (Fox and Jones, 2009, p.72). Nevertheless, relatively low response rates might lead to a final sample that is not nationally representative of the population. To address this concern, I compare the mean socio-demographic characteristics in the HINTS and the American Community Survey (ACS) samples (Table 1). The ACS has much higher response rates and is more likely to be nationally representative.

[Table 1 about here.]

The most notable difference between the two samples is the gender distribution: males are under-represented in the HINTS sample. In 2005 and 2008, the HINTS sample appears to be on average older. However, part of the difference in the mean age might be explained by the top-coding of the age variable in the ACS. The proportion of white respondents is similar in both samples, except for the year 2005, when it is higher by 5ppt in the HINTS sample. There are no substantial differences in the fraction of married (or living as married) individuals between the HINTS and the ACS samples with an exception of the year 2003, when this figure is lower by 4ppt in the HINTS sample. The differences in some of the demographic characteristics between the two samples suggest that the HINTS sample is

[1]No over-sampling was done in HINTS 2005.

[2]The National Cancer Institute refers to 2008 data as HINTS 2007, although it were collected in the beginning of 2008; I refer to this data set as HINTS 2008.

not entirely representative of the U.S. population. Given these differences, I include socio-demographic characteristics in the model. As long as the assumption that the selection into the HINTS sample is not based on the endogenous variables is maintained, the reported parameter estimates are consistent. To address a threat to homoskedasticity assumption, standard errors are specified to be robust in all estimations. To present the nationally representative descriptive statistics of variables, I use the sampling weights included in the HINTS data set. The sampling weights account for the fact that the probability of being in the sample varies across the socio-demographic groups.

3.2 Health care utilisation and Internet health information variables

In this analysis, health care utilisation is measured by an individual's number of visits to a health professional within a 12 month period. In the survey, respondents are asked: "During the past 12 months, not counting times when you went to an emergency room, how many times did you go to a doctor, nurse, or other health professional to get care for yourself?" Thus, the definition of a health professional is broad: it includes doctors, nurses, and other health professionals of any specialisation with the possible exception of mental health. In the 2003 and 2008 surveys, the question about an individual's visits to a health professional follows a question "*Not including psychiatrists and other mental health professionals*, is there a particular doctor, nurse, or other health professional that you see most often?" Therefore, some respondents of these surveys may have believed that visits to a mental health professional should be excluded when reporting the number of visits to a health professional. In the 2005 survey, respondents are not asked this question and could be more likely to include visits to a psychiatrist when reporting the number of health professional visits.

Possible answers to the survey question on the number of health professional visits are "None", "1 time", "2 times", "3 times", "4 times", "5-9 times", and "10 or more times". To perform the estimations, the health care utilisation variable needs to be recoded by assigning values to the last two categories. In the baseline specification, the variable takes the value 7 if the answer is "5-9 times", which is a midpoint of the interval. It takes the value 12 if the answer is "10 or more times", which corresponds to visiting a doctor once a month on average. Later, I check the sensitivity of the results to the dependent variable coding.

Figure 1 presents the estimated distribution of the number of health professional visits for the U.S. population in 2003, 2005, and 2008. Most of the population do not use health professional services frequently: around two thirds of the population make 3 or less visits a year. Slightly less than 20% of the population visit a doctor, a nurse, or other health professional less than once a year. On the other hand, around a quarter of the population are frequent users of health professional services; slightly less than 15% visit a health professional 5-9 times per year and above 10% make 10 or more visits per year.

[Figure 1 about here.]

Figure 1 also shows that the proportion of infrequent users of health professional service (none or 1 visits) has decreased and the fraction of the frequent users (5 or more visits) has increased over time. The year 2005 stands out with a sharp increase in a proportion of the population with 10 or more health professional visits. In 2008, however, the proportion of the population in this category decreases again. A possible explanation for this trend is that the definition of a health professional is interpreted differently in the 2005 survey compared to the 2003 and 2008 surveys, as mentioned above. The estimated mean number of health professional visits in the population is 3.50 for 2003, 4.10 for 2005, and 3.74 for 2008. These estimates are based on the assumption that the mean of the category "5 to 9 visits" is 7 and the mean of the category "10 or more visits" is 12.

To capture the effect of Internet health information, I use a binary variable indicating whether a respondent has recently searched for health information on the Internet for himself. The availability of this variable enables me to answer the question if access to online health information affects individuals' decision to visit a health professional. I do not have data on the quantity and content of health information or how often a respondent has searched the Internet; therefore, I do not aim to answer the question if and how the effect of online health information on health professional visits varies with information volume and type.

The survey question that I use to construct the Internet health information variable varies over the survey years. The 2003 and 2005 surveys ask respondents directly whether they have used the Internet to look for health or medical information for themselves in the past 12 months. If the answer is "yes", the Internet health information variable takes the value 1; otherwise, it takes the value 0. The 2008 survey asks respondents whether they have used the Internet in their most recent search for information about health or medical topics and whether this information was for themselves or someone else. The Internet health information variable takes the value 1 if a respondent reports that he has used the Internet in his most

recent health information search and indicates that this information was for himself, and the value 0 otherwise.

Figure 2 shows what percentage of the U.S. population is estimated to have used the Internet to search for health information for themselves in 2003, 2005, and 2008. In 2003, the proportion of the population who have used the Internet as a health information source is slightly less than a third. This proportion increases to 35% in 2005 and further rises to 37% in 2008. These estimates are lower than those of the Pew Internet and American Life Project (61% in 2008), since I focus only on the Internet use that is likely to affect health professional visits. My estimates exclude the people who have used the Internet to look for health information for somebody else. I also exclude information about diet, nutrition, exercise, and quitting smoking from the definition of "health or medical information" in this analysis. Furthermore, the Pew Internet and American Life Project estimates include people who have looked for information about a particular doctor, health insurance, and Medicare or Medicaid. It is unlikely that the respondents of the HINTS would consider these issues as "health or medical information".

[Figure 2 about here.]

Table 2 compares individuals who have and have not used the Internet to search for health information according to a number of characteristics. People who have looked for health information on the Internet are relatively younger and more likely to be female, white, and married. Individuals in this group assess their health as better: the proportion of people with good and very good self-assessed health is larger and the proportion of people with fair or poor health is smaller. There are no big differences in other health outcomes between the two groups. Individuals who have used the Internet to search for health information are more likely to live in metropolitan counties and be members of higher income households, which could be associated with better Internet access. They are more likely to have health care coverage. A larger proportion of these individuals are employed and the proportion of students, retirees, homemakers, and disabled people is smaller. People who have searched for health information online are better educated and more likely to have a college degree, which could be related to better Internet and information search skills. The two groups are different in terms of health behaviour. People who have used the Internet as a health information source are more likely be non-smokers, exercise, and consume larger quantities of vegetables and fruits, which could indicate that they are more concerned about their health.

[Table 2 about here.]

Figure 3 presents the distribution of the number of visits to a health professional for people who have used the Internet to search for health information and for those who have not. The proportion of individuals with no visits is substantially lower in the first group (10.0% compared to 20.8%). This group also has a higher proportion of frequent users with 5-9 or 10 or more health professional visits. There are no major differences in the other categories. The mean number of visits for people who have searched for health information on the Internet is 4.42. The mean for those who have not used the Internet for this purpose is approximately 1 visit lower. The difference in the means is statistically significant at 1% level. The comparison of the distributions suggests that health care consumption, as measured by the number of visits to a health professional, is higher among people who have used the Internet to look for health information. To conclude that this relationship is causal, however, we need to account for the heterogeneity of the two groups, as the selection into the user group is not random.

[Figure 3 here.]

3.3 Instrumental variable

To find an instrument for the Internet health information variable, I have looked at the factors that influence an individual's probability of searching for health information online. Access to affordable high-speed Internet service is one of these factors. According to Pew Internet and American Life Project estimates, 88% of people with broadband access looked for the information about health or medical issues online in 2008, and the corresponding figure for dial-up Internet users is only 72% (Fox and Jones, 2009, p.8). The type of Internet connection is not, however, a valid instrument, since it does not satisfy the exogeneity assumption, which requires that the probability of having high-speed Internet connection would be uncorrelated with the omitted heterogeneity. The HINTS 2008 data shows that controlling for socio-demographic characteristics, individuals with higher education and higher household income are more likely to have a high-speed Internet connection (DSL, cable, or wireless) at home. Another instrument I have considered is an indicator whether an individual has children under 18 in the household, which might be associated with a higher demand for the high-speed Internet. However, the probability of having children under 18 is also correlated with education and household income, which violates the instrument exogeneity requirement.

Variables that affect high-speed Internet supply are less likely to be correlated with the omitted heterogeneity and, therefore, are potential instruments. The ease of access to public rights-of-way[3] is one of the factors influencing high-speed Internet penetration. The participants of the Broadband Forum, conducted by the National Telecommunications and Information Administration, have cited rights-of-way issues as having a major impact on broadband deployment (NTIA, 2003b). The National Association of Regulatory Utility Commissioners has recognised that "the rights-of-way practices of certain governmental entities have emerged as a barrier to the deployment of advanced telecommunications and broadband networks" (The Study Committee on Public Rights-of-Way, 2002, p. i). The right-of-way regulations vary across the U.S. states. Wallsten (2005) empirically analyses how different state policies are associated with broadband penetration (as measured by the number of broadband subscribers per capita in a state). He finds that the states that specifically grant telecommunication firms access to public rights-of-way have higher broadband penetration (Wallsten, 2005, p. 11).

I use information on U.S. states' right-of-way policies to construct the instrument for the Internet health information variable. It is a binary variable that takes the value 1 if an individual resides in a state where right-of-way regulations are more favourable to telecommunication providers and the value 0 otherwise. I consider that a state has more favourable right-of-way regulations if it explicitly grants telecommunication firms access to local or state public rights-of-way or restricts local governments' authority or both. Given the discussion above, I expect the instrument to be positively associated with the probability of using the Internet to look for health information. Information about right-of-way regulations comes from a survey of U.S. state laws conducted by the National Telecommunications and Information Administration (NTIA, 2003a). I verify and supplement this information using the official state statutes (FindLaw, 2009). Based on my definition, 38 states and the District of Columbia have right-of-way policies that are relatively more favourable to telecommunication providers. Appendix 1 lists these states.

The first stage results, presented in Table 3, support the instrument relevance assumption. Holding other factors fixed, people residing in the states with more favourable right-of-way policies are indeed more likely to search for health information on the Internet. The coefficient on the instrument is statistically significant. The probability of looking for health

[3]Right-of-way is the privilege of someone to pass over land belonging to someone else.

information on the Internet is also positively associated with younger age and being female, white, and married, holding other factors fixed.

Although the instrument is highly statistically significant, the coefficient on the Internet health information variable is not estimated precisely in the full sample. To increase the precision of this estimate, I exclude from the sample the states with low proportion of urban population and low population density, where the right-of-way policies are unlikely to have an effect on high-speed Internet supply. The justification for this exclusion is that the access to public rights-of-way is a more important issue in cities and towns than in rural areas. Although Wallsten (2005) finds a significant relationship between right-of-way policies and broadband penetration in general, he concludes that these policies do not matter for rural broadband penetration (as measured by the number of rural dwellers in zip codes with at least one broadband provider). In areas with low population density, the major obstacle to broadband deployment is cost. Therefore, the ease of right-of-way access is likely to be a secondary issue in these areas.

The restricted sample includes the states with a higher than median proportion of urban population (71.6%) and higher than median population density (88.6 persons per square mile) in the year 2000. The sample contains 14,518 observations (2,891 less than the full sample). Appendix 2 shows which states are excluded from the sample. Table 3 demonstrates that the instrument is more relevant in the restricted sample, as expected. The coefficient estimate is 0.057 in the restricted sample compared to 0.050 in the full sample. In addition, the F-statistic is larger (46.08 compared to 36.03). The restriction of the sample results in a precise estimate of the effect of Internet health information. I also present results for the full sample and check the sensitivity of the results to using different cut-off points (mean and lower quartile).

[Table 3 about here.]

Although we cannot test for instrument exogeneity, we can address potential threats to exogeneity. This assumption would be violated if the right-of-way policies were associated with other policies affecting health care service use. In this case, the states with more favourable right-of-way policies would have lower or higher expenditures on health services. Therefore, I compare the personal health care expenditures (as a percentage of gross state product) in the two groups of states using the medical expenditure estimates for 1991, provided by The Centers for Medicare and Medicaid Services (CMS, 2007). There is no

statistically significant difference in the mean expenditures between the two groups of states (t-statistic = 0.360), which provides support for the exogeneity assumption.

If we believe that the effect of Internet health information is heterogenous (it varies across individuals), a 2SLS estimate is a local average treatment effect (LATE). In this analysis, it is an average effect of Internet health information on health care utilisation for people who search for such information because their state makes it easier for Internet providers to access rights-of-way. The LATE literature refers to them as compliers. This group is smaller than the total population. It excludes individuals who would use the Internet to look for health information irrespective of right-of-way policies. This group also excludes people who would not search for health information online regardless of right-of-way policies. Nevertheless, this sub-population is interesting for policymakers. Since right-of-way regulations affect behaviour of individuals in this group, other policies with similar incentives might influence their choices. For example, it is likely that people in this group are price sensitive. Therefore, the results of this study could be used for predicting the effects of policies that aim to influence the use of Internet health information by means of price incentives.

4 Results

4.1 Main results

Table 4 reports the 2SLS estimates. As the instrument is a state-level variable, I cluster errors by state in all estimations. The results show that Internet health information has a positive quantitatively large effect on health care utilisation. Using the Internet to look for health information increases the expected number of health professional visits by around 3.6 visits per year, holding other factors fixed. The effect is precisely estimated. The Internet health information variable is statistically significant at the 5% level. We are 95 % confident that the true coefficient value lies between 0.674 and 6.490.

[Table 4 here.]

How can we explain the estimated positive effect of Internet health information? Assuming that health knowledge changes the marginal products of health care and own time by the same percentage, Grossman's model predicts that health knowledge would have a positive effect on health care demand in 2 cases:

1. If the elasticity of the health demand curve were greater than unity and health knowledge increased productivity;

2. If the elasticity of the health demand curve were less than unity and health knowledge decreased productivity (Grossman, 1972, p.246).

Grossman (1972, p.239) suggests that the elasticity of the health demand curve is less than unity. Consequently, the estimated positive effect of Internet health information on health care consumption implies that health knowledge reduces perceived efficiency of health production: more informed consumers believe that an additional health professional visit produces less health investment relative to less informed consumers. The study of Singaporean Internet users somewhat confirms this: people who use the Internet to search for health information believe that "health care professionals can make mistakes" and "do not necessary know enough", whereas people who do not use Internet as a health information source "place great trust on health care professionals" (Tang and Lee, 2006, p.120). As a result, more informed consumers demand more health care. For example, they could visit another health professional for a second opinion.

The estimated positive effect of Internet health information on health care utilisation supports the hypothesis that better access to health information by consumers increases their reliance on health professionals because of the uncertainty and error in Internet health information. Therefore, consumers need health professionals to interpret and verify the information that they find on the Internet. Tang and Lee (2006, p.115) have found that the majority of people who use the Internet to search for health information approach health professionals to verify such information. According to another survey, 60% of people who go online to look for information about health topics rely on such information only if their doctor tells them to do so (Harris Interactive, 2002, p.3).

Moreover, information about health conditions that people find on the Internet might require diagnostic tests or more advanced treatments, and consumers have limited capabilities and resources to treat and test for medical conditions. There is evidence that individuals who more frequently use the Internet for health-related purposes are more likely to suggest diagnosis to their physicians, and request specific treatments (Rice and Katz, 2006, p.159). Although the Internet can substitute for health professionals in some cases, it is not widespread, as the results show. This finding is supported by the survey data: only 11% of people say that they use the Internet instead of speaking with their physician (Rice and Katz, 2006, p.160).

Overall, the results of this analysis suggest that Internet health information is a complement to rather than a substitute for health professionals.

Table 4 shows that the OLS estimate of the effect of Internet health information on health care utilisation is more than 3 times smaller than the 2SLS estimate. If we do not address the endogeneity of the Internet health information variable, we underestimate its effect on health care consumption. The positive difference between the 2SLS and OLS estimates implies that the omitted heterogeneity is positively correlated with the Internet health information variable and negatively correlated with the number of health professional visits or vice versa. A potential source of unobserved heterogeneity is self-efficacy, defined as the confidence of one's ability in performing a task. Tang and Lee (2006, p.120) have observed that Internet users who go online to search for health information feel more confident in handling, discerning, and verifying large amounts of information. On the other hand, greater self-efficacy decreases reliance on health professionals for the interpretation of health information and reduces the need to visit a health professional.

The socio-demographic variables have the expected effects on health care utilisation. The age effect is positive, but small in magnitude. On average, males have almost 1 visit to a health professional per year less than females. Race is statistically insignificant once we control for the other factors. Being married or living as married decreases the number of visits to a health professional slightly. According to the 2SLS estimates, the average number of health professional visits has increased from 2003 to 2005, but there are no significant differences between years 2003 and 2008. Nevertheless, the increase in 2005 could be explained by the different interpretation of "a health professional" in 2005 survey, as discussed in Section 3.2.

Table 5 presents Mullahy's non-linear IV and Poisson estimates for the Internet health information variable. To compare the results to the linear model estimates, I also report the average predicted effect for the individuals who currently do not use the Internet to look for health information. The average predicted effect is qualitatively similar to the 2SLS estimate: it is positive and quantitatively large. The non-linear IV estimate is, however, less precise. Consistent with the linear model results, the Poisson model underestimates the effect of Internet health information, as it fails to account for endogeneity.

[Table 5 about here.]

4.2 Sensitivity analysis

In this section, I check the sensitivity of the results to the assumptions made in the analysis. First, I report the results for the unrestricted sample and the samples constructed using different cut-off points. The baseline results are based on the sample of U.S. states with a proportion of urban population and population density higher than median, as the right-of-way policies are more effective in such states. Alternative cut-off points are the lower quartile and the mean. There are 4 states where the proportion of urban population and population density is lower than the lower quartile. If we use the mean as a cut-off point, we exclude 19 states, many of which are the same as those excluded from the baseline sample. The full sample includes the residents of all states (17,409 observations).

[Table 6 about here.]

Table 6 presents the 2SLS estimates for the different samples. The estimated Internet health information effects are qualitatively similar. The estimates in the last 3 columns are smaller in magnitude, but all fall into the 95% confidence interval of the baseline estimate, reported in the first column. The coefficient is statistically significant at the 10 % level in all samples, but the effect of Internet health information is estimated most precisely in the baseline sample. The instrument performs well in all samples (F-statistic is greater than 34).

Second, I check the sensitivity of the results to the coding of the dependent variable. The baseline specification assigns the value 12 to the observations with 10 or more visits to a health professional in the past 12 months. The estimated Internet health information effect increases at a constant rate, as the value assigned the "10 or more visits" category changes from 10 to 15. The coefficient on the Internet health information variable is statistically significant at 5% level in all specifications. Changing the value assigned to the category "5-9 visits" from 7 to 5, 6, 8, or 9 leads to similar observations: a higher value is associated with a larger estimated effect. Thus, assigning the lowest values to both categories gives the lower bound of the Internet health information effect on health care utilisation. Similarly, assigning the highest values to both categories, gives the upper bound of the effect (assuming that the mean number of visits is 15 or less for the observations with 10 or more visits). As Table 7 shows, both bounds are precisely estimated. Thus, recoding the dependent variable does not change the conclusion that Internet health information has a positive economically and statistically significant effect on health care consumption.

[Table 7 about here.]

Next, I verify the assumption that the effect of Internet health information is constant over time, which is implied by equation (1). Table 8 reports the 2SLS estimates for the 2003-2005 and 2008 samples. I pool the 2003 and 2005 data, because using them separately leads to large standard errors and makes it difficult to compare the estimates. The instrument is the same in both estimations. The estimated effect of the Internet health information variable is larger in magnitude in 2003-2005 sample. The difference between the estimates could indicate that the effect of Internet health information on health care utilisation has decreased over time. It could also reflect the questionnaire changes over the years, as explained in Section 3.2.

The estimates for the 2003-2005 and 2008 samples are not, however, qualitatively different. 2003-2005 estimate falls into the 95% confidence interval of the 2008 estimate and vice versa. Moreover, we have to regard the estimates in the last 2 columns with care. The effect of online health information is imprecisely estimated in the 2008 sample. Although the 2003-2005 estimate is statistically significant at the 5 % level, the first stage F-statistic is less than 10, indicating that the instrument is quite weak and the inferences can be misleading. Overall, the data does not provide enough evidence to reject the assumption of the time constant effect.

[Table 8 about here.]

Lastly, I check how the results would differ if we attempted to account for the individual heterogeneity by including the available control variables, rather than using IV estimation. Table 9 presents OLS estimates including the observed individual characteristics and variables that could proxy for some of the unobserved characteristics. Health variables include self-reported physical and mental health status, cancer history, and body mass index. Household income, employment status, health care coverage, and whether respondent lives in metropolitan county describe the access to health care. Health behaviour variables include smoking, exercising, and vegetable and fruit consumption and proxy for the unobserved concern about an individual's own health. Table 2 presents the full list and means of these variables. The sample size in this specification is smaller than in the baseline specification because of the missing values of the added characteristics.

The OLS estimate of the Internet health information effect on health care utilisation changes only slightly, once we control for a number of individual characteristics in addition to main socio-demographic variables (from 1.109 in Table 4 to 0.986 Table 9). It is much lower than the 2SLS estimate (3.58): the difference between the two estimates shows that the available

variables cannot account for the heterogeneity between Internet health information seekers and non-seekers. As a result, the OLS estimator is inconsistent. This exercise shows that reliable estimates of the effect of Internet health information on health care utilisation cannot be obtained without IV or other methods that account for the endogeneity of this variable.

[Table 9 about here.]

5 Discussion and conclusion

In this paper, I aim to determine whether Internet health information has a causal effect on health care consumption. More specifically, I examine if using the Internet to search for health or medical information affects an individual's number of visits to a health professional per year. I find that the effect of Internet health information on health care utilisation is positive, quantitatively large, and precisely estimated.

The estimated positive effect implies that Internet health information is more likely to be a complement rather than a substitute for formal health care. The raw data indicate that health care consumption is higher by approximately 1 visit for people who search for health information online. When I address the endogeneity of the Internet health information variable by using the IV method, the estimated effect increases to approximately 3.6 visits. The difference between the two estimates indicates that people who use the Internet to search for health information have characteristics that negatively correlate with the number of health professional visits. Failing to account for the individual heterogeneity, underestimates the effect of Internet health information on health care utilisation.

For policymakers, these findings suggest a means of influencing the consumption of health care. A possible application of the results is preventive care, an area where higher use of services is desirable, because it reduces the need for resources in the future. A program designed to increase access and exposure to relevant information on the Internet would contribute to achieving this goal. On the other hand, areas where health care is overused would need policies that control the information available on the Internet, because it may add to the problem of overuse of health care. The results of this analysis are especially applicable to policies that use price incentives to influence health information search behaviour.

The major challenges of this analysis have been: (1) obtaining the data that contains both Internet health information and health care utilisation variables; and (2) finding an

instrument for the Internet health information variable. I address the first issue by using the U.S. Health Information National Trends Survey data, which is managed by the National Cancer Institute. Although the main purpose of the survey is to collect cancer related information, it includes questions applicable to this study. The Health Information National Trends Survey could be of interest to researchers in health economics, as it contains some questions that other health surveys do not cover.

It is difficult to find individual-level instrument for the Internet health information variable that satisfies both exogeneity and relevance assumptions. For example, the type of Internet connection at home is a candidate instrument, but it violates the exogeneity assumption: households with high-speed Internet connection are wealthier, more educated and possibly more technology-savvy, which affects their health care consumption. I use a state-level instrument, which is plausibly exogenous. It is a dummy variable describing states' right-of-way policies. High-speed Internet supply is higher in the states with more favourable policies, which leads to a larger proportion of the state population using the Internet to search for health information online. Other studies that analyse effects of the Internet on various outcomes could also use this variable as an instrument.

Looking at how online health information affects different types of health care could be an extension to this study. Reported estimates measure the average effect over various types of health professional visits. It is likely that information obtained from the Internet differently affects visits for preventive care and visits for advice as well as visits for minor and major treatments. For example, finding information about the beneficial effect of cancer screening should encourage an individual to see a doctor; on the other hand, finding the answer to a health question would reduce the need to see a health professional. Analysis that differentiates between the types of health care would provide more detailed picture of the situation.

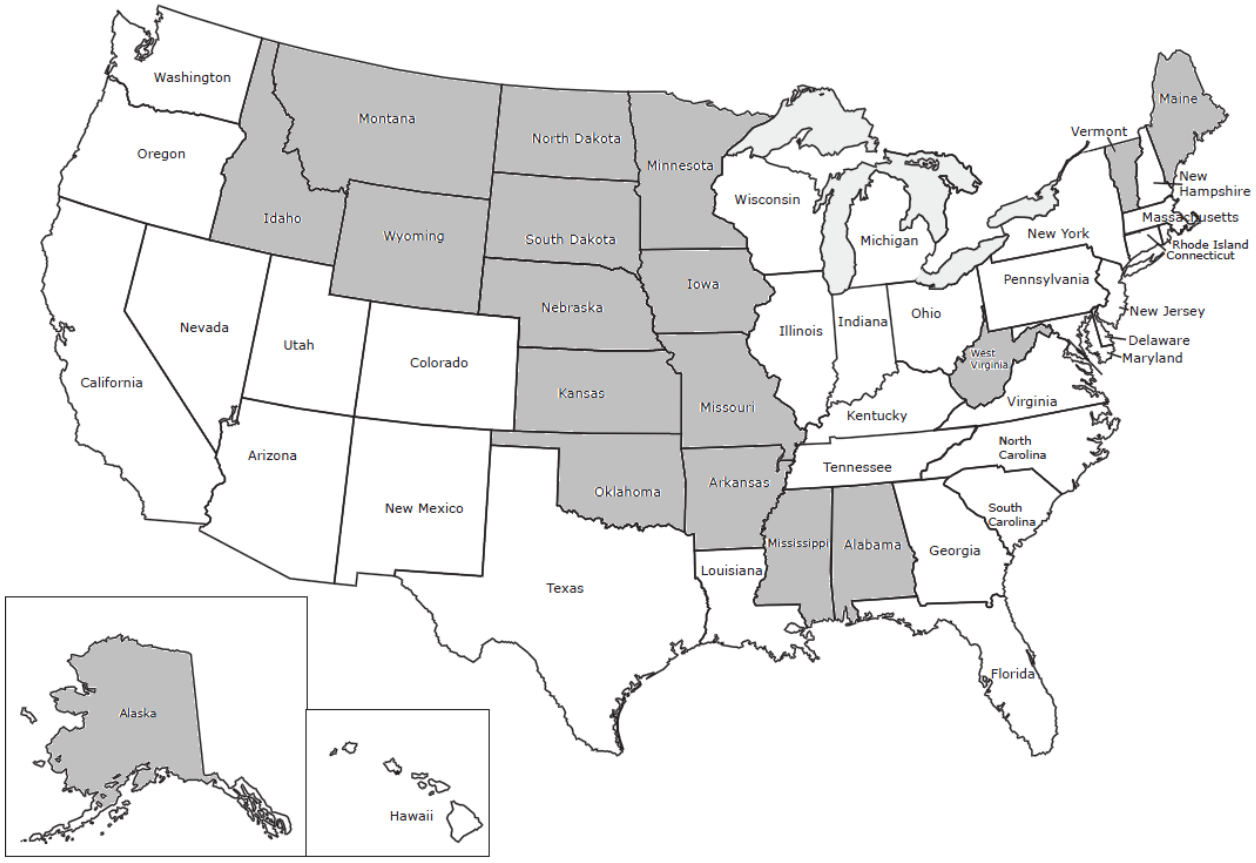
Acknowledgements

The author would like to thank Denise Doiron, Denzil Fiebig, Mark Rosenzweig, Kevin Lang, Hong Il Yoo, and the participants of the 7th Summer Workshop in Health Economics at the University of New South Wales for their helpful comments and advice.

Appendix 1. State right-of-way policies

States with more favourable right-of-way policies	States with less favourable right-of-way policies
Arizona	Alabama
Arkansas	Alaska
California	Hawaii
Colorado	Illinois
Connecticut	Indiana
Delaware	New Hampshire
District of Columbia	New Mexico
Florida	North Dakota
Georgia	Utah
Idaho	West Virginia
Iowa	Wisconin
Kansas	Wyoming
Kentucky	
Louisiana	
Maine	
Maryland	
Massachusetts	
Michigan	
Minnesota	
Mississippi	
Missouri	
Montana	
Nebraska	
Nevada	
New Jersey	
New York	
North Carolina	
Ohio	
Oklahoma	
Oregon	
Pennsylvania	
Rhode Island	
South Carolina	
South Dakota	
Tennessee	
Texas	
Vermont	
Virginia	
Washington	

Appendix 2. States excluded from the restricted sample



References

CMS, 2007. Health expenditures by state of provider: Summary tables, 1980-2004.Centers for Medicare and Medicaid Services, <http://www.cms.gov/NationalHealthExpendData/downloads/nhestatesummary2004.pdf>, accessed on 16 June 2010.

FindLaw, 2009. Official state codes. FindLaw, a Thomson Reuters business, <http://law.findlaw.com/state-laws/state-codes.html>, accessed on 15 June 2010.

Fox, S., Jones, S., 2009. The social life of health information. Pew Internet & Americal Life Project, <http://pewinternet.org/~/media//Files/Reports/2009/PIP_Health_2009.pdf>, accessed on 10 June 2010.

Grossman, M., 1972. On the concept of health capital and the demand for health. Journal of Political Economy 80 (2), 223.

Harris Interactive, 2002. Four-nation survey shows widespread but different levels of internet use for health purposes. Health Care News 2 (11), 1–4.

Hsieh, C.-R., Lin, S.-J., 1997. Health information and the demand for preventive care among the elderly in taiwan. The Journal of Human Resources 32 (2), 308–333.

Kenkel, D., 1990. Consumer health information and the demand for medical care. The Review of Economics and Statistics 72 (4), 587–595.

Lee, C.-J., 2008. Does the internet displace health professionals? Journal of Health Communication: International Perspectives 13 (5), 450 – 464.

Mullahy, J., 1997. Instrumental-variable estimation of count data models: Applications to models of cigarette smoking behavior. Review of Economics and Statistics 79 (4), 586–593.

Nichols, A., 2007. ivpois: Stata module for iv/gmm poisson regression.<http://ideas.repec.org/c/boc/bocode/s456890.html>, accessed on 18 June 2010.

NTIA, 2003a. Rights-of-way laws by state. National Telecommunications and Information Administration, U.S. Department of Commerce, <http://pewinternet.org/~/media//Files/Reports/2006/PIP_Online_Health_2006.pdf.pdf >, accessed on 15 June 2010.

NTIA, 2003b. State and local rights of way success stories. National Telecommunications and Information Administration, U.S. Department of Commerce, <http://www.ntia.doc.gov/ntiahome/staterow/ROWstatestories.htm >, accessed on 23 August 2010.

Rice, R. E., Katz, J. E., 2006. Internet use in physician practice and patient interaction. In: Murero, M., Rice, R. E. (Eds.), Internet and Health Care. Lawrence Erlbaum Associates, Publishers, New Jersey, pp. 149–176.

Tang, E., Lee, W., 2006. Singapore internet user's health information search: Motivation, perception of information sources, and self-efficacy. In: Murero, M., Rice, R. E. (Eds.), Internet and Health Care. Lawrence Erlbaum Associates, Publishers, New Jersey, pp. 107–126.

The Study Committee on Public Rights-of-Way, 2002. Promoting broadband access through public rights-of-way and public lands. The National Association of Regulatory Utility Commissioners, <http://www.naruc.org/Publications/row_summer02.pdf >, accessed on 23 August 2010.

Wagner, T. H., Hibbard, J. H., Greenlick, M. R., Kunkel, L., 2001a. Does providing consumer health information affect self-reported medical utilization? Evidence from the healthwise communities project. Medical Care 39 (8), 836–847.

Wagner, T. H., Hu, T.-w., Hibbard, J. H., 2001b. The demand for consumer health information. Journal of Health Economics 20 (6), 1059–1075.

Wallsten, S., 2005. Broadband penetration: An empirical analysis of state and federal policies. AEI-Brookings Joint Center Working Paper 05-12, <http://www.heartland.org/custom/semod_policybot/pdf/17468.pdf>, accessed on 11 June 2010.

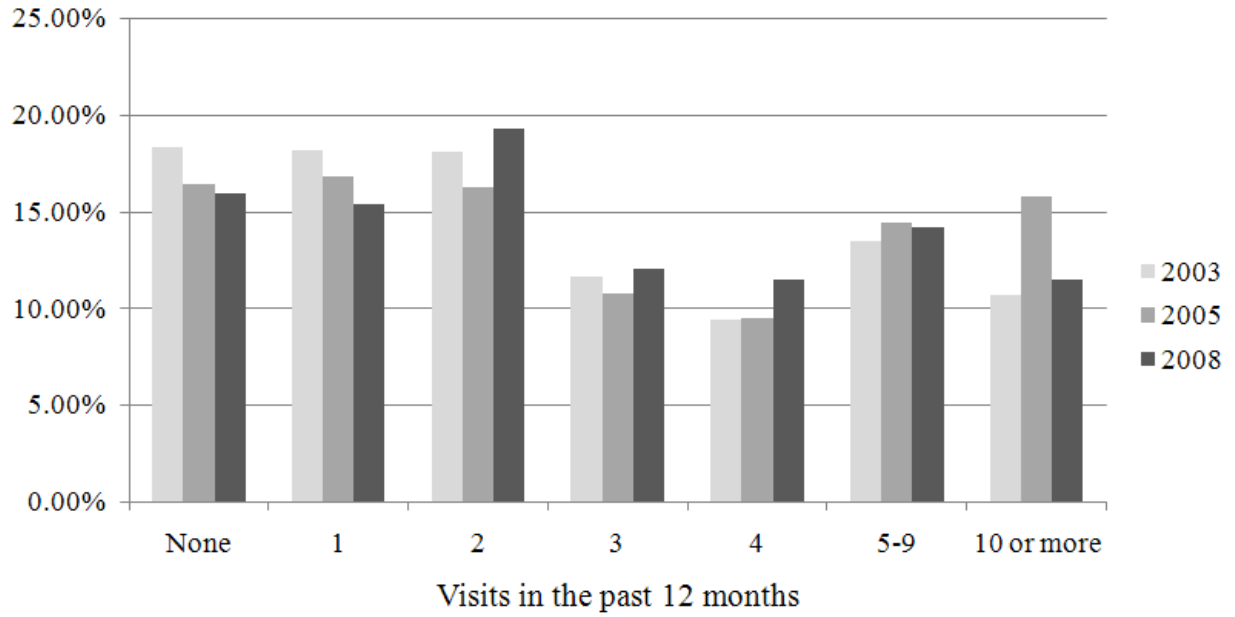


Figure 1: The distribution of the number of visits to a health professional by year. *Notes*: The estimates are calculated using sampling weights. Sample size is 6,339 in 2003, 5,380 in 2005, and 7,595 in 2008.

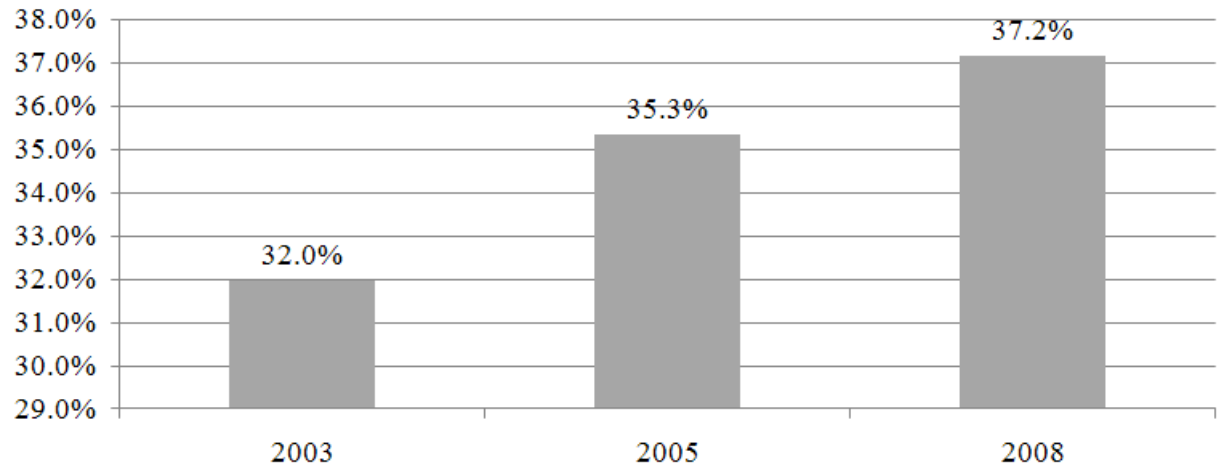


Figure 2: Percentage of the U.S. population who have used the Internet to search for health information for themselves, by year. *Notes*: The estimates are calculated using sampling weights. Sample size is 6,350 in 2003, 5,489 in 2005, and 7,355 in 2008.

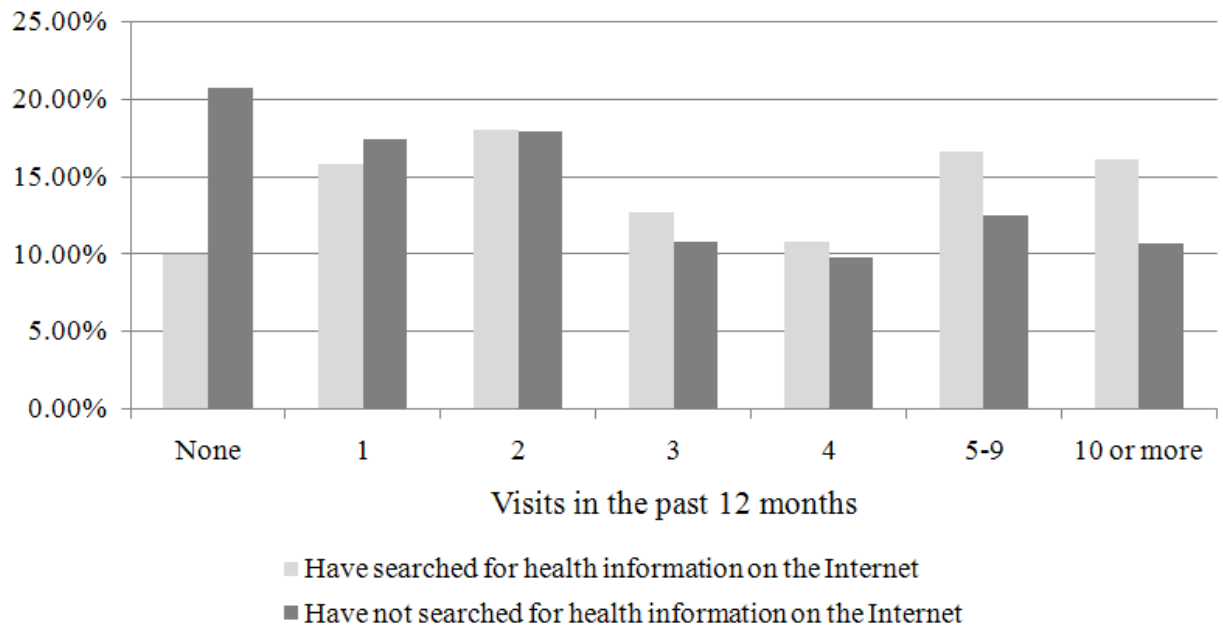


Figure 3: Distribution of the number of visits to a health professional by search for health information on the Internet. *Notes*: The estimates are calculated using sampling weights. Sample size is 6,684 for the first group and 12,203 for the second group.

Table 1: Comparison of the mean socio-demographic characteristics in the Health Information National Trends Survey and the American Community Survey samples

	2003		2005		2008	
	ACS	HINTS	ACS	HINTS	ACS	HINTS
Age	47.76	47.74	48.17	52.17	48.54	54.16
Male	0.47	0.40	0.47	0.35	0.48	0.39
White	0.82	0.80	0.81	0.86	0.80	0.83
Married or living as married	0.60	0.56	0.60	0.58	0.57	0.59

Table 2: Mean characteristics by search for health information on the Internet.

	Looked for health information on the Internet	
	Yes	No
Age	40.82	48.13
Male	0.44	0.53
White	0.85	0.79
Married or living as married	0.65	0.63
Health		
Excellent health	0.12	0.12
Very good health	0.35	0.32
Good health	0.37	0.34
Fair or poor health	0.16	0.21
Psychological distress score (0-24)	4.68	4.82
Had cancer	0.09	0.10
Body mass index	27.04	27.48
Access to health care		
Metro county	0.84	0.78
Household income < $20,000	0.10	0.22
Household income $20,000 to $50,000	0.28	0.38
Household income $50,000 to $75,000	0.23	0.18
Household income > $75,000	0.39	0.22
Employed	0.68	0.58
Unemployed	0.05	0.05
Not in labour force	0.28	0.36
Has health care coverage	0.90	0.84
Education		
Less than high school	0.04	0.17
High school	0.19	0.35
Some college or technical school (1-3yrs)	0.38	0.29
College (4 or more years)	0.38	0.19
Health behaviour		
Smokes everyday or some days	0.19	0.25
Participates in physical activity at least once a week	0.76	0.66
Consumes more than median quantity of fruits	0.45	0.39
Consumes more than median quantity of vegetables	0.38	0.27
N	5457	8781

Note: The estimates are calculated using sampling weights.

Table 3: First stage results, OLS estimates (dependent variable: Internet health information)

	Full sample		Restricted sample	
	Coefficient	Standard error	Coefficient	Standard error
IV	0.050***	0.008	0.057***	0.008
Age	−0.007***	0.000	−0.007***	0.000
Male	−0.053***	0.006	−0.051***	0.006
White	0.093***	0.009	0.099***	0.009
Married	0.061***	0.009	0.063***	0.010
2005	0.046***	0.009	0.042***	0.010
2008	0.074***	0.007	0.073***	0.007
Constant	0.570***	0.017	0.565***	0.018
F-stat (IV)	36.03		46.08	
N	17409		14518	

Notes: Standard errors are clustered by state. A constant is included. Year 2003 dummy is omitted.
*** denotes statistical significance at the 1% level.

Table 4: 2SLS and OLS estimates (dependent variable: the number of visits to a health professional in the past 12 months)

	2SLS		OLS	
	Coefficient	Standard error	Coefficient	Standard error
Internet health information	3.582**	1.484	1.109***	0.053
Age	0.062***	0.011	0.044***	0.003
Male	−0.875***	0.112	−1.001***	0.067
White	−0.286	0.177	−0.048	0.089
Married	−0.268**	0.104	−0.114**	0.055
2005	0.436***	0.086	0.540***	0.074
2008	0.004	0.130	0.184**	0.080
N	14518		14518	

Notes: Standard errors are clustered by state. A constant is included. Year 2003 dummy is omitted.
** denotes statistical significance at the 5% level.
*** denotes statistical significance at the 1% level.

Table 5: Non-linear IV and Poisson estimates for the Internet health information variable (dependent variable: the number of visits to a health professional in the past 12 months)

	Coefficient	Standard error	Average predicted effect
Non-linear IV	0.703	0.506	4.019
Poisson	0.272***	0.013	1.231

Notes: Sample size is 14518. Standard errors are clustered by state and bootstrapped with 200 replications. Both regressions include age, gender, race, marital status, year variables, and a constant. Average predicted effect is calculated for *Internet health information*=0.
*** denotes statistical significance at the 1% level.

Table 6: Sensitivity of results to sample selection, 2SLS estimates (dependent variable: the number of visits to a health professional)

	Cut-off point			
	Median	Lower quartile	Mean	Full sample
Internet health information	3.582**	3.361**	2.872*	2.998*
	(1.484)	(1.642)	(1.646)	(1.692)
No. of states excluded	18	4	19	0
N	14518	17125	14465	17409

Notes: Standard errors are clustered by state and reported in parentheses. All regressions include age, gender, race, marital status, year variables, and a constant.
* denotes statistical significance at the 10% level.
** denotes statistical significance at the 5% level.

Table 7: Sensitivity of results to dependent variable coding , 2SLS estimates (dependent variable: the number of visits to a health professional)

	Lower bound	Baseline estimate	Upper bound
Internet health information	2.811**	3.582**	4.540**
	(1.137)	(1.484)	(1.989)

Notes: Sample size is 14518. Standard errors are clustered by state and reported in parentheses. All regressions include age, gender, race, marital status, year variables, and a constant.
** denotes statistical significance at the 5% level.

Table 8: Variation of the Internet health information effect over time, 2SLS estimates (dependent variable: the number of visits to a health professional)

Sample	Pooled	2003-2005	2008
Internet health information	3.582**	4.388**	2.870
	(1.484)	(2.074)	(1.760)
N	14518	8938	5580

Notes: Standard errors are clustered by state and reported in parentheses. All regressions include age, gender, race, marital status, year variables, and a constant.
** denotes statistical significance at the 5% level.

Table 9: OLS estimates of a regression with full set of individual characteristics versus 2SLS estimates (dependent variable: the number of visits to a health professional)

	OLS		2SLS	
	Coefficient	Standard error	Coefficient	Standard error
Internet health information	0.986***	0.079	3.582**	1.484
Control variables				
Socio-demographic	Yes		Yes	
Year dummies	Yes		Yes	
Health	Yes			
Access to health care	Yes			
Education	Yes			
Health behaviour	Yes			
No. of parameters	28		8	
N	11808		14518	

Notes: Standard errors are clustered by state. Both regressions include a constant.
** denotes statistical significance at the 5% level.
*** denotes statistical significance at the 1% level.